NABARD RESEARCH STUDY - 45

Enabling Farmers with Game Theory and Artificial Intelligence Based Services

Indian Institute of Science, Bengaluru

आर्थिक विश्लेषण एवं अनुसंधान विभाग
Department of Economic Analysis & Research

राष्ट्रीय कृषि और ग्रामीण विकास बैंक, मुंबई
National Bank for Agriculture and Rural Development, Mumbai

2024

Enabling Farmers with Game Theory and Artificial Intelligence Based Services by Indian Institute of Science (IISc), Bengaluru

Research & Development Project funded by

NABARD

Submitted By

Dr. Yadati Narahari
Computer Science and Automation
Indian Institute of Science
Bengaluru

भारतीय विज्ञान संस्थान

Enabling Farmers with Game Theory and Artificial Intelligence Based Services by Indian Institute of Science (IISc), Bengaluru

About NABARD Research Study Series

The NABARD Research Study Series has been started to enable wider dissemination of research conducted/sponsored by NABARD on the thrust areas of Agriculture and Rural Development among researchers and stakeholders. The current study titled 'Enabling Farmers with Game Theory and Artificial Intelligence Based Services' conducted by Indian Institute of Science (IISc), Bengaluru is the forty-fifth in the series.

Artificial Intelligence (AI) combines computer science and robust datasets, to enable problem-solving. AI was first used in agriculture in 1985 by McKinion and Lemmon to develop a cotton crop simulation model named GOSSYM that used AI to optimize cotton production by harnessing the vast agricultural data and applying advanced analytics techniques to find patterns and discover novel insights. Today AI is playing a crucial role in agriculture to determine optimal irrigation schedules, nutrient application timings, monitor plant health, detect diseases, identify and remove weeds, and recommend effective pest control methods and suitable agronomic products. Within crop management, these solutions can be further divided into areas such as crop disease diagnosis, yield prediction, crop recommendation, price forecasting, and market design. However, the adoption of AI and Machine Learning (ML) in the agricultural landscape of India remains limited due to the complexity of these technologies and lack of user-friendly platforms that cater specifically to the agricultural domain.

This study aims to develop AI-based technology for accurate prediction of market prices for agricultural produce, optimal crop recommendation to farmers and setting up of sustainable marketplace for farmers as well as consumers. The study also attempts to create a data analysis platform that streamlines and automates problem analysis and evaluation related to various agricultural data problems by creating a convenient Artificial Intelligence – Machine Learning (AI-ML) pipeline.

Hope this report would make a good reading and help in generating debate on issues of policy relevance. Let us know your feedback.

Kuldeep Singh
Chief General Manager
Department of Economic Analysis and Research

Acknowledgments

First and foremost, I wish to express my gratitude to the leadership of NABARD for providing us with an opportunity to work on a dream project like this one. I would especially like to highlight the regular support and feedback provided by Shri Kuldeep Singh (Chief General Manager), and his colleagues at NABARD including Dr. Ashutosh Kumar (General Manager), Dr. Sohan Premi (Deputy General Manager), and Ms. Anshumala (Assistant Manager).

I wish to place on record the initial push for working with NABARD provided by Mr. Ramana Tadepalli. I am grateful to him. The encouragement provided by Dr. P.V.S. Surya Kumar, former Deputy Managing Director, NABARD, has been tremendous. I wish to gratefully thank Dr. P.V.S. Surya Kumar.

I have benefited immensely from several thought leaders in the domain of agriculture: Prof. N. Viswanadham (Department of Computer Science and Automation, IISc, Bengaluru); Prof. Lalith Achoth (University of Agricultural Sciences, Bengaluru); Prof. Veda Murthy (Dairy Science College, Bengaluru); Mr. Ravi Trivedi (Nudge Foundation, Bengaluru); and Dr. Nipun Mehrotra (Agri Collaboratory, Bengaluru). My thanks to all of them.

This project would not have been possible without the support from students, research staff, and research interns. They are prominently and befittingly listed in a separate section in this report and also featured in one of the back covers of this report. My thanks to all of them. In particular, I must mention Dr. Mayank Ratan Bhardwaj whose Ph.D. work was crucial to a major part of this project work. I also like to specially mention Mr. Inavamsi Enaganti for coming up with innovative ideas many a time.

I must thank the Department of CSA and IISc for providing excellent infrastructural support and research ecosystem. In particular, I thank Ms. Padmavathi and the CSA office staff for their support. Thanks go out to the Chair and staff, Centre for Scientific and Industrial Consulting (CSIC, IISc) for providing logistical support.

We sincerely hope that the outcomes of this project will make at least a small or marginal improvement to the small and marginal farmers of India who provide the lifeline for India.

Y. Narahari

DISCLAIMER

This study has been supported by the National Bank for Agriculture and Rural Development (NABARD) under its Research and Development (R&D) Fund.

Contents

Executive Summary

This project work is motivated by the critical need to address a perennial global problem, namely, how to mitigate the distress of the small and marginal agricultural farmers in emerging economies. Key reasons behind the low returns, and losses, faced by the farmers include the inherent uncertainty in agriculture, unaffordability of advanced technologies, and lack of access to markets. This project formulates and attempts to, at least partially solve, a few of these problems in agriculture, using artificial intelligence and game theory techniques. Novel solutions are proposed that assist the farmers and the state administration during various stages of the agricultural crop cycle, starting from the pre-sowing and sowing decisions and going right up to the harvesting of the produce. These solutions are: PREPARE (Prediction of Prices in Agriculture), ACRE (Agricultural Crop Recommendation Engine), and PROSPER (Protocol for Selling Produce for Enhanced Revenue). A computational platform for agricultural data analytics, AGRI-VAAHAN, has also been built as a part of this project.

PREPARE: Accurate prediction of agricultural crop prices is a crucial input for decision-making by various stakeholders in agriculture: farmers, consumers, retailers, wholesalers, and the Government. PREPARE accurately predicts crop prices using historical price information, climatic conditions, soil type, location, and other key determinants. The proposed approach uses graph neural networks (GNNs) in conjunction with a standard convolutional neural network (CNN) model to exploit geospatial dependencies in prices. PREPARE works well with noisy legacy data and produces a performance that is at least 20% better than the state-of-the-art results in the literature.

ACRE: A key challenge faced by small and marginal farmers is to determine which crops to grow to maximize their utility. ACRE provides a rigorous, data-driven backend for designing farmer-friendly mobile applications for assisting farmers in choosing crops. ACRE uses available data such as soil characteristics, weather conditions, and historical yield data, and uses machine learning/deep learning models to compute an estimated utility to the farmer. The main idea of ACRE is to generate several recommendations of portfolios of crops, with a ranking of portfolios based on the Sharpe ratio, a popular risk metric used for evaluating financial investments.

PROSPER: This contribution is motivated by the need to design a robust market mechanism to benefit farmers (producers of agricultural produce) as well as buyers of agricultural produce (consumers). We design a volume discount auction with a farmer collective as the seller and consumers (high volume or retail customers) as

buyers. Our auction mechanism satisfies properties such as incentive compatibility, individual rationality, Nash social welfare maximization, and realistic business constraints. Since an auction satisfying all of these properties is a theoretic impossibility, we take the route of designing deep learning networks that learn such an auction with minimal violation of the desired properties.

AGRI-VAAHAN: Artificial Intelligence (AI) and Machine Learning (ML) have emerged as powerful tools with the potential to revolutionize the agriculture sector. By harnessing the vast amounts of agricultural data and applying advanced analytics techniques, AI and ML can provide valuable insights and predictive models that help farmers optimize their crop production, manage resources more efficiently, and mitigate risks. The adoption of AI and ML in agriculture is still limited due to the complexity of these technologies and the lack of user-friendly platforms that cater specifically to the agricultural domain. AGRI-VAAHAN aims to address this challenge by creating a convenient computational platform that streamlines and automates problem analysis and evaluation related to various agricultural data problems.

The suite of AI based and game theory based solutions offered in this project, namely PREPARE, ACRE, and PROSPER, constitute a bouquet of innovative approaches towards mitigating the problems faced by small and marginal farmers in emerging economies. A detailed design of mobile apps for these applications has been completed, to enable technology transfer.

Another key contribution of this project work is the preparation of curated datasets that would be available for public access and could be used by the research community.

The project has resulted in five publications in flagship international conferences with two more in the pipeline. A major outcome of this project is capacity generation in this strategic area. The project has been anchored by the Ph.D. thesis of Dr. Mayank Ratan Bhardwaj [10] and has resulted in as many as 10 M.Tech. project theses [86, 57, 18, 19, 47, 61, 28, 64, 20, 78], including an award-winning thesis. In addition, as many as 8 research assistants and research interns have received rigorous training as a part of this project.

List of Figures

List of Tables

1 Introduction

1.1 Background and Motivation

Agriculture has a significant role to play in any emerging economy and provides the source of income and employment for a large portion of the population. Minor changes in crop patterns and their supply chains can lead to a huge impact on the profits of the farmers, as well as, affect the nation's economy as a whole. It is thus imperative for the central and state administrations to properly plan and execute the operations of, and those related to, the agricultural sector.

In India, for example, agriculture plays a pre-eminent role in employment generation as it is a major source of income for around 58 percent of the population [60] [36]. Agriculture in India accounts for 17-18% of the gross domestic product of the nation [82]. India is the second largest food producer in the world [70]. It produces around 1 Billion Tonnes of food commodities per year. It is the second largest producer of wheat and rice [91] [90]. However, unfortunately, India is ranked 71^{st} in the global food security index and 101^{st} in the global hunger index. So there is a critical and urgent need to address the inefficiency of planning and operations in India.

Figure 1: A typical small and marginal farmer is challenged by many problems and is always in a state of distress. Image courtesy: Google Images.

The primary reason for the above dichotomy in Indian agriculture is small landholding sizes. The average landholding size in India is about one hectare (approximately 2.47 acres) [83]. Majority of land-owning Indian farmers are small or marginal farmers (farmers owning less than 5 acres of land). There are numerous

challenges faced by small and marginal farmers. According to the 2016-17 Economic Survey in India, a farmer's average monthly income in 17 selected states is around a meagre ₹1700, resulting in farmer distress and even suicides (Figure 1). There is an increasing trend of conversion of agricultural land for non-agricultural purposes. Furthermore, 48 percent of farmers are not in favour of their children taking up agriculture as their profession, preferring to live in urban areas. The articles and reports by Kovvali and Bharti [71], Chand [15], NITI Aayog, Government of India [4], Ministry of Agriculture, Cooperation, and Farmers Welfare, Government of India [56], and Aatre et al. [3] have brought out the multiple challenges facing the agriculture sector in India and the need for urgent action. Such farmer distress is a common phenomenon in most emerging economies.

Set in this backdrop, this dissertation work is motivated by the critical need to address the perennial global problem mentioned above, i.e., how to mitigate the distress of the small and marginal farmers in emerging economies. Key reasons behind the low returns and losses faced by the farmers include the inherent uncertainty in agriculture, unaffordability of advanced technologies, and lack of access to markets. This dissertation formulates and attempts to, at least partially solve, a few of these problems in agriculture, using artificial intelligence and game theory techniques. Novel solutions are proposed that assist the farmers and the state administration during various stages of the agricultural crop cycle, starting from the pre-sowing and sowing decisions and going right up to the harvesting of the produce.

The solutions proposed in this work have been developed while keeping in mind the common adoption barriers - connectivity and access; affordability; literacy and skill levels; timely availability of relevant information; and data security [16].

1.2 Problems Addressed in the Report

As already stated, this dissertation makes an attempt at solving some of the most pressing problems of small and marginal farmers in emerging economies. Pain points in the planning and operations, arising from the deficiencies and restrictions of the small and marginal farmers, have been identified and solutions have been proposed with the help of Game Theory and Artificial Intelligence techniques. The specific problems and issues addressed in this project are described in the rest of this section.

1.2.1 Crop Price Prediction

The farmer procures inputs (such as seeds, fertilizers, and pesticides) and toils hard for weeks or months, even years for particular crops, to ensure a good harvest. The harvested crop is transported, to be sold, to an agricultural produce market, popularly called a *mandi* in India (Figure 2 shows a typical mandi in India). However, when the crop reaches the mandi, if the crop price is low on that day, the farmer suffers losses. The future price of the crop is detrimental to the success of all the important decisions that a farmer makes, like who to sell, what to sow, how much

Figure 2: A typical mandi. Image courtesy: Newsclick.

to sow, how much to harvest, how much to store, how much to sell, where to sell, how to sell, when to sow, when to harvest, when to sell, what implements to be used and why they are to be used for sowing, harvesting, etc. (The seven fundamental circumstances of Boethius. See [1]). Unfortunately, the farmer does not know what the crop price would be in the future.

The price of a crop depends on a variety of factors. Climatic conditions and water availability dictate the yield, and hence, the price of the crop. Consumption of certain crops is also dependent on the climatic conditions. Demand of the crop, which may vary for different reasons, also determines the crop price. The consumption and demand depend on myriad factors such as geo-politics, incentives, and natural calamities. Local factors such as hoarding and market flooding also contribute to volatility in prices. Since most of these factors are themselves difficult to predict, there is significant uncertainty in crop prices. It is not an uncommon sight to see farmers dumping produce due to prices falling below the cost of transportation to the market [38]. Figure 3 shows such extreme situations in India. Occasionally, there are also situations where the prices of crops skyrocket as it happened in the case of tomato prices very recently [66].

1.2.2 Crop Selection by Farmers

While the state administration is faced with the problem of crop acreage allocation, individual farmers are faced with the problem of determining which specific crops to grow to maximize their utility. With a wrong choice of crops, farmers could end up

(a) Farmer dumping tomatoes in a lake due to low prices in mandis [8].

(b) Farmer dumping onions on the road due to low prices in mandis. Image courtesy: India Today.

Figure 3: Farmers dumping crops due to low prices in mandis.

with sub-optimal yields and low revenues, and possibly suffer losses. For example, in India, most farmers, being small or marginal, do not have access to information or expensive computing power. This makes it difficult for them to choose crops. It is noticed that farmers generally grow the same crops that they had grown previously or grow the crops that neighbouring farmers are planning to grow. Even if the government representatives disseminate advice based on district-level or region-level advisories, some farmers might not follow this advice due to personal reasons or due to local factors such as input availability, procurement costs, soil quality, cultural factors, peer pressure, etc. The farmer will benefit by a tool that would advise a set of the best crops based on local factors, predicted crop price, etc. Depending on their personal preferences the farmers may choose to sow a subset of these crops. There are many efforts in this direction in the literature and the startup world. However, there is significant room to improve the available methodologies.

1.2.3 Crop Selling by Farmers

There is an urgent need to design a robust market mechanism to benefit farmers (producers of agricultural produce) as well as buyers of agricultural produce (consumers). We design a volume discount auction with a farmer collective as the seller and consumers (high volume or retail customers) as buyers. A farmer collective is a group of farmers coming together to gain from the power of aggregation. Our auction mechanism satisfies properties such as incentive compatibility, individual rationality, Nash social welfare maximization, and realistic business constraints. Since an auction satisfying all of these properties is a theoretic impossibility, we take the route of designing deep learning networks that learn such an auction with minimal violation of the desired properties. The proposed auction, which we call PROSPER (PROtocol for Selling agricultural Produce for Enhanced Revenue), is superior in

many ways to the classical VCG (Vickrey-Clarke-Groves) mechanism in terms of richness of properties satisfied and further outperforms other baseline auctions as well. We demonstrate our results for a realistic thought experiment on selling perishable vegetables.

1.2.4 Development of an AIML Platform for Agriculture

In the present day, Artificial Intelligence (AI) and Machine Learning (ML) are the biggest game changers in all sectors. However, their application in the field of agriculture is still in the nascent stage. Vast amounts of data is being collected in the agricultural sector, but it is not being utilized to the maximum possible extent. Using AI and ML predictive models can be developed to help farmers optimize crop production, efficiently manage resources, and mitigate risks. Machine learning algorithms can help farmers and agri-businesses make better-informed decisions and optimize their operations. AI and ML are being used to assist farmer communities in various areas of agriculture, including crop management, water management, soil management, and livestock management. Within crop management, these solutions can be further divided into areas such as crop disease diagnosis, yield prediction, crop recommendation, price forecasting, and market design. The adoption of AI and ML in agriculture is still limited due to the complexity of these technologies and the lack of user-friendly platforms that cater specifically to the agricultural domain. Agri-Vaahan aims to address this challenge by creating a convenient AI-ML pipeline that streamlines and automates problem analysis and evaluation related to various agricultural data problems. Agri-Vaahan provides a platform that streamlines each stage of the model development process.

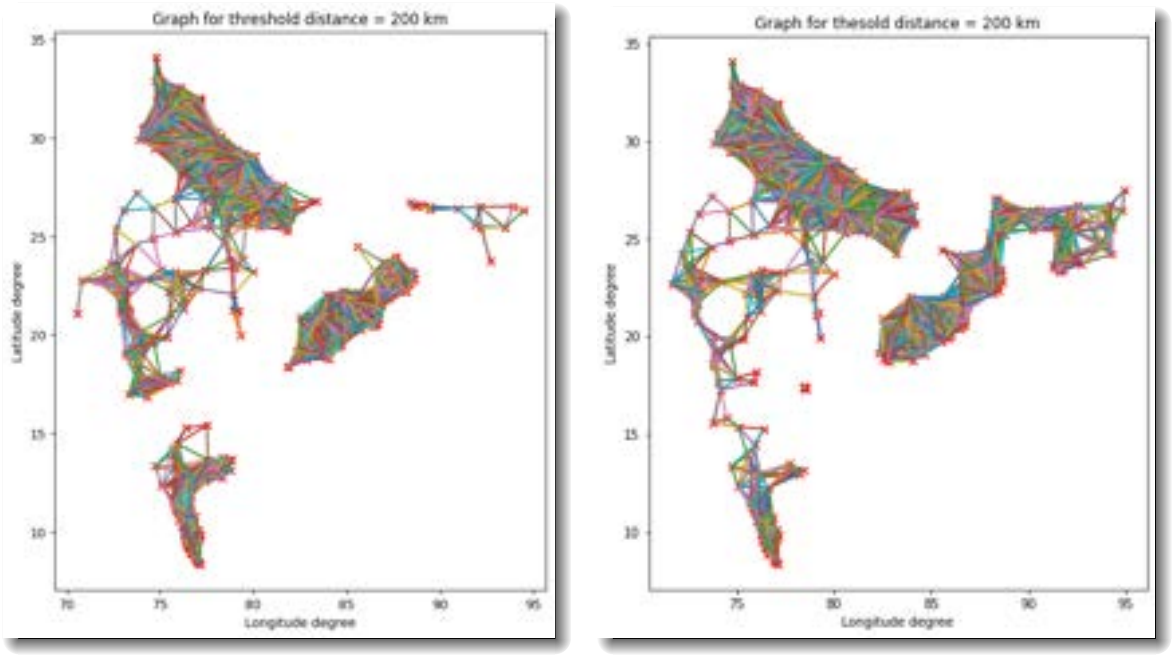
1.3 Contributions and Outline

This dissertation attempts to formulate and solve contemporary problems in agriculture, of the kind highlighted in Section 1.2, using artificial intelligence and game theory techniques. Novel solutions are proposed that assist the state administration and the farmers during various stages of the agricultural crop cycle, starting from the pre-sowing and sowing decisions and going right up to the harvesting of the produce.

The specific solutions offered by this dissertation are: PREPARE (Prediction of Prices in Agriculture), ACRE (Agricultural Crop Recommendation Engine), and PROSPER (Protocol for Selling Produce for Enhanced Revenue). We describe these next and provide an outline of the chapters that contain these contributions.

1.3.1 PREPARE: Prediction of Prices in Agriculture

Accurate prediction of crop prices at an agricultural market or mandi (we will use the word mandi henceforth) is a crucial input for decision-making by various stakeholders in agriculture: farmers, consumers, retailers, wholesalers, and the Government.



(a) Graph of mandis where tomato is sold. (b) Graph of mandis where potato is sold.

Figure 4: Graphs for Tomato and Potato. An edge exists between two mandis if they are within 200 Km of each other.

The prices prevalent in a mandi tend to be highly stochastic and the farmer may not have any control on, or even idea of, the future prices. Our proposed methodology, which we call PREPARE, accurately predicts crop prices using historical price information, climatic conditions, soil type, location, and other key determinants. In this direction, an innovative deep learning based approach is proposed, which achieves increased accuracy in price prediction compared to existing results in the literature. The technical novelty in this methodology is that it captures geo-spatial dependencies between prices of agricultural markets or mandis that are in close proximity to each other. The proposed approach uses graph neural networks (GNNs) in conjunction with other deep learning models to exploit the geospatial dependencies in prices. Figure 4 shows the graphs of mandis for tomato and potato respectively, which were constructed by adding an edge between each pair of mandis that were located within (a threshold distance of) 200 Km of each other.

PREPARE works well with noisy legacy data that is characteristic of the public data available in India and produces a performance that is at least 20% better than the best results available in the literature. Accurate price prediction using PREPARE will significantly enhance the farmer's and state administration's decision-making abilities. Additionally, PREPARE can be used as an input for one other proposed solution in this project, namely, ACRE.

Figure 5: An intercropped field. Image courtesy: agraryo.com.

1.3.2 ACRE: Agricultural Crop Recommendation Engine

While CROP-S can be used by the state administration to make recommendations to the farmers, individual farmers tend to have their own preferences and restrictions in selecting the crops to be grown. These could be due to cultural practices, contractual obligations, input availability, soil suitability, requirement for fodder or self-consumption, or even peer pressure. ACRE, which forms the subject of Chapter 3, can be used to recommend to the individual farmer, which crop (or combination of crops) to grow. The farmer may choose to grow a single crop or grow a combination as either mixed cropping or intercropping. Figure 5 shows a typical field where intercropping is being practiced.

ACRE is a tool that provides a scientific method to choose a crop or a portfolio of crops, to maximize the utility to the farmer. ACRE uses available data such as soil characteristics, weather conditions, and historical yield data, and uses machine learning/deep learning models to compute an estimated utility to the farmer. The main idea of ACRE is to generate several recommendations of portfolios of crops, with a ranking of portfolios based on the Sharpe ratio, a popular risk metric used for evaluating financial investments. Although several papers exist in the current literature that address the problem of crop selection, most of them select the crop based on a single criterion such as water consumption, soil quality, etc. ACRE considers all inputs, including the risk-averse nature of the farmer, before recommending the crop (or portfolio of crops). ACRE provides a rigorous, data-driven back-end for designing farmer-friendly mobile applications for assisting farmers in choosing crops.

1.3.3 PROSPER

This contribution is motivated by the need to design a robust market mechanism to benefit farmers (producers of agricultural produce) as well as buyers of agricultural produce (consumers). We design a volume discount auction with a farmer collective as the seller and consumers (high volume or retail customers) as buyers. A farmer collective is a group of farmers coming together to gain from the power of aggregation. Our auction mechanism satisfies properties such as incentive compatibility, individual rationality, Nash social welfare maximization, and realistic business constraints. Since an auction satisfying all of these properties is a theoretic impossibility, we take the route of designing deep learning networks that learn such an auction with minimal violation of the desired properties. The proposed auction, which we call PROSPER (PROtocol for Selling agricultural Produce for Enhanced Revenue), is superior in many ways to the classical VCG (Vickrey-Clarke-Groves) mechanism in terms of richness of properties satisfied and further outperforms other baseline auctions as well. We demonstrate our results for a realistic thought experiment on selling perishable vegetables.

1.3.4 AGRI-VAAHAN

The main objective of Agri-Vaahan is to design a pipeline that addresses various agricultural challenges using artificial intelligence, machine learning, and deep learning techniques. The pipeline follows the standard machine learning stages but is tailored to the agriculture domain. The pipeline includes: (1) Input data module (2) Feature selection module (3) Data pre-processing and data imputation (4) Data visualisation (5) Model selection (6)Model training and testing, and (7) Output generation, visualisation, and model evaluation.

1.3.5 Other Key Outcomes

The suite of AI based and game theory based solutions offered in this project, namely PREPARE, ACRE, and PROSPER, constitute a bouquet of innovative approaches towards mitigating the problems faced by small and marginal farmers in emerging economies.

A key contribution of this project is the preparation of curated datasets that will be available for public access and could be used by the research community.

The project has resulted in 6 publications in international conferences (1 ACM and 3 IEEE) with one other manuscript in submission.

A major outcome of this project is capacity generation in this strategic area. The project has been anchored by the Ph.D. thesis of Dr. Mayank Ratan Bhardwaj [10] and has resulted in as many as 10 M.Tech. project theses [86, 57, 18, 19, 47, 61, 28, 64, 20, 78]. In addition, 8 research assistants and research interns have received rigorous training as a part of this project.

2 PREPARE: Crop Price Prediction

A very crucial input in the field of agricultural planning is the expected future price of the crop. Hence, accurate crop price prediction is essential to effective crop planning. In this chapter, we explore different combinations of deep learning models such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and graphical neural networks (GNNs) for accurately predicting crop prices. Our experiments reveal that, amongst the various combinations, the CNN-GNN combination works best, performing at least 20% better than the results available in the literature. Our models use historical price information, climate conditions, soil type, location, and other key determinants of crop prices as inputs. Our approach, PREPARE, works well with noisy legacy data and we are able to predict prices up to 30 days ahead. We choose two vegetables, potato (stable price behavior) and tomato (volatile price behavior) and work with noisy public data available from Indian agricultural markets. The results of this chapter have appeared in [12].

2.1 Introduction

Accurate predictions of crop yield and crop price provide valuable inputs for decision making by various stakeholders in agriculture: farmers, consumers, retailers, wholesalers, dealers, and the Government. Some of these decisions have far-reaching implications for the economic well-being of farmers, ensuring food security, stability of supplies, breeding of seeds, and for economic planning in general. This chapter focuses on the problem of crop price prediction.

There are several factors affecting crop prices. These include the expected yield, expected demand, export projections, import decisions, supply chain factors, weather conditions, geospatial factors, unanticipated events such as a pandemic or a flood, etc. Compounding this is the fact that the data that are available in many emerging economies about historical crop prices and crop price variations have several issues such as missing values, outliers, and even data entry errors.

Accurate prediction of crop prices is therefore a grand challenge problem, but at the same time an important one to help secure the economic prosperity of farmers. This chapter compares various combinations of deep learning models and reveals that geospatial (related to a geographic location) dependencies can be harnessed to obtain improved accuracy in predictions of crop prices.

2.1.1 Use-Cases for Price Prediction

There are many potential benefits of knowing what the prices are likely to be and we describe two use-cases below.

2.1.1.1 A Use-Case for the Farmer: When to Harvest

In India, 85% of the farmers belong to the class of small and marginal farmers, with land holdings of less than 5 acres. Small and marginal farmers face many challenges in understanding and interpreting the market and its price dynamics to use them to their advantage. Often the farmers are forced to sell at low prices for lack of information about price forecasts and also due to limited access to cold storage and other facilities. Given any crop, the correct time to harvest is crucial in preventing crop losses. A key determinant in deciding *when to harvest*, is what the market price is going to be when the produce reaches the market. If accurate price predictions are available for the next several days and weeks, the farmer will be able to make an informed decision. Otherwise, the farmer may end up incurring heavy cold storage costs, transportation costs, and losses due to low market price. With price forecasts and simple advisories available, the farmers will be able to decide when to harvest, when to sell their crops, how to plan for the next season, etc.

2.1.1.2 A Use-Case for the Government: Interventions to Support Farmers

Price prediction will help the Government or regulatory authority in making informed decisions on what minimum support prices (MSP) to set for the different crops so that farmers can get benefited. When the predicted price is high, the Government may encourage the exporters to sell locally, rather than overseas, by imposing a minimum export price (MEP). If the one-week or two-week price prediction for a certain crop forecasts attractive prices, then the farmers are elated. However, if low prices are predicted, the farmers could get into panic mode. In such situations, the Government can intervene and assure the farmers that it would buy the produce (or at least a part of it) at a price that is favorable to the farmer. The Government can also initiate various measures for supporting the farmers such as connecting the farmers to food processing units, enabling affordable storage of the produce, regulating import and export, etc.

2.1.2 Contributions and Outline

Agricultural crop price prediction is a technically challenging problem with an extensive body of literature. This chapter advances the state of the art by improving the accuracy of predictions further by introducing innovations in modeling. It uses geospatial proximity in addition to temporal data for making better price predictions. In particular, we use graph neural networks (GNNs) in conjunction with standard deep learning models such as convolutional neural networks (CNNs) to exploit any geospatial dependencies in crop prices. GNNs have powerful representation learning capabilities for graph structured data and have been applied in wide-ranging applications [96]. It is observed that neighbouring mandis experience similar weather patterns, farming practices, and soil conditions due to physical proximity. In addition

to these factors, produce movement between nearby mandis causes a semblance of price stabilisation in mandis within a certain geographical distance of each other. To capture their proximity we represent markets as the nodes of the graph and we use the edges to connect the nodes representing markets close to each other.

The data that we use pertains to all the markets in India and covers two crops from the opposite ends of the spectrum of price volatility, tomato and potato.

We begin by a comparison of CNN, CNN-GNN, RNN, GNN-RNN and CNN-GNN-RNN for predicting the price of tomato in various mandis of UP. CNN-GNN (henceforth referred to as CGNN) comes up as the clear winner of this competition. Having chosen a winner, we now compare our champion technique's results with those of the state of the art, PECAD [31].

Our first experiments are with tomato. The paper [31] only reports coefficient of variation (CoV) results for 4 days, 6 days, and 9 days. It does not report CoV for other time horizons or any other performance metric for any time horizon. The results clearly show that the CGNN approach outperforms the PECAD approach in all the cases reported in [31].

We then turned our attention to potato and obtained results with two approaches - PECAD [31] and CGNN (our approach). The paper [31] does not report any results for potato, so we ran the PECAD code (available from the authors of [31]) on our potato dataset. We computed five different performance measures: root mean square error, mean absolute error, coefficient of variation, R2 value, and Pearson's correlation coefficient. The results clearly show that the CGNN approach outperforms the PECAD approach on all performance metrics for all time horizons.

In Section 2.2, we provide a review of relevant work in crop price prediction. In Section 2.3, we describe the details of our methodology: nature of data used by us; data imputation and data curation; deep learning models used; and the architecture of the price prediction network. In Section 2.4, we report the results from our experiments: a comparison of various combinations of deep learning techniques, followed by a comparison of our best technique with the state of the art. Section 2.5 provides the conclusions drawn from these experiments.

We have used tomato and potato as two representative crops. The same methodology is applicable to other crops as well.

2.2 Review of Relevant Work

There have been numerous research efforts toward crop price prediction. The problem offers a number of technical challenges since the crop price is determined by a large number of factors. It should be noted that the price prediction problem is related to but also different from the crop yield prediction problem. There is abundant literature on the crop yield prediction problem. [85] is a comprehensive survey on yield prediction. Here, we review recent papers on price prediction and a paper on yield prediction that are most relevant to our work.

Ma, Nowocin, Marathe, and Chen [51] present a crop price forecasting system using data from *agmarknet* [23] (a Government website) in India. They train a classification model of prices using 1352 markets in India and produce interpretable price forecasts. The pricing data available from *agmarknet* is sparse and the authors impute missing entries using collaborative filtering to obtain a dense dataset. Using these data, a decision-tree-based classifier is trained to predict the direction of movement in crop prices at different markets. The system uses adaptive nearest neighbor methods to obtain interpretable forecasts. The forecasting system is used to predict price changes for six crops (brinjal, cauliflower, pointed gourd, mango, tomato, and green chilli). The authors do not take into account spatio-temporal dependencies of crop prices. Additionally, they classify the price as either going up or down, while we predict the actual prices.

The paper by Guo, Woodruff, and Yadav [31] presents a deep learning based algorithm PECAD (Price Estimation for Crops using the Application of Deep learning), for accurate prediction of future crop prices based on past pricing and volume patterns. The paper uses real-world daily price and volume data of different crops across India using the same database as above, *agmarknet*. From the available markets, they eliminate the markets that have less than 10% data available and carry out experimentation on the remaining markets. The data are batched into n-day non-overlapping snippets and pre-processed using imputation techniques such as SoftImpute to account for missing data entries. PECAD works with a wide and deep neural network architecture consisting of two separate convolutional neural network models (trained for pricing and volume data respectively). PECAD outperforms existing baseline methods by achieving 25% less Coefficient of Variation (CoV). This model takes into account the latitude and longitude of the large number of markets to capture the effect of geospatial factors at a high level. It does not specifically exploit the dependence of crop prices at one market on the prices at other markets.

The paper by Madaan et al. [52] looks at onion and potato trading in India and presents an evaluation of a price forecasting model, and an anomaly detection and classification system to identify incidents of hoarding of stock by the traders. The dataset consists of a time series of wholesale prices and arrival volumes of the agricultural commodities at several village-level marketplaces, and retail prices of the commodities at the city centers. The paper presents a qualitative analysis of the effect on these time series of events such as hoarding, weather disturbances, and external shocks. The authors employ ARIMA and SARIMA models in conjunction with an LSTM model for their prediction tasks. The pricing models are useful to reduce information asymmetries and to detect anomalies that can help regulate agricultural markets to operate more fairly. This paper does not take into account factors such as geospatial dependencies.

Zhang et al. [94] propose a model selection framework that includes time series features and forecast horizons for predicting the price of agricultural commodities.

As many as 29 features are considered as possible influencers of prices. Three models are explored: neural networks, support vector regression, and extreme learning machine. Random forests and support vector machines are applied for learning the relationships between the features. Techniques are proposed to remove feature redundancies. This model does not capture any geospatial dependencies.

Jain et al. [39] present the architecture of an end-to-end pipeline for robust crop price prediction through the analysis of historical marketplace data and weather data. They also discuss data quality-related features. The paper presents a framework that facilitates context-based model selection strategies with data quality, model stability, and historical price trends as the context determinants. The authors experiment with various regression models and show the results for tomato and maize crops for 14 markets in the state of Karnataka, India. This modeling effort also does not capture geospatial factors.

The paper by Fan et al. [27] explores the use of graph neural networks (GNN) in conjunction with recurrent neural networks (RNN) to capture geospatial as well as temporal knowledge in crop yield prediction. The work is able to take advantage of the spatial structure in the data to produce better yield predictions. Their model first extracts year-wise embeddings using a CNN model. These are then fed as input to the GNN, which enables inductive representation learning on graphs. These outputs are then fed into the RNN to obtain the final yield results. The graph here is an unweighted graph that consists of more than 2000 counties as nodes and the edges represent neighborhood relationships among the counties. Detailed experiments on large-scale datasets covering 41 states in the United States demonstrate that the approach outperforms state-of-the-art machine learning methods across multiple datasets by almost 10 %. The graph captures the neighborhood relationship among the counties, thereby bolstering the results. It should be noted here that [27] works on yield prediction in USA, which has rich and perfect data. Our work, on the other hand, deals with crop price prediction in India, where the data available are sparse as well as noisy.

As already pointed out, the crop yield prediction is a different problem than the crop price prediction problem. The set of factors affecting crop price form a superset of factors affecting crop yield. Crop price crucially depends on demand and supply (which in turn depends on yield) while yield depends very little on price or demand. Price prediction is clearly a more involved problem and taking into account geospatial dependencies is in itself an interesting, independent problem.

2.3 Methodology

In this section, we describe the methodology used to predict the daily price for agricultural markets or mandis. It should be noted that a mandi can be characterized by its latitude and longitude. We start by describing the data used and pre-processing techniques applied on the data.

2.3.1 Crop and Weather Data

The website run by the Directorate of Marketing & Inspection (DMI), Ministry of Agriculture and Farmers Welfare, Government of India [23] provides daily Price (in rupees per quintal) and Arrival data (in tonnes) for all mandis in India. The Price data includes the daily maximum price, minimum price, and modal price for each crop in each mandi. However, data availability is highly variable across crops and mandis. For our experimentation, the mandis with very sparse data were weeded out. Only those mandis were considered that had price and arrival data for at least 4 days in each year between 2014 and 2018. Of the 1320 and 730 mandis where data for tomato and potato was available, 557 and 676 mandis, respectively, were considered for price prediction.

Weather data consisting of hourly values of Rainfall, Temperature, Surface net solar radiation, and Humidity were taken from Copernicus, the European Union's Earth Observation Programme [26], for a period of five years (2014-2018). This data could be used directly as all the missing values and anomalies were corrected before the data was uploaded. The data are available for every 0.25 degree change in latitude and longitude. Each mandi's latitude and longitude were compared to the available pairs of latitude and longitude and the weather data for that mandi were taken to be the same as the weather data of the nearest available point. For mandis that were almost equidistant from multiple available points, averages of the available data were taken. In this manner, hourly weather features were computed for each mandi.

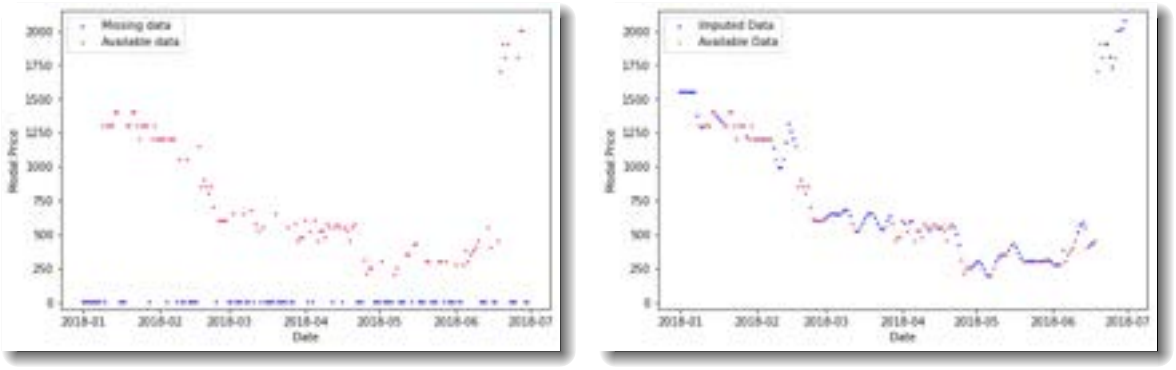
2.3.2 Pre-Processing

The price and arrival data have many missing values. An illustrative example of sparsity in tomato price data in Sardhana mandi of Meerut district is provided in Figure 6a. All over India, 46% of the price data are missing. While some mandis like Kharar have 1.81% data missing, mandis like K.R.Pet have 98.14% data missing.

In addition to the missing values, we found several instances where the data were misreported, probably due to the manual data-entry process. In these cases, the entry for a particular day would typically have an additional zero or a missing digit. To catch these and other outliers, we checked if the current value was either six times higher, or lower than a sixth, of the average of the previous week. If there were no values present in the previous week, we backtracked until the most recent ground-truth value available.

We then performed year-wise spline imputation [5] on these data. Since spline does not work well for sparse data, this technique was used only for mandis that had data available for more than half of the year. We found that cubic splines work best for these data.

The spline imputation also gives rise to some outliers, which were caught using a moving-window approach. For each imputed data-point, a window of size 15 (one



(a) Before Imputation (b) After Imputation

Figure 6: Data imputation on Tomato prices for the first half of 2018 in Sardhana mandi. The zero-prices correspond to the dates where datum was not available.

week before, that particular day, and one week after) was taken and all the ground truth values in that window were considered. If no ground-truth values were present in that window, it was expanded on both sides until sufficient values were obtained. Then, the imputed value was reported as an outlier if it was beyond +/-15% of the window's extrema.

Finally, we used linear interpolation to impute the remaining missing values. Figure 6b displays the tomato prices in Sardhana mandi after imputation.

All numerical values were normalised. The date feature was broken down into two cyclic features, namely month and day, both of which were transformed into two dimensions using a sine and cosine transformation. This is done using the following (circular embedding) transformation for the features month as well as day of month :

$$x_{sin} = \sin\left(2\pi x/\max(x)\right) \quad and \quad x_{cos} = \cos\left(2\pi x/\max(x)\right)$$

For example, the date June 10 would be broken down into the 10^{th} day of the 6^{th} month. Further, $day = 10$ would be transformed into $day_{sin} = \sin\left(2\pi 10/30\right) = \sin\left(2\pi/3\right)$ and $day_{cos} = \cos\left(2\pi 10/30\right) = \cos\left(2\pi/3\right)$, while $month = 6$ would be transformed into $month_{sin} = \sin\left(2\pi 6/12\right) = \sin\left(\pi\right)$ and $month_{cos} = \cos\left(2\pi 6/12\right) = \cos\pi$.

Since all the mandis are closed on Sundays, resulting in no activity, all Sundays were removed from the data set. The latitude and longitude of each mandi were also provided as additional features.

2.3.3 Data Analytics

Price data are susceptible to minor fluctuations even over a short span of a few days. These fluctuations occur due to a variety of factors including measuring errors, human behavior, and randomised natural phenomena, which cannot be measured due to lack of data and set mechanisms. To smoothen the data, the features were av-

Temporal Features	Features from Copernicus	Features from Agmarknet
Day	Temperature	
Month	Total Precipitation	Arrival Quantity
Year	Relative Humidity	Modal Price
	Surface net Solar Radiation	
Spatial Features	**Computed Features**	
	Day of month - Cyclic encoding: sine	
Latitude	Day of month - Cyclic encoding: cosine	
Longitude	Month - Cyclic encoding: sine	
	Month - Cyclic encoding: cosine	
	Average of previous 7 days modal price	

Table 1: Input features

eraged over the duration of a few days. The corresponding data entry obtained by averaging all the features over n days is called an *n-day snippet*. A single n-day snippet is categorised by crop, mandi, and the first day of the snippet. Thus the data corresponding to a crop in a particular mandi was broken into continuous intervals of n-day snippets after the removal of Sundays. 4, 6, 9, 15, and 30 were considered as possible values for n. The consideration was to look at representative values from a few days to a month, as well as to compare our results to the state of the art. All 16 features for each n-day snippet were taken as input. A list of all the features is provided in Table 1.

While the temporal (direct and computed) features help catch seasonal trends, the spatial features help catch geospatial dependencies. Congeniality of the climatic features has an effect on the yield and arrivals, thereby affecting the crop price.

Since the data is temporal in nature, to ensure that causality is maintained during the modeling process, future data was avoided while building the models. Hold out cross validation was used to validate the models. Model training was carried out over the 3-year period of 2014-2016. Validation was done using the data for year 2017 and the data for year 2018 were used for testing.

2.3.4 Deep Learning Models

Experiments were carried out on various combinations of deep learning models, viz., CNN, CNN-GNN, RNN, GNN-RNN and CNN-GNN-RNN. Given below are very brief descriptions of the three basic deep learning models that were combined together to form various complex deep learning architectures for Agricultural Price Prediction.

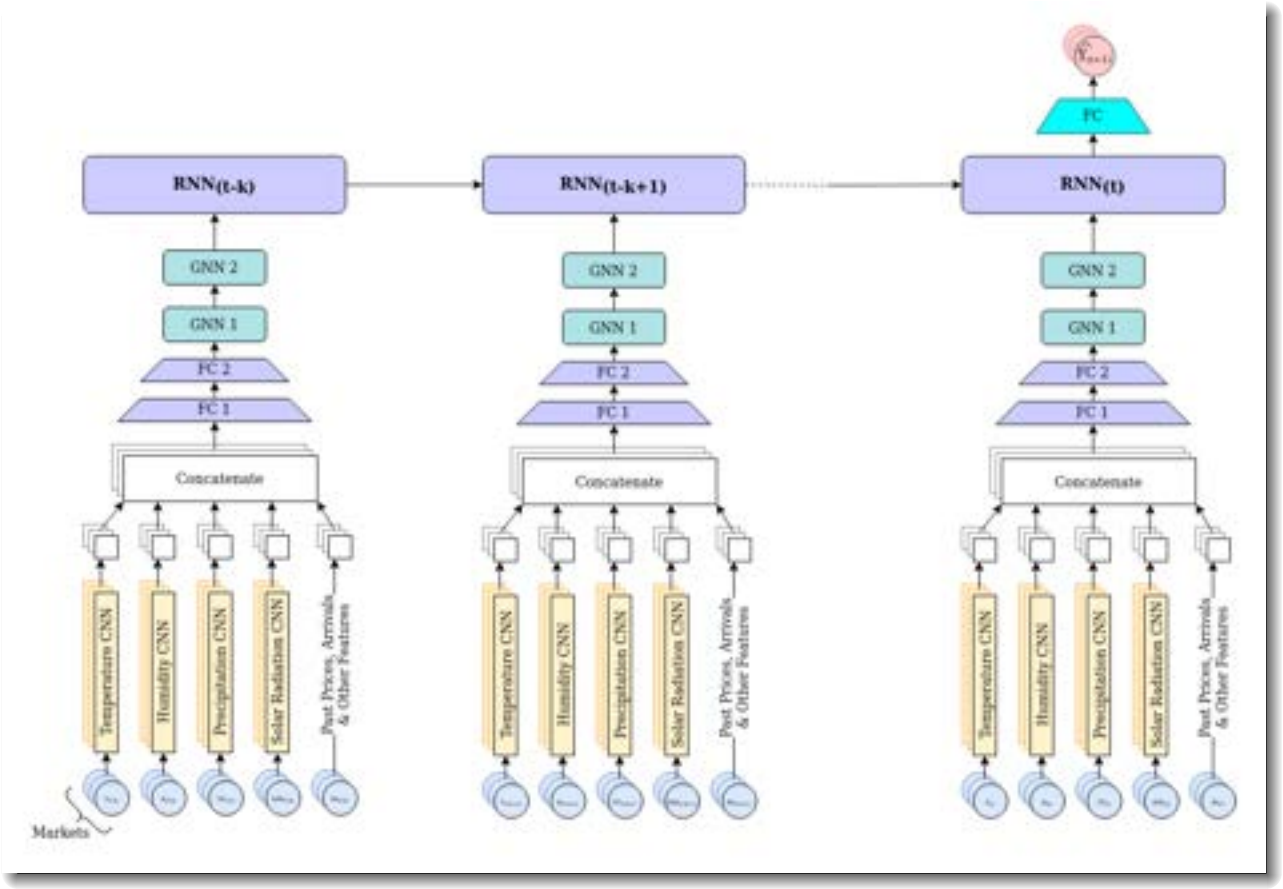


Figure 7: CNN-GNN-RNN Model

Convolutional Neural Networks (CNNs)

To capture temporal variations in weather patterns and extract informative representations, we employ individual 1-dimensional convolution operations on each weather feature. The application of 1D convolution operations proves beneficial in extracting relevant features from shorter, fixed-length sequences, such as hourly weather data.

Recurrent Neural Networks (RNNs)

Recurrent Neural Networks were used to capture the temporal patterns in mandi prices. The LSTM variant of RNN gave best results when a Fully Connected layer was applied on top of the RNN, with the inclusion of suitable non-linear activation functions.

Graph Neural Networks (GNNs)

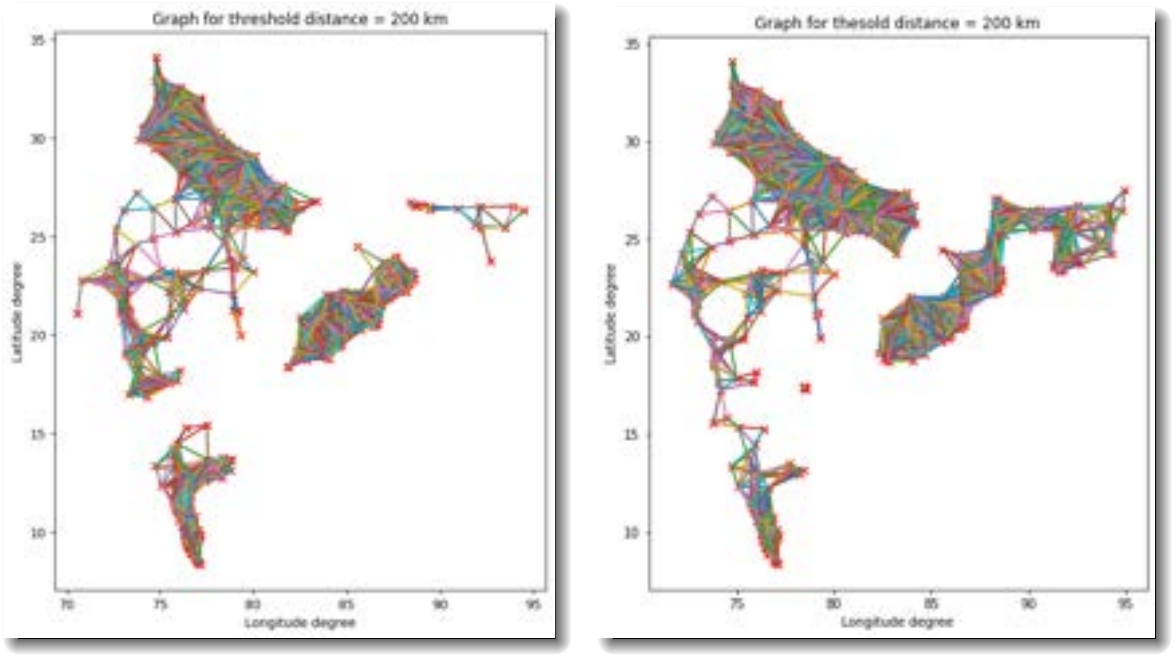
Graph Neural Network is a neural network that operates directly on a graph and induces embedding vectors for nodes based on their neighborhood. Various ways to incorporate neighborhood information lead to various types of GNNs like Graph Convolution Networks, GraphSAGE, Graph Attention Networks, etc.

Figure 7 gives a quick glimpse of CNN-GNN-RNN, the most complex architecture made from these basic deep learning models. Our experiments reveal that for price prediction in Indian mandis, CNN-GNN (CGNN) is the most apt combination of deep learning models. In the following section we describe in detail the architecture and training process of the CGNN model.

2.3.4.1 CNN-GNN Model

Historical data in the form of n-day snippets were used to predict the average price of the subsequent n-day snippet. The weather features are available as hourly values for each mandi (See Section 2.3.1 for details). A 1-dimensional convolution (1D CNN) operation was applied to capture temporal changes and generate a compact embedding. A 1D CNN is very effective when one expects to derive embeddings from shorter, fixed-length sequences. The 1D-CNN was paired with a ReLU activation function and a Max pool layer. Four such instantiations were created; one for each weather feature. The outputs of the 4 models were then concatenated with the remaining data. As the number of intermediate features obtained at this stage was large, the concatenated result was passed through two successive Fully-Connected layers. The resulting embedding vector of size 5 for each mandi constitutes the input to the GNN model.

For the GNN model, the graph network was built for all the mandis in the data set that were shortlisted (on the basis of data availability) for the crop being considered. Each vertex of the graph represents a shortlisted mandi and is categorised by the mandi's geographic location given by its latitude and longitude. If the direct distance



(a) Graph of mandis where tomato is sold. (b) Graph of mandis where potato is sold.

Figure 8: Graphs for Tomato and Potato. An edge exists between two mandis if they are within 200 Km of each other.

(as the crow flies) between two mandis was found to be within a certain threshold, an edge was added to the two vertices in the graph corresponding to these mandis. Various such thresholds were considered. Of these, the 200 Km threshold provided the best prediction results at the all-India level. It should be noted that each crop would have a different subset of shortlisted mandis (depending on data availability) and thus, a separate graph would be required for each crop. Furthermore, the perishability, growth requirements, storage and price dynamics of each crop could result in a different threshold being selected for each crop. Figure 8 shows the final graphs obtained for the threshold of 200 Km for tomato and potato respectively.

A range of experiments was performed with different variants of GNN like GCN [45], GraphSAGE [32], GAT [87], and GAT V2 [13]. We found that the GraphSAGE variant with two layers works best for our setting. The final GNN model consists of two GraphSAGE layers followed by ReLU activation layers. The detailed architecture of the model is shown in Figure 9.

2.3.5 Performance Metrics

As two distributions cannot be compared using a single parameter, it is impossible to measure the goodness of fit of predicted values using a single metric. Similarity or prediction accuracy is measured in different dimensions by different metrics. In

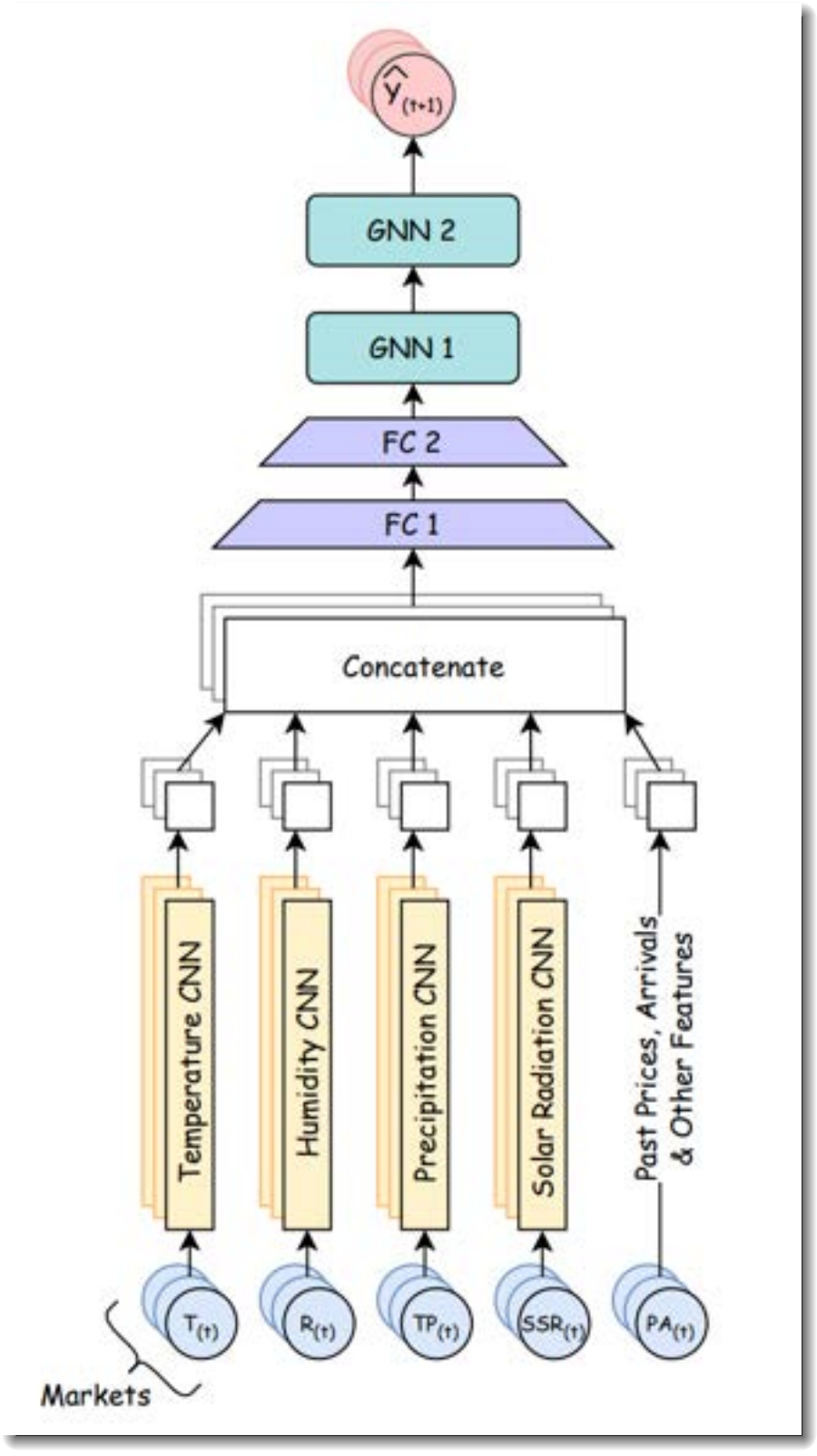


Figure 9: CGNN Model

the absence of a set of metrics that can be used for a universal comparison, it is best to use as many metrics as possible. We found that all papers on crop price prediction in the literature were reporting a single or at most two metrics. In order to confirm the veracity of our models, we have computed and reported all of the metrics mentioned below. The error metrics are calculated over a set of n data points where $y = \{y_1, y_2, ...y_n\}$ are the predicted prices and $x = \{x_1, x_2, ...x_n\}$ are the corresponding actual prices. $\bar{y} = (\Sigma_{i=1}^{n} y_i)/n$ and $\bar{x} = (\Sigma_{i=1}^{n} x_i)/n$ are the respective mean values of the predicted and actual prices.

- Mean Absolute Error (**MAE**) is calculated as the sum of absolute errors between predicted price and target price, divided by the sample size.

$$MAE = \frac{\Sigma_{i=1}^{n} |y_i - x_i|}{n} \tag{1}$$

  MAE is a widely used error metric, which treats positive and negative deviations equally. MAE is resilient to outliers when compared to other metrics like the Root Mean Square Error.

- Root Mean Square Error (**RMSE**) is the standard deviation of the prediction errors.

$$RMSE = \sqrt{\frac{\Sigma_{i=1}^{n} (y_i - x_i)^2}{n}} \tag{2}$$

  Due to the squaring of the error, RMSE is sensitive to outliers and can be used as a replacement of (or in conjunction with) MAE when outliers are not acceptable.

- Coefficient of Variation (**CoV**) is calculated as the RMSE divided by the mean of the sample.

$$CoV = \frac{RMSE}{\bar{x}} = \frac{\sqrt{\Sigma_{i=1}^{n} (y_i - x_i)^2 / n}}{\bar{x}} \tag{3}$$

  The denominator in the formula helps make CoV a very effective metric, especially when making comparisons between prediction techniques applied to different crops, or different mandis, or both. It can also be used to compare the performance of a model for different crops or for different mandis. For example, predictions made by a model for prices of different crops in a mandi might

have the same RMSE values in-spite of the average prices of both the crops being different. This implies that the model works better for one crop than the other (since the RMSE values are same, we can safely conclude that the model works better for the crop with the higher mean price), which is reflected in their respective CoV values that are different from one another.

- Coefficient of Determination or R2 score (R^2) is defined as the proportion of the variance in the dependent variable that is predictable from the independent variables.

$$R^2 = 1 - \frac{\Sigma_{i=1}^{n}(y_i - x_i)^2}{\Sigma_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

  The ideal R2 score is 1. An R2 score of 0 means that the model is unable to explain the relationship between the actual and predicted price, while an R2 score of 1 implies that the model perfectly explains the relationship.

- Pearson's Correlation Coefficient (**r**) is a measure of linear correlations between the predicted prices and target prices. It is calculated as the covariance between the two sets divided by the product of the standard deviations of the two sets.

$$r = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\Sigma_{i=1}^{n}(y_i - \bar{y})^2}} \tag{5}$$

  $r \in [-1, 1]$. The Pearson's Correlation Coefficient is invariant under separate changes in location and scale in the two prices, i.e., transforming X to $a + bX$ and Y to $c + dY$, where $a, b, c,$ and d are constants such that $b > 0$ and $d > 0$, does not alter the value of the correlation coefficient.

2.4 Experimental Results

In this section we present the results of our experiments. Section 2.4.1 presents the results of the comparisons between different combinations of some popular deep learning models. The model with the best performance is then compared, in Section 2.4.2, with the state of the art.

2.4.1 Comparison of Various Combinations of DL Models

Experiments were conducted with various deep learning models and their combinations. These experiments revealed that, in general, for predicting prices in Indian agricultural produce mandis, a combination model of CNN-GNN works best.

Development of a common model for predicting the price in all mandis of Uttar Pradesh was attempted. Various threshold values were tried for different crops. In

	CNN	CNN-GNN	RNN	GNN-RNN	CNN-GNN-RNN
4 days	126.10	166.72	853.32	323.29	368.29
6 days	274.67	208.24	785.87	395.62	361.01
9 days	387.61	173.22	782.04	350.47	313.96

Table 2: RMSE values of different combinations of DL techniques for Tomato price prediction in UP. 150 Km threshold was used for constructing the graph for GNN.

general it was noticed that the use of GNNs in conjunction with other popular deep learning techniques predicts prices much better than when the deep learning technique is applied in isolation. Table 2 presents one such comparison of the RMSE values. These results are for the prediction of Tomato prices in the mandis of UP. It was found that all GNN techniques give best results for Tomato price prediction in UP when the threshold distance is taken as 150 Km. As is evident from the table, in general, CNN-GNN performs better than CNN. The performance of GNN-RNN is superior to that of RNN. Although CNN-GNN-RNN performs better than RNN, it is unable to beat the performance of the other models.

In the following section, the performance of our best performing model, CNN-GNN (CGNN) is compared with that of the state of the art.

2.4.2 Comparison of CGNN with the State of the Art

In this section, we present the results of our experimentation on two popular crops in India: tomato and potato. We chose tomato and potato for a clear reason: prices of potato are somewhat stable while the prices of tomato are volatile. We train a model each for tomato and potato to predict the prices in all the mandis of India where these crops are sold. In the following sections, we shall see a comparison of the performance of CGNN (our approach) with that of PECAD (Price Estimation for Crops using the Application of Deep learning) [31], the best model in the current literature.

The CGNN model was trained for 4000 epochs and the best checkpoints were saved based on the performance of the model on the validation set. A learning rate of 0.1 was used, coupled with a scheduler that reduced the learning rate to 0.01 after 3000 epochs.

2.4.2.1 Comparisons for Tomato

Table 3 shows the results obtained for tomato with two approaches - PECAD [31] and CGNN (our approach). The paper [31] only reports coefficient of variation (CoV) results for 4 days, 6 days, and 9 days. It does not report the coefficient of variation for other time horizons or any other performance metric for any time horizon. Hence, the corresponding values are not available in Table 3. We have computed all five different

	RMSE	MAE	CoV	R2 value	Pearson's
4 days PECAD	-	-	0.216	-	-
4 days CGNN	231.62	142.92	**0.168**	0.924	0.962
6 days PECAD	-	-	0.242	-	-
6 days CGNN	269.14	171.33	**0.195**	0.897	0.948
9 days PECAD	-	-	0.285	-	-
9 days CGNN	308.12	207.08	**0.223**	0.863	0.937
15 days PECAD	-	-	-	-	-
15 days CGNN	460.65	347.75	0.334	0.689	0.885
30 days PECAD	-	-	-	-	-
30 days CGNN	532.29	411.25	0.383	0.569	0.838

Table 3: Performance metrics for PECAD [31] and CGNN (our approach) for tomato crop price prediction ('-' means results not available)

	RMSE	MAE	CoV	R2 value	Pearson's
4 days PECAD	134.82	92.07	0.111	0.958	0.983
4 days CGNN	119.58	66.69	0.098	0.967	0.983
6 days PECAD	174.18	138.13	0.144	0.929	0.979
6 days CGNN	131.37	77.78	0.107	0.960	0.980
9 days PECAD	182.29	142.12	0.150	0.922	0.973
9 days CGNN	141.89	89.49	0.115	0.952	0.976
15 days PECAD	228.83	184.73	0.186	0.875	0.961
15 days CGNN	149.88	100.66	0.120	0.946	0.973
30 days PECAD	343.50	269.09	0.279	0.714	0.893
30 days CGNN	201.77	148.87	0.159	0.898	0.952

Table 4: Performance metrics for PECAD [31] and CGNN (our approach) for potato crop price prediction

performance measures. The results in Table 3 clearly show that the CGNN approach outperforms the PECAD approach in all the cases reported in [31] by achieving 21% less coefficient of variation.

2.4.2.2 Comparisons for Potato

Table 4 shows the results obtained for potato with two approaches - PECAD [31] and CGNN (our approach). The paper [31] does not report any results for potato, so we ran the PECAD code (available from the authors of [31]) on our potato dataset. We computed five different performance measures root mean square error, mean absolute error, coefficient of variation, R2 value, and Pearson's correlation coefficient. The results in Table 4 clearly show that the CGNN approach outperforms the PECAD approach on all performance metrics for all time horizons. Our models are achieving an average of 27% less CoV than the PECAD model.

As expected, across all models, the CoV for tomato price prediction is much higher than the CoV for potato price prediction since tomato prices are more volatile.

2.5 Summary

In this chapter, various popular deep learning techniques and combinations thereof were applied towards the goal of predicting crop prices in the mandis. It was found that, in general, the application of graph neural networks along with other deep learning techniques leads to more precise predictions. Our innovative deep learning model, combining convolutional neural networks and graph neural networks, provides highly accurate estimates of crop prices. Experiments were carried out on tomato and potato crops across all the markets in India. PREPARE produces a performance that is at least 20% better than the results available in the literature.

3 ACRE: Agricultural Crop Recommendation

Crop selection at the sowing stage heavily affects the farm output and the well being of the farmer. With a wrong choice of crops, farmers could end up with sub-optimal yields and low, and possibly even loss of, revenue. This chapter seeks to design and develop ACRE (Agricultural Crop Recommendation Engine), a tool that provides a scientific method to choose a crop or a portfolio of crops, to maximize the utility to the farmer. CROP-S, described in the previous chapter, can be used by the state government to make generalized recommendations for all farms lying in a particular geographical area. However, there may be certain restrictions and preferences specific to individual farmers, which may cause the farmer to choose a different crop (or set of crops) from what was advised by the state administration. In such scenarios, ACRE aids the individual farmer in choosing the optimum crops. Towards this end, already available data such as soil characteristics, weather conditions, and historical yield data is used along with state-of-the-art machine learning/deep learning models to compute an estimated utility to the farmer. These utilities are used to generate several recommendations of portfolios of crops. The different portfolios are ranked based on a popular risk metric in financial investments, the Sharpe ratio. We use publicly available data from *Copernicus*, *VDSA*, and the *agmarknet* portal to perform several thought experiments with ACRE. ACRE provides a rigorous, data-driven back-end for designing farmer-friendly mobile applications for assisting farmers in choosing crops. The results of this chapter are published in [65].

3.1 Introduction

A crucial reason behind the struggles faced by small and marginal farmers is that they frequently make poor crop-selection decisions. Crop selection is an extremely important decision for farmers and there are numerous factors to be taken into account while choosing a crop or a combination of crops.

Factors Influencing Crop Selection

The handbook by Chandra Shekara et al. [76] enunciates a number of factors influencing the choice of crops. We provide a brief summary here. The factors include [76] [88]:

1. **Climatic factors**: Temperature, rainfall, sun shine hours, relative humidity, wind velocity, wind direction, seasons, and agro-ecological situations.
2. **Soil factors**: Soil type, pH value, and soil fertility.
3. **Water availability**: Tanks, wells, dams, ground water, rainfall, water quality, water suitability, resources for lifting water, and availability of micro-irrigation systems.

4. **Cropping system options**: Inter-cropping, mixed cropping, multi-storeyed cropping, relay cropping, and crop rotation.
5. **Socio-Cultural factors**: Traditional best practices; farmer beliefs and superstitions; opinion of family members, neighbours, and friends.
6. **Expected profit**: Profit expected from selling the crops post harvest. There is inherent risk due to market conditions.
7. **Expected yield**: Estimated yield of a crop during harvest. There is inherent risk due to crop damage or loss.
8. **Economic conditions**: Land holding of the farmer; financial resources (including credit availability); labour availability and affordability; availability and affordability of farm mechanization.
9. **Technological factors**: Access to modern technology; feasibility of technology options.
10. **Market demand and access**: Demand for the crop in neighbouring areas; accessibility to markets and real-time market information.
11. **Government policies and schemes**: Availability incentives, schemes, accessibility to Government call centres, access to farmer producer organisations (FPOs), etc.
12. **Agricultural inputs**: Availability and access to high quality inputs (seeds, fertilizers, pesticides, etc.).
13. **Post harvest storage and processing**: Access to storage facilities, food processing units, technologies for adding value to crops, etc.

It is difficult to expect small and marginal farmers to weigh-in all the factors described above in an informed or algorithmic way and choose the best crops. Lack of knowledge, lack of awareness, and interference of local intermediaries are major obstacles for small and marginal farmers in choosing the best portfolio of crops. This is the major motivation behind the development of ACRE (Agricultural Crop Recommendation Engine), a tool that provides a scientific method for choosing a crop or a portfolio of crops to maximize the utility to the farmer. ACRE provides a powerful back-end (computational engine) that embeds rigorous algorithms to provide decision support to the farmer in crop selection. To use ACRE effectively, there is a need to provide a farmer-friendly interface (for example a mobile application) that gathers only essential information from the farmer and provides a crop portfolio recommendation using ACRE. In this chapter, our focus is on the computational engine. Clearly, the farmer should be required to provide minimal information with a simple user interface and should not worry at all about the computational engine.

3.1.1 Contributions and Outline

ACRE works with information such as location of the farmer, feasible subsets of crops, season (Kharif, Rabi, whole year), utility functions of farmers, irrigation facilities, crop rotation cycle, area (in hectares), human labor cost, etc. Some of these are provided by the farmer and ACRE infers the rest of the information. ACRE either sources or infers the parameters such as temperature, rainfall, sunlight, humidity, soil type, and soil nutrients. The computational engine in ACRE computes estimates of yield and cost, as well as, the standard deviations of these estimates. The system then computes an appropriate farmer utility for each crop and provides individual crop recommendations for Kharif and Rabi seasons, based on these utilities. The system also provides inter-crop recommendations, considering several portfolios of crops that could be grown in a given season. A technical novelty of ACRE is the use of Sharpe ratio [74] [75], a portfolio investment metric widely used in financial portfolio selection, for evaluating crop portfolios.

One of the significant components of ACRE is crop yield estimation. Crop yield is a complex variable influenced by several factors, including genotype, environment, and interactions. For accurate yield prediction, a fundamental understanding of the functional relationship between yield and these interaction components is required [42]. For yield prediction, we use an ensemble technique consisting of standard machine learning and deep learning regression models. Previous efforts reported in the state-of-the-art literature compute an average yield for different crops. They do not compute the variance of yield which is a good indicator of the risk involved in selecting a given crop. To accurately capture the variance in yield, we use the triangular distribution to capture yield data by considering three parameters: mode, maximum, and minimum values of the crop yield. After considering the predicted yield and price, estimated cost, and the crop cycle duration of each crop, we recommend optimal crop portfolios to the farmers based on the involved risk. Our recommendation is in terms of the proportion of each crop a farmer should grow in order to maximize the utility while accounting for risk.

Section 3.2 provides a review of relevant work on crop recommendation and crop yield estimation. Section 3.3 provides an overview of the Sharpe ratio and brings out its relevance for agricultural crop portfolio selection. Section 3.4 provides details of the public datasets that were used in this study. Section 3.5 describes the various building blocks of ACRE. Section 3.6 presents our experimental results on yield estimation and crop recommendation. Section 3.7 mentions some factors that ACRE does not currently consider and suggests how these factors can be incorporated while using ACRE for recommending crops. Section 3.8 concludes the chapter.

3.2 Review of Relevant Work

In this section, we provide a review of the relevant literature. Shi, Wang, and Fang [77] have selected 1176 papers in the area of *Artificial Intelligence for Social Good*

from leading publication venues during the years 2008-2019 and provide a summary of the research. One of the key application areas explored by them is digital agriculture. Many papers surveyed by them address the problems of yield prediction and crop recommendation. A significant amount of research has been reported on the effects of soil characteristics, climatic conditions, and geography on agricultural productivity.

3.2.1 Relevant Work in Crop Recommendation Systems

Von Lücken and Brunelli [50] use multi-objective evolutionary algorithms to select the best crop to plant for sustainable land use based on soil data. The costs of fertilizing and liming, cultivation, and the expected fluctuation of total return are among the optimization criteria. Their work provides a method based on multi-objective evolutionary algorithms to help select an optimal cultivation plan by taking into account five crop options and five objectives. It considers a multi-objective crop selection problem having the following objectives: minimize cost of fertilization and lime application, minimize total cost of production, maximize average return, maximize worst-case return, and minimize standard deviation of returns.

Priyadharshini et al. [68] propose a system to assist the farmers in crop selection by considering factors such as sowing season, soil, and geographical location. To recommend a suitable crop to the user, the proposed system considers environmental data such as rainfall and temperature, as well as soil characteristics such as soil type, pH value, and nutrient concentrations. The paper seeks to develop a robust model that can accurately estimate crop sustainability in a given state for a given soil type and meteorological circumstances, and make recommendations for the best crops in the area to benefit the farmer. They use the following datasets: (1) Yield Dataset: contains yield for 16 major crops grown across all the states in kg per hectare; (2) Cost of Cultivation Dataset: provides the cost of cultivation for each crop in rupees per hectare; (3) Modal price of crops: gives the market prices for these crops over a period of two months; (4) Price of Crops: gives the current market price of the crops in rupees per hectare; (5) Soil Nutrient Content Dataset: contains Nitrogen content, Phosphorous content, Potassium content, and average pH value; and (6) Rainfall Temperature Dataset: contains maximum and minimum rainfall, maximum and minimum temperature, and pH values. The proposed system is implemented with linear regression and neural network and compared with K-Nearest Neighbours (KNN), KNN with cross validation, Decision Tree, and Naive Bayes techniques.

Pudumalar et al. [69] propose a recommendation system through an ensemble model with majority voting technique using different machine learning tools to recommend a crop for the site specific parameters. Their dataset contains the soil specific attributes collected for Madurai district tested at soil testing lab, Madurai, Tamil Nadu, India. The crops considered in this paper include millets, groundnut, pulses, cotton, vegetables, banana, paddy, sorghum, sugarcane, and coriander. Dif-

ferent machine learning models give the best suitable crop, based on the given input, and then the crop which is selected by most of the models is recommended to the farmer.

Madhuri and Indiramma [53] introduce a recommendation system that uses an artificial neural network (ANN) with four layers for recommending the best suitable crop. The crops considered are maize, finger-millet, rice, and sugarcane. Two locations have been considered in this work, namely, Hadonahalli and Durgenahalli of Doddaballapur, Karnataka, India. Predefined conditions favourable for the selected crops are used to train the neural network.

Only representative papers on crop recommendation have been presented in this section. The reader is referred to papers cited in the above papers and in the paper by Shi, Wang, and Fang [77] to probe further.

3.2.2 Relevant Work in Crop Yield Prediction

There is abundant literature on the crop yield prediction problem and the reader is pointed to a comprehensive survey on crop yield prediction by Klompenburg, Kassahun, and Cata [85].

The paper by Sharma, Rai, and Krishnan [73] presents a method to predict crop yields from publicly available satellite imagery. The idea is to learn a deep neural network model for predicting the yield of wheat crop for tehsils (also called taluk – a local unit of administration covering a town and villages within a radius of 20 to 30 Km) in India. The method uses raw satellite imagery and notably, there is no need to extract any hand-crafted features or perform dimensionality reduction on the images. A byproduct of this work is to create a new dataset comprising a sequence of satellite images and the exact crop yield for the years 2001-2011 covering a total of 948 tehsils. This dataset is used to train and evaluate the proposed approach on tehsil level wheat predictions. The model outperforms existing methods by over 50%. Additional contextual information such as the location of farmlands, water bodies, and urban areas improves the yield estimates.

You et al. [93] develop a scalable, accurate, and low-cost technique for predicting agricultural yields using publicly available remote sensing data. They suggest a strategy based on deep learning ideas rather than the hand-crafted features commonly utilized in the remote sensing sector. They also present a dimensionality reduction strategy that enables them to train a convolutional neural network or a long-short term memory (LSTM) network and learn valuable features automatically even when labeled training data is scarce. They include a Gaussian Process component to model the data's spatio-temporal structure, thus increasing the accuracy.

The paper by Shahhosseini et al. [72] shows improvements in yield prediction accuracy when ML models are used in conjunction with additional side information obtained from a simulation cropping systems model. Soil water related variables (particularly growing season average drought stress and average depth to water ta-

ble), when supplied as input to the ML model, lead to better predictions. The study is based on data from the central US Corn Belt.

Khaki, Wang, and Archontolis [43] propose a deep learning framework for agricultural yield prediction based on environmental data. They use convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Using historical data, the proposed CNN-RNN model is used in conjunction with other popular methods such as random forest (RF), deep fully connected neural networks (DFNN), and LASSO (Least Absolute Shrinkage and Loss Operator) to forecast corn and soybean yield across the entire Corn Belt (including 13 states) in the United States for the years 2016, 2017, and 2018. The CNN-RNN model is created to capture the temporal dependence of environmental factors and the genetic improvement of seeds over time. Their model is a combination of CNNs, fully connected layers, and RNNs. They have predicted average yield for corn and soybean using an RNN layer that has a time length of 5 years.

An optimization model created by Awad [9] is able to estimate crop yield by increasing remote sensing data availability. The optimization model is created using a well-known algorithm, Trust-Region Methods for Nonlinear Minimization, that fits available data to an exponential equation. The experimental results prove the accuracy and reliability of the new model in estimating the crop yield in the case of missing remote sensing data. Using potato as the crop, the paper shows that the lack of remote sensing data is not a serious handicap for yield prediction.

Fan et al. [27] have introduced a novel graph-based recurrent neural network (GNN-RNN) for crop yield prediction, which incorporates both geographical and temporal structure. Their method is trained, validated, and tested on over 2000 counties from 41 states in the United States of America, covering years from 1981 to 2019. The paper compares 11 representative machine learning models, including GNN and GNN-RNN, on US county-level crop yields for corn and soybean. They are able to achieve an impressive improvement over existing models. The paper clearly brings out the importance of exploiting geospatial context in making yield predictions.

A pilot study by the International Crops Research Institute for the Semi-Arid Tropics [34] in India instructed farmers to delay planting by three weeks using predictive models based on climate and weather data. In the Indian state of Andhra Pradesh, they tested a new sowing application for farmers combined with a personalized village advisory dashboard, with the results showing a 30 percent greater average yield per hectare. During the pilot, farmers received 10 sowing advisories containing important information such as sowing recommendations, seed treatment, optimum sowing depth, preventive weed management, land preparation, farmyard manure application, harvesting recommendations, shade drying of harvested pods, and storage.

3.2.3 Positioning of our Work

In most research efforts in the current literature, only certain subsets of crops have been taken into account while making recommendations. Furthermore, the recommendation has been for a single crop to the farmer based on some objective function. The papers do not consider inter-cropping and do not recommend portfolios of multiple crops after assessing risk to the farmer. ACRE is a more generalized solution that can recommend a portfolio of crops from any given set of crops that the farmer is comfortable growing. ACRE computes the expected utility to the farmer taking into account profit and risk. ACRE also provides several options for utility functions. Most importantly, ACRE uses ideas from the financial investments literature to generate crop portfolio recommendations. In particular, ACRE uses Sharpe ratio [74, 75], a widely used financial portfolio selection metric, to obtain a ranking of crop portfolios. Section 3.3 presents an overview of the Sharpe ratio.

In terms of yield prediction, we have seen that many methods have been employed in the literature. We have used ensemble methods that combine different machine learning and deep learning models to produce accurate yield estimates.

3.3 Sharpe Ratio

The *reward-to-variability ratio*, popularly known as the Sharpe Ratio [74] [75] is a popular investment portfolio performance measure. It compares the return of an investment with its risk. This is done by comparing the performance of the investment, such as a portfolio, to a risk-free asset after adjusting for its risk. It is also used in other similar contexts where one has to choose between a number of options that may yield different returns. In ACRE, we have used the Sharpe ratio to quantify the risk-reward of potential crop-portfolios that may be suggested to the farmer. The investment portfolio in the agricultural context is the portfolio of crops grown by the farmer. The return from the portfolio is the revenue from sale of the harvest. The risk-free asset is the cost saved by the farmer by not growing anything at all. It can additionally include a fixed risk free rent begotten by leasing out the land. The risk here involves the risk in the yield, the duration of growing period and the price of the crops in the portfolio, which translates to a direct uncertainty in profit. The Sharpe Ratio is a single representative number that is drawn from the mean and variance of the differential with a standard benchmark portfolio, making it a very useful tool in portfolio recommendation. A major advantage of the Sharpe Ratio is that it is scale independent. When we choose the portfolio with the highest Sharpe Ratio we ensure that additional risk is compensated with additional reward. This ensures that an optimal portfolio of crops that maximizes profit while accounting for risk is found.

The Sharpe Ratio can be computed Ex-Post (that is, as late as possible when most information required is available) or as Ex-Ante (that is, as early as possible, working with probabilistic estimate of required information). In ACRE, we use the historical values of various input parameters to predict the future yield and price. These are

used to calculate the Ex-Ante Sharpe Ratio, which helps us make recommendations to the farmer depending on the farmer's risk preferences. The Ex-Ante Sharpe Ratio is calculated as the ratio of the mean and the standard deviation of the differential:

$$S \equiv \frac{\bar{d}}{\sigma_d}$$

where d is the differential between the returns generated by the recommended crop portfolio, P and the benchmark (e.g., minimum rental income), B. If the returns generated by portfolio P and the benchmark are represented by R_P and R_B respectively, then the differential is calculated as

$$\tilde{d} \equiv \tilde{R_P} - \tilde{R_B}.$$

We use tildes in the equation above to signify that the exact values of the returns, R_P and R_B might not be known before hand (Ex-Ante). The Sharpe Ratio can also be defined as the ratio of expected added return per unit of added risk to the benchmark.

In the absence of a benchmark, rather than taking a differential return ratio, one may also take a direct ratio of the return's mean and standard deviation. We have used this approach to compare the various crop portfolios.

3.4 Data Collection and Curation

We have used a variety of datasets in this project. The datasets cover yield, weather, and soil characteristics. All these data are drawn from Indian Government websites. The yield data is from [22], [23], and [35]; the price data is from [23]; and the weather data is from [26].

3.4.1 Yield Data

We are using the yield data provided by VDSA [35] and ICRISAT. This data is available for various crops from 1966 to 2011. We worked on Uttar Pradesh yield data since that was more complete when compared to the other states. The granularity of the data is district-wise.

The agmarknet website [23] provides arrival data, which has the production quantity of different crops in the neighbourhood of each mandi. The granularity of this data is mandi-wise in every district. The arrival data for different products unfortunately does not contain the area (in hectares) from which the products have arrived.

We have also used data for yield available at [22], and its granularity is district-wise. It includes state, district, area in hectares, and production in tonnes. It contains the average yield of a crop for a particular year.

However, this data suffers from many missing values. We have used standard techniques in data imputation [40] [92] to fill the missing values.

3.4.2 Weather Data

The weather data was downloaded from [26]. This data includes environmental parameters such as temperature, rainfall, humidity, and sunlight incident on the ground or crop. This weather data is available for every 0.25-degree change in latitude and longitude. We have taken the weather data, month-wise, for Kharif and Rabi seasons, separately.

3.4.3 Soil Data

We have taken the data available at [26] for the type of soil. This parameter is the texture (or classification) of soil used by the land surface scheme of the ECMWF (European Centre for Medium-Range Weather Forecasts) Integrated Forecasting System (IFS) to predict the water holding capacity of soil in soil moisture and runoff calculations. It is derived from the root zone data (30-100 cm below the surface) of the FAO/UNESCO Digital Soil Map of the World, DSMW (FAO, 2003), which exists at a resolution of 5' X 5' (about 10 Km). The soil types are 1: Coarse, 2: Medium, 3: Medium fine, 4: Fine, 5: Very fine, 6: Organic, and 7: Tropical organic. The Agro-Ecological Sub-Regions (AESR) index represents which type of land and which type of soil are present in a particular region based on latitude and longitude. This parameter does not vary in time.

3.5 Building Blocks of ACRE

The basic idea of ACRE is to make crop portfolio recommendations to farmers. Towards this end, we collect various input data from the farmer such as location and crop preferences, which could be based on their traditional or cultural practices.

3.5.1 Input Parameters

ACRE uses many input parameters, which can be classified into direct parameters and inferred parameters.

- **Direct Parameters**: The farmer can directly select these parameters. These include state and district, subset of crops, season (Kharif, Rabi, Whole Year), utility functions, irrigation facilities, crop rotation cycle, area (in hectares), human labour cost, etc. Each of these would have default options that will apply if the farmer is unable to select some of the parameters. In fact, some of these could even be inferred if the farmer is unable to select them.

- **Inferred Parameters**: These parameters are derived from direct parameters. These include crop price, temperature, rainfall, sunlight, humidity, soil type and nutrients in the soil. ACRE infers these parameters. Some of the direct parameters that are not selected by the farmer are also inferred by the system.

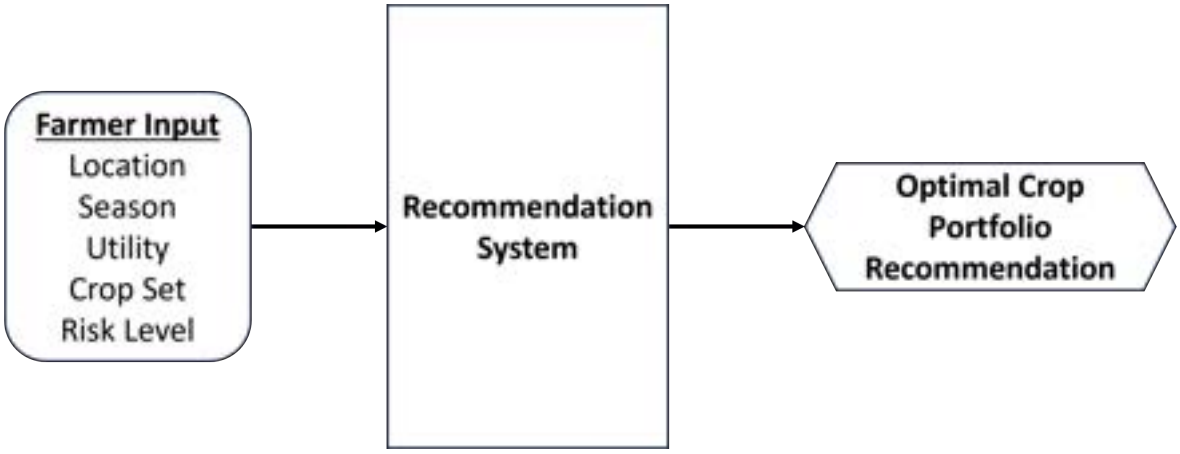


Figure 10: Farmer's view of ACRE.

ACRE takes all available inputs and builds models to predict yield, price, cost, and duration of crop growing period. The farmer can choose the type of utility (yield, profit, or risk). The models are used to compute the utility distribution (triangular distribution) based on the utility function chosen by the farmer. Based on the utilities, ACRE recommends a possible portfolio of crops, in accordance with the Sharpe ratio, considering expected return and risk associated with the portfolio. Figure 10 shows the farmer's viewpoint of ACRE. The farmer sees ACRE as a black box, which takes various inputs from the farmer and outputs a crop portfolio.

3.5.2 Utility Calculator

Figure 11 shows the basic architecture of the utility calculator. The utility calculator computes the utility distribution for the farmer based on the parameters provided and the utility function chosen by the farmer. The utility calculator outputs the expectation and standard deviation of the utilities, which are used in recommending a suitable portfolio of crops to the farmer.

3.5.2.1 Utility Functions

The farmer can choose from the following utility functions.

Profit (Considering Growing Period)

$$\text{Utility} = \frac{Y * P - C}{T} \tag{6}$$

Here, Y denotes yield in tonnes per hectare, P denotes crop price in rupees per tonne, C denotes the cost of cultivation per hectare, and T denotes duration of the

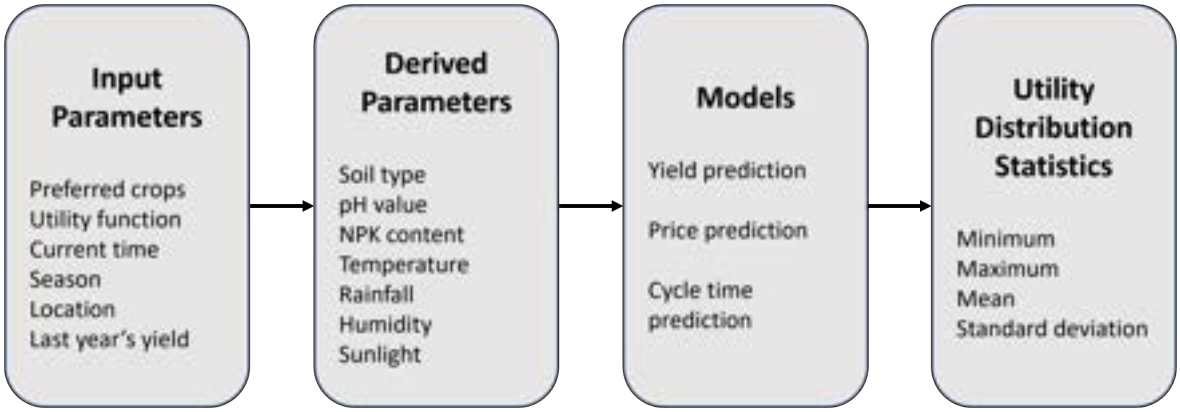


Figure 11: Architecture of the utility calculator

growing period for a crop in months. This utility function computes the profit for the farmer on a particular crop. It helps maximize the farmer's profit by considering the crop logistics cost and growing period on a monthly basis. This utility function assumes that farmers can grow any crop in any combination. We can also consider the variance in the yield and price while generating the profit utility distribution for various crops.

Profit (Without Considering Growing Period)

$$\text{Utility} = Y * P - C \tag{7}$$

This utility function does not consider the duration of the growing period, because, sometimes a farmer could grow only one crop in the entire year due to specific constraints like scarcity of resources, local customs, administrative policies, etc. In this case the actual duration of crop growing period would not matter as long as it is less than one year. One may also account for rate of interest to be more accurate.

Yield

This utility function simply considers the expected yield per hectare. This can be used for scenarios where only the yield of a particular crop is the determining factor. The farmer may not be interested in the profit from these crops and might be growing them for reasons such as personal consumption, adding nutrients to the soil, cattle feed, etc.

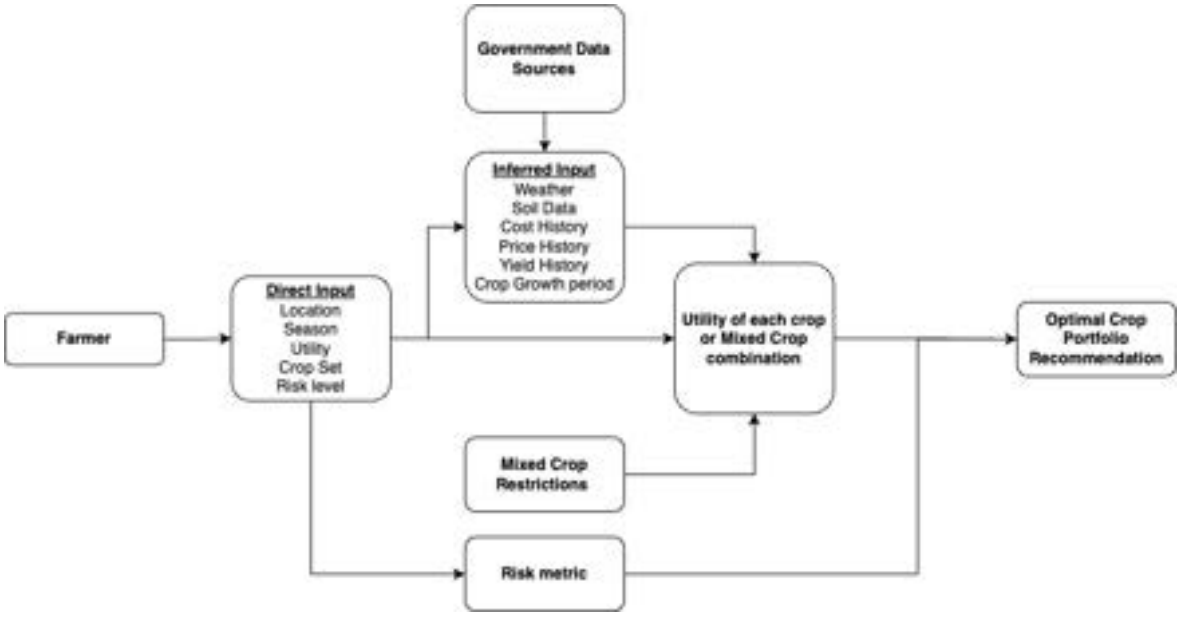


Figure 12: Data flow model

3.5.2.2 Crop Portfolio Recommendation

Based on the utility distribution and risk profile of the farmer, ACRE recommends a convex combination of crops to the farmer.[1] Cultural and traditional practices of the farmer can be taken into account in this portfolio of crops by eliciting from the farmer which crops the farmer is interested in growing. Also, we could take last year's yields for the farmer's crops and check whether or not the recommendation given by ACRE is aligned with the farmer's traditional practices or cultural preferences. Further, instead of providing a single recommendation, ACRE can provide a ranked list of multiple best portfolio recommendations and the farmer could choose based on any personal or cultural preferences.

Figure 12 contains a depiction of the data flow in ACRE. The direct inputs such as location of the farm, set of crops under consideration, etc., could be provided by the farmer (with expert assistance, if required). This is used to download the required inferred inputs such as local weather, price history in nearby mandis, etc. There are some crops that cannot be grown together and the mixed crop restrictions take such constraints into account. These constraints, the direct inputs, and the indirect inputs are used to calculate the utility to the farmer provided by various crop portfolios. Depending on the farmer's risk preferences, one or more of these crop portfolios are recommended to the farmer as the final output.

[1] A convex combination of numbers is a weighted sum of the numbers with weights in the range $[0, 1]$ and the weights summing to 1.

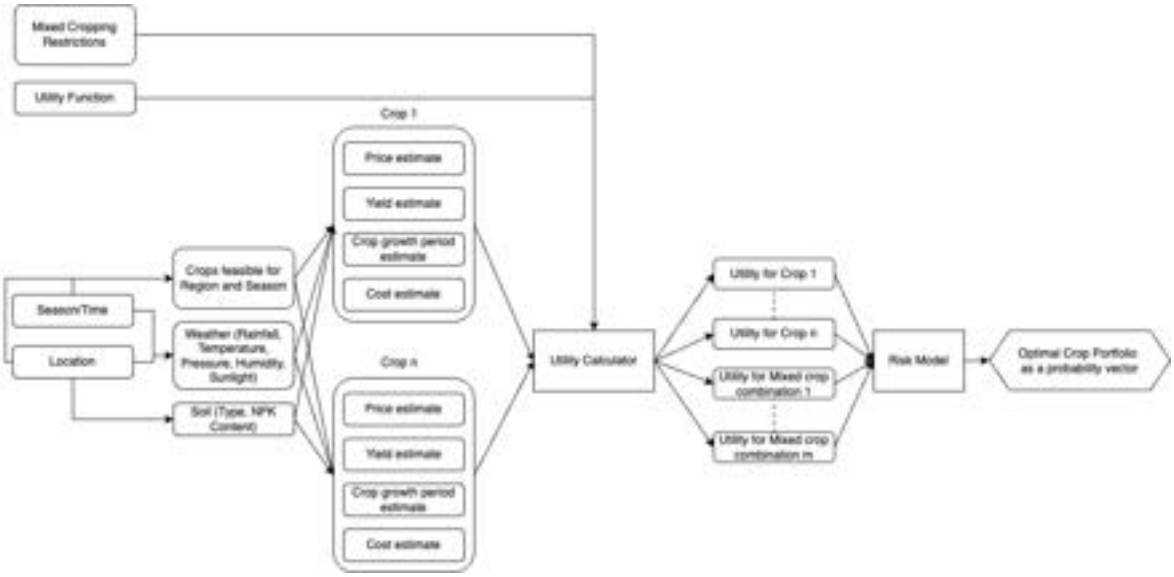


Figure 13: Detailed model of ACRE

The various building blocks of the recommendation system are shown in Figure 13. The utility calculator computes the utility distribution, from the yield and price predictions, for different feasible combinations of the crops provided as input. The risk model takes into account the utility distributions for all these specified crops and the risk profile of the farmer, and, recommends a suitable crop or combination of crops. The crop portfolio is a convex combination that specifies the recommended proportions in which the different crops should be grown.

3.6 Experiments and Results

In the experiments reported here, we have worked with profit maximization utility for farmers. To obtain the utility distribution of a particular crop, we need yield, price, cost, and growing period of a crop. Currently, we have considered the yield distribution and price distribution of a crop. Duration of growing period and costing models have not been considered in this work, but can be looked into for more accurate predictions. We have considered the price of a crop during the harvesting period of the particular crop. We have obtained the price distribution using the price variations in these months. For getting the yield distribution of a crop, we have used the ensemble technique in which we have trained multiple machine learning models and considered a 10% deviation from the predicted yield to get maximum and minimum values (based on our reading of historical data).

Regression Models	RMSE	Confidence Level for 5% Deviation	Confidence Level for 10% Deviation	R2 Score	Correlation Coefficient	Coefficient of Variation
Polynomial Regression	0.33	35.53	63.16	0.76	0.89	0.14
Random Forest	0.22	45.39	76.64	0.89	0.94	0.09
DNN	0.25	46.38	74.34	0.86	0.93	0.10
CNN-DNN	0.27	44.41	71.38	0.84	0.92	0.11
LSTM	0.21	44.69	72.67	0.89	0.95	0.09
CNN-LSTM	0.22	47.91	74.60	0.88	0.94	0.09

Table 5: Yield prediction results for Wheat for various regression models.

3.6.1 Crop Yield Prediction

For our pilot experiments, we have considered three crops - rice, maize, and groundnut - for the Kharif season and three crops - wheat, barley, and masur-dal - for the Rabi season. We have experimented with several standard machine learning and deep learning models [48] to predict yield. These include: Polynomial Regression, Random Forest Regression, Deep Neural Network Regression (DNN), Convolutional Neural Network on Deep Neural Network Regression (CNN-DNN), Long Short Term Memory Regression (LSTM), and Convolutional Neural Network on Long Short Term Memory Regression (CNN-LSTM) where a convolutional neural network is used for extracting the embeddings of the features. The features used for predicting the yield include latitude, longitude, land characteristics such as the AESR index, and weather parameters such as rainfall, temperature, sunlight, and humidity.

The results are measured and compared using different metrics. These metrics include Root Mean Square Error (RMSE), Confidence Level for 5% Deviation, Confidence Level for 10% Deviation, R^2 Score, Pearson Correlation Coefficient, and Coefficient of Variation. The confidence interval represents the percentage of the predicted values within this range of actual test values. The correlation coefficient represents how much the predicted and actual test values are related. The coefficient of Variation is the ratio of the standard deviation of predicted values and the expectation of actual test values.

For training the regression models, we have taken the yield data for Uttar Pradesh from 1982 to 2008, and for testing we have taken the data from 2009 to 2011. We have developed all the regression models mentioned above for all the crops separately. Wheat has the highest stability in crop yield over the years and the regression models expectedly predicted its yield very accurately.

Table 5 provides the metrics for all the regression models for wheat. The random forest, DNN, and CNN-LSTM regression models perform better than the other

Crops	RMSE	Confidence Level for 5% Deviation	Confidence Level for 10% Deviation	R2 Score	Correlation Coefficient	Coefficient of Variation
Barley	0.28	31.35	54.13	0.81	0.90	0.16
Wheat	0.22	45.39	76.64	0.89	0.94	0.09
Masur-Dal	0.14	33.93	57.14	0.49	0.71	0.17
Rice	0.23	40.79	66.78	0.81	0.90	0.12
Maize	0.29	18.92	39.53	0.55	0.74	0.24
Groundnut	0.29	20.65	33.77	0.48	0.70	0.44

Table 6: Yield prediction results for all crops using the random forest model.

regression models. The above models have been explored for other crops as well. The random forest regression model is found to outperform other regression models for almost all the crops. Table 6 provides the results for all crops using the random forest model. The estimates for groundnut and maize are not as good as other crops due to the inherent variability in the yields of these two crops.

3.6.1.1 Results Using Ensembling

Ensemble techniques integrate multiple base models to create a single best-fit predictive model. Ensemble methods, if properly deployed, provide higher predictive accuracy compared to any individual model. We have experimented with the ensemble technique in which we have taken all the combinations of random forest, DNN, CNN-DNN, and LSTM regression models for ensembling. For almost all the crops, ensemble techniques are found to perform better than the individual models. The ensemble technique which consists of random forest and DNN has the highest Confidence Levels for 5% and 10% Deviation and is found to perform better than other ensemble techniques for almost all the crops. Table 7 shows the different ensembles of several regression models and their performance for yield prediction of wheat. The ensembles (RF, DNN), (RF, CNN-DNN), and (RF, LSTM) perform better than the others. Table 8 shows the results of the ensemble (RF, DNN) for all the crops.

Models	RMSE	Confidence Level for 5% Deviation	Confidence Level for 10% Deviation	R2 Score	Correlation Coefficient	Coefficient of Variation
RF, DNN	0.13	66.15	89.74	0.96	0.98	0.06
RF, CNN-DNN	0.14	61.03	84.87	0.95	0.98	0.06
RF, LSTM	0.12	64.10	89.74	0.96	0.98	0.06
DNN, CNN-DNN	0.18	53.33	76.41	0.92	0.96	0.08
DNN, LSTM	0.15	53.85	82.56	0.94	0.97	0.07
CNN-DNN, LSTM	0.16	51.79	78.97	0.93	0.97	0.08
RF, DNN, CNN-DNN	0.14	59.74	84.87	0.95	0.97	0.07
RF, DNN, LSTM	0.13	62.56	88.21	0.96	0.98	0.06
RF, CNN-DNN, LSTM	0.13	61.28	85.38	0.95	0.98	0.06
DNN CNN-DNN, LSTM	0.16	53.85	80.00	0.94	0.97	0.07
RF, DNN, CNN-DNN, LSTM	0.14	59.74	86.41	0.95	0.98	0.06

Table 7: Yield prediction results for wheat: Ensemble technique.

Crops	RMSE	Confidence Level for 5% Deviation	Confidence Level for 10% Deviation	R2 Score	Correlation Coefficient	Coefficient of Variation
Barley	0.14	43.55	72.93	0.90	0.94	0.09
Wheat	0.13	66.15	89.74	0.96	0.98	0.06
Masur-Dal	0.14	32.73	53.57	0.43	0.66	0.17
Rice	0.13	52.82	84.35	0.92	0.96	0.07
Maize	0.14	33.34	62.69	0.78	0.90	0.13
Groundnut	0.28	12.45	30.81	0.49	0.70	0.43

Table 8: Yield prediction results for the ensemble technique consisting of RF and DNN for all the crops.

3.6.2 Results on Profit Utilities

As mentioned in Section 3.6.1, we have chosen three possible crops each for the Rabi and Kharif seasons. We have computed the profit utility distributions for the farmer consisting of average profit, maximum profit, minimum profit, and actual profit of the crops. We have also computed the coefficient of risk in profit (standard deviation

Crop	Season	Predicted Profit	Maximum Profit	Minimum Profit	Actual Profit	Variance (Risk)
Rice	Kharif	5717.89	6444.15	1592.598	6172.70	0.19
Maize	Kharif	5875.08	6312.06	5456.69	6423.62	**0.03**
Groundnut	Kharif	5444.50	**21104.88**	1431.89	5844.84	0.78
Barley	Rabi	4632.27	9139.18	3939.41	4579.91	0.25
Wheat	Rabi	6571.12	7430.05	5577.08	6743.59	0.06
Masur-Dal	Rabi	16456.15	**17218.37**	15492.85	18889.67	**0.02**

Table 9: Profit utility distribution for different crops for the year 2011.

Crop	Season	Predicted Profit	Maximum Profit	Minimum Profit	Actual Profit	Variance (Risk)
Rice	Kharif	5434.73	6118.97	4842.20	6402.49	0.05
Maize	Kharif	5449.76	5880.31	5036.29	6930.78	**0.03**
Groundnut	Kharif	4507.94	**20187.35**	494.427	4884.12	0.94
Barley	Rabi	3945.59	4449.53	3171.27	3876.45	0.07
Wheat	Rabi	6281.94	7427.60	5473.42	7030.19	0.06
Masur-Dal	Rabi	8067.66	**8668.42**	7366.42	19000.67	**0.03**

Table 10: Profit utility distribution for different crops for the year 2010.

in profit divided by the expected profit value for the crops).

The maximum profit and minimum profit are computed using the 10% deviation in predicted yield. We have chosen 10% based on our observation of ground truth data. Note that different crops have different durations of growing periods and for normalisation, the profit is computed in rupees per hectare per month. In the Kharif season, groundnut has the maximum profit but also shows the highest variation in profit, representing the highest risk. In contrast, maize has the lowest maximum profit but the lowest risk. Similarly, in the Rabi season, masur-dal shows the highest maximum profit and also the lowest risk and turns out to be the best crop for Rabi season.

Table 9 shows the profit utility distributions for the Kharif and Rabi seasons for the year 2011. Table 10 shows the profit utility distributions for the Kharif and Rabi seasons for the year 2010. Here, again, groundnut has the highest maximum profit as well as the highest risk. Maize has the lowest maximum profit but it is the most stable crop in the Kharif season. In the Rabi season, masur-dal is again showing the highest maximum profit with the lowest risk in comparison to all the other crops in the Rabi season.

Year	Season	Maximizing Profit	Minimizing Risk
2011	Kharif	Groundnut	Maize
2011	Rabi	Masur-Dal	Masur-Dal
2010	Kharif	Groundnut	Maize
2010	Rabi	Masur-Dal	Masur-Dal

Table 11: Individual crop recommendation based on maximum profit and minimum risk.

3.6.3 Recommendation of Individual Crops

Based on the profit utility distribution, the variance or risk associated with the profit, and the farmer's risk profile, ACRE recommends the best possible crop to the farmer for the particular season. Some farmers are risk-loving; they would prefer the crop which maximizes their profit irrespective of the risk associated with that crop. ACRE would typically suggest such farmers to grow groundnut in the Kharif season. On the other hand, there are also farmers who are very risk-averse; these farmers prefer stable, even if somewhat low, returns from the crop. A good suggestion for such farmers in the Kharif season would be to grow maize. For the Rabi season, masur-dal would be best suited to both these types of farmers. Table 11 shows single crop per season recommendations for the years 2010 and 2011 for Kharif and Rabi seasons. As the risk profile varies between these two extremes, the suggestions made by ACRE vary over different proportions of the crops to be sown.

3.6.4 Sharpe Ratio Based Crop Portfolio Recommendation

To make recommendations to farmers whose risk profile lies somewhere between the two extremes (risk loving and risk averse), ACRE analyzes the results of sowing different convex combinations of crops. The Sharpe ratio is used to calculate the suitability of various combinations or portfolios of crops for the farmers for a particular season.

To begin with, the farmer's field or entire growing area is divided into 10 equal parts or units. All possible combinations are considered for allocating the 10 units to different crops suitable for that season. For each combination, using the values of expected costing, yield, price, etc., the expected profit and the standard deviation in profit is calculated. Triangular distributions have been used for modeling the yield and price of crops. Using these values, the Sharpe ratios are calculated for all the crop combinations.

In ACRE, we use the Sharpe ratio on profit maximization and it is computed by dividing the expected profit value by the standard deviation in the profit. For easier comparison, the values of the Sharpe ratio for different portfolios have been scaled from 0 to 100, where 0 is the lowest and 100 is the highest Sharpe ratio for a crop profile. There are four possible extreme categories for a portfolio: (a) the highest

return and least risk (b) lowest return and least risk (c) highest return and highest risk (d) lowest return and highest risk. The categories (a) and (b) would have high Sharpe ratio values, while (c) and (d) would have low Sharpe ratio values. Clearly, the categories (a) and (b) are favourable for small and marginal farmers.

Table 12 shows the Sharpe ratios for different portfolios of crops for the Kharif season for the years 2009, 2010, and 2011. The highest Sharpe ratio is calculated in the years 2009 and 2010 for the portfolio of rice, maize, and groundnut with the convex combination of (0.4,0.6,0.0). The convex combination (0.0, 1.0, 0.0) yields the highest Sharpe ratio for the year 2011. Thus, in 2011, growing only maize yields the highest Sharpe ratio.

Table 13 shows the Sharpe ratios for different portfolios of crops for the Rabi season for the years 2009, 2010, and 2011. In the years 2009, 2010, and 2011 the highest Sharpe ratio for Rabi season is achieved respectively by the convex combinations (0.0, 0.4, 0.6), (0.2, 0.2, 0.6), and (0.0, 0.2, 0.8). It is evident that the optimal portfolios are a mixture of more than one crop. Instead of growing the crop which gives highest expected profit, it is better to take an optimal portfolio to reduce risk.

3.7 Administrative Policies and Socio-Cultural Factors

Thus far, only factors such as weather and soil conditions have been considered in ACRE. There could be other factors such as water, availability of labour and availability of farm equipment that could be taken into account, subject to data availability. There are, however, certain parameters, such as socio-cultural factors, that are hard to enumerate. Sometimes, even finding data relevant to these parameters is difficult. Some of these parameters are:

- Do the Government policies favour the recommended crops?
- Are there any incentives available for the recommended crops?
- Do the traditional beliefs and family conditions (and even superstitions) favour the recommended crops?
- What do the family members, friends, neighbours, and relatives think about the recommended crops?
- What were the previous experiences of the farmer with the recommended crops?

The best way to handle such socio-cultural factors would be to get a ranked list of multiple recommendations from ACRE and provide assistance to the farmer in choosing the best recommendation while keeping these factors in mind.

Portfolios **(Rice, Maize, Groundnut)**	Sharpe Ratio **2009**	Sharpe Ratio **2010**	Sharpe Ratio **2011**
(0.0, 0.0, 1.0)	0	0	0
(0.0, 0.2, 0.8)	0.84	0.89	1.07
(0.0, 0.4, 0.6)	2.23	2.37	2.85
(0.0, 0.6, 0.4)	5	5.33	6.4
(0.0, 0.8, 0.2)	13.13	14.03	16.83
(0.0, 1.0, 0.0)	73.42	84.91	100
(0.2, 0.0, 0.8)	1.15	0.89	1.03
(0.2, 0.2, 0.6)	2.64	2.37	2.79
(0.2, 0.4, 0.4)	5.63	5.33	6.27
(0.2, 0.6, 0.2)	14.44	14.09	16.22
(0.2, 0.8, 0.0)	93.29	99.71	66.77
(0.4, 0.0, 0.6)	3.05	2.36	2.68
(0.4, 0.2, 0.4)	6.24	5.32	5.99
(0.4, 0.4, 0.2)	15.63	14.05	14.61
(0.4, 0.6, 0.0)	100	100	36.86
(0.6, 0.0, 0.4)	6.83	5.29	5.59
(0.6, 0.2, 0.2)	16.69	13.93	12.57
(0.6, 0.4, 0.0)	90.7	85.42	23.74
(0.8, 0.0, 0.2)	17.61	13.72	10.56
(0.8, 0.2, 0.0)	77	68.47	16.81
(1.0, 0.0, 0.0)	65.32	54.93	12.57
(0.33, 0.33, 0.33)	7.9	7.09	8.06
(0, 0.5, 0.5)	3.34	3.56	4.27
(0.5, 0, 0.5)	4.57	3.54	3.92
(0.5, 0.5, 0)	96.66	93.76	29.13

Table 12: Sharpe ratio for different crop portfolios for Kharif season for the years 2009, 2010, and 2011.

Portfolios **(Barley, Wheat, Masur-Dal)**	Sharpe Ratio **2009**	Sharpe Ratio **2010**	Sharpe Ratio **2011**
(0.0, 0.0, 1.0)	87.31	69.7	93.62
(0.0, 0.2, 0.8)	99.79	85.99	100
(0.0, 0.4, 0.6)	100	79.6	96.69
(0.0, 0.6, 0.4)	84.22	52.75	77.65
(0.0, 0.8, 0.2)	63.96	24.64	51.58
(0.0, 1.0, 0.0)	47.19	3.01	29.32
(0.2, 0.0, 0.8)	56.3	82.06	76.12
(0.2, 0.2, 0.6)	56.96	100	73.83
(0.2, 0.4, 0.4)	53.14	80.63	63.07
(0.2, 0.6, 0.2)	45.64	43.2	45.31
(0.2, 0.8, 0.0)	36.86	13.12	26.87
(0.4, 0.0, 0.6)	26	84.52	41.96
(0.4, 0.2, 0.4)	24.38	94.23	35.07
(0.4, 0.4, 0.2)	22	62.16	26
(0.4, 0.6, 0.0)	19.04	24.4	15.95
(0.6, 0.0, 0.4)	12	65.51	20.32
(0.6, 0.2, 0.2)	10.63	60.7	14.38
(0.6, 0.4, 0.0)	9.06	30.44	7.95
(0.8, 0.0, 0.2)	4.55	31.77	7.8
(0.8, 0.2, 0.0)	3.46	20.93	3.09
(1.0, 0.0, 0.0)	0	0	0
(0.33, 0.33, 0.33)	30.32	83.56	39.28
(0, 0.5, 0.5)	93.54	67.43	88.94
(0.5, 0, 0.5)	17.78	77.88	29.61
(0.5, 0.5, 0)	13.28	28.79	11.46

Table 13: Sharpe ratio for different crop portfolios for Rabi season for the years 2009, 2010, and 2011.

3.8 Summary

ACRE generates several recommendations of portfolios of crops with a ranking of portfolios based on the Sharpe ratio. The farmer can choose to grow the crops in accordance with the portfolio that best matches the farmer's risk profile. ACRE recommends portfolios of crops by computing various utility values for farmers using data on crop characteristics such as yield, price, costing, and crop growing period. It also considers other factors such as soil and weather parameters. The recommendation system computes the estimated profit of each crop and recommends portfolios that maximize profit while accounting for the risk in yield, price, and duration of growing period of the crop.

ACRE provides a rigorous, data-driven back-end for designing farmer-friendly mobile applications for assisting farmers in choosing crops.

Chapter 2 provided an accurate model for crop price prediction that helps in, among other things, deciding when to harvest the crop. Using the predicted price, predicted yield, and several other factors, ACRE, discussed in this chapter, helps the farmers choose the optimal portfolio of crops with respect to their specific constraints and availability of resources. Having chosen the crops to be grown, the farmer now faces the next hurdle - procuring the necessary inputs required for sowing, irrigation, fighting pests, harvesting, etc.

4 PROSPER: A Marketplace for Selling Agricultural Produce to Maximize Social Welfare

A robust market mechanism is essential for the farmer to sell the produce. This market would effectively and efficiently connect the farmers (producers of agricultural produce) to the buyers of agricultural produce (consumers). We design a volume discount auction with a farmer collective as the seller and consumers (high volume or retail customers) as buyers. A farmer collective is a group of farmers coming together to gain from the power of aggregation. Our auction mechanism satisfies properties such as incentive compatibility, individual rationality, Nash social welfare maximization, and realistic business constraints. However, it is theoretically impossible to design an auction satisfying all of these properties. Therefore, we design deep learning networks that learn such an auction with minimal violation of the desired properties. The proposed auction, which we call PROSPER (PROtocol for Selling agricultural Produce for Enhanced Revenue), is superior in many ways to the classical VCG (Vickrey-Clarke-Groves) mechanism in terms of richness of properties satisfied and further outperforms other baseline auctions as well. We demonstrate our results for a realistic thought experiment on selling perishable vegetables.

4.1 Introduction

Small and marginal farmers generally own 2 to 5 acres of land and grow 1 or 2 crops at a time. In terms of selling the produce that they grow, they may not always be able to reach consumers and even if they do, they may not be able to negotiate a competitive price. Consumers can be classified into three categories based on the volumes they purchase. The consumers who purchase the largest volumes are companies such as Reliance Retail Ltd. and Supermarket Grocery Supplies Pvt. Ltd. (commonly known as BigBasket) who buy the produce and redistribute it through several outlets or online platforms. The second category of consumers comprises those who own standalone retail stores and generally sell to individual consumers in a certain locality. The third category consists of the individual customers themselves who generally purchase a smaller quantity that suffices to satisfy their household needs. In reality, it is not practical for an individual farmer to be matched directly with consumers as several logistical issues arise. Some examples of these are a farmer not being able to reliably satisfy all the needs of the matched consumers, large companies looking to purchase produce of a particular grade, etc. For individual farmers to reach individual consumers, several other issues may arise such as transportation logistics and the preference to sell all of the harvested crop. It is important to find a way of

selling the produce that maximizes the social welfare (combined utilities of farmers and consumers).

For these reasons, it is essential for an organisation or official intermediary to step in and solve the problem of selling produce to all potential consumers in a way that is competitive for the farmers as well as the consumers. Farmer Collectives (FCs) or Farmer Producer Organisations (FPOs) have been set up with this being one of the primary objectives. An FPO can be thought of as an entity that represents a large group of small and marginal farmers (generally between 100 to a few thousand farmers) of a certain geographical region. Grouping of small farmers into FPOs can greatly support small farmers in modern agricultural markets, through functions that small farmers cannot perform individually, such as supplying high-quality inputs closer to villages, aggregation and marketing of produce, facilitating access to infrastructure and technology, and providing training [58]. From now on, we use the acronym FC (rather than FPO) to maintain uniformity.

There exist more than 1500 FCs in Brazil, more than 7500 FCs in Germany, more than 63000 FCs in India, and more than 1700 FCs in USA. A recent report [84] mentions a staggering 1.2 million agricultural cooperatives across the globe today. Harnessing the potential of FCs through collective action to make agricultural marketing more attractive for farmers and consumers has high potential for huge impact.

Since scientifically designed auction mechanisms can promote honest behavior and healthy competition among buyers and sellers [55], we propose to develop a suitable auction for selling agricultural produce through intermediation by FCs. In this contribution, we propose an auction mechanism that can help small and marginal farmers sell their produce while obtaining the best price that also remains competitive to the consumers. Currently, auction mechanisms are not very popular for selling agricultural produce and we show in this chapter that an intermediary such as an FC will make auctions a viable option to use. Although platforms like the National Agricultural Market (eNAM) (https://www.enam.gov.in/web/) exist, their effective usage requires the farmers to be highly technologically proficient.

The mechanism proposed is a volume discount auction where bids are invited from consumers, possibly seeking discounts based on volumes. The primary objective of the auction is to maximize the social welfare, that is, maximize the combined utilities of all the players (FC and consumers). We call our mechanism PROSPER auction (Protocol for Optimal Selling of Produce for Enhanced Revenue). The PROSPER auction satisfies the following properties: incentive compatibility, individual rationality, social welfare maximisation, fairness, and business constraints. Such an auction, according to mechanism design theory [46, 59] is a theoretical impossibility, hence, we propose a deep learning based approach that minimizes a loss function that captures the violation from the desired properties. We show that our mechanism almost achieves the performance of a standard VCG (Vickrey-Clarke-Groves) mechanism even while satisfying several additional properties.

4.2 A Review of Relevant Work

Selling the agricultural output produced by the farmers (producers) to various consumers is an age old problem that has been attempted in a variety of ways. There have been several efforts to match farmers to consumers to maximize the utilities for the farmers and the consumers.

Several countries have experimented with online platforms for selling agricultural produce so as to benefit farmers and consumers. Examples include exchange platforms in Ethiopia, Kenya, Nairobi, and Uganda, the eNational Agriculture Market (eNAM) in India, and the Unified Market Platform (UMP) in the state of Karnataka, India [49]. The mechanisms used in these platforms are from a wide spectrum: warehouse-based negotiation, ascending auctions, first-price sealed-bid auctions, etc. [49]. It is not clear which mechanism is most beneficial for farmers [80]. Empirical evidence shows that the mechanisms may or may not be beneficial. As a case in point, the Ethiopia Commodity Exchange is very popular for coffee and sesame seeds and not popular for any other crops [33, 49]. In India, for example, only 14% of all farmers have registered on eNAM, and over half of these registered farmers are reported to not have benefited from the platform [49].

Kudu is an agricultural marketplace in Uganda [62] in which farmers and traders use their mobile devices to place bids (requests to buy) and asks (requests to sell) using a centralized nationwide database. Kudu identifies profitable trades, which are proposed to the participants. The platform also gathers price data and broadcasts it back to farmers and traders using SMS, drawing from a large set of national, regional, and local markets and providing a uniquely tailored information set to each user. An automatic matching algorithm takes as input a set of bids and asks and algorithmically proposes trades. Manual matching is also provided as an option. The matching algorithm runs three times a day. At run time, the algorithm simultaneously considers all bids and asks in the system and proposes a feasible set of trades that maximizes the total gains from trade, according to the scoring function. The matching algorithm uses a maximum weight matching algorithm in a bipartite graph. The method tries to find a fair price by setting the recommended price of a transaction to the minimum competitive (i.e., Walrasian) prices for the matching market. This makes truthful bidding a dominant strategy for buyers. The mechanism is not incentive compatible for sellers (farmers). The Kudu marketplace has been fairly successful but its widespread deployment has been hampered by many logitical issues.

Viswanadham, Chidananda, Narahari, and Dayama [89] provide a good overview of the working of the Indian Agricultural markets, which are called mandis. They propose that mandis should be transformed into electronic exchanges and present a mixed integer programming model that the electronic exchange needs to solve in an iterative way to optimally match buyers with sellers. They also present a stylized case study to illustrate the functioning of such an electronic exchange.

Prasanna Devi, Narahari, Viswanadham, Kiran, and Manivannan [21] propose a matching algorithm that innovatively uses the Gale Shapley algorithm [30]. The results obtained using this approach outperform the results obtained using an English auction based method. It is found that the proposed method produces stable matching, which is preference-strategy proof and it also reduces the need for number of rounds of allocation.

Levi, Rajan, Singhvi, and Zheng [49] introduce a behavior-centric, field-based, data-driven methodology to propose and design auction mechanisms to enhance the revenue of agricultural farmers in online agricultural platforms. They propose and implement a new two-stage auction for the agri platform for the Karnataka State in India for a major lentils market. Their implementation saw the participation of more than 10,000 small and marginal farmers in the market in three months time. The Karnataka state Government is set to select suitable commodities and markets to implement the two-stage auction on a larger scale.

In this chapter, our idea is to harness the potential of having a powerful intermediary to match farmers and consumers. Farmer collectives (FCs) which are quite popular in many countries serve this purpose in a natural way. Our idea is simple: A farmer collective will aggregate all the output from its farmer-members and sell the aggregated commodities to potential consumers using a suitable mechanism such as auctions. In particular, the mechanism we propose is a volume discount forward auction which we call PROSPER (Protocol for Selling Produce for Enhanced Revenue) auction.

4.3 PROSPER Auction

In initial interactions with an FC, we determined that PROSPER auctions would work best for perishables such as vegetables. Therefore, for our example, let us consider an FC whose farmers largely grow vegetables and hence, only vegetables are considered for sale by the FC. Further, for each vegetable that is grown, a separate auction is conducted, thus, for simplicity, let us consider the sale of one particular vegetable - tomato. While the description of the auction is given only for tomatoes, it can be replicated for any and all vegetables of a perishable nature. The consumers in the auction are either big companies like Reliance and BigBasket or standalone retail store owners. We do not include small individual customers as consumers in the auction marketplace since an auction is not required for the small individual customers. Instead, a simple e-commerce platform would be most convenient there.

The setup of the auction is as follows. Individual farmers approach the FC with the produce that they have grown. The FC accumulates all the produce grown and sorts it into categories based on their quality. For each farmer, the FC keeps track of the amount of tomato in each category. Following this accumulation, the FC will announce to all potential consumers, the total volume of tomato available in each grade. Each consumer is asked to place a bid along with the requirement for each

grade. A bid is the maximum price per unit at which the consumer is willing to buy that particular grade of tomato and the requirement is the number of units they want of that grade, for which they are willing to pay the price that they bid. Large scale consumers may usually obtain the daily market values based on current supply and demand in the market and submit their bids based on these values. Figure 1 captures the workflow in the PROSPER auction.

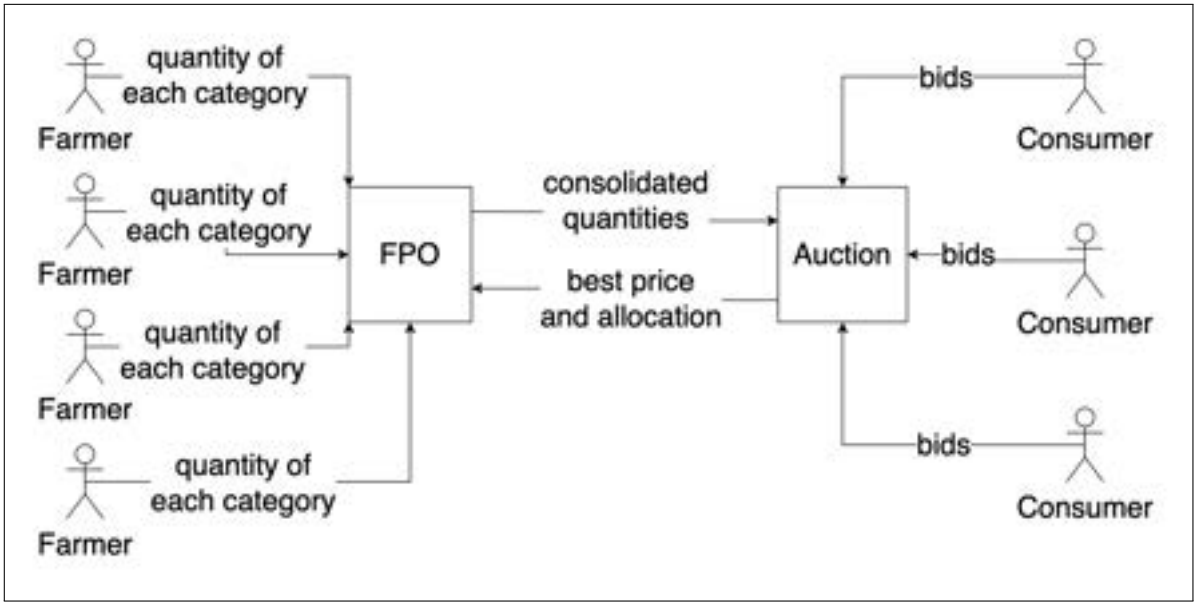


Figure 14: PROSPER auction for maximizing social welfare

4.3.1 An Example

Consider there to be two grades of tomato, Grade A and Grade B (A is superior to B). Let us assume that 100 Kg of each category has been brought by the farmers to be sold by the FC. Let us say there are two large consumers, C1 and C2. C1 may bid ₹5/Kg for 100 Kg of A and ₹4/Kg for 100 Kg of B. C2 may bid ₹5.5/Kg for 80 Kg of A and ₹3.5/Kg for 100 Kg of B.

The bids may also demand volume discounts. Consider the following as an example of volume discount bids by C1 for 100 Kg of category A: ₹5.25/Kg for up to 50 Kg and ₹5/Kg if the quantity is more than 50 Kg. This means that if 75 Kg of A is purchased by C1, we consider the bid price as (₹5/Kg * 75). There are several factors to be considered while determining who is allocated how much of the produce, and at what price. These will be elaborated in the rest of the chapter.

4.3.2 A Desiderata of Properties for PROSPER Auction

We bring out the most desirable properties for such an auction mechanism. Our analysis is based on our familiarity with and study of FCs and the dynamics of selling agricultural produce.

Incentive compatibility (IC)

Incentive compatibility ensures truthful bidding by the consumers and is a fundamental requirement of any auction mechanism. The most powerful version of IC is dominant strategy incentive compatibility (DSIC), which means bidding true values is best irrespective of the bids of the other players.

Individual rationality (IR)

Individual rationality ensures that the FC and consumers obtain non-negative utility by participating in the auction. The most powerful version of IR is ex-post IR, which implies that the utility to each participating player will be non-negative irrespective of the bids of the other players.

Social welfare maximization (SWM)

This implies maximizing the sum of utilities of the participants in the auction. In the PROSPER auction, the participants involved are the FC and the consumers. The utility of the FC captures the joint utility of all the farmers whose produce is aggregated by the FC. The utility of the FC or the consumer is defined as the amount of money they gain through the auction mechanism. Assume a consumer submits a truthful bid of ₹6/Kg for some grade of tomato. Through the auction mechanism, let us assume the consumer is asked to pay the FC ₹5/Kg, then, the consumer's utility is ₹1/Kg. If the average cost to produce 1 Kg of a certain grade of tomato is ₹3.5 and the FC is paid ₹5/Kg for this grade of tomato, the utility of the FC is ₹1.5/Kg. SWM aims to maximise the sum of utilities of the FC and all the consumers. The profit made by the FC are distributed to the farmers either based on the quantities of each grade of tomato brought by the individual farmers or based on the amount of stake the farmer holds in the FC.

Nash Social Welfare Maximization (NWM)

Although social welfare maximization satisfies the utilitarian rule, it allows a particular player's utility to take a hit if another player is receiving a commensurate or greater utility boost. A more egalitarian approach would be to maximize the product of utilities of all the participants in the auction. Hence, we focus on NSW maximization rather than SWM since NSW maximization is known to benefit both the seller (farmer collective on behalf of farmers) and the consumers [14].

Revenue maximization for farmers (FOPT)

This means the expected total revenue generated for the farmers is maximized. FOPT implies a farmer-friendly auction.

Cost minimization for consumers (COPT)

This means the expected total cost for the consumers is minimized. COPT implies a consumer-friendly auction.

Fairness (FAIR)

Fairness implies that the winning consumers are chosen in a fair way. An index of fairness would be envy-freeness – no consumer can increase her utility by adopting another consumer's outcome. If envy-freeness is not achievable, the next best option is envy minimization (which is what we pursue in this chapter).

Business constraints (BUS)

Satisfaction of business constraints refers to constraints such as having a minimum number of winning consumers (to avoid monopoly by a single or small number of consumers), a maximum number of winning consumers (to minimize logistics costs), a maximum fraction of business to be awarded to any consumer, etc.

Which set of Properties to Satisfy?

Ideally, we would like the PROSPER auction to satisfy IC, IR, NWM, FAIR, and BUS. The properties FOPT and COPT make the auction farmer friendly or consumer friendly, respectively. They have been included for the sake of completeness and for comparison purposes. Satisfying the set of properties (IC, IR, NWM, FAIR, BUS) is clearly a tall order. The classical Vickrey-Clarke-Groves (VCG) mechanisms [59] satisfy IC, IR, and SWM but may not always satisfy FAIR, NWM and BUS. Mechanism design theory [46, 59] is replete with impossibility theorems which make it clear that these properties cannot all be simultaneously satisfied. The ambitious goal of this chapter is to devise innovative auctions that achieve these properties with as little compromise as possible, using a deep learning approach.

4.3.3 Technical Platform for PROSPER Auction

For the above auction to be effectively implementable, a platform needs to exist for the farmers to have an effective communication with the FC. The most convenient way is to have a mobile app that is capable of a variety of tasks such as collecting crop data, land data, harvesting data, etc. from each farmer and creating a database that can be used by the farmers of the FC for multiple purposes.

In terms of the PROSPER auction, each farmer needs to input data such as the amount of each crop they have of each grade. The quality or grade can be determined by the FC at any stage before selling and the data can be entered accordingly. The FC can have a separate login for the PROSPER app, which will allow the FC to view the

consolidated list of produce as described already. The FC can then announce this list, after due consideration and discretion, to a variety of consumers. The consumers in turn will have a separate login for the PROSPER app and can place their bids along with the required quantities. Once the bids are placed, the FC can execute the auction after logging in. On successful completion of the auction execution, the FC will receive a recommendation for the quantity of produce, of each grade, to be allocated to each customer, together with the price that should be charged. The FC could commit an initial price to the farmers and if the FC suffers a loss, it is assumed that they have enough surplus to bear this loss. Any profits made by the FC through these sales are to be redistributed to the farmers based on the stakes that the farmers hold in the FC. A mechanism could be potentially devised which redistributes the profits back to the farmers based on the amount of produce they have contributed that has been sold. It is to be noted that the FC will have to factor in the service charges that the FC will incur because of the auction and other logistics.

Such a mobile app can also be extended to a variety of use cases such as weather prediction, crop price prediction, crop recommendation, and an auction for the procurement of items such as seeds, fertilisers, pesticides, etc. The FCs can also make use of this app to export produce and set up retail outlets of their own.

4.4 A Deep Learning Approach for PROSPER Auction

We propose a novel methodology in this chapter which is an extension of the works of [29, 24, 95, 11] which explore data-driven approaches to auction design. It is to be noted that all the mentioned mechanisms guarantee Individual Rationality. [24] maximizes the revenue of the seller in the auction while minimizing the violation of IC. [11] is a procurement auction (in the reverse direction of the auction proposed by [24] and us) which aims to minimize the cost to the buyer while minimizing properties such as IC, envy, and business constraints. The architecture we use for the PROSPER auction is similar to the ones used by [24] and [11], however, it is to be noted that we cannot use them as is, the reasons for which are as follows.

In the PROSPER auction, we are dealing with the selling of homogeneous units with volume discounts whereas the work of [24] considers auctions with additive (or unit-demand) valuations. Our volume discount auction setting is not additive and is not unit-demand either. To see this, observe that the value of $2x$ units with volume discounts is not the same as twice the value of x units. Additionally, the consumers in our setting wish to buy any number of units, unlike the unit-demand buyers considered by [24]. Auctions with additive valuations and unit-demand valuations make it possible to use allocation networks whose outputs are simply (stochastic) allocation matrices. Since our setting does not allow this, we produce an allocation tuple as output - with each element in the tuple being the allocation for the corresponding consumer. This complicates the computation of the payments and makes theoretical analysis non-trivial.

[11] considers the case of procurement auctions with the primary aim of minimizing the procurement cost. Ours is a forward auction whose primary aim is to maximise the NSW subject to IR while minimizing the violations of IC, envy, business constraints. Further, we explore the case where we maximise the revenue to the FC and the case where we maximise the revenue to the consumers, while minimizing IC violation.

4.4.1 The Volume Discount Auction Setting

In this section, we formally describe the volume discount auction setting. The notations used here are similar to the ones used by [29, 24, 95, 11]. There is a single seller (FC) who intends to sell m homogeneous units of a certain item to n consumers using a forward auction with volume discount bids. Let $\ell := \lfloor \frac{m}{k} \rfloor$ for some predefined k. The volume discount bidding is implemented as follows. First, each consumer i submits a volume discount bid in the form of a vector $b^{(i)} = (b_1^{(i)}, b_2^{(i)}, \cdots, b_k^{(i)})$ of k intervals. Then, given the vector of bids $b = (b^{(1)}, \cdots, b^{(n)})$ as input, the mechanism outputs allocation and payment vectors denoted by the tuple $\langle a(b), p(b) \rangle$. Here, $a(b) = (a_1(b), \cdots, a_n(b))$ denotes the allocation vector and $p(b) = (p_1(b), \cdots, p_n(b))$ denotes the payment vector with each $a_i(b)$ being the number of units sold to consumer i and $p_i(b)$ being the payment made by consumer i.

The consumers have their own private willingness to buy (WTB), which determines the maximum price at which the consumer is ready to buy. For consumer i, denote the WTB by $v^{(i)} = (v_1^{(i)}, ..., v_k^{(i)})$. These valuations are assumed to be drawn from some prior distribution $\mathcal{F}$. In our setting, $\mathcal{F}$ is common knowledge among all the consumers and the FC, whereas the realized vector of valuations $v^{(i)}$ is known privately only to the individual consumer i.

The utility for a consumer is defined as a function of her private valuations $v^{(i)}$, allocation a, and payment p, and is given by

$$u_i(v^{(i)}; b) = \sum_{j=1}^{a_i(b)} v^{(i)}_{\lceil j/\ell \rceil} - p_i(b) \tag{8}$$

We now recall some definitions. A mechanism is DSIC if no agent can gain utility by misrepresenting her valuations, regardless of the strategies adopted by the other agents. That is,

$$u_i(v^{(i)}; (v^{(i)}, b^{(-i)})) \geq u_i(v^{(i)}; b) \qquad \forall v, b, i. \tag{9}$$

An auction mechanism is called ex-post IR if every agent earns non-negative utility by participating in the auction. We assume that there is no participation/entry cost. Our proposed neural network architecture is designed to ensure the ex-post IR condition which is equivalent to

$$u_i(v^{(i)}; v) \geq 0 \qquad \forall v, i. \tag{10}$$

The goal is to maximise the nash social welfare, i.e., to maximise the product of utilities of all the players subject to ex-post IR and minimizing the violation of DSIC. Here, we consider the cumulation of all consumers to be one player and the FC to be the other player. Hence, the nash social welfare is the product of the utility of the FC and the sum of utilities of all the consumers. The utility of a consumer i is given in Eq. 8. Hence, the sum of the utilities of all consumers is given as

$$u_C = \sum_{i=1}^{n} u_i(v^{(i)}; b) \tag{11}$$

To define the utility of the FC, we must first define the term reserve price. The reserve price of the FC is the lowest price the FC is willing to sell the crop at, for example, this price may be the total cost of producing the crop. Hence, the utility of the FC can be given as

$$u_{FC} = \sum_{i=1}^{n} p_i(b) - m \cdot p_{res} \tag{12}$$

where p_{res} is the reserve price of a single unit of a particular grade of a particular crop.

Nash social welfare is therefore defined as

$$nsw = u_{FC} \cdot u_C \tag{13}$$

Note that the revenue to the FC is different from the utility of the FC. The revenue to the FC is defined as

$$revenue_{FC} = \sum_{i=1}^{n} p_i(b) \tag{14}$$

Our approach considers additional, practically motivated constraints (multiple units, volume discount bids, envy minimization, and business constraints). Following section 2.2.2 of [24], one can guarantee DSIC property by ensuring that the expected ex-post regret for every consumer, r_i, is 0. The expected ex-post regret of the PROSPER mechanism is defined as

$$r_i = \mathbb{E}_{v \sim \mathcal{F}}[\max_{b}[u_i(v^{(i)}; b) - u_i(v^{(i)}; (v^{(i)}, b^{(-i)}))]]. \tag{15}$$

The regret is computed empirically, which adequately approximates the real regret [24].

Remarks:

- The mechanism solicits volume discount bids from each consumer. These bids represent each consumer's valuation for a single unit from each 'lot' of units. There are k lots of (almost) equal size. The consumer i's bid $b_1^{(i)}$ is applied to all the goods if the seller sells at most one lot; i.e. ℓ goods. For the goods from the second lot, the bid $b_2^{(i)}$ is used, i.e., a price of $b_1^{(i)}$ per unit for the first ℓ goods and a price of $b_2^{(i)}$ for the remaining goods (up to ℓ). And so on.
- All the goods from a given lot are valued equally. Thus, if the FC sells more lots, their valuation per good decreases. In the agricultural domain, this corresponds to savings from the use of bulk transport, warehouse clearance, mass production, etc.
- In many practical settings, the value of k is determined endogenously. This value depends on factors such as packaging method, carton size, nature of the goods, etc. However, when k is a design parameter, it presents an interesting challenge in auction design. A larger value of k introduces more granularity, which is better for the buyer. But it also introduces complex bidding procedures possibly leading to a less effective implementation. We leave the study of this aspect as an interesting future work.

4.5 Deep Learning Based Formulation

We propose a neural network based formulation to satisfy IR while minimizing the violation of DSIC, along with some other desirable and practical constraints.

The goal is to minimize a composite loss function that consists of the following parts; the nash social welfare (to be maximised), the regret penalty, the envy penalty, the business penalty, and the Lagrangian term (as we use the method of differential multipliers [67]) for regret and envy. We also provide models that maximize the revenue of the FC and maximize the revenue of the consumers as the primary goal while minimizing IC violations.

The following are the various components of the loss function which are used in different combinations based on the requirement. They are each described in further detail through this section.

$$\texttt{revenue}_{\texttt{FC}} = \sum_{i=1}^{n} p_i(b)$$

$$\texttt{nsw} = u_{FC} \cdot \sum_{i=1}^{n} u_i(v^{(i)}; b)$$

$$\texttt{penalty}_{\texttt{regret}} = \rho_{\texttt{regret}} \sum_{i=1}^{n} \tilde{r}_i^2$$

$$\texttt{penalty}_{\texttt{envy}} = \rho_{\texttt{envy}} \sum_{i=1}^{n} e_i^2$$

When trying to minimize envy, the Lagrangian loss used is

$$\texttt{LagrangianLoss} = \sum_{i=1}^{n} \lambda^{(i)}_{\texttt{regret}} \tilde{r}_i + \lambda^{(i)}_{\texttt{envy}} e_i$$

and when we are not trying to minimize envy, the Lagrangian loss used is

$$\texttt{LagrangianLoss} = \sum_{i=1}^{n} \lambda^{(i)}_{\texttt{regret}} \tilde{r}_i$$

Here, $\tilde{r_i}$ is the empirical regret. We compute $\tilde{r_i}$ by using another optimizer over the bids, coming from the same distribution as $\mathcal{F}$, which maximizes the utility for agent i. To approximate the expectation over the distribution $\mathcal{F}$, we maximize the sample mean of regret over the batch.

Business Constraints

The seller may wish to impose various business constraints in the PROSPER auction - for example, the seller may require that at least 3 consumers buy at least 20% of the items each. We take care of this by adding a penalty term for violating various business constraints while training the network. For having a minimum of s consumers each with an allocation of at least a_{min}, the penalty would be

$$\texttt{penalty}_{\texttt{business}} = \begin{cases} 0 & a^{(s)} \geq a_{min} \\ \frac{\rho_{\texttt{business}}}{a^{(s)}} & a^{(s)} < a_{min} \end{cases} \tag{16}$$

where $a^{(s)}$ is the s^{th}-highest allocation. Other business constraints are also possible. For instance, no consumer is allocated more than 50% of the units. The loss function in this case also adds the penalty for violating business constraints.

4.5.1 Envy Minimization

In auctions, it is often desirable to have some additional fairness constraints, specifically, minimization of envy. Envy (or dissatisfaction) for an agent is defined as the maximum utility they could gain if they were given the allocation and payment of some other agent. So the envy for consumer i, given the valuation tuple $v = (v^{(1)}, ..., v^{(n)})$ is

$$e_i(v) = \max_{h \in [n]} [(p_h(b) - \sum_{j=1}^{a_h(b)} v^{(i)}_{\lceil j/\ell \rceil})] - u_i(v^{(i)}; v) \tag{17}$$

We minimize envy by adding a term for envy in our Lagrangian loss, along with an envy penalty.

4.5.2 Allocation Network and Payment Network

The model consists of two feed-forward networks - an allocation network and a payment network (See Fig. 15 and Fig. 16 for details). The input for both networks is the $n \times k$ matrix where the i^{th} row is the bid $b^{(i)}$ for supplier i, which is assumed to be equal to the valuation. The output of the allocation network is the allocation tuple described in Section 4.4.1. The allocation network uses the softmax function to ensure that the allocation tuple is a probability vector. This is multiplied by m to ensure that the allocations across the agents sum up to exactly m. The output of the payment network is a payment multiplier tuple, $\hat{p} = (\hat{p}_1, ..., \hat{p}_n)$, which, when multiplied by the total WTB of the allocation, gives the payment tuple, i.e.,

$$p_i(b) = \hat{p}_i(b) \sum_{j=1}^{a_i(b)} v^{(i)}_{\lceil j/\ell \rceil} \tag{18}$$

Each $\hat{p}_i$ is guaranteed to be within the range $[0, 1]$ in order to ensure IR. The network architecture takes care of this by using a sigmoid layer.

4.5.3 Training Procedure

In all our experiments, 5 layers with 80 to 100 neurons in each layer, were used for both the payment and allocation networks. The Adam optimizer was used for training the network weights and stochastic gradient descent was used to learn the Lagrangian parameters and minimize the loss function.

During the training phase, we performed nested optimizations. For one step of optimization over the network weights, we executed R steps of optimization over the bids to compute the empirical regret. Here, we used gradient ascent to maximise the value of empirical regret had the consumers bid mistruthfully. We gradually

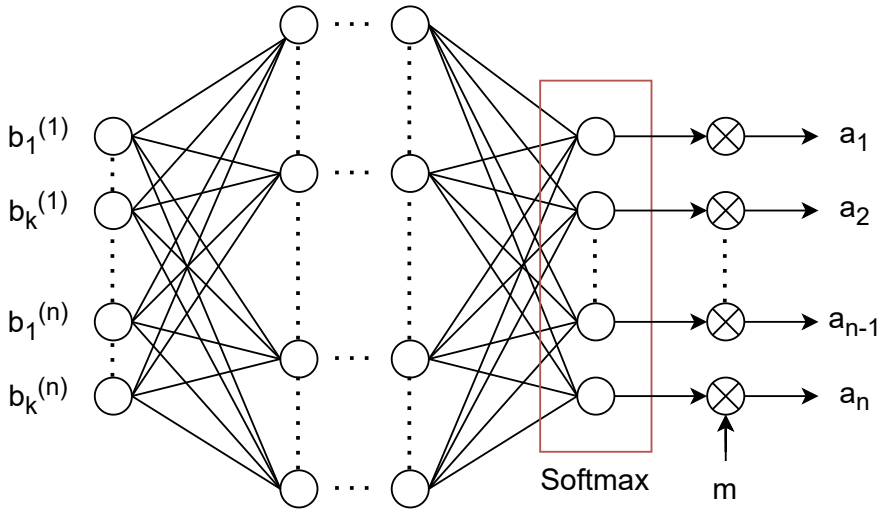


Figure 15: Allocation Network

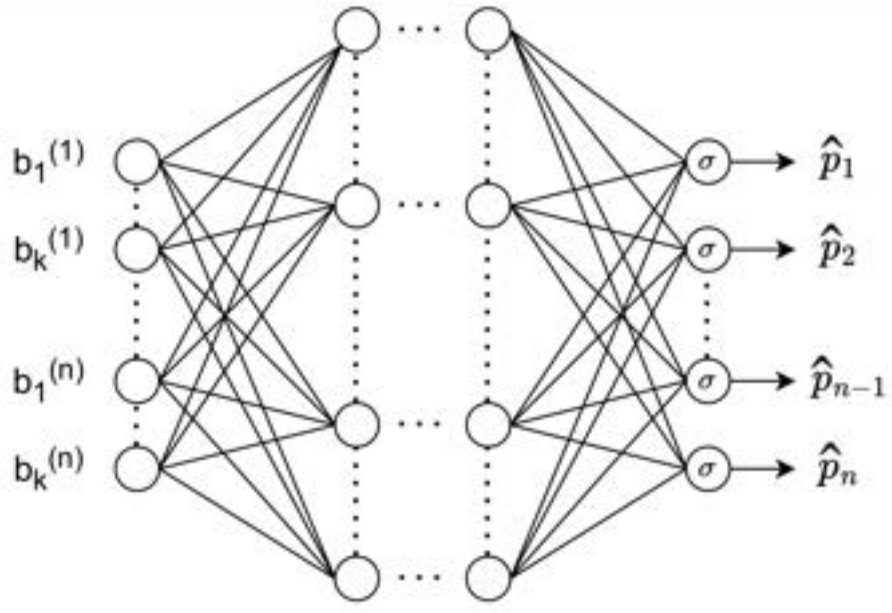


Figure 16: Payment Network

increase the learning rate as well as the penalties associated with regret and envy for each epoch. The above idea for empirical regret computation is the same as the one proposed in [24].

4.5.4 Some Notes on the Methodology

Versatility: The methodology presented in this chapter is versatile in the sense of its ability to model the minimization or maximization of a wide variety of performance metrics. For example, we can use this methodology to maximize the Nash social welfare subject to incentive compatibility, individual rationality, and business constraints. Nash social welfare maximization is widely known for its fairness properties [14].

Computational Complexity: The methodology proposed here essentially transforms a mechanism design problem into an optimization problem. So the question would arise as to why an efficient optimization procedure cannot be used to solve the problem, at least approximately. The constraints we deal with such as business constraints and envy minimization are nonlinear and linear approximations do not work well with such constraints. Moreover, the linear approximation will have an exponential number of variables. The deep learning technique has the advantage that a single substantial effort of training will amortize the computational complexity over a large number of experiments.

4.6 Experimental Results

In this section, we describe our experimentation with an FPO-mediated market for a perishable vegetable such as tomato or brinjal or chili pepper or seasonal fruits. As described in Section 1, the FPO gathers the produce to be sold from a number of registered farmers and invites bids from potential consumers (buyers). The consumers submit volume discount bids. For our experimentation, we consider 1000 units of the produce (for example, 1000 Kg of Chili pepper) and we assume that all the prices are in US $ (we will not mention these units henceforth). We assume five consumers – they could be major customers like retail chains or medium level stores like community grocery stores. It is important to satisfy the requirements of the consumers as far as possible. Further, it is important to benefit the consumers as well as the farmers to ensure sustainability of the market mechanism. We determine the valuation of a single unit of the produce for each consumer to be drawn from the uniform distribution U[350,450]. The values 350 and 450 are in cents, however, the rest of the chapter including Table 14 refers to the prices in US $. The consumers place volume discount bids in the following manner. These bids have been formed to cover all representative scenarios.

- Consumer 1 specifies a flat per unit price for the entire range 1 to 1000 units and does not bid with any discount.

- Consumer 2 bids with a discount of 5% if the purchase volume is between 501 and 1000 units. This discount is over the base price that applies for the purchase volume from 1 to 500.
- Consumer 3 bids with a discount of 3% for a volume between 301 and 600 units and 6% the volume is between 601 and 1000. The discount is over the base price that applies for the volume range 1-300.
- Consumer 4 bids with a discount of 2%, 4% and 6% for the volume ranges 251-500; 501-750; and 751-1000 units, respectively.
- Consumer 5 bids with a discount of 2%, 4%, 6% and 8% for the volume ranges 201-400; 401-600; 601-800; and 801-1000 units, respectively.

An important parameter to be selected is the reserve price. A lower reserve price hurts the farmer while a higher reserve price hurts the consumers. We have logically chosen 300 cents per unit after observing the results for different values of reserve prices. Our model can be used to pick a suitable value for reserve price that serves the farmers and consumers the best. Refer to Table 14 for all the numerical results. We compute the following performance metrics:

- Utility to the FPO, which is the total revenue to the FPO minus the valuation
- Utility to the consumers, which is the sum of the utilities of all consumers (utility of each consumer is valuation minus the payment made by the consumer)
- Social welfare (sum of utilities of the FPO and the consumers)
- Nash social welfare, which is the nth root of the product of all the utilities where $n-1$ is the number of consumers and FPO is regarded as the nth player
- Revenue to the FPO from the sale (which is also the total payments made by all the consumers)
- Raise in envy per unit utility, which indicates the raise in the envy whenever the utility raises by 1 unit. This measures the degree of envy freeness of the allocation.
- **Envy:** Maximum possible fractional (per unit) increase in utility of a customer if they were to be allotted another customer's allocation and payment.

The above six metrics correspond to the columns in the table. The values shown are averages over 6000 runs where in each run, we generate the valuations of the FPO and the consumers according to the uniform distribution U[350,450]. There are five rows corresponding to five different logically relevant auctions for this scenario: VCG

Table 14: Various utility measures (in US $) on sale of 1000 units using different Auction Mechanisms

	FPO Utility	Total Consumer Utility	Social Welfare	Nash Social Welfare	FPO Revenue	Envy
VCG	763	264	1027	193016	3763	0.1817
FPO Optimal	859	96	955	79292	3859	0.2274
Consumer Optimal	0.26	875	875.26	243	3000.26	0.1985
NSW Optimal	397	603	1000	243444	3397	0.1378
NSW Optimal with Envy Minimization	251	748	999	190990	3251	0.0129

Auction; FPO Optimal Auction; Consumer Optimal Auction; NSW Optimal Auction; and NSW Optimal Auction with Envy Minimization. The first three are standard auctions in the literature and we regard them as baseline auctions. While the VCG auction can be implemented using standard techniques, the rest of the auctions are implemented using the deep learning approach described in Section 3.

4.6.1 Baseline Auctions

VCG Auction

Our first baseline auction is the standard VCG mechanism which satisfies dominant strategy incentive compatibility, and individual rationality, and maximizes social welfare (sum of utilities of FPO and the consumers). The results show that the utility of the FPO is much higher than the combined utility of the consumers, thus the mechanism appears to be more FPO friendly (that is farmer friendly) and the consumers may not be excited by this auction mechanism.

FPO Optimal Auction

Our second baseline auction is to maximize the expected revenue of the FC subject to satisfying incentive compatibility and individual rationality via loss function minimization ($loss = -(revenue_{FC}) + penalty_{regret} + LagrangianLoss$). This auction produces the highest possible revenue to the FPO and the farmers will be delighted with such an auction (see Table 14). At the same time, such an auction may turn away consumers from participating in the auction.

Consumer Optimal Auction

Our third baseline auction is to maximize the revenue of the consumers (by minimizing the revenue to the FC) subject to satisfying incentive compatibility and individual rationality via loss function minimization ($loss = revenue_{FC} + penalty_{regret} + LagrangianLoss$). This auction produces the highest possible revenue to the con-

sumers (see Table 14). However, such an auction will prove to be unattractive to the farmers since the FPO revenue takes a beating.

4.6.2 NSW Auction

Nash social welfare has the attractive property of satisfying fairness of allocation as well as guaranteeing a high level of social welfare. We now consider an auction that maximizes NSW as well as satisfies certain business constraints ($loss = -(nsw) + penalty_{regret} + penalty_{business} + LagrangianLoss$). The business constraints we consider are in the form of allocating the produce to a minimum number of consumers and a maximum number of consumers. The motivation for having a constraint on minimum number of consumers is to spread the business among competing consumers providing equal opportunities to small and big consumers. The constraint on maximum number of consumers is to optimize logistics costs. Section 3 has brought out the learning process for a deep neural network that satisfies individual rationality, incentive compatibility, and NSW maximization, subject to business constraints. Let us call this auction NSW auction. We find from the table that the social welfare of the VCG auction is 1027 while that of the NSW auction is 1000; however, there is a perceptible difference in the utilities of the FPO and consumers between these two auctions. Recall that in the case of the VCG auction, the utilities were loaded in favour of the FPO. In the case of the NSW auction, the utilities are 397 for the FPO and 603 for the consumers which is more balanced than the values of 763 and 264, respectively, in the case of the VCG auction. The NSW for the NSW auction is 243,440 which is clearly superior to 193,016 of the VCG auction. The FPO revenue is 3763 in the case of the VCG auction while it is 3397 in the case of the NSW auction. The decreased revenue for the FPO is compensated by increased revenue for consumers. The envy of NSW auction is 0.1378 compared to 0.1817 of the VCG auction. To summarise, the NSW auction achieves a better balance in the utilities for the FPO and consumers, achieves better NSW, results in more revenue for the consumers, and leads to less envy, at the cost of some decrease in the revenue for the FPO. It is clear that the NSW auction will be more acceptable to both farmers and consumers than the VCG auction which has a bias towards the farmers.

NSW Auction with Envy Minimization

Though NSW balances social welfare maximization with fairness of allocation, in many situations, fairness may have to be accorded a high priority. In agricultural market situations, there is a critical need to ensure that nobody goes out of business due to aggressive bidding by power players. Envyfreeness is a way of ensuring this. An auction that tries to maximize NSW as well as minimize envy would be most desirable ($loss = -(nsw) + penalty_{regret} + penalty_{envy} + penalty_{business} + LagrangianLoss$). Table 14 shows the results for this new auction. As expected, this achieves a significantly small value of envy, however, at the cost of reduced utility and revenue to the

FPO and reduced value of NSW. The utility and revenue are higher than that of the NSW auction, which again tilts this auction in favour of the consumers more than the FPO. Overall, this auction has less desirable properties than the NSW auction.

4.7 Conclusions

In this chapter, we have presented a powerful mechanism for selling harvested produce of farmers to potential consumers through intermediation by farmer collectives using volume discount auctions. The designed auctions maximize the social welfare subject to fairness constraints and business constraints. Detailed experimentation on these auctions show the efficacy of the mechanisms designed. Our work provides clear evidence that the proposed mechanisms will be more attractive than existing traditional methods. Additionally, the proposed mechanisms also bring many other benefits such as ensuring farmer welfare, consumer delight, inducing honesty in bidding, utilizing scale economies, selecting deserving consumers, and the possibility to ensure fairness of allocation.

It is important to see how such mechanisms can be deployed. Deploying these on the ground does pose a few challenges. A key challenge is to convince the farmer collective, farmers, and consumers that these mechanisms will indeed work. This is directly connected to the explainability of these mechanisms, which is an important direction for future work.

5 AGRI-VAAHAN: An AIML Pipeline for Agricultural Data Analytics

AI and ML algorithms can help farmers and agri-businesses make better-informed decisions and optimize their operations. The use of AI and ML in agriculture offers the potential to drive innovation and improve outcomes in the industry by providing more accurate and efficient ways to predict and optimize crop growth, yield, and resource utilization. Agri-Vaahan provides a convenient AI-ML pipeline that streamlines and automates problem analysis and evaluation related to various agricultural data problems. Agri-Vaahan provides a platform that streamlines each stage of the model development process.

5.1 Agri-Vaahan Pipeline

The main objective of Agri-Vaahan is to design a pipeline that addresses various agricultural challenges using machine learning and deep learning techniques. The pipeline follows the standard machine learning stages but is tailored to the agriculture domain. Figure 17 provides an overview flowchart of Agri-Vaahan, and the following description outlines the requirements and brief use cases of each module:

1. **Input Data:** Data is sourced from two different channels: direct data and indirect data. The direct data source contains essential features such as crop type, mandi location attributes, crop price, arrival quantities, and more. The indirect data source includes relevant pointers like temperature, rainfall, humidity, and soil data. This module effectively provides a GUI solution for handling common data-related issues, such as redefining data types and combining multiple files.

2. **Feature Selection:** The effectiveness of machine learning algorithms relies on selecting the most relevant features. Statistical methods like correlation coefficients and predictive algorithms like decision trees are used to identify important features that have a strong relationship with the target variable[17]. Feature selection helps enhance algorithm performance.

3. **Data Pre-processing:** This stage involves transforming the raw dataset into the appropriate format for model input. It can be further divided into three sub-modules: Data Cleaning, Data Imputation, and Outlier Detection. Data Cleaning removes irrelevant or erroneous data rows and columns. Data Imputation techniques, described in Section 5.6, are applied to handle missing data. Outliers and invalid values, such as negative prices or rapid price changes, are detected and handled through removal or re-imputation.

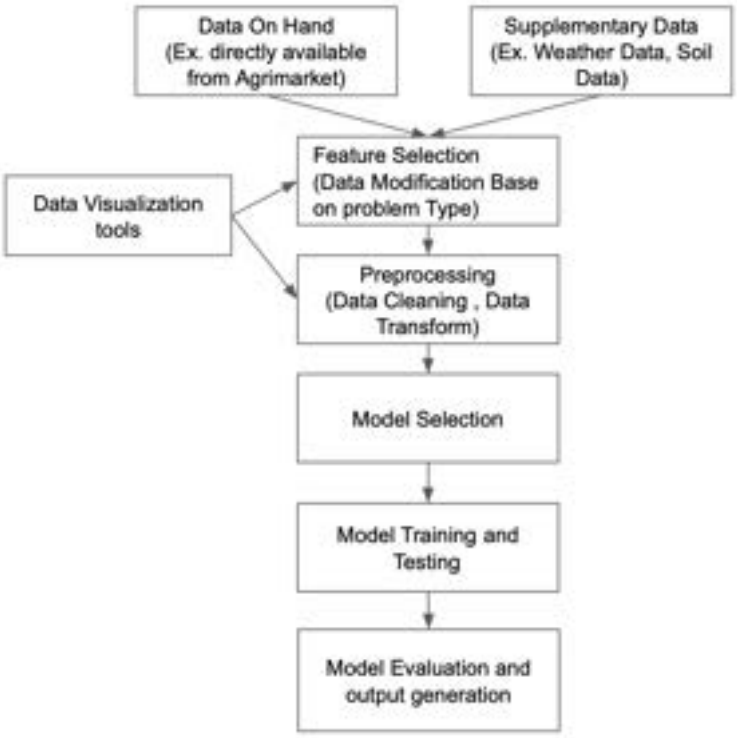


Figure 17: Problem Formulation Flowchart

4. **Data Visualisation:** This step aids in tuning hyperparameters for feature selection, model selection, and data pre-processing. Basic visualizations such as time series plots, value distribution plots, counts, and correlation matrices are employed [44].

5. **Model Selection:** This stage enables users to select multiple models and hyperparameters based on the specific problem they are addressing. Agri-Vaahan's primary focus is to provide state-of-the-art machine learning models and deep learning models, which can be used for solving most agriculture-related problems. Our particular focus is on developing forecasting and recommendation models.

6. **Model Training and Testing:** The data is split into training and testing sets, typically 75%-80% for training and the remaining for testing. The splitting algorithm depends on the data distribution and the problem at hand. Time series analysis employs a look-ahead strategy, while tabular datasets often use random splitting. Adaptive Synthetic Sampling is applied in scenarios with imbalanced target classes to avoid bias in evaluation. Agri-Vaahan offers conditional base datasplit [17] and a look-ahead strategy for the time series dataset.

7. **Output Generation and Model Evaluation:** This final step generates the output and evaluates the efficacy of the techniques employed in the pipeline. The results are displayed to the user, providing insights into the performance of the models and the accuracy of the predictions.

In addition to the machine learning pipeline, Agri-Vaahan also offers support for linear/non-linear objectives with linear constraints. This is particularly helpful for problems such as crop recommendation systems.

5.2 Agri-Vaahan Value Proposition

After conducting a comprehensive review of existing use cases and mapping AI and ML tools to various classes of data and agricultural problems, we have designed Agri-Vaahan, which offers the following unique contributions to the agriculture domain:

1. **Single stop framework for agriculture solutions:** Agri-Vaahan provides a comprehensive framework that caters to the needs of agriculture researchers by offering a wide range of state-of-the-art models for training and evaluation. It serves as a centralized platform where researchers can access various tools and models, making it easier to adapt and implement AI and ML solutions in the agriculture domain.

2. **Data Analytic tools:** Agri-Vaahan includes a variety of user-friendly data analytics tools that are essential in the prediction pipeline. These tools allow users to perform data exploration, visualization, and analysis, providing valuable insights and facilitating better decision-making. The platform offers visualization tools with a graphical user interface, enabling users to gain a deeper understanding of their datasets.

3. **User Immersive:** Agri-Vaahan is built as a web application using Streamlit, making it highly user-friendly and easily accessible. Users can install the framework locally and access it through a web interface. Additionally, Agri-Vaahan provides options to transform and analyze Python code snippets. It also provides optimization solving frameworks for advanced use cases like crop recommendation systems.

4. **Streamlines Model Training Testing:** Agri-Vaahan streamlines the process of problem analysis by providing a structured pipeline that guides users through each stage of the model development process. The framework integrates data preprocessing, feature selection, model training, and evaluation, making it easier for researchers to analyze agricultural data and develop accurate predictive models. Further, Agri-Vaahan offers users the ability to define flexible models. These sequential models can have different layers of CNN1d, GNN, LSTM, RNN, GRU, and Feed Forward.

Overall, Agri-Vaahan aims to simplify the adoption of AI and ML solutions in the agriculture domain by offering a user-friendly platform, comprehensive tools, and state-of-the-art models, ultimately fostering increased research and innovation in the field of agriculture.

5.3 Architecture of Agri-Vaahan Pipeline

This subsection provides an in-depth explanation of the architecture of the underlying Agri-Vaahan pipeline. The pipeline's flow, stages, and features are illustrated in Figure 18. While the majority of the sections align with the typical workflow of an AI-ML pipeline, there are certain unique aspects within Agri-Vaahan.

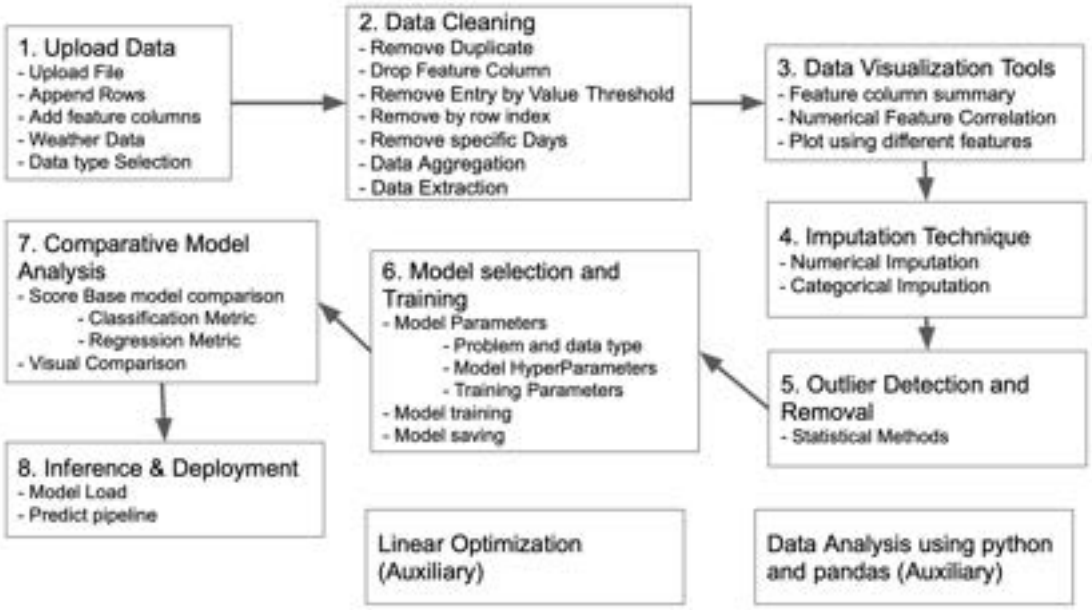


Figure 18: Architecture of Agri-Vaahan Pipeline

At the start of the pipeline, the "Upload Data" section serves as the entry point for input data. Following that, the "Data Preprocessing and Transformation" subsection encompasses the stages of data cleaning, imputation techniques, and outlier detection. These stages aim to ensure the quality and completeness of the dataset. The subsequent sections, "Modeling" and "Inference and Deployment," are similar to their counterparts in traditional AI-ML pipelines. These stages involve the selection and training of appropriate models based on the problem type and dataset. The trained models are then deployed for inference and utilized to generate predictions or recommendations. Additionally, within the data preprocessing stage, data visualization plays a crucial role in aiding method selection and hyperparameter tuning.

The “Data Visualization” section provides tools to visualize the output of the trained models, facilitating a deeper understanding of the data and model performance.

Overall, the architecture of Agri-Vaahan 2.0 incorporates essential components from standard AI-ML pipelines, while also introducing specific modules and functionalities tailored to the unique requirements of the agriculture domain.

5.4 Data Uploading in Agri-Vaahan

This section focuses on the data upload process, which involves combining and formatting datasets to meet the specific requirements of Agri-Vaahan. The provided methods are designed to handle typical data combination scenarios using a user-friendly graphical user interface (GUI). During dataset upload, users have the option to merge multiple files, add new feature columns, and represent the data in a desired format or type.

In the context of price prediction, the upload section offers tools to extract weather data from the existing dataset based on parameters such as “District” and “Date”. The following subsection provides detailed information about the data sources required for price prediction and crop recommendation tasks.

5.4.1 Price Prediction Dataset

- **Direct Data**: This is considered the main data source for training, containing the most useful features. An example from the agmarknet dataset [23] consists of features like crop type, mandi location attributes, crop price, arrival quantities, etc.
- **Indirect Data**: This includes indirect data like temperature, rainfall, humidity, soil data, etc.

Some good references for downloading agriculture-related data are provided below.

1. *agmarknet.gov.in* [23]: For the daily price data and arrival quantity in the market.
2. *cds.climate.copernicus.eu* [26]: For longitude and latitude based weather data such as rainfall, temperature, humidity, etc.
3. *data.gov.in* [22]: For datasets related to soil, temperature, crop yield, etc.

5.4.2 Crop Recommendation Data set

- **Direct Data**: This problem requires price prediction and yield prediction for calculating profit. The *agmarknet.gov.in* [23] provides all mandis price data and arrival data. However, the data needs to be curated for model training.

- **Indirect Data**: In order to calculate profits and distributions, additional data such as cost of cultivation, cost of production, and cycle time are required. Further details are presented in the case study presented in Chapter **??**.

5.5 Data Cleaning and Curation in Agri-Vaahan

In real-world datasets, it is common to encounter missing or inaccurate values, as well as features that may not be useful or require standardization. The data cleaning stage in Agri-Vaahan provides a GUI-based interface that offers various standard methods for data cleaning, as described in the subsections below.

5.5.1 Remove Duplicate Values

Duplicate values or data points in a dataset can introduce bias in model training, as the model may become skewed towards the distribution of duplicate instances. Agri-Vaahan provides two methods to address this issue:

1. Remove duplicate values without key: This method removes identical rows from the dataset without considering any key or trend identification. It ensures that the dataset contains only unique data points.

2. Remove duplicate value with keys: Consider the case where multiple entries of unique items might erroneously exist. In such a scenario, using the unique item as a key, we can drop the duplicates. For example: In the Market Daily dataset one could use “Market Name”, “Date”, and “Crop” as keys to weed out duplicates.

5.5.2 Drop Feature Column, Remove Entry by Value Threshold, Remove by Row Index, Remove Specific Days

This section gives a brief introduction of each feature requirement. As the data was extracted from multiple sources, it has lots of erroneous data and missing values.

1. **Drop feature column**: This feature allows the user to create a clean dataset by dropping unnecessary feature columns. It helps in removing redundant or irrelevant information from the dataset, focusing only on the essential features.

2. **NAN value removes by threshold:** In some cases, imputation techniques may not be effective when a large number of datapoints have missing values. It becomes crucial to remove datapoints that do not meet a minimum threshold of available features. The threshold can vary based on the availability of datapoints, ensuring the dataset retains a sufficient amount of information.

3. **Remove by row index**: Similar to dropping feature columns, this feature allows the user to select specific rows and remove them from the dataset. It can be useful when certain rows contain irrelevant or erroneous data that needs to be excluded from the analysis.

4. **Remove specific days**: In time series datasets, there may be instances where the data contains entries for invalid days. For example, in a market price dataset, there might be data points for Sundays, even though the market is closed on Sundays. This feature enables the removal of all entries corresponding to a specific day, ensuring the dataset reflects accurate and valid information.

5.5.3 Data Aggregation

In data analytics, working with large datasets or a high number of features can often be overwhelming and may not provide useful insights. The data aggregation tool in Agri-Vaahan allows users to reduce the dataset based on specific conditions, providing a more manageable and focused dataset for further analysis. There are two methods of data aggregation available, tailored for time-series and tabular datasets.

Time-Based Aggregation

This feature enables users to aggregate data based on a given condition column and a time token such as "n-day," "yearly," or "monthly." The "n-day" option allows users to specify a window for aggregation, such as weekly, every two days, or every four days. The aggregated data is generated by combining the selected features within each time window, providing a comprehensive view of the dataset. Additionally, this feature ensures the continuity of time-series data by adding time tokens that may be missing in the original dataset.

Condition-Based Aggregation

The condition-based aggregation feature allows users to analyze statistics of the dataset, such as mean, median, mode, and count, based on specified conditions. For example, in a crop yield dataset organized by district and year, users can perform analyses such as calculating the daily total yield arrival per state, the mean production per year, the average monthly production, and more. The parameters for this feature are determined based on the conditional columns and the desired function operations for each column. Currently, the tool offers numerical feature operations such as mean, median, mode, and sum.

5.5.4 Data Extraction

In the data preprocessing stage, it is often necessary to extract or drop specific data to create a targeted dataset for model training. The data extraction tool in Agri-Vaahan facilitates this process by allowing users to extract data based on specific criteria or drop irrelevant data. For example, if the focus is on analyzing data related to "Bangalore" and "wheat" from the AGMARKET dataset, users can utilize the tool's implementation of SQL query-based data extraction. Users can iteratively build query inputs by adding subparts of the query using logical operators such as "AND" and

"OR". By previewing the extracted data, users can make informed decisions to either drop rows or extract the data for further analysis.

5.6 Imputation Techniques in Agri-Vaahan

This section provides the tools to analyze and impute missing values using various standard techniques, such as linear interpolation, spline interpolation, etc. An additional feature has been implemented to drop or impute values that fall below the missing threshold. The section is divided into three subsections: Overview of Missing Value Based on Condition, Numerical Feature Imputation, and Categorical Feature Imputation

Based on an extensive review of AI and ML techniques [25, 37], we have identified a list of data imputation techniques, which are mentioned below.

1. **Mean/Median/Mode Imputation:** This involves replacing the missing value of the feature or variable with a measure of the central tendency of the feature. This has the advantage of maintaining the measure of the central tendency of the sample. If the missing rate is high, this method might introduce a bias and hence, should be avoided.

2. **Substitute:** Here, a specific/external value is used as an imputation. This method is dependent on an external subject and its source validation.

3. **Hot-deck Imputation:** In this technique, a randomly chosen value from an individual in the sample that shares similar characteristics with other variables is used. It ensures identical characteristics and randomness attribute values of missing data points. Due to the unique sampling method, evaluation error has the same standard error.

4. **Cold-deck Imputation:** This method involves choosing a value systematically from an individual variable that shares similar values with other variables. This method differs from hot deck imputation as it removes the random variation.

5. **Regression Imputation:** This technique uses regression on the available data to impute the missing data.

6. **Stochastic Regression Imputation:** Similar to the Regression Imputation technique, the value is predicted based on the available data and some additional noise/random residue.

In addition to the methods mentioned above, prediction-based algorithms are also used for imputation [6]. For numerical time series datasets, the methods described below are very helpful:

1. **Back Fill and Forward Fill:** Useful where the missing values are few and far in between, forward fill imputes values based on the previous values while backfill imputes values based on the values that follow.

2. **Spline Interpolation:** This imputation technique considers intervals around the missing points and approximates a polynomial interpolation equation around them. For example, a cubic spline fits a 3rd-order equation to find the missing value [79].

3. **Time-series based Prediction Algorithm:** In between time series missing point allow us to create a window around the missing point Let $x_{k-t}, ...x_{k-1}, x_k, x_{k+1}, ..x_{k+t}$ is time series on subpart here x_k is missing point so,$x_{k-t}, ...x_{k-1}$ past values and $x_{k+1}, ..x_{k+t}$ future value can act as feature to train model with this dataset this allows to capture how series evolve around time leading to batter imputation. The model/imputer can be used as K-Nearest Neighbour or Random Forest. [6, 81]

4. **Matrix Factorization base imputation(Soft Impute):** This method is applied where multiple similar characteristic time-series data points are available. Samples from the nearest market(mandi) are considered as data points for predicting the missing values. A matrix is constructed using all the individual time series. It is then decomposed and reconstructed using the matrix's highest energy component [7].

We have explored different methods specific to the agriculture domain. Currently, Agri-Vaahan offers mean imputation, substitution methods, and polynomial interpolation techniques.

5.6.1 Outlier Detection for Time Series Dataset

Outlier detection is a crucial step in identifying and handling anomalies in a dataset. When it comes to price prediction in the agricultural domain, outliers can be defined as values that deviate significantly from the expected behavior, such as negative prices, sudden drastic changes, decimal mistakes, and other irregularities. Agri-Vaahan incorporates a window-based statistical method combined with re-imputation techniques to identify and address common outliers in time series datasets.

The method employed by Agri-Vaahan takes advantage of the characteristic behavior of prices, where sudden changes are less likely to occur within smaller windows of time. By setting a window size and a scaling factor as hyperparameters, the method establishes lower and upper bounds for price values within the window. Any values outside this range are considered outliers. The upper bound is calculated as $Q_3 + IQR \times$ scale, and the lower bound is calculated as $Q_1 - IQR \times$ scale, where

Q_1 and Q_3 are the first and third quartile values, and IQR is the interquartile range. The scaling factor provides an additional margin based on user input.

However, it's important to note that this method may not be effective when dealing with datasets that have a high number of missing values, which are imputed using techniques like BackFill, ForwardFill, or linear methods. In such cases, the relationships between the missing values and their nearest neighbors in the window may result in a constant or small variation, making outlier detection challenging.

In the case of yield or arrival datasets, where the data points represent quantities available per day, the occurrence of anomalies is more complex. The discontinuous nature of these phenomena makes it difficult to identify outliers using traditional statistical methods designed for continuous data. In such a scenario, the user can use statistical analysis, such as mean and standard deviation on global data using data aggregation, that can be useful base in query base extraction or dropping rows.

5.7 Data Visualization in Agri-Vaahan

Data visualization is a crucial task in determining data preprocessing techniques, selecting imputation techniques, and choosing models. It provides users with an overview of data analysis and aids in decision-making. In this section, we will discuss visualization techniques in the context of price prediction and crop recommendation systems. These techniques are also applicable to similar types of problems. Agri-Vaahan offers various visualization tools, including feature value analysis, correlation plots, custom visualization plots based on conditions, and plots with common parameters.

5.7.1 Feature Analysis

Feature Analysis provides a basic overview of feature distribution and value counts. The first section of Feature Analysis displays the count of missing values, which helps identify the need for further imputation techniques or potential feature modifications. For example, in crop price data, there may be missing values for various crop varieties and grades. By combining the variety and grade information, users can reduce the number of missing values. The second section of Feature Analysis provides value counts and distributions of features.

5.7.2 Correlation Plot for Numerical Features

Correlation plots visualize the relationships among numerical features. Users can select at least two numerical features to create a correlation matrix. This matrix is helpful for identifying correlations between different features. A "Select all" checkbox is provided to calculate correlations among all numerical features. Users also have the option to remove specific features from the correlation matrix.

5.7.3 Visualization Plot (Conditional / Basic Parameters)

Visualization plots play a crucial role in data analysis and decision-making. They provide a visual representation of data that can be easily perceived and analyzed by humans. Agri-Vaahan offers two types of visualization plots:

1. Conditional Plot: This feature allows users to create abstract plots based on specific conditions. The first data that satisfies the condition column is extracted, and the remaining plot parameters are applied. This helps users visualize data based on specific criteria.

2. Basic Plot (Using Plotly Library): This feature provides an easy-to-use method for creating plots, especially using categorical values or indicators. For example, crop price plots can be grouped by district, grade, or variety. The Plotly library offers a wide range of options for plotting complex data in a simple and intuitive manner.

Both these methods provide various types of plots, such as bar charts, line charts, histograms, scatter plots, etc. However, using these features requires a basic understanding of the parameters required to create the desired charts. In case any errors or warnings occur during the plotting process, users can refer to the provided use cases and documentation on the reference site [2] for troubleshooting and further guidance.

5.8 Model Selection in Agri-Vaahan

Agri-Vaahan provides AI-ML solutions for problem categories such as classification and regression, specifically tailored for time series and tabular datasets. Our focus is primarily on time series forecasting applications. In this section, we present a brief overview of the models implemented in Agri-Vaahan. Further information on common parameters applied and the GUI reference can be found in the user guide supplied in the annexure.

5.8.1 Machine Learning Models (For Regression and/or Classification)

In this subsection, we describe the basic machine learning models that are commonly used in agricultural machine learning problems, including detection and forecasting [41, 54].

1. **K-Nearest Neighbourhood**: The KNN algorithm classifies new data points based on the majority class of their k nearest neighbors. It can be used for both regression and classification tasks. The equation is represented below:

$$\textit{Euclidean distance} = \sqrt{\sum_{N}(y-x)^2}\textit{ , where N is vector dimension} \tag{19}$$

2. **Linear, Ridge, Lasso, and elastic Regressor**: Linear Regression is a widely used model for regression tasks. It assumes a linear relationship between the input features and the target variable. The model is trained to find the best-fit line that minimizes the sum of squared errors.

   The linear regression equation can be represented as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{20}$$

   The Ridge regression equation can be represented as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \lambda \sum_{i=1}^{n} \beta_i^2 \tag{21}$$

   The Lasso regression equation can be represented as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \lambda_2 \sum_{i=1}^{n} |\beta_i| \tag{22}$$

   The Elastic regression equation can be represented as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \lambda_1 \sum_{i=1}^{n} |\beta_i| + \lambda_2 \sum_{i=1}^{n} \beta_i^2 \tag{23}$$

   where y is the target variable, $x_1, x_2, ..., x_n$ are the input features, and $\beta_0, \beta_1, \beta_2, ..., \beta_n$ are the coefficients. The model estimates the values of these coefficients during the training process to create a linear equation that predicts the target variable based on the input features. Linear regression is suitable when the relationship between the features and the target variable is linear, and it provides interpretable coefficients that indicate the impact of each feature on the target variable. As shown in Equations 23, 22, and 21, an additional term is added to regularize the trainable parameter where the Ridge penalty provides equal importance to each feature. Lasso penalty makes model parameter(β_i) near to zero for the independent feature with respect to the target feature. λ_1 and λ_2, that control the amount of Ridge and Lasso penalties applied, respectively. This allows for a balance between Ridge and Lasso regularization techniques.

3. **Logistic Regression:** Logistic regression is a classification model used when the target variable is binary or categorical. It estimates the probability of a data point belonging to a specific class. The logistic regression equation for the binary class is given by:

$$P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)}} \tag{24}$$

The Logistic regression for multi-class classification with soft-max function is given by:

$$P(y=i) = \frac{e^{(\beta_0^{(i)} + \beta_1^{(i)} x_1 + \beta_2^{(i)} x_2 + ... + \beta_n^{(i)} x_n)}}{\sum_{j=1}^{K} e^{(\beta_0^{(j)} + \beta_1^{(j)} x_1 + \beta_2^{(j)} x_2 + ... + \beta_n^{(j)} x_n)}} \tag{25}$$

where $P(y=i)$ represents the probability of the data point belonging to class i, and K is the number of classes.

Above equation 25 24 can be further improved by introducing lasso and ridge penalty as discussed in equation 22 21

4. **Random Forest:** Random Forest is a powerful machine learning model that combines an ensemble of decision trees to make predictions. Each decision tree in the Random Forest independently suggests a class prediction based on a subset of the features and data. The final prediction of the Random Forest model is determined by aggregating the predictions of all the individual trees through a voting process. The class with the highest number of votes is selected as the overall prediction.

   For regression tasks, Random Forest takes the mean of all the decision tree predictions to provide a continuous output. This averaging process helps in reducing the variance and improving the overall prediction accuracy.

5. **Support Vector Machine:** SVM is a powerful algorithm used for both classification and regression tasks. It aims to find the optimal hyperplane that separates data points of different classes while maximizing the margin. The mathematical objective function of SVM for classification is:

$$minimize : \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i \tag{26}$$

$$\textit{subject to: } y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0,\ i = 1, 2, ..., n \tag{27}$$

   where $\mathbf{w}$ represents the weight vector, b is the bias term, ξ_i is the slack variable that allows for misclassification, y_i is the true class label of sample $\mathbf{x}_i$, and C is the regularization parameter that controls the trade-off between maximizing the margin and allowing misclassifications.

6. **XGBoost:** XGBoost stands for "Extreme Gradient Boosting," and it is based on the gradient boosting framework. It uses an ensemble of weak prediction models, typically decision trees, to iteratively improve the overall predictive power. Each weak model is trained on the residuals of the previous models, focusing on the data points that were not predicted properly.

   One of the key features of XGBoost is its ability to handle missing data by automatically learning how to best handle those missing values during the training

process. It also incorporates regularization techniques to prevent over-fitting, such as controlling the complexity of the individual weak models and adding penalty terms to the loss function.

In addition to all models implemented for time series forecast tasks, creating dataset by using past/lag values as feature columns while predicting nth future value. In our analysis, as a temporal component is increased, model performance diverges largely. This is shown in Chapter 4, Price forecasting studies.

5.8.2 Deep Learning Models and Frameworks

In Agri-Vaahan, deep learning models are implemented based on the data type and problem type. Currently, there are two categories of deep learning models: tabular dataset models and time series forecasting models. For tabular datasets, Agri-Vaahan offers a simple feed-forward neural network model. Users have the flexibility to determine the number of layers and parameters for the model based on their requirements. For time series forecasting, Agri-Vaahan provides several state-of-the-art models such as Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), and Gated Recurrent Unit (GRU). These models are designed specifically to handle temporal data and capture sequential patterns. In addition, an advanced model called RNN-GNN (Recurrent Neural Network-Graph Neural Network) has been developed specifically for price prediction tasks. This model combines the power of recurrent neural networks with graph neural networks to leverage both temporal and relational information in the data [12].

The architecture and workings of each model are described in detail in the following subsections, providing insights into their implementation and usage within the Agri-Vaahan pipeline.

Recurrent Neural Network (RNN)

Recurrent Neural Network (RNN) is a type of neural network that is designed to process sequential data by considering the temporal dependencies between data points. It is especially useful for tasks such as time series analysis, natural language processing, speech recognition, and machine translation. In the time series forecasting task, the model takes past value/feature as input to predict nth future value.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{28}$$

where W_f is network parameter and x_t is t_{th} input and h_{t-1} is hidden latent representation. The f_t is the t^{th} state representation,.

$$y_i = O_t(f_t) \tag{29}$$

Using i^{th} latent representation, the output layer will transform into a target variable (y_i), RNNs have the ability to capture and model complex temporal patterns in the

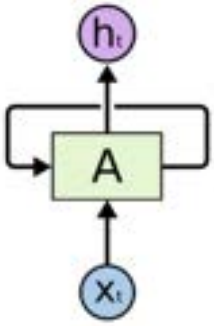


Figure 19: Recurrent Neural Network Block diagram [63]

data, making them suitable for tasks that involve sequences. However, they can be computationally expensive and suffer from vanishing or exploding gradients, which can affect training stability. Techniques such as gradient clipping and regularization methods like dropout are often employed to address these challenges.

Long-Short Term Memory (LSTM) and Gated Recurrent Unit (GRU)

LSTM (Long-Short Term Memory) and GRU (Gated Recurrent Unit) are both types of recurrent neural networks (RNNs) used for sequence modeling tasks. LSTM is specifically designed to capture long-term dependencies in the data by utilizing a memory cell and gating mechanisms. It is commonly employed in tasks such as time series forecasting. GRU, on the other hand, is a modified version of LSTM that simplifies the gating mechanism by removing the forget gate state. This modification helps address the issues of gradient exploitation and vanishing gradients commonly encountered in training RNNs.

In LSTM, the calculations involve several gates including the input gate (f_i), forget gate (f_f), output gate (o_i), memory cell (c_i), and hidden state (h_i). These gates regulate the flow of information within the LSTM unit. On the other hand, GRU simplifies the gating mechanism by representing the input gate as $f_i = 1 - f_f$, The rest of the gates and flow remain the same as in the standard LSTM architecture. The working of LSTM is shown in Figure 20

RNN-GNN

The RNN-GNN [12] architecture contains mainly two components, RNN and GNN, where RNN extracts the sequential/temporal component from the input data, and GNN extracts the neighborhood information-based feature representation. In this section, we explain the requirements and working of both components in the context of price prediction. The architecture in Fig. 22 can also be used for similar problems where the target feature is dependent on neighborhood and temporal information.

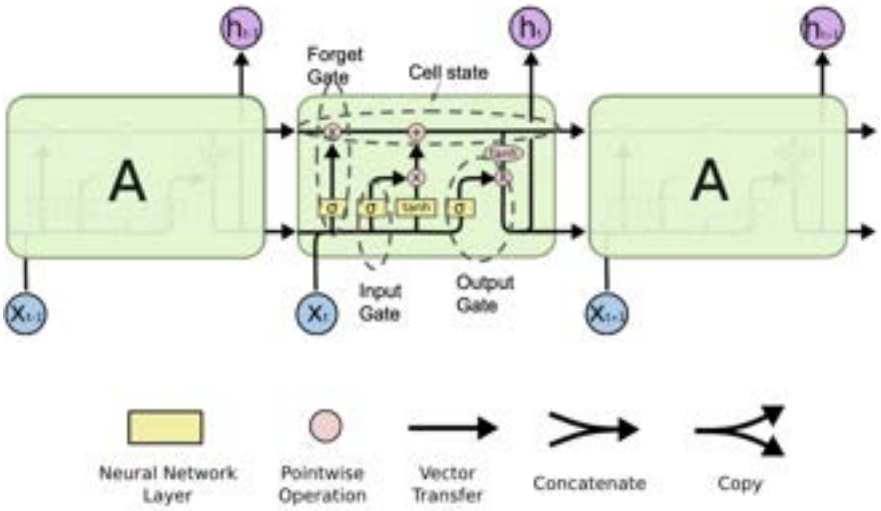


Figure 20: Long short term memory block diagram [63]

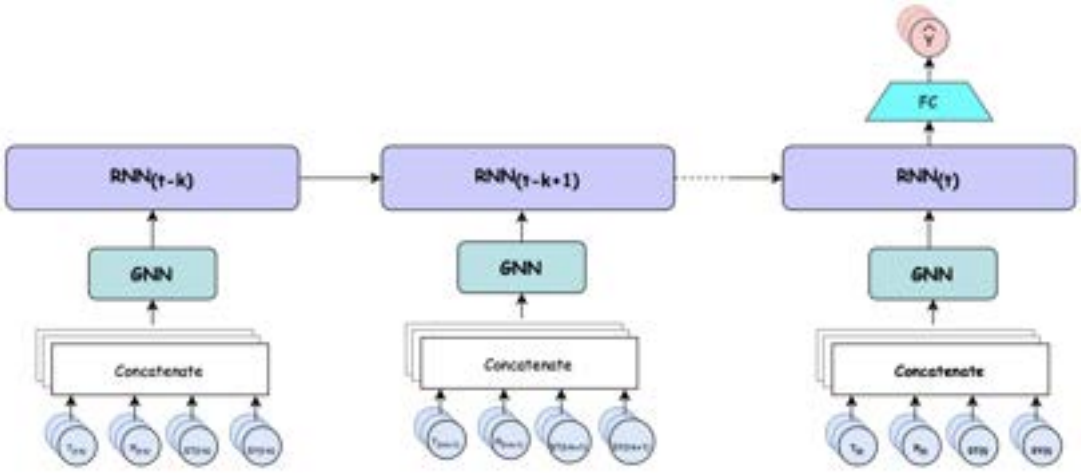


Figure 22: GNN-RNN unrolled through time. S_t,R_t,ST_t,SY_t denote Temperature, Relative Humidity, Soil type, and price data at time step t for all markets, respectively. $\hat{Y}$ denotes the predicted price for all the markets

1. **GNN**

   The Graph Neural Network (GNN) is specifically designed to process and analyze data that is structured as a graph. It focuses on learning embedding vectors for nodes by taking into account the information available in their local neighborhoods. In a graph $G = (V, E)$, where V represents the set of nodes and E represents the set of edges, GNN utilizes an adjacency matrix A to represent the

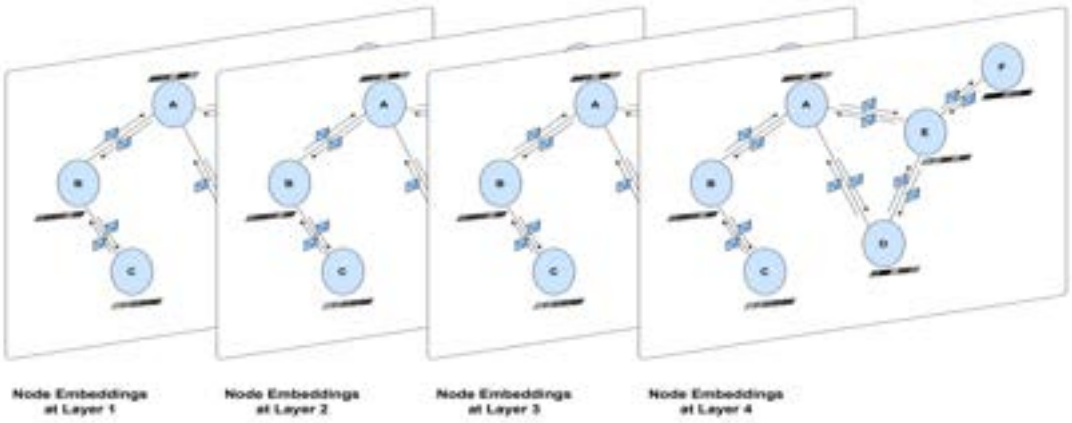


Figure 21: Update of each node embedding using Neighbourhood representation in GNN [12]

connections between nodes. The adjacency matrix A determines whether there is an edge between two nodes, with an entry $A_{i,j}$ being set to 1 if there is a connection from node j to node i. In price prediction, the Adjacency matrix entries are taken as the inverse of the distance between the two nodes (Mandis). Moreover, the edge between the two markets is present if the distance between the two is lesser than a threshold value.

GraphSAGE Convolution is used for extracting the neighborhood information.The visual explanation of the working of Graph Neural Networks is explained in Figure 21. Mean operations are used for aggregating Node information, followed by updation using GraphSAGE convolution operation[32, 12].

2. **RNN**

   As shown in Figure 22, the hidden latent representation after the GNN model consists of neighborhood information for all time series instances. For time-series datasets, the RNN component extracts new hidden representations with temporal information from the inputs. The working of the RNN layer is described in Section 5.8.2. Furthermore, the RNN layer is followed by an output head feed-forward neural network for the final prediction.

5.9 Comparative Model Analysis

In the first section, we trained model analysis using different metrics as per problem type and a brief introduction of different metrics. In the second stage, we provided a scatter plot for better visualization of prediction output compared to the reference. Further, users can use the section 5.7 for better visualization tools.

5.9.1 Performance Metrics

Based on the problem type (Classification and Regression), Agri-Vaahan offers the following performance metrics to evaluate the models. These metrics help in assessing the model's performance and making informed decisions. Agri-Vaahan allows users to calculate these metrics using data indicator columns, enabling the evaluation of train, test, and validation metrics for all trained models.

5.9.2 Regression Metrics

1. Mean Absolute Error (MAE): The mean absolute error measures the average absolute difference between the predicted and actual values. It is calculated as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

   where y_i represents the actual value and $\hat{y}_i$ represents the predicted value for the i-th data point.

2. Root Mean Square Error (RMSE): The root means square error is a commonly used metric to measure the average deviation between the predicted and actual values, giving higher weight to larger errors. It is calculated as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

3. R2 Score: The R2 score, also known as the coefficient of determination, indicates the proportion of the variance in the target variable that can be explained by the model. It is calculated as follows:

$$R2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

   where $\bar{y}$ represents the mean of the actual values.

4. Pearson Correlation Coefficient: The Pearson correlation coefficient measures the linear correlation between the predicted and actual values. It ranges from -1 to 1, where 1 represents a perfect positive correlation, -1 represents a perfect negative correlation, and 0 represents no correlation. It is calculated as follows:

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}$$

   where $\bar{y}$ and $\bar{\hat{y}}$ represent the means of the actual and predicted values, respectively.

5. Max Error: The maximum error represents the largest difference between the predicted and actual values:

$$MaxError = \max_{i=1}^{n} |y_i - \hat{y}_i|$$

6. Mean Percentage Absolute Error (MPAE): The mean percentage absolute error measures the average percentage difference between the predicted and actual values. It is calculated as follows:

$$MPAE = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{|y_i - \hat{y}_i|}{|y_i|} \right) \times 100\%$$

Classification Metrics

1. Accuracy: Accuracy measures the percentage of correctly predicted instances out of the total instances in the dataset. It is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

   Where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

2. Precision: Precision is the proportion of correctly predicted positive instances out of the total predicted positive instances. It is calculated as:

$$Precision = \frac{TP}{TP + FP}$$

3. Recall: it is also known as sensitivity or true positive rate, measures the proportion of correctly predicted positive instances out of the total actual positive instances. It is calculated as:

$$Recall = \frac{TP}{TP + FN}$$

4. F1-Score: The F1-score is the harmonic mean of precision and recall. It provides a balanced measure between precision and recall. It is calculated as:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

5. Confusion Matrix: The confusion matrix is a tabular representation that shows the counts of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions. It helps in evaluating the performance of a classification model by providing insights into the type of errors made.

5.10 Inference and Deployment in Agri-Vaahan

Agri-Vaahan provides users with the capability to evaluate trained models. Users can upload their datasets, include a prediction column, and compare the predictions with reference data for analysis. This allows users to make predictions on new data and assess the performance of the models. If reference target data is available, users can gain valuable insights by comparing the predictions with the reference values. It is important to note that the deployment feature in Agri-Vaahan is currently limited to evaluation and retraining purposes and does not offer full functionality for production deployment.

6 Mobile Apps for PREPARE, ACRE, and PROSPER

This chapter presents a specification for the three mobile apps developed as a part of this project. These include: PREPARE (Prediction of Prices in Agriculture); ACRE (Agricultural Crop Recommendation Engine); and PROSPER (Protocol for Selling Produce for Enhanced Revenue).

The three main applications developed as a part of this project, namely PREPARE, ACRE, and PROSPER have been designed with the intent of deploying them as mobile apps that could be used by the farmers and other stakeholders.

Mobile App for PREPARE

6.1.1 Purpose

The "PREPARE - Price Prediction for Agriculture" mobile app aims to provide farmers and stakeholders in the agricultural sector with accurate price predictions for agricultural products. Accurate prediction of agricultural crop prices is a crucial input for decision-making by various stakeholders in agriculture: farmers, consumers, retailers, wholesalers, and the Government. These decisions have significant implications regarding the economic well-being of the farmers, which includes helping them make informed decisions regarding planting, harvesting, and selling their crops.

6.1.2 Target Audience

The app is designed for

- Farmers
- Agricultural workers
- Traders
- Anyone who is interested in agricultural crop pricing.

6.1.3 Key Features

User Registration and Authentication

- User registration via email, phone number, or social media.
- Secure user authentication to protect personal information.

- Also requires the user to input basic information including address, location, and other specific details.

User Profile

- Users can manage their profiles, update information, and customize app settings.

Dashboard

- A user-friendly dashboard displaying key information and data visualizations.
- Popular crops and their price trends.
- Location-specific trends and new practices.
- Factors affecting price fluctuations (weather, supply, demand).
- Graphical representation of price trends.
- Overview of crop pricing trends.
- Relevant information about the user: Land area owned, available equipment, location, etc.
- Access to saved preferences.
- Quick access to favorite commodities.

Crop Selection

- Users can select the specific agricultural crop they are interested in, using checkboxes.

Crop Price Prediction

- Utilizes historical price information, climate conditions, soil type, location, and other key determinants of crop prices, market trends, and machine learning algorithms to provide price predictions for selected crops.
- Display the name of the crop selected.
- Let the user choose the location of the mandi, market, etc.
- View historical and current market prices.
- Users can receive and view price forecasts for different time periods (daily, weekly, monthly) over short-term and long-term.
- The mechanism could use graph neural networks (GNNs) in conjunction with a standard convolutional neural network (CNN) model to exploit geospatial dependencies in prices.

Price Alerts

- Users can set price alert notifications to receive updates when the predicted price for a chosen crop crosses a certain threshold.

Market News and Analysis

- Provide detailed market analysis and factors affecting crop prices.
- Include weather forecasts and their potential impact on crop pricing.
- Aggregated news feed related to agriculture and commodities.
- Analysis and insights from industry experts.

Historical Data

- Access to historical pricing data for crops for trend analysis.
- Interactive charts and graphs for visual representation.

Weather Data Integration

- Real-time weather data for the user's location or chosen agricultural region.
- Weather forecasts for the upcoming days.
- Notifications for severe weather events.

Notifications

- Push notifications for price alerts and relevant market news.

Support and FAQs

- Access to support resources.
- Frequently Asked Questions section.

6.1.4 Design and User Experience

User Interface

- Intuitive and user-friendly design with easy navigation.
- Clean and responsive layout for both mobile phones and tablets.

Data Visualization

- Interactive charts, graphs, and infographics to represent pricing trends.
- Incorporate color-coding for user-friendly data interpretation.

Personalization

- Customizable dashboards and notifications to suit individual user preferences.

6.1.5 Technical Requirements

Platforms

- Develop the app for both iOS and Android platforms.

Development Stack

- Select a suitable technology stack, considering factors like scalability and performance.
- Front-end: React Native or Flutter
- Back-end: Python

Data Sources

- Secure and reliable data sources for real-time pricing information and weather forecasts.
- AgMarkNet for Price and Arrival Datasets.
- Copernicus for geo-specific weather data.
- Machine learning models for price predictions.

Security

- Implement robust security measures to protect user data and privacy.
- Encryption for data transmission.

Scalability

- Ensure the app can handle a growing user base and data volume.

Testing

- Rigorous testing for functionality, usability, and security.
- Beta testing with a selected group of users before the official launch.

6.1.6 Feedback and Iteration

- Incorporate user feedback for continuous improvement.

6.1.7 Conclusion

This specification serves as a comprehensive guide for the development of the "PREPARE - Price Prediction for Agriculture" mobile app. It outlines the purpose, features, user experience, technical requirements, and other critical aspects of the app's development. It should be continuously updated and refined throughout the development process as necessary.

6.2 Mobile App for ACRE

6.2.1 Definition

The "ACRE - Crop Recommendation for Agriculture" app is a mobile application designed to assist farmers and agricultural experts in making informed decisions about crop selection. The application leverages the Sharpe Ratio, a risk metric from financial portfolio management, to rank various crop portfolios. ACRE aims to optimize crop selection to maximize utility and minimize the risk of sub-optimal yields and revenue loss for farmers.

The primary purpose of ACRE is to provide users with crop recommendations based on historical agricultural data, local climate conditions, and market trends. By utilizing the Sharpe Ratio, ACRE evaluates and ranks crop portfolios, helping users make data-driven decisions about which crops to cultivate.

6.2.2 Scope

The application will provide a platform for users to input their specific agricultural data, receive crop recommendations, create diversified crop portfolios, and access a comprehensive crop information database.

6.2.3 Target Audience

- Farmers
- Agricultural experts and consultants
- Agricultural students
- Agricultural extension workers

6.2.4 Key Features

User Registration and Authentication

- User registration via email, phone number, or social media.
- Secure user authentication to protect personal information.
- Requires user to input basic information including address, their geographical location, and other specific details including area of owned growable land (in hectares), human labor available, basic cost of production, etc.

- Users need also to input soil type and nutrients and available resources and facilities like irrigation.

User Profile

- Users can manage their profiles, update information, and customize app settings.
- Users can create profiles with information about their farming practices, preferences, and resources.
- The app considers these profiles when generating crop recommendations, tailoring them to individual needs.
- Based on the input parameters, the Utility Function is calculated and other parameters can be inferred such as Temperature, Rainfall, Sunlight, Humidity, Soil Type and Nutrients and Crop Price.

Dashboard

- A user-friendly dashboard displaying key information and data visualizations.
- Popular crops and their price trends.
- Best crops to be grown in the region specified.
- Location-specific trends and new practices.
- Factors affecting price fluctuations (weather, supply, demand).
- Graphical representation of price trends.
- Relevant information about the user: Land area owned, available equipment, location, etc.
- Access to saved preferences.
- Quick access to favorite commodities.

Crop Selection

- Users can select the desired agricultural commodities they are interested in growing, using checkboxes.
- The user needs to also put in details of last year's yields for his crops.

Crop Portfolio Recommendations

- Users can view a set of combinations of portfolios based on their set preferences and crop selection.
- ACRE will provide recommendations for crop portfolios with detailed information on suggested crops and their recommended proportions based on historical data.
- The recommendation system utilizes historical yield and price information, climate conditions, soil type, location, and other key determinants of predicted demands, transportation costs, and compliance ratios to arrive at a nearly optimal assignment of crop acreages to districts.

Sharpe Ratio Analysis

- ACRE calculates the Sharpe Ratio for each crop portfolio, providing a risk-adjusted return measure to assess the trade-off between crop yield and risk.
- Users can view the Sharpe Ratio for each recommended portfolio to make informed decisions.

Weather and Climate Data

- The app integrates with local weather data sources to provide real-time weather and climate information.
- Users can access historical weather data to understand seasonal patterns and make crop decisions accordingly.

Crop Information Database

- Maintain a comprehensive database of crop information.
- Include details on crop characteristics, growth cycles, and best practices.

Market Insights

- ACRE offers market information, including crop prices and demand trends, to help users make decisions that align with market conditions.

Notifications and Alerts

- ACRE can send users timely alerts about weather changes, market fluctuations, or other factors affecting their crop portfolios.

Support and FAQs

- Access to support resources.
- Frequently Asked Questions section.

6.2.5 User Interface

- Intuitive and user-friendly design with easy navigation.
- Clean and responsive layout for both mobile phones and tablets.

6.2.6 Data Visualization

- Interactive charts, graphs, and infographics to represent pricing trends.
- Incorporate color coding for user-friendly data interpretation.

6.2.7 Personalization

- Customizable dashboards and notifications to suit individual user preferences.

6.2.8 Technical Requirements

Platforms

- Develop the app for both iOS and Android platforms and the website for FPO.

Development Stack

- Select a suitable technology stack, considering factors like scalability and performance.
- Front-end: React Native or Flutter
- Back-end: Python

Data Sources

- Secure and reliable data sources for real-time pricing information and weather forecasts.
- AgMarkNet for Price and Arrival Datasets.
- Area, Production, and Yield data from Directorate of Economics and Statistics.
- Copernicus for geo-specific weather data.
- Machine learning models for price predictions.

Security

- Implement robust security measures to protect user data and privacy.
- Encryption for data transmission.

Testing

- Rigorous testing for functionality, usability, and security.
- Beta testing with a selected group of users before the official launch.

6.3 Mobile App for PROSPER

6.3.1 Definition

The “Protocol for Optimal Selling of Agricultural Produce for Enhanced Revenue” or “PROSPER” mobile app aims to provide a platform that connects individual farmers with Farmer Producer Organizations (FPOs), facilitating efficient selling of agricultural produce through an auction mechanism hence enabling farmers and other stakeholders to maximize their revenue.

This platform aims to ensure that small and marginal farmers can obtain the best prices for their produce while maintaining fairness for consumers. Additionally, it will offer various other features to support the agricultural ecosystem.

6.3.2 Scope

The app will allow farmers to list their produce, FPOs to conduct auctions, and consumers to participate in auctions, and provide administrators with tools for system management. The app will include secure user registration, listings, bidding, payment processing, and notification features including:

- User registration and profile management.
- Listing and searching for agricultural produce.
- Communication between buyers and sellers.
- Pricing recommendations based on market data.
- Secure payment processing.
- Integration with GPS for location-based services.
- Analytics and reporting for users.

6.3.3 Target Audience

- Farmers: Individuals or small agricultural producers looking to sell their produce.
- Farmer Producer Organizations (FPOs): Organizations responsible for conducting auctions and ensuring fair prices for both farmers and consumers.
- Consumers: Individuals and businesses interested in purchasing fresh produce, including small-volume buyers and large companies.
- Admin: System administrators responsible for managing the platform.

6.3.4 User Roles and Permissions

Farmers

- Create and manage profiles.
- List their agricultural produce and requirements.
- Participate in auctions.
- Receive notifications on auction outcomes and transactions.

Farmer Producer Organizations (FPOs)

- Create and manage profiles.
- Conduct auctions.
- Manage listings and bids.
- Process payments.
- Access reports and analytics.

Consumers

- Create and manage profiles.
- Participate in auctions.
- Make payments.
- Receive notifications.

Admin

- Manage user accounts and profiles.
- Monitor and moderate listings and auctions.
- Access reporting and analytics.
- Ensure system security and integrity.

6.3.5 Functional Requirements

User Registration and Profile Authentication

- User registration via email, phone number, or social media.
- Secure user authentication to protect personal information.
- It also requires the user to input basic information including address, location, and other specific details.
- Verification mechanisms for farmers and FPOs.

Listings and Auctions

- Farmers can create listings for their produce.
- FPOs can schedule and conduct auctions.
- Real-time listing updates and auction status.

Bidding and Auction Management

- Bidding functionality for consumers.
- FPOs can manage auctions, including setting starting prices and closing auctions.
- Real-time updates on bids.

Payments and Transactions

- Secure payment processing for winning bids.
- Transaction history and receipts.
- Integration with payment gateways.

Notifications

- Real-time notifications for bid status, auction results, and payments.

Reports and Analytics

- Reports on auction performance, sales, and revenue.
- Data analytics for decision-making.

Admin Panel

- User account management.
- Listing and auction moderation.
- Analytics dashboard.
- System configuration and maintenance.

6.3.6 Technical Requirements

Platforms

- Develop the app for both iOS and Android platforms.

Development Stack

- Select a suitable technology stack, considering factors like scalability and performance.
- Front-end: React Native or Flutter
- Back-end: Python with FastAPI Integration

Data Sources

- Secure and reliable data sources for real-time pricing information and weather forecasts.
- AgMarkNet for Price and Arrival Datasets.
- Copernicus for geo-specific weather data.
- Machine learning models for price predictions.

Security

- Implement robust security measures to protect user data and privacy.
- Encryption for data transmission.

Scalability

- Ensure the app can handle a growing user base and data volume.

Testing

- Rigorous testing for functionality, usability, and security.
- Beta testing with a selected group of users before the official launch.

6.3.7 Feedback and Iteration

- Incorporate continuous app improvement through suggestions and updates from all stakeholders and other users.

6.3.8 Summary

This “PROSPER” mobile app specification document provides a comprehensive overview of the app’s purpose, features, user roles, technical requirements, and more. It serves as a foundation for the development and successful launch of the app, aiming to connect farmers and FPOs to enhance revenue and fairness in agricultural produce sales. Finally, any profits made by the FPOs through these sales are normally redistributed to the farmers based on the stakes that the farmers hold in the FPO.

7 Curated Datasets for Indian Agriculture

This section describes the data collected from various sources for the different problems addressed in this project. While some of the datasets are clean and could be used directly, several other datasets needed a significant amount of data cleaning and curation. This chapter provides details on various datasets that are available. We provide details on the curated datasets as well.

7.1 List of Data Repositories

7.2 Agmarknet
7.3 Copernicus
7.4 Crop Production and Land Use Statistics
7.5 Directorate of Economics and Statistics
7.6 ICRISAT
7.7 VDSA

7.2 Agmarknet

Source: https://agmarknet.gov.in/

7.2.1 List of Mandis

Duration: State-wise List of Mandis from 2008 to 2014 and 2014 to 2018.

Link to OneDrive data: Click here or go to

`https://indianinstituteofsciencey.sharepoint.com/:f:`
`/g/personal/mayankb_iisc_ac_in/`
`EnG-eeZEAZ1Mtnx7ABehXwgBNc0MtJS6mtmgmEj-2onsBQ?e=hsbgHu`

Data Granularity: Mandi level.

Crops: Brinjal, Cauliflower, Chillies, Gram, Mango, Pgourd, Potato, Tomato, Wheat.

Data Heads: State, State Code, District, Market, Commodity, Commodity Code.

Code to Download Data: Mandi List can be downloaded from AgMarknet by running the code file named download mandis.py available at

`https://indianinstituteofscience-my.sharepoint.com/:u:/g/personal`
`/mayankb_iisc_ac_in/EZvO_tCIU11DhQGLbq9fmV4BZo8HCjdWNtLwCGEvm6jpDQ`

7.2.2 Price and Arrival Data Extraction

Python scripts are used to download the relevant Price and Arrival Data in CSV format.

File format: Python files (.py)

Format of downloaded files: Comma Separated Values files (.csv)

Link to downloaded CSV files: Click here or go to

`https://indianinstituteofscience-my.sharepoint.com/:f:/g/personal/mayankb_iisc_ac_in/EqSIpRwCIY9NpgYvNVN8IlUBP5WvJRPyd88cjX9oUQjmlQ?e=nQ2Oz8`

Code to Download Data: Price and Arrival Data can be downloaded from agmarknet by running the code in the folder named *Download from agmarknet*

`https://indianinstituteofscience-my.sharepoint.com/:f:/g/personal/mayankb_iisc_ac_in/EhLe4OVs-MFBtzgim7aZWXoB90SY8q6grvvPzqJEjgUhSw?e=balG8c`

Data Imputation: Required. See following section.

7.2.3 Curated Datasets

Duration: Price and Arrival Data from 2008 to 2022.

File format: Excel and Comma Separated Values (.xlsx, .csv)

Link to OneDrive data: Click here or go to

`https://indianinstituteofscience-my.sharepoint.com/:f:/g/personal/mayankb_iisc_ac_in/Euj2Q-boXD9DtLhkmyp2EiIBTAyc3-9cOkLP6-btpfiFdw?e=uMPx8D`

Data Granularity: Mandi (Market) level.

Crops:

- **All India Data with Spline imputation:** Brinjal, Onion, Potato, Tomato and Wheat
- **Partial Crop Data without imputations:** Arhar, Bajra, Barley, Chana (Bengal Gram Dal), Cotton, Green Chillies, Green Gram (Moong Whole), Ground Nut, Jowar, Maize, Masur, Mustard, Paddy, Ragi, Red Chilli, Rice, Soyabean, Urad (Black Gram Dal)

- **CROP-S Data:** Selected Kharif and Rabi Crops with Area, Production and Yield Values.

Data Heads:

- **Arrival Data:** State, State Code, District, Market, Commodity, Commodity Code, Arrivals, Date
- **Price Data:** Date, State Name, State Code, District Name, District Code, Mandi Name, Mandi Code, Crop Name, Crop Code, Crop Variety, Arrival Quantity, Minimum Price, Maximum Price, Modal Price

Data Imputation: Filled NaN values using Spline interpolation and linear interpolation. The imputed data was uploaded in the folder Crop Data available at

`https://indianinstituteofscience-my.sharepoint.com/:f:/g/personal/mayankb_iisc_ac_in/Eh871fpdRX5ArtR33GgJ4xABT9_AxoT5SKu639VPRxhYeA?e=qOjcyO`

For the detailed step by step process utilized for obtaining the curated data sets refer to the text file named *Steps.txt* available at

`https://indianinstituteofscience-my.sharepoint.com/:t:/g/personal/mayankb_iisc_ac_in/ETccF17qF9ZEpNPIhpwu9bIB2Q87EO9Y3gTBwcXdoC9Znw?e=BA1QFz`

The imputation methods included the following:

- Data Cleaning was performed using Data-Cleaning.py

  `https://indianinstituteofscience-my.sharepoint.com/:u:/g/personal/mayankb_iisc_ac_in/ES-2oRFuFcVNk3wWSGRhTsAB28QEZJRFeFsJWy6IituNDA?e=UENNYT`

- Duplicate Mandi-dates were removed using Remove-Duplicates.py

  `https://indianinstituteofscience-my.sharepoint.com/:u:/g/personal/mayankb_iisc_ac_in/EWf9kJibts9Nm5BN-IdwkSgBPEeb9_HwMOr9HSWwcPBx7w?e=q4jIbZ`

- Total number of entries were counted using Initial-Count-Crop.py

  https://indianinstituteofscience-my.sharepoint.com/:u:/g/personal
  /mayankb_iisc_ac_in/EaOEnmZ_w8dJiOpxQdplsjIB2jMlw6uxm12gV5UVuwgQnw?e=8YoPEA

- Mandi selection was performed and data files with data of selected mandis were generated using Selected-Mandi-Data.py

  https://indianinstituteofscience-
  my.sharepoint.com/:u:/g/personal/mayankb_iisc_ac_in/
  EWQFD5-siw9Ji1aS1GWES-wBOVqCm0U7LiJIgO8vfj2ARQ?e=bTrAV4

- Missing dates and mandi-wise missing data were added using Add-Missing-Dates.py

  https://indianinstituteofscience-
  my.sharepoint.com/:u:/g/personal/mayankb_iisc_ac_in/
  Ecjf6iAMnrJFvJPlWsI3LdQBE2pQYnnXEhCPhpF0Swl-jQ?e=PWODwD

- Outliers were detected using Find-Outliers.py

  https://indianinstituteofscience-
  my.sharepoint.com/:u:/g/personal/mayankb_iisc_ac_in/
  EdssRxVqfuZEu1ewoZGbN9oBBq8hd0GO-f3vcLq1qCB24Q?e=9Xrb2I

7.3 Copernicus

Climate data with latitudes and longitudes.

Source: Depending on the required parameters and data granularity (Hourly or Monthly), the data was downloaded from the following links:

Pressure levels Monthly

https://cds.climate.copernicus.eu/cdsapp#!/dataset
/reanalysis-era5-pressure-levels-monthly-means?tab=form

Single levels Monthly

https://cds.climate.copernicus.eu/cdsapp#!/dataset
/reanalysis-era5-single-levels-monthly-means?tab=form

Pressure levels Hourly

```
https://cds.climate.copernicus.eu/cdsapp#!/dataset
/reanalysis-era5-pressure-levels?tab=form
```

Single levels Hourly

```
https://cds.climate.copernicus.eu/cdsapp#!/dataset
/reanalysis-era5-single-levels?tab=form
```

Duration: Annual data from 2008 to 2021

File format: NetCDF and Comma Separated Values (.nc, .csv)

Link to OneDrive path: Click here or go to

```
https://indianinstituteofscience-my.sharepoint.com/:f:/g/personal
/mayankb_iisc_ac_in/Eo2ojoyMYVFKtIVVqfbKNoUBL75s0pEi8Zqjhl5g5hwjTQ?e=boRQc1
```

Data Granularity: Data for all the regions of India is available with step-size being a change of 0.25 in Latitude or Longitude. Separate folders are maintained for Hourly and Monthly values. These folders contain one file each for every parameter-year pair.

Data Heads:

- Total Precipitation
- Surface net solar radiation(Sunlight)
- Relative Humidity
- Temperature
- Soil type

Data Imputation: Not required

7.4 Crop Production and Land Use Statistics

Source: https://aps.dac.gov.in

Data Granularity: District level, seasonal (Rabi/Kharif) data

Duration: 2008 to 2021.

File format: Excel (.xls,.xlsx)

One drive path: Click here or go to

https://indianinstituteofscience-my.sharepoint.com/:f:/g/personal
/mayankb_iisc_ac_in/EilRtp3oTdpMoerhnWabgvEBmfy_B-_N5oLGdKRZ00knsg?e=x2zLn3

Data Heads:

- Area (Hectare)
- Production (Tonnes)
- Yield (Tonnes/Hectare)

Crops: Arecanut, Arhar/Tur, Bajra, Banana, Barley, Black pepper, Cardamom, Cashewnut, Castor seed, Coconut, Coriander, Cotton(lint), Cowpea(Lobia), Dry chillies, Garlic, Ginger, Gram, Groundnut, Guar seed, Horse-gram, Jowar, Jute, Khesari, Linseed, Maize, Masoor, Mesta, Moong(Green Gram), Moth, Niger seed, Oilseeds total, Onion, Other Rabi pulses, Other Cereals, Other Kharif pulses, Other oilseeds, Other Summer Pulses, Peas and beans (Pulses), Potato, Ragi, Rapeseed and Mustard, Rice, Safflower, Sannhamp, Sesamum, Small millets, Soyabean, Sugarcane, Sunflower, Sweet potato, Tapioca, Tobacco, Turmeric, Urad, Wheat

Data Imputation: Not required

7.5 Directorate of Economics and Statistics

Under the Department of Agriculture and Farmers Welfare, Ministry of Agriculture and Farmers Welfare, Government of India.

Source: https://eands.dacnet.nic.in/Cost_of_Cultivation.htm

Duration: Annual data from 2004-05 to 2020-21 (plus summary file from 1996-2004)

File format: Excel (.xls,.xlsx)

One drive path: Click here or go to

https://indianinstituteofscience-my.sharepoint.com/:f:/g/personal
/mayankb_iisc_ac_in/Eix8MbOqTvFBuoixAmCS5jsBYSZBC_Ycf13PuikmYtbQmQ?e=ZThFZJ

Data Granularity: State-level, year-wise

Crops: Arhar, Bajra, Cotton, Groundnut, Jowar, Maize, Moong, Nigerseed, Paddy, Ragi, Sesamum, Soyabean, Sunflower, Urad, Barley, Gram, Lentil, R&M, Safflower, Wheat, Jute, Onion, Sugarcane, Potato & Coconut.

Data Heads:

- Cost of Cultivation (Rs./Hectare)
  - a) **Operational Cost**
    - Human Labour
      - Family
      - Attached
      - Casual
    - Animal Labour
      - Hired
      - Owned
    - Machine Labour
      - Hired
      - Owned
    - Seed
    - Fertilizer
    - Manure
    - Insecticides
    - Irrigation Charges
    - Miscellaneous
    - Interest on Working Capital
  - b) **Fixed Cost**
    - Rental Value of Owned Land
    - Rent Paid For Leased-in-Land
    - Land Revenue, Taxes, Cesses
    - Depreciation on Implements & Farm Building
    - Interest on Fixed Capital
- Cost of Production (Rs./Qtl)

- Value of Main Product (Rs./Hectare)
- Value of By-Product (Rs./Hectare)
- Material & Labour Input/Hectare of
  - Seed (Kg.)
  - Fertilizer (Kg. Nutrients)
  - Manure (Qtl.)
  - Human Labour (Man Hrs.)
  - Animal Labour (Pair Hrs.)
- Rate per Unit (Rs.)
  - Seed (Kg.)
  - Fertilizer (Kg. Nutrients)
  - Manure (Qtl.)
  - Human Labour (Man Hrs.)
  - Animal Labour (Pair Hrs.)
  - Implicit Rate (Rs./Qtl.)
  - Number of Holdings in Sample
  - Number of Tehsils in Sample
  - Derived Yield (Qtl./Hectare)

Data Imputation: Not required

7.6 ICRISAT

Duration: Annual data from 1966 to 2018 (data until 2011 is also available as a subset of the VDSA dataset mentioned in Section 7.7)

File format: Excel (.csv,.xlsx)

Link to OneDrive data: Click here or go to

`https://indianinstituteofscience-my.sharepoint.com/:f:/g/personal/mayankb_iisc_ac_in/Ep4jdPXJnuhMj0_qifpI8k8B5QRxTIiGDFsFYvcJpSDpMg?e=XCmiLb`

Data Granularity: District-level

Data Heads:

- Area, Production and Yield
- Farm Harvest Price
- Fertilizer Consumption
- Land Utilization
- Population

Data Imputation: Not required.

7.7 VDSA

Organization - ICRISAT-ICAR-IRRI Collaborative Research Project

Maintained by ICRISAT and updated under the Village Dynamics in South Asia (VDSA) Project for data related to key agricultural and socioeconomic variables.

Source: http://vdsa.icrisat.ac.in/

Duration: Annual data from 1966 to 2011

File format: Excel (.xls,.xlsx)

Link to OneDrive data: Click here or go to

`https://indianinstituteofscience-my.sharepoint.com/:f:/g/personal/mayankb_iisc_ac_in/EsPCTw7Ci-NJmoT2p8maScYBwSB76cyPQ43Jf12l7ouv8g?e=GWtB3u`

Data Granularity: District-level

Data Heads:

- Area and Production: Cereals, Pulses, Oil seeds and Selected Cash Crops
- Crop-wise Irrigated Area
- HYV Area (Cereal Crops)

- Land Use (Geographical Area, Forest Area, etc.)
- Net Cropped Area, Gross Cropped Area, Net Irrigated Area and Gross Irrigated Area
- Source-wise Irrigated Area
- Farm Harvest Prices: All Crops
- Fertilizer Consumption
- Fertilizer Prices
- Fodder Area and Fodder Irrigated Area
- Field Labor Wages
- Markets and Roads
- Annual and Monthly Actual Rainfall
- Annual and Monthly Normal Rainfall
- Length of Growing Period
- Soil Type
- Annual and Monthly Normal Potential Evapotranspiration (PE)
- Annual Moisture Availability Index (MAI)
- Agroecological Subregions (AESR)
- Population Census Data (Population, Literacy, Cultivators, Agric. Laborers)
- Livestock (Census Data)
- Operational Holdings (Census Data)
- Farm Implements and Machinery (Census Data)

Data Imputation: Not required

8 Publications and Manuscripts from the Project

1. Mayank Ratan Bhardwaj, Azal Fatima, Inavamsi Enaganti, and Y. Narahari. Incentive Compatible Mechanisms for Efficient Procurement of Agricultural Inputs for Farmers through Farmer Collectives. In ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (ACM COMPASS), 2022 Jun 29 (pp. 696-700).
   URL: *https://dl.acm.org/doi/pdf/10.1145/3530190.3534842*

2. Rohit Patel, Inavamsi Enaganti, Mayank Ratan Bhardwaj, and Y. Narahari. A Data-driven, Farmer-oriented Agricultural Crop Recommendation Engine (ACRE). In International Conference on Big Data Analytics, 2022 Dec 19 (pp. 227-248).
   URL: *https://link.springer.com/chapter/10.1007/978-3-031-24094-2_16*

3. Mayank Ratan Bhardwaj, Bazil Ahmed, Prathik Diwakar, Ganesh Ghalme, and Y. Narahari. Designing Fair, Cost-optimal Auctions based on Deep Learning for Procuring Agricultural Inputs through Farmer Collectives. In International Conference on Automation Science and Engineering (IEEE CASE), 2023 Aug 26 (pp. 1-8).
   URL: *https://doi.org/10.1109/CASE56687.2023.10260598*

4. Mayank Ratan Bhardwaj, Jaydeep Pawar, Abhijnya Bhat, Deepanshu, Inavamsi Enaganti, Kartik Sagar, and Y. Narahari. An innovative Deep Learning Based Approach for Accurate Agricultural Crop Price Prediction. In International Conference on Automation Science and Engineering (IEEE CASE), 2023 Aug 26 (pp. 1-7).
   URL: *https://doi.org/10.1109/CASE56687.2023.10260494*

5. Mayank Ratan Bhardwaj, Abhishek Chaudhary, Inavamsi Enaganti, Kartik Sagar, and Y. Narahari. A Decision Support Tool for District Level Planning of Agricultural Crops for Maximizing Profits of Farmers. In International Conference on Automation Science and Engineering (IEEE CASE), 2023 Aug 26 (pp. 1-6).
   URL: *https://doi.org/10.1109/CASE56687.2023.10260581*

6. Mayank Ratan Bhardwaj, Gogulapati Sreedurga, Vishisht Rao, Y. Narahari. PROSPER-NB: A social welfare maximizing mechanism based on Nash Bargaining for selling agricultural produce. Manuscript under preparation.

7. Mayank Ratan Bhardwaj, Vishisht Rao, Kartik Sagar, Bazil Ahmad, Y. Narahari. Deep Learning Meets Auction Design for Sale of Agricultural Produce through Farmer Collectives to Maximize Nash Social Welfare. Submitted for publication.

9 Ph.D., M.Tech. Project Students. Research Engineers, and Research Interns

- **Mayank Ratan Bhardwaj**. Novel Algorithms for Improving Agricultural Planning and Operations using Artificial Intelligence and Game Theory. Ph.D. Dissertation. Department of Computer Science and Automation, Indian Institute of Science. July 2023.

- **Jaydeep Pawar Vasudev**. An innovative deep learning based approach for accurate agricultural crop price prediction. M.Tech. Project Report. Department of Computer Science and Automation, Indian Institute of Science. June 2023. (Awarded the Best M.Tech. Project Prize among 70+ M.Tech. Project dissertations in the Department of CSA, IISc).

- **Kishan Mittal**. ACRE 2.0: Agricultural Crop Recommendation Engine. M.Tech. Project Report. Department of Computer Science and Automation, Indian Institute of Science. June 2023.

- **Abhishek Kumar Chaudhary**. CROP-S: A Decision Support Tool for District Level Planning of Agricultural Crops for Maximising Profits of Farmers. M.Tech. Project Report. Department of Computer Science and Automation, Indian Institute of Science. June 2023.

- **Chaitanya Chennam**. PROSPER: Protocol for Optimal Selling of Agricultural Produce for Enhanced Revenue. M.Tech. [AI] Project Report. Departmemt of Computer Science and Automation, Indian Institute of Science. June 2023.

- **Kaushik Hareshbhai Kukadia**. AGRI-VAAHAN 2.0: An AIML Platform for Data Analytics for Agriculture. M.Tech. [AI] Project Report. Department of Computer Science and Automation, Indian Institute of Science. June 2023.

- **Sneha Negi**. Designing Fair, Cost Optimal Combinatorial auctions based on Deep Learning for Procuring Agricultural Inputs through Farmer Collectives. M.Tech. Project Report. Department of Computer Science and Automation, Indian Institute of Science. June 2023.

- **Azal Fatima**. Mechanism Design for Efficient Procurement of Agricultural Inputs to Farmers. M.Tech. Project Report. Department of Computer Science and Automation. Indian Institute of Science, Bangalore, June 2022.

- **Rohit Patel**. ACRE – Agricultural Crop Recommendation Engine. M.Tech. Project Report. Department of Computer Science and Automation. Indian Institute of Science, Bangalore, June 2022.

- **Deepanshu**. Improved Methods for Agricultural Crop Price Prediction. M.Tech. Project Report. Department of Computer Science and Automation. Indian Institute of Science, Bangalore, June 2022.
- **P. Sowjanya**. Agri-Vaahan: An AIML Pipeine for Agricultural Data Analytics. M.Tech. Project Report. Department of Computer Science and Automation. Indian Institute of Science, Bangalore, June 2022.
- **Inavamsi Enaganti**. Research Engineer.
- **Sanath Patil**. Research Assistant.
- **Kartik Sagar**. Research Assistant.
- **Vishisht Rao**. Research Assistant and Research Intern.
- **C.M. Arun**. Research Assistant.
- **Abhijnya Bhat**. Research Intern.
- **Vaidehi Bhaskara**. Research Intern.
- **S. Ravikumar**. Project Assistant.

References

[1] Five Ws. URL https://en.wikipedia.org/wiki/Five_Ws.

[2] Plotly express arguments in python, 2023. URL https://plotly.com/python/px-arguments/.

[3] V.K. Aatre, V.V.R. Sastry, K.B. Umesh, Jaywant Arakeri, Nipun Mehrotra, Y.Narahari, C.T. Ramachandra, G. Senthil Kumaran, M.S. Sheshshayee, and Ravi Trivedi. Technologies for Transformation of Indian Agriculture. In *Proceedings of INAE (Indian National Academy of Engineering) Workshop*, January 2022. URL https://drive.google.com/file/d/1higJtp0yTWthtLZVRky6PMLcpyO0lY5_/view.

[4] NITI Aayog. National strategy for artificial intelligence. *Paper. June*, pages 2019–01, 2018.

[5] J Harold Ahlberg, Edwin Norman Nilson, and Joseph Leonard Walsh. *The Theory of Splines and Their Applications: Mathematics in Science and Engineering: A Series of Monographs and Textbooks, Vol. 38*, volume 38. Elsevier, 2016.

[6] Hyun Ahn, Kyunghee Sun, and Kwanghoon Pio Kim. Comparison of missing data imputation methods in time series forecasting. *CMC-COMPUTERS MATERIALS & CONTINUA*, 70(1):767–779, 2022.

[7] AH Alamoodi, BB Zaidan, AA Zaidan, OS Albahri, Juliana Chen, MA Chyad, Salem Garfan, and AM Aleesa. Machine learning-based imputation soft computing approach for large missing scale and non-reference data imputation. *Chaos, Solitons & Fractals*, 151:111236, 2021.

[8] P Alli and A Narayanamoorthy. Why do farmers resort to dumping produce?, Apr 2020. URL https://www.thehindubusinessline.com/opinion/why-do-farmers-resort-to-dumping-produce/article31365453.ece.

[9] Mohamad M Awad. Toward precision in crop yield estimation using remote sensing and optimization techniques. *Agriculture*, 9(3):54:1–54:13, 2019.

[10] Mayank Ratan Bhardwaj. *Novel Algorithms for Improving Agricultural Planning and Operations using Artificial Intelligence and Game Theory*. PhD thesis, Indian Institute of Science, Bangalore, 2023.

[11] Mayank Ratan Bhardwaj, Bazil Ahmed, Prathik Diwakar, Ganesh Ghalme, and Y Narahari. Designing fair, cost-optimal auctions based on deep learning for

procuring agricultural inputs through farmer collectives. *2023 IEEE 19th International Conference on Automation Science and Engineering (CASE)*, pages 1–8, 2023.

[12] Mayank Ratan Bhardwaj, Jaydeep Pawar, Abhijnya Bhat, Inavamsi Enaganti, Kartik Sagar, Y Narahari, et al. An innovative deep learning based approach for accurate agricultural crop price prediction. *2023 IEEE 19th International Conference on Automation Science and Engineering (CASE)*, pages 1–7, 2023.

[13] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491*, 2021.

[14] Ioannis Caragiannis, David Kurokawa, Hervé Moulin, Ariel D Procaccia, Nisarg Shah, and Junxing Wang. The unreasonable fairness of maximum Nash welfare. *ACM Transactions on Economics and Computation (TEAC)*, 7(3):1–32, 2019.

[15] Ramesh Chand. Doubling farmer's income: Rationale, strategy, prospects, and action plan. Technical report, NITI Aayog (National Institution for Transforming India), March 2017.

[16] Ranveer Chandra and Stewart Collis. Digital agriculture for small-scale producers: challenges and opportunities. *Communications of the ACM*, 64(12):75–84, 2021.

[17] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.

[18] Abhishek Kumar Chaudhary. CROP-S: A Decision Support Tool for District Level Planning of Agricultural Crops for Maximising Profits of Farmers, 2023.

[19] Chaitanya Chennam. PROSPER: Protocol for Optimal Selling of Agricultural Produce for Enhanced Revenue, 2023.

[20] Deepanshu. Improved methods for agricultural crop price prediction, 2022.

[21] S Prasanna Devi, Y Narahari, Nukala Viswanadham, S Vinu Kiran, and S Manivannan. E-mandi implementation based on gale-shapely algorithm for perishable goods supply chain. In *2015 IEEE International Conference on Automation Science and Engineering (CASE)*, pages 1421–1426. IEEE, 2015.

[22] Digital India Initiative. Agriculture. URL `https://data.gov.in/sector/Agriculture`.

[23] Directorate of Marketing & Inspection (DMI), Ministry of Agriculture and Farmers Welfare, Government of India. Agmarknet, 2023. URL `https://agmarknet.gov.in/`.

[24] Paul Dütting, Zhe Feng, Harikrishna Narasimhan, David C. Parkes, and Sai S. Ravindranath. Optimal auctions through deep learning. *Communications of the ACM*, 64(8):109–116, 2021.

[25] Tlamelo Emmanuel, Thabiso Maupong, Dimane Mpoeleng, Thabo Semong, Banyatsang Mphago, and Oteng Tabona. A survey on missing data in machine learning. *Journal of Big Data*, 8(1):1–37, 2021.

[26] European Centre for Medium-Range Weather Forecasts. The climate data store. URL `https://climate.copernicus.eu/climate-data-store`.

[27] Joshua Fan, Junwen Bai, Zhiyun Li, Ariel Ortiz-Bobea, and Carla P Gomes. A GNN-RNN approach for harnessing geospatial and temporal information: application to crop yield prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11873–11881, 2022.

[28] Azal Fatima. Mechanism design for efficient procurement of agricultural inputs to farmers, 2022.

[29] Zhe Feng, Harikrishna Narasimhan, and David C Parkes. Deep learning for revenue-optimal auctions with budgets. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018)*, pages 354–362, 2018.

[30] D. Gale and L. S. Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962.

[31] Hangzhi Guo, Alexander Woodruff, and Amulya Yadav. Improving lives of indebted farmers using deep learning: predicting agricultural produce prices using convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13294–13299, 2020.

[32] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[33] Manuel A Hernandez, Shahidur Rashid, Solomon Lemma, and Tadesse Kuma. Market institutions and price relationships: The case of coffee in the ethiopian commodity exchange. *American Journal of Agricultural Economics*, 99(3):683–704, 2017.

[34] ICRISAT. Microsoft and ICRISAT's Intelligent Cloud pilot for Agriculture in Andhra Pradesh increase crop yield for farmers. URL `https://www.icrisat.org/`

microsoft-and-icrisats-intelligent-cloud-pilot-for-agriculture-in\
-andhra-pradesh-increase-crop-yield-for-farmers/.

[35] ICRISAT. VDSA: Village Dynamics Studies in South Asia. http://vdsa.icrisat.org/vdsa-database.aspx.

[36] India Brand Equity Foundation. Agriculture in India: Information about Indian agriculture & its importance. URL http://www.ibef.org/industry/agriculture-india.aspx.

[37] Amelia Ritahani Ismail, Nadzurah Zainal Abidin, and Mhd Khaled Maen. Systematic review on missing data imputation techniques with machine learning algorithms for healthcare. *Journal of Robotics and Control (JRC)*, 3(2):143–152, 2022.

[38] Krishi Jagran. Farmers throw tomatoes in trash after price dropped to Rs. 3 per Kg, Aug 2022. URL https://krishijagran.com/agriculture-world/farmers-throw-tomatoes-in-trash-after-price-dropped-to-rs-3-per-kg/.

[39] Ayush Jain, Smit Marvaniya, Shantanu Godbole, and Vitobha Munigala. Towards context-based model selection for improved crop price forecasting. In *5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD)*, pages 195–203, 2022.

[40] Sebastian Jäger, Arndt Allhorn, and Felix Biebmann. A benchmark for data imputation methods. *Frontiers in Big Data*, July 2021.

[41] Andreas Kamilaris and Francesc X Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147:70–90, 2018.

[42] S. Khaki and L. Wang. Crop yield prediction using deep neural networks. *Frontiers in Plant Science*, 10, 2019.

[43] Saeed Khaki, Lizhi Wang, and Sotirios V Archontoulis. A CNN-RNN framework for crop yield prediction. *Frontiers in Plant Science*, 10:1750, 2020.

[44] Zhwan M Khalid, Subhi RM Zeebaree, et al. Big data analysis for data visualization: A review. *International Journal of Science and Business*, 5(2):64–75, 2021.

[45] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[46] Vijay Krishna. *Auction theory*. Academic press, 2009.

[47] Kaushik Hareshbhai Kukadia. AGRI-VAAHAN 2.0: An AIML Platform for Data Analytics for Agriculture, 2023.

[48] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521 (7553):436–444, 2015.

[49] Retsef Levi, Manoj Rajan, Somya Singhvi, and Yanchong Zheng. Improving farmers' income on online agri-platforms: Evidence from the field. *Available at SSRN 3486623*, 2022.

[50] Christian Von Lücken and Ricardo Brunelli. Crops selection for optimal soil planning using multiobjective evolutionary algorithms. In *AAAI-2008, 22nd International Conference of the American Association for Artificial Intelligence*, pages 1751–1756, 2008.

[51] Wei Ma, Kendall Nowocin, Niraj Marathe, and George H Chen. An interpretable produce price forecasting system for small and marginal farmers in India using collaborative filtering and adaptive nearest neighbors. In *Proceedings of the Tenth International Conference on Information and Communication Technologies and Development*, pages 1–11, 2019.

[52] Lovish Madaan, Ankur Sharma, Praneet Khandelwal, Shivank Goel, Parag Singla, and Aaditeshwar Seth. Price forecasting & anomaly detection for agricultural commodities in India. In *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 52–64, 2019.

[53] J Madhuri and M Indiramma. Artificial neural networks based integrated crop recommendation system using soil and climatic parameters. *Indian Journal of Science and Technology*, 14(19):1587–1597, 2021.

[54] Vishal Meshram, Kailas Patil, Vidula Meshram, Dinesh Hanchate, and SD Ramkteke. Machine learning in agriculture domain: A state-of-art survey. *Artificial Intelligence in the Life Sciences*, 1:100010, 2021.

[55] P. Milgrom. Auctions and bidding: A primer. *Journal of Economic Perspectives*, 3(3):3–22, 1989.

[56] Ministry of Agriculture, Cooperation, and Farmers Welfare, Government of India. Transforming Indian Agriculture : Consultation Paper on IDEA (India Digital Ecosystem of Agriculture). June 2021.

[57] Kishan Mittal. ACRE 2.0: Agricultural Crop Recommendation Engine, 2023.

[58] Divya Nair, Rupika Singh, Sumedha Jalote, Vinod Kumar Sharma, and Will Thompson. Enabling farmer producer companies' success in marketing, 11 2020. URL https://www.idinsight.org/publication/enabling-farmer-producer-companies-success-in-marketing/.

[59] Y. Narahari. *Game Theory and Mechanism Design*. IISc Press (Bengaluru, India) and The World Scientific (Singapore), 2014.

[60] National Portal of India. Agriculture. URL http://www.india.gov.in/topics/agriculture.

[61] Sneha Negi. Designing fair, cost optimal combinatorial auctions based on deep learning for procuring agricultural inputs through farmer collectives, 2023.

[62] Neil Newman, Lauren Falcao Bergquist, Nicole Immorlica, Kevin Leyton-Brown, Brendan Lucier, Craig McIntosh, John Quinn, and Richard Ssekibuule. Designing and evolving an electronic agricultural marketplace in uganda. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 1–11, 2018.

[63] Christopher Olah. Understanding LSTM Networks, 2015. URL https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[64] Rohit Patel. ACRE – Agricultural Crop Recommendation Engine, 2022.

[65] Rohit Patel, Inavamsi Enaganti, Mayank Ratan Bhardwaj, and Y Narahari. A Data-Driven, Farmer-Oriented Agricultural Crop Recommendation Engine (ACRE). In *Big Data Analytics: 10th International Conference, BDA 2022, Hyderabad, India, December 19–22, 2022, Proceedings*, pages 227–248. Springer, 2023.

[66] Tushar Pawar and Sandip Dighe. Dumped a month ago, tomatoes fetch Maharashtra farmers 'record prices', 6 2023. URL https://timesofindia.indiatimes.com/city/nashik/dumped-a-month-ago-tomatoes-fetch-maha-farmers-record-prices/articleshow/101351584.cms.

[67] John Platt and Alan Barr. Constrained differential optimization. In *Neural Information Processing Systems*, pages 612–621, 1987.

[68] A Priyadharshini, Swapneel Chakraborty, Aayush Kumar, and Omen Rajendra Pooniwala. Intelligent crop recommendation system using machine learning. In *2021 5th international conference on computing methodologies and communication (ICCMC)*, pages 843–848. IEEE, 2021.

[69] S Pudumalar, E Ramanujam, R Harine Rajashree, C Kavya, T Kiruthika, and J Nisha. Crop recommendation system for precision agriculture. In *2016 Eighth International Conference on Advanced Computing (ICoAC)*, pages 32–36. IEEE, 2017.

[70] Sean Ross. 4 countries that produce the most food, April 2023. URL `https://www.investopedia.com/articles/investing/100615/4-countries-produce-most-food.asp`.

[71] KJS Satyasai and Sandhya Bharti. Doubling farmers' income: Way forward. *Rural Pulse*, 14:1–4, 2016.

[72] Mohsen Shahhosseini, Guiping Hu, Isaiah Huber, and Sotirios V Archontoulis. Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. *Scientific reports*, 11(1):1–15, 2021.

[73] Sagarika Sharma, Sujit Rai, and Narayanan C Krishnan. Wheat crop yield prediction using deep LSTM model. *arXiv preprint arXiv:2011.01498*, 2020.

[74] William F. Sharpe. Mutual fund performance. *Journal of Business*, 39(1):119–138, 1966.

[75] William F. Sharpe. The Sharpe Ratio. *Journal of Portfolio Management*, 21(1): 49–58, 1994.

[76] P Chandra Shekara, N Balasubramani, R Sharma, C Shukla, A Kumar, BC Chaudhary, and M Baumann. Farmer's handbook on basic agriculture. *Desai Fruits & Vegetables Pvt. Ltd, Navsari*, 2016.

[77] Zheyuan Ryan Shi, Claire Wang, and Fei Fang. Artificial intelligence for social good: A survey. *arXiv preprint arXiv:2001.01818*, 2020.

[78] P. Sowjanya. Agri-vaahan: An aiml pipeine for agricultural data analytics, 2022.

[79] Helmuth Späth. *One dimensional spline interpolation algorithms*. AK Peters/CRC Press, 1995.

[80] Dariusz Strzkebicki. The electronic marketplace as the element of the agricultural market infrastructure. *Problems of Agricultural Economics*, (1_2015), 2015.

[81] Bin Sun, Liyao Ma, Wei Cheng, Wei Wen, Prashant Goswami, and Guohua Bai. An improved k-nearest neighbours method for traffic time series imputation. In *2017 Chinese Automation Congress (CAC)*, pages 7346–7351. IEEE, 2017.

[82] S Sunder. Financial Express, Jan 2018. URL https://www.financialexpress.com/budget/india-economic-survey-2018-for-farmers-agriculture-gdp-msp-1034266/.

[83] PVS Suryakumar. Bridging tech and agriculture, Jan 2022. URL https://www.thehindubusinessline.com/opinion/bridging-tech-and-agriculture/article38156586.ece.

[84] Earth Observing System. Agricultural cooperatives: Specifics, role, pros, and cons. Technical report, https://eos.com/blog/agricultural-cooperatives/, 2022.

[85] Thomas Van Klompenburg, Ayalew Kassahun, and Cagatay Catal. Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, 177:105709, 2020.

[86] Jaydeep Pawar Vasudev. An innovative deep learning based approach for accurate agricultural crop price prediction, 2023.

[87] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *stat*, 1050:20, 2017.

[88] Vikaspedia. Critical factors to be considered for selection of crops, 2022. URL https://vikaspedia.in/agriculture/crop-production/critical-factors-to-be-considered-for-selection-of-crops.

[89] Nukala Viswanadham, Sridhar Chidananda, Yadati Narahari, and Pankaj Dayama. Mandi electronic exchange: Orchestrating Indian agricultural markets for maximizing social welfare. In *2012 IEEE International Conference on Automation Science and Engineering (CASE)*, pages 992–997. IEEE, 2012.

[90] WorldAtlas. Largest rice-producing countries, 2023. URL https://www.worldatlas.com/articles/largest-rice-producing-countries.html.

[91] WorldAtlas. Top wheat producing countries, 2023. URL https://www.worldatlas.com/articles/largest-rice-producing-countries.html.

[92] Katarzyna Woźnica and Przemysław Biecek. Does imputation matter? benchmark for predictive models. *arXiv preprint arXiv:2007.02837*, 2020.

[93] Jiaxuan You, Xiaocheng Li, Melvin Low, David Lobell, and Stefano Ermon. Deep gaussian process for crop yield prediction based on remote sensing data. In *Thirty-First AAAI conference on artificial intelligence*, pages 4559–4565, 2017.

[94] Dabin Zhang, Shanying Chen, Ling Liwen, and Qiang Xia. Forecasting agricultural commodity prices using model selection framework with time series features and forecast horizons. *IEEE Access*, 8:28197–28209, 2020.

[95] Zhanhao Zhang. A survey of online auction mechanism design using deep learning approaches. Technical Report arXiv:2110.06880v1, arXiv Preprint, 2021.

[96] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1: 57—81, 2020.

राष्ट्रीय कृषि और ग्रामीण विकास बैंक, मुंबई

NATIONAL BANK FOR AGRICULTURE AND RURAL

www.nabard.org /nabardonline