Price Prediction of Potato

Harshita Gupta – 210911224
Aastha Sinha – 210911320
Charu Yadav – 210911396

Abstract - This report evaluates the efficacy of various machine learning models including LSTM, KNN, and a pre-trained decision tree-based Random Forest regressor in predicting the prices of potatoes in India. By utilizing historical datasets encompassing potato prices, regional statistics, and rainfall patterns, the study aims to provide accurate price trend forecasts to aid stakeholders in the agricultural sector. The models leverage distinct aspects of the data, such as temporal sequences, regional variations, and environmental influences, to predict future pricing dynamics. The findings are intended to support decision-making processes for farmers, traders, and policymakers by improving their understanding of market trends and enabling more informed economic choices in the volatile agriculture market.

Introduction

The volatility of vegetable prices presents significant challenges for stakeholders in the agricultural sector, particularly in regions like India where farming constitutes a primary economic activity. Accurate price predictions are crucial for ensuring profitability and sustainability in this sector. Recent advancements in machine learning offer promising tools for enhancing the accuracy of such predictions. This study focuses on evaluating the effectiveness of various predictive models, including Long Short-Term Memory (LSTM), K-Nearest Neighbors (KNN), and a Random Forest regressor utilizing pre-trained decision trees, specifically tailored to forecast potato prices. Utilizing historical data, the research integrates factors like regional economic conditions, climate variations, and historical pricing trends. The objective is to explore the comparative effectiveness of these models in capturing and predicting the dynamic and often nonlinear interactions among these factors. Through rigorous training and testing of these models, this report aims to uncover insights that could lead to more robust forecasting methods, thereby aiding stakeholders from farmers to policymakers in making informed decisions.

Literature Review:

Zhang et al. [1] compare the performance of various classifiers for ECG quality assessment, focusing on Random Forest and SVM variants. While this study is unrelated to agricultural commodities, it exemplifies the use of machine learning in different domains.

Madaan et al. [2] present a comprehensive framework for forecasting agricultural commodity prices, incorporating 29 time series features across several models including ANN, SVR, and ELM. The model selection framework aims to enhance forecast accuracy through systematic feature analysis and reduction. Despite advancements in forecast accuracy, the limitation to three models suggests the potential for exploring more powerful classifiers.

Vijayalaxmi et al. [3] explore the correlation between temperature changes and vegetable prices using machine learning techniques. By employing web scraping to collect comprehensive data, they apply various regression models to predict price fluctuations, providing valuable insights for farmers in managing crop planning and mitigating the impact of hyperinflation.

Warnakulasooriya et al. [4] tackle the issue of vegetable wastage and fluctuating demand in Sri Lanka's retail sector with a comprehensive framework that integrates blockchain technology for supply chain verification. This approach enhances transparency and allows for better supplier selection, while demand and price prediction models embedded in the system help manage supply stability and reduce wastage, promoting overall sustainability in the retail sector.

Gamage et al. [5] discuss the "Smart Agriculture Prediction System for Vegetables Grown in Sri Lanka," focusing on mitigating the lack of knowledge about yield and price that affects farmers' decision-making. Utilizing a mobile application, SMS, and API, the system employs machine learning algorithms like elastic net, ridge, and multilinear regression for yield and price predictions, and a genetic algorithm for crop optimization. This integrated approach aims to provide farmers with actionable insights for improved agricultural productivity.

Ramesh et al. [6] employ Seasonal ARIMA modeling in their study to forecast agricultural prices of fruits and vegetables in Bangalore, India. Utilizing the Box-Jenkins technique, they focus on identifying and capturing seasonal trends and variations, thus aiding stakeholders in decision-making and risk management through accurate future price forecasts.

Nalwanga et al. [7] introduce a fuzzy logic-based model within the IoT framework for predicting tomato prices in Kenya. Utilizing four years of climatic data, the model shows promising accuracy in capturing the complex dynamics of price fluctuations and environmental factors, although it focuses solely on one vegetable type.

Sharma et al., [8] propose a novel price prediction model using supervised machine learning algorithms, focusing on fruits, vegetables, and pulses according to weather patterns. Starting with basic KNN and progressing to a more refined pre-trained Decision Tree Regressor, the model shows notable accuracy improvements, achieving up to 91.70% in predictive performance, emphasizing its potential for enhancing financial forecasting in agriculture.

Fan et al. [9] describe a robust LSTM model for forecasting onion and potato prices in India, emphasizing resistance against MSP fluctuations. The paper, however, notes a drawback in its excessive length and vague details on its anomaly detection model, which compares mandi and retail prices.

Ma Wei et al. [10] propose a GNN-RNN framework for crop yield prediction, addressing the challenges posed by the interplay of weather, soil quality, and geographical factors. The model aims to improve the generalizability of yield predictions that have traditionally been limited to short timespans or specific regions.

Bhardwaj et al. [11] focus on enhancing produce price forecasting for small and marginal farmers in India using collaborative filtering and adaptive nearest neighbors. Their method addresses the shortcomings of current forecasting systems by predicting exact prices and offering longer-term forecasts relevant to agricultural cycles.

Klompenburg et al. [12] utilize a deep learning-based method incorporating GNNs and CNNs for agricultural crop price prediction. By leveraging key determinants such as historical price data and climatic conditions, the approach demonstrates a significant improvement in forecasting accuracy, particularly for potatoes and tomatoes.

Zhang et al. [13] conduct a systematic literature review on crop yield prediction, analyzing 50 studies and identifying key features and algorithms used in the field. The review highlights the prevalent use of ANNs and the potential of CNNs and LSTMs in deep learning applications for crop yield forecasting.

Jin et al. [14] review an innovative STL-LSTM method for forecasting vegetable prices, notably cabbage and radish, in Chinese and Korean markets. By integrating Seasonal Trend Loess preprocessing with LSTM deep learning, the study leverages meteorological and market data, achieving accuracies of 92% and 88% respectively. The approach effectively addresses market volatility and offers insights into vegetable price patterns, though it acknowledges the need for further refinement in instances of lower accuracy.

Madaan et al., [15] again present an LSTM model for forecasting agricultural commodity prices in India, particularly onions and potatoes, emphasizing the model's robustness against MSP fluctuations and including an anomaly detection component.

Pre-Processing:

Data Sources

The analysis utilized three primary data sources:

1. Potato Prices (potato.csv): This dataset contains records of potato prices from various centers on specific dates. Each record includes the date, center name, commodity name, and price.
2. State and District Information (states.csv): This dataset maps each district to its corresponding state, providing a geographical hierarchy useful for aggregating data at the state level.

3. Rainfall Data (rainfall_new.csv): Detailed monthly and annual rainfall statistics are provided for each state/union territory across different years, allowing for temporal and spatial rainfall analysis.

Data Processing Steps

Data Loading and Initial Cleaning:

- Each CSV file was loaded into a Pandas DataFrame.
- The states.csv file was cleaned to remove any unnecessary columns, focusing solely on the state and district information.

Data Integration:

- The Centre_Name from potato.csv was matched with the district from states.csv to attribute each potato price record to a specific state.
- This required converting both columns to lowercase to ensure accurate matching regardless of text case discrepancies.

Temporal Alignment:

- The date in potato.csv was parsed to extract the year and month, facilitating a direct comparison with the rainfall data.
- This temporal data was crucial for correlating potato prices with specific monthly rainfall measures.

Rainfall Data Matching:

- Rainfall data was joined with the potato price data using the state, year, and month as keys.
- A dynamic column selection method was implemented to fetch the correct monthly rainfall amount based on the extracted month from the potato data.

Final Dataset Compilation:

- A consolidated DataFrame was created, containing the state, date, matched monthly rainfall, and potato price.
- Rows containing null values in any of the crucial columns (State, Rainfall, Price) were removed to ensure data quality and accuracy.

Output Generation: The cleaned and integrated data was saved back to a CSV file for further analysis or reporting.

Tools Used

- Python: The primary programming language used for data manipulation and processing.

- Pandas: A Python library employed for its powerful data structures and functions, facilitating easy data manipulation and analysis.

Methodology:

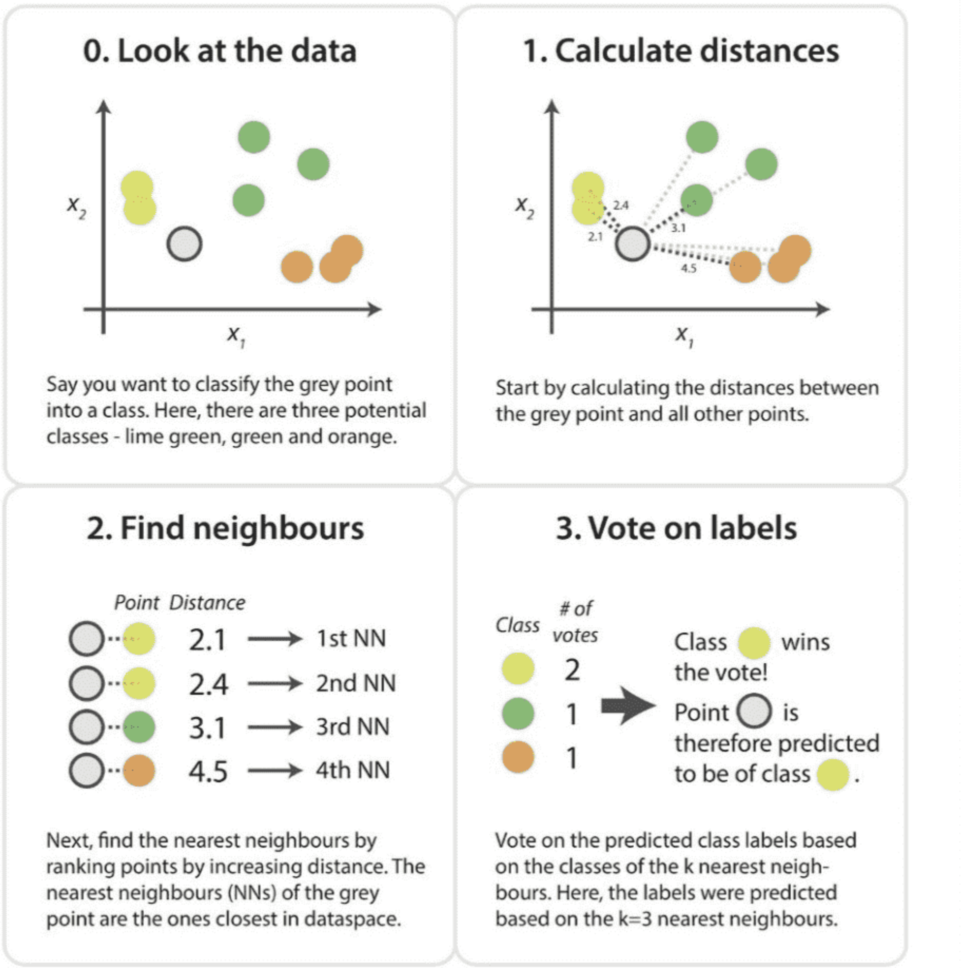
K-Nearest Neighbors:

To predict potato prices, we employed the K-Nearest Neighbors (KNN) regression model explained via Figure A, a non-parametric method well-suited for handling complex relationships without assumptions about data distribution.

Data Preprocessing: The dataset comprised features such as state, date, rainfall, and potato prices. We encoded the 'state' categorical data numerically and extracted year and month from the 'date' feature. The data was then split into training and testing sets, with 80% used for training and 20% for testing.

Feature Scaling: Given KNN's sensitivity to the scale of data due to its reliance on distance calculations, we applied standard scaling to the features to ensure all had equal influence.

Model Training: We used a pipeline approach in sklearn to streamline the scaling and model training process. The KNN model was initialized with (k = 5), a commonly chosen starting point. The model was trained on the scaled features and price data.



0. Look at the data

Say you want to classify the grey point into a class. Here, there are three potential classes - lime green, green and orange.

1. Calculate distances

Start by calculating the distances between the grey point and all other points.

2. Find neighbours

Next, find the nearest neighbours by ranking points by increasing distance. The nearest neighbours (NNs) of the grey point are the ones closest in dataspace.

3. Vote on labels

Vote on the predicted class labels based on the classes of the k nearest neighbours. Here, the labels were predicted based on the k=3 nearest neighbours.

Figure A: Explanation of KNN

Pre – Trained Decision Tree Regressor based Random Forest Regression

Data Preprocessing: The dataset comprises observations of potato prices from various states in India, alongside associated rainfall data and dates. Initial data preprocessing involved:

- Converting the date column from string format to Python's datetime, enabling extraction of temporal features such as year and month.
- Calculating lagged rainfall features to capture past rainfall data (one, two, and three months prior) and computing a rolling average of rainfall over the previous three months to incorporate historical climate trends.
- Handling categorical data by employing one-hot encoding for the state column, facilitating the inclusion of geographical variations in the model.

Feature Engineering:

- Temporal features (year and month) were extracted to account for seasonal variations in potato prices.

- Lagged rainfall features and rolling averages were integrated to reflect the delayed effects of weather conditions on agricultural outputs.

Model Development:

Two models were developed and evaluated:

1. Decision Tree Regressor: Served as a preliminary model, offering a baseline by capturing non-linear relationships between features and potato prices.
2. Random Forest Regressor: An ensemble model using multiple decision trees to reduce overfitting and improve predictive accuracy. This model was further refined with parameters informed by the initial Decision Tree model.

Figure B explains how the model works as a whole.

Model Training and Evaluation:

- The dataset was split into training (80%) and testing (20%) sets to evaluate model performance.
- Models were trained on the training set and evaluated using the (R^2) metric, which describes the proportion of variance in the dependent variable that is predictable from the independent variables.
- Feature importance was analyzed to identify the most influential variables impacting potato prices.

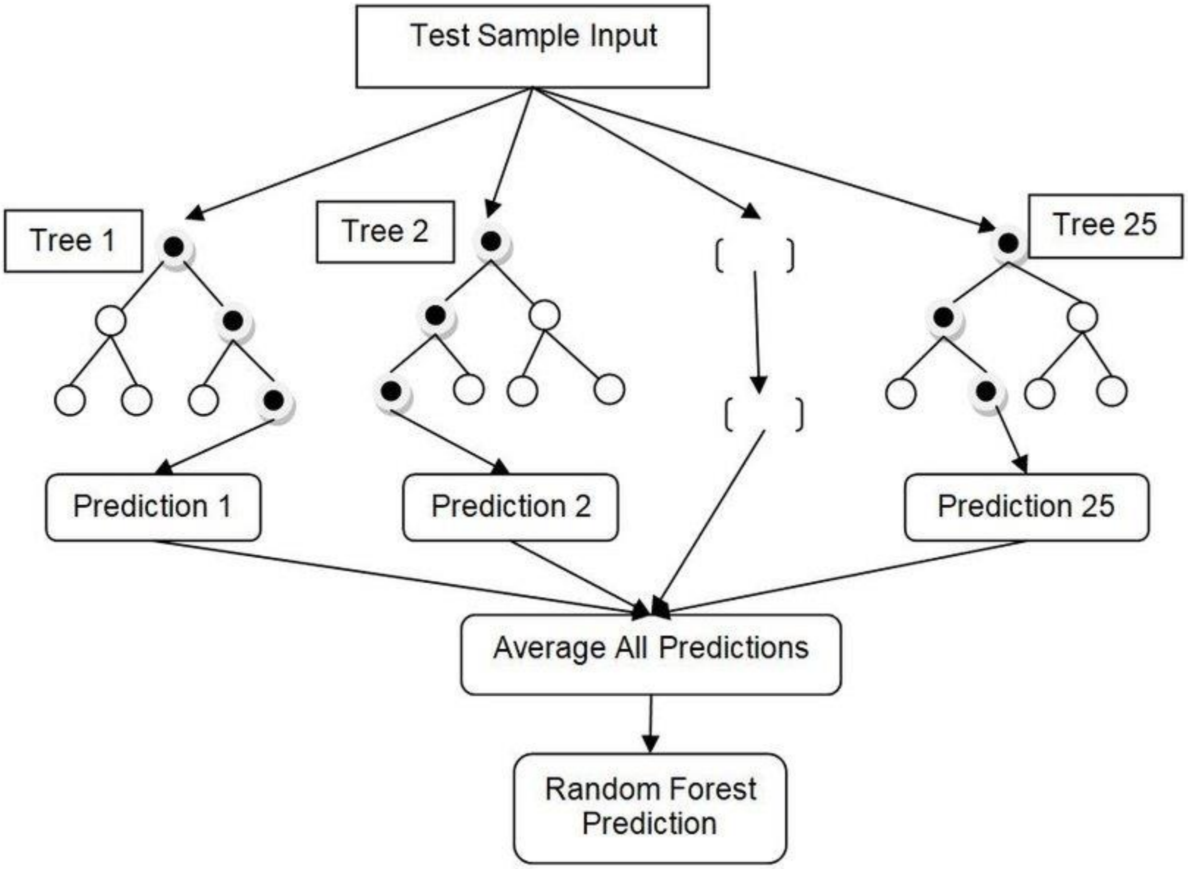


Figure B: Explanation of Random Forest Regressor

Long Short-Term Memory Regression Model

The objective of this study was to predict daily potato prices using historical data on prices and rainfall. The methodology encompassed data preprocessing, model development, training, and evaluation, implemented using Python and TensorFlow. Figure C explains how the model works.

Data Preprocessing: The dataset comprised entries of daily rainfall and potato prices across various states. The 'date' field was parsed and formatted correctly to ensure chronological ordering. Categorical data, such as 'state', was encoded numerically. To handle any redundancy, duplicate records were removed. Data was then aggregated by date to calculate daily averages of rainfall and prices, making the dataset uniform and easier to model. The features 'rainfall' and 'price' were normalized using MinMaxScaler to scale the values between 0 and 1, facilitating a more stable training process.

Model Development: An LSTM model was chosen for its ability to capture temporal dependencies in time series data, which is crucial for accurate forecasting in dynamic conditions such as weather and economic factors affecting potato prices. The model architecture consisted of:

- Two LSTM layers with 50 units each to learn the sequence representation, interspersed with Dropout layers (20% rate) to prevent overfitting.
- A Dense output layer to predict the continuous value of potato prices.

Training: The model was trained using an 80-20 split of the data into training and testing sets. The sequences of 30 days of data were used to predict the price on the next day. The model was trained for 50 epochs with a batch size of 32, using the Adam optimizer and mean squared error as the loss function.

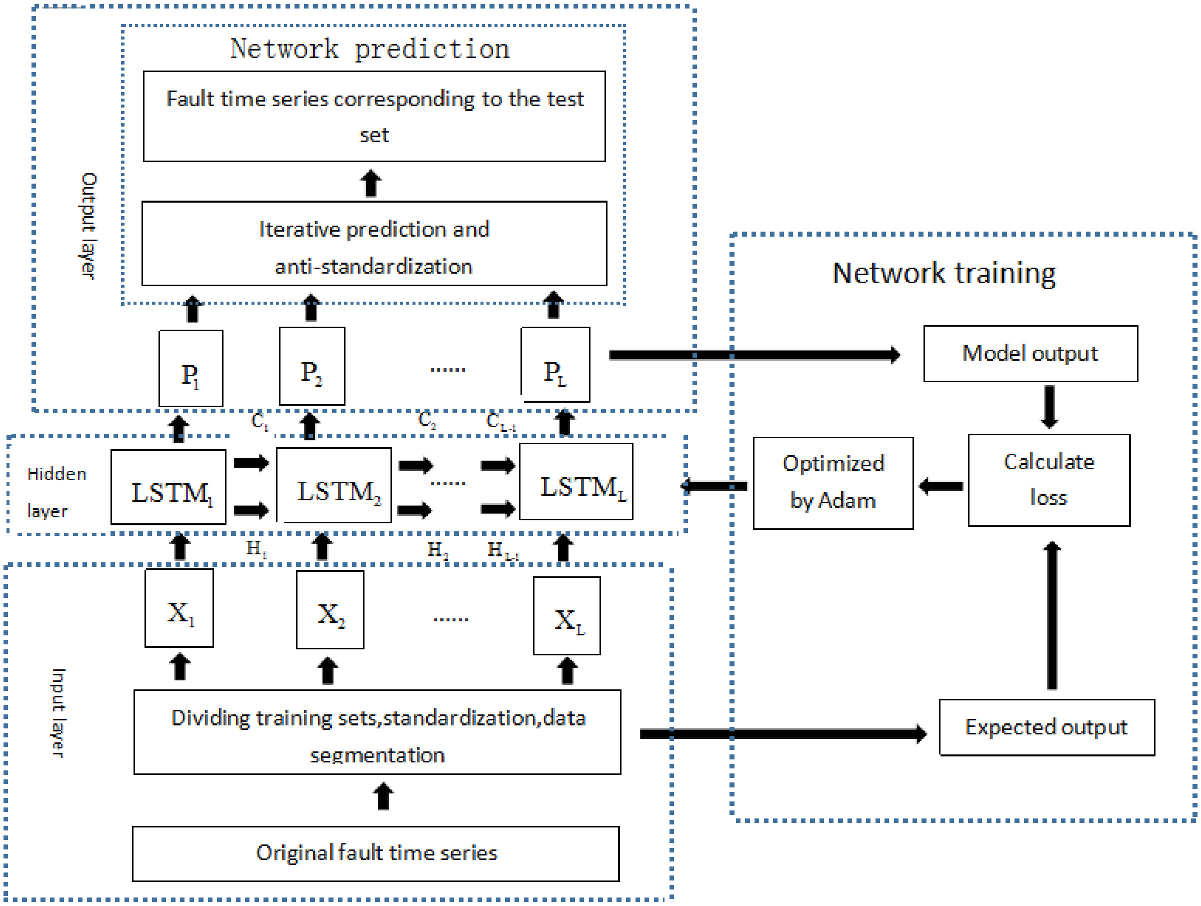


Figure C: Explanation of LSTM model

Results

K-Nearest Neighbors

Model Evaluation: The model's performance was assessed using Mean Squared Error (MSE) and R-squared metrics where R^2=0.79, indicating accuracy of 79%. Further insights were gained through visual assessments using scatter plots of actual vs. predicted prices and histograms of residuals represented by Figure1 and Figure 2 respectively. These visualizations helped evaluate the accuracy and distribution of errors.

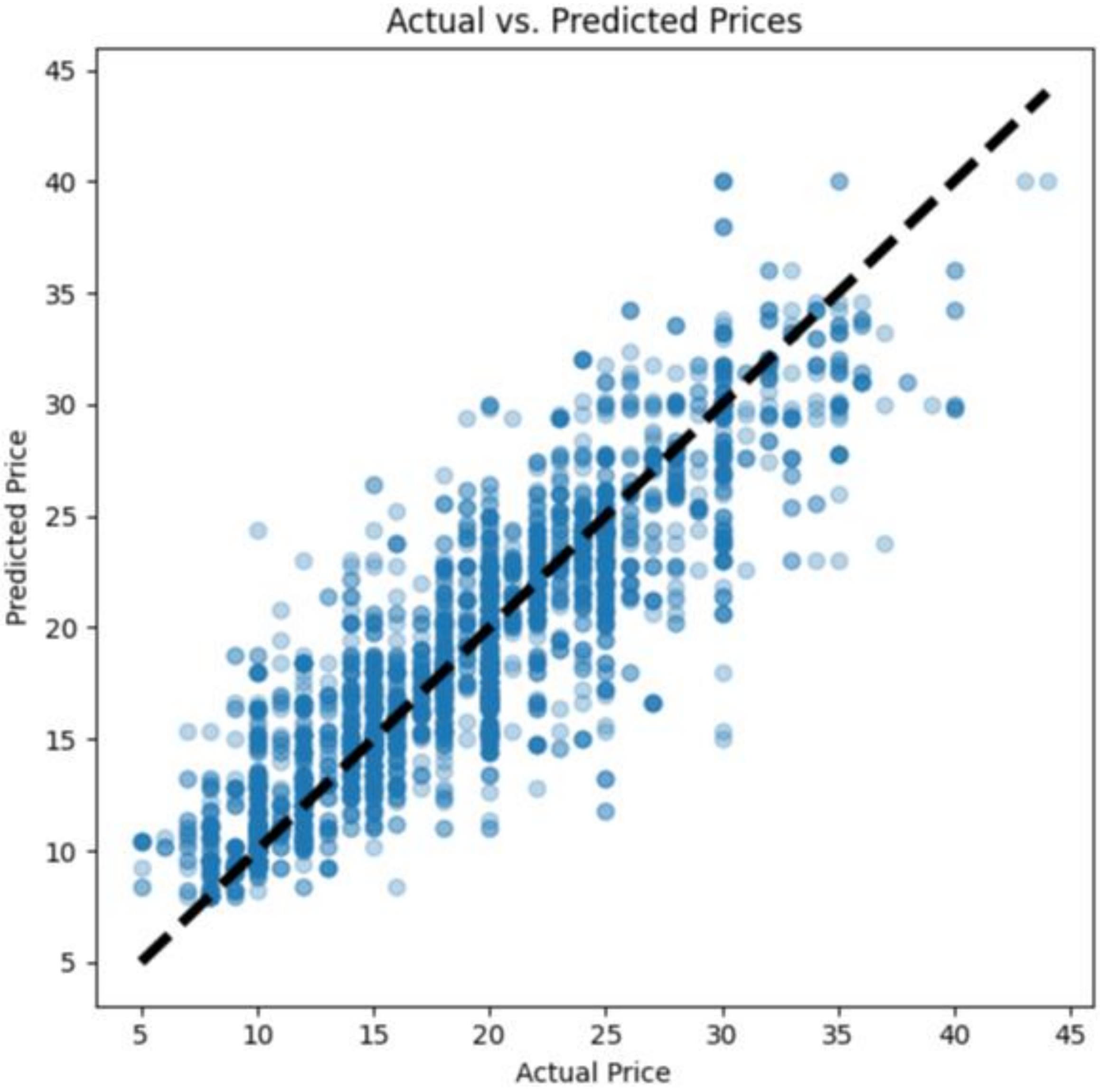


Figure 1: Scatterplot for KNN

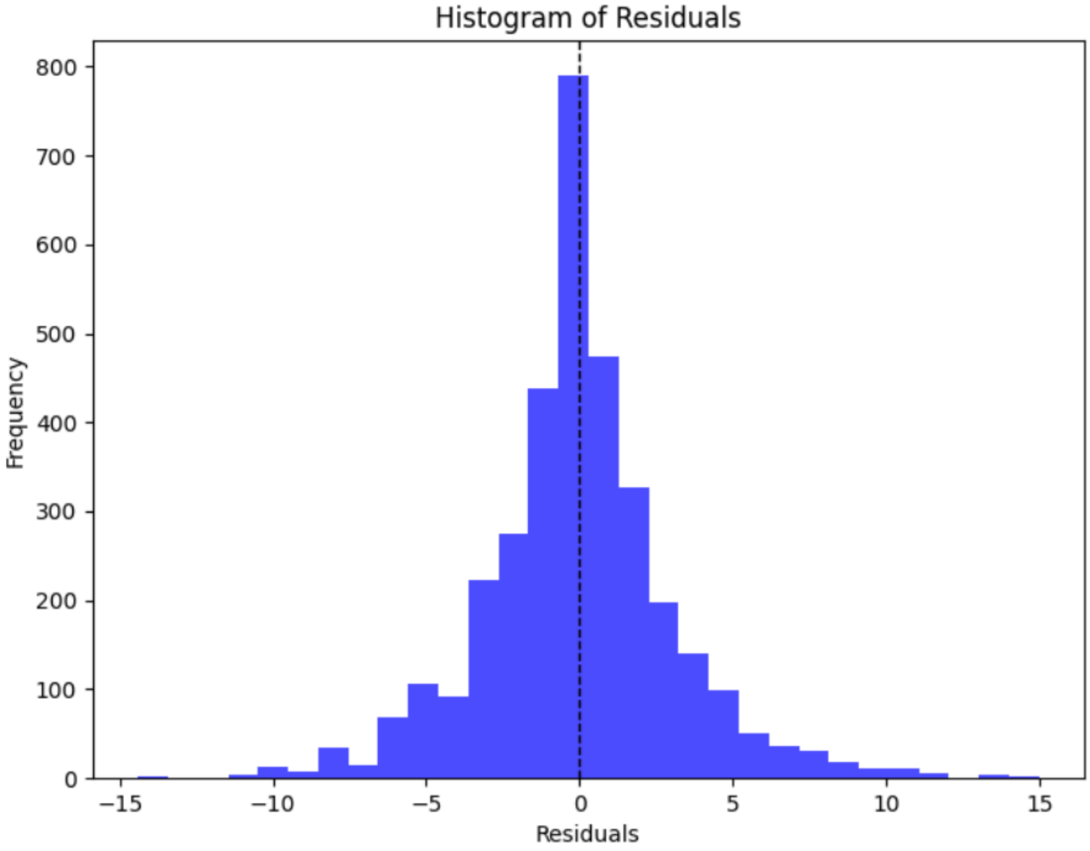


Figure 2: Residual Histogram for KNN

Pre – Trained Decision Tree Regressor based Random Forest Regression

Visualization: Key visualizations included histograms to compare the distribution of actual and predicted prices and scatter plots to illustrate the correlation between predicted and actual values respected by Figure 3 and Figure 4. The model's performance was assessed using Mean Squared Error (MSE) and R-squared metrics where R^2=0.82 indicating accuracy of 82%.

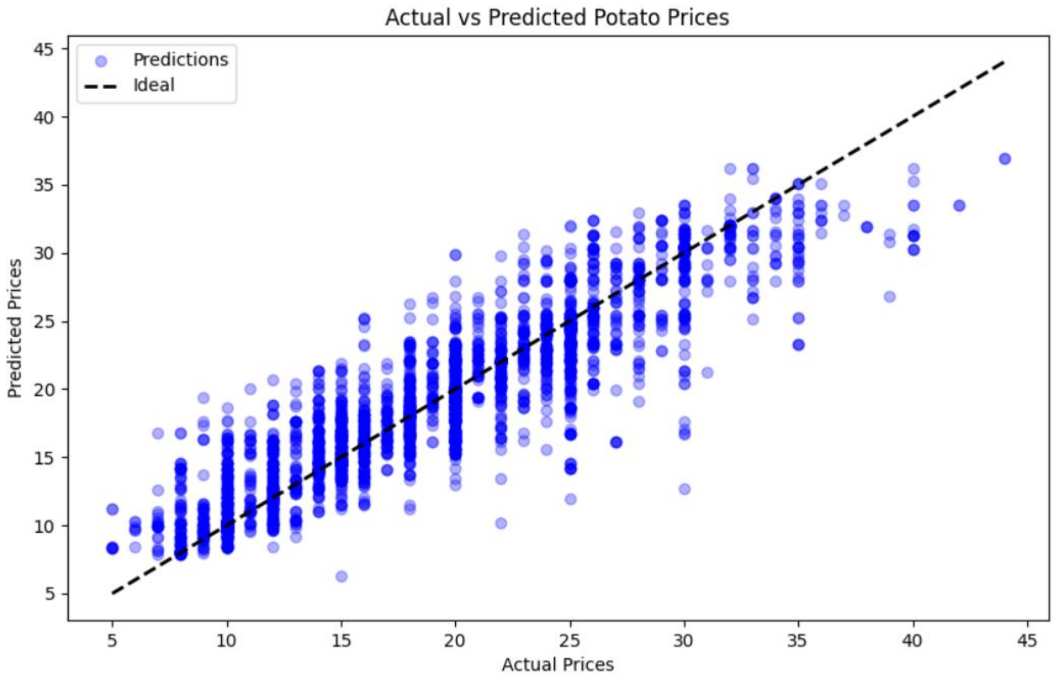


Figure 3: Scatterplot for pre-trained DT based Random Forest

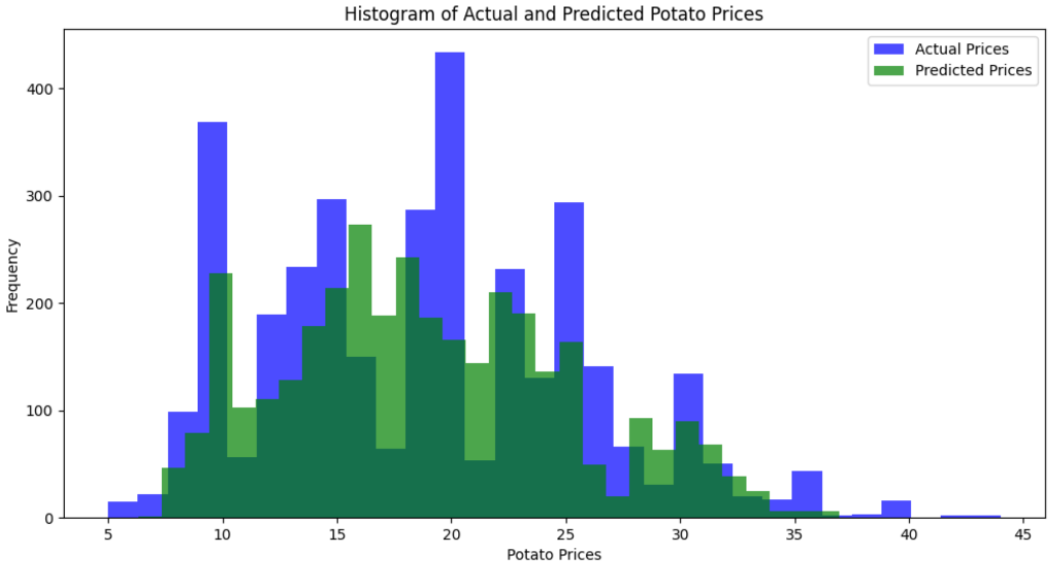


Figure 4: Histogram for pre-trained DT based Random Forest

Long Short-Term Memory Regression Model

Evaluation: Model performance was assessed using the test dataset. Key metrics included Mean Squared Error (MSE) and R^2 where R^2=0.97, indicating accuracy of 97%, providing insights into the model's predictive accuracy and error rate in percentage terms.

Results Visualization: The results were visualized through scatter plots comparing actual and predicted prices and histograms showing the distribution of both represented by Figure 5 and Figure 6. These visualizations helped in understanding the model's performance and error characteristics visually.

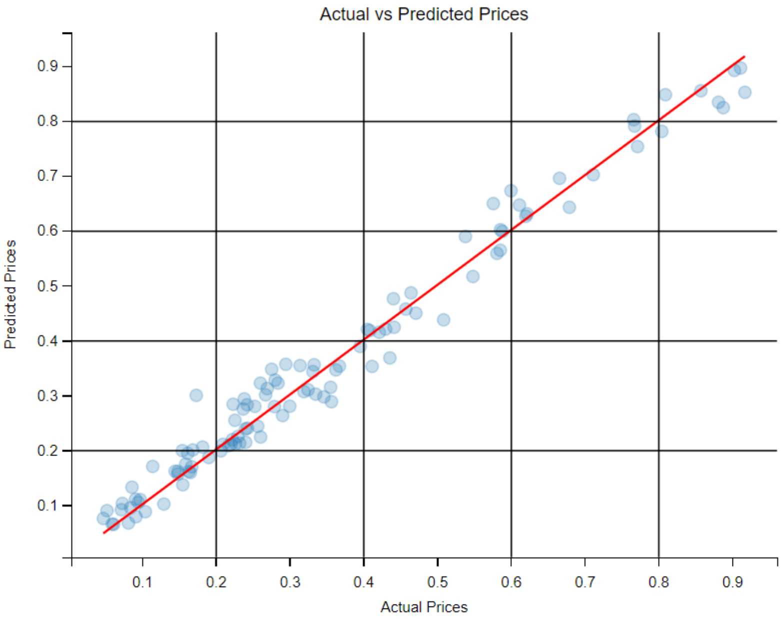


Figure 5: Scatterplot for LSTM

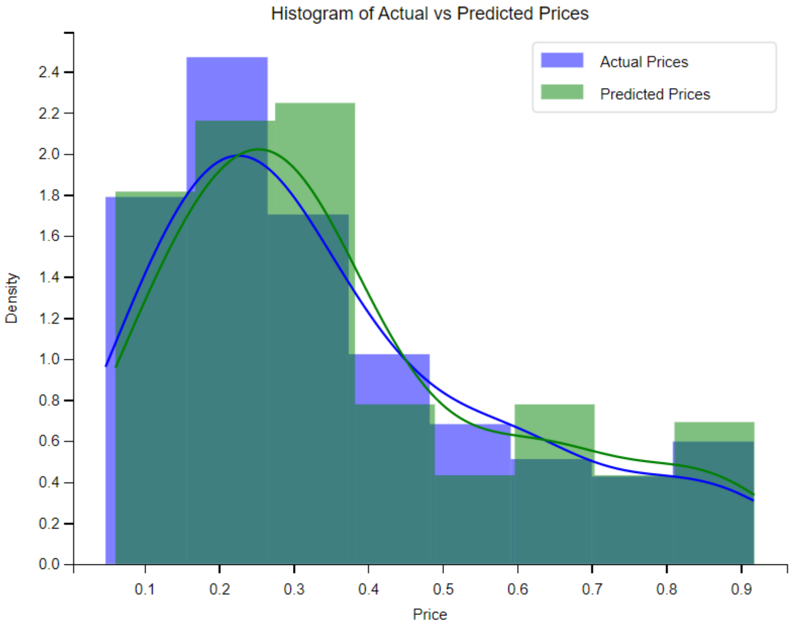


Figure 6: Histogram for LSTM

Conclusion

The comparative study of LSTM, KNN, and Random Forest models in predicting potato prices in India highlights the potential of machine learning in transforming agricultural economics. Each model demonstrated unique strengths in handling specific types of data and forecasting challenges. The Random Forest model, with its decision tree basis, was particularly effective in managing the non-linearity of price influences, while LSTM excelled in capturing temporal price dynamics influenced by external factors such as weather conditions. Meanwhile, KNN provided valuable insights when dealing with localized data variations. These findings suggest that an integrated approach employing multiple models could further enhance predictive accuracy. Moving forward, expanding the datasets to include more diverse variables and extending the forecast horizon could potentially improve model robustness, offering stakeholders valuable tools for navigating the complexities of agricultural markets.

References:

[1] Zhang, Yatao & Wei, Shoushui & Zhang, Li & Liu, Chengyu. (2019). Comparing the Performance of Random Forest, SVM and Their Variants for ECG Quality Assessment Combined with Nonlinear Features. Journal of Medical and Biological Engineering. 39. 10.1007/s40846-018-0411-0.

[2] Madaan, Lovish & Sharma, Ankur & Khandelwal, Praneet & Goel, Shivank & Singla, Parag & Seth, Aaditeshwar. (2019). Price forecasting & anomaly detection for agricultural commodities in India. 52-64. 10.1145/3314344.3332488.

[3] Vijayalaxmi, K., et al. "Vegetable Price Prediction Against Temperature Changes Using Machine Learning Techniques." Sreenidhi Institute of Science & Technology(A), Hyderabad.

[4] Warnakulasooriya, Hashini, et al. "Supermarket Retail – Based Demand and Price Prediction of Vegetables." Faculty of Computing, Sri Lanka Institute of Information Technology, Colombo, Sri Lanka.

[5] Gamage, Rashmika, et al. "Smart Agriculture Prediction System for Vegetables Grown in Sri Lanka." Department of Computer Science and Software Engineering, Sri Lanka Institute of Information Technology, Malabe 10115, Sri Lanka.

[6] Ramesh, Dharavath, and Ekaansh Khosla. "Seasonal ARIMA to forecast fruits and vegetable agricultural prices." Department of Computer Science & Engineering, Indian Institute of Technology (ISM), Dhanbad, Jharkhand, India.

[7] Nalwanga, Rosemary, and Ayalew Belay. "Fuzzy Logic based Vegetable Price prediction in IoT." African Center of Excellence in IoT, University of Rwanda, Rwanda.

[8] Sharma, Chandan, et al. "Price Prediction Model of fruits, Vegetables and Pulses according to Weather." Amity School of Engineering and Technology, Amity University, Noida, Uttar Pradesh, and Dayalbagh Educational Institute, Agra.

[9] Fan, Joshua, et al. "A GNN-RNN Approach for Harnessing Geospatial and Temporal Information: Application to Crop Yield Prediction." Department of Computer Science, Cornell University, USA; Department of Applied Economics & Management, Cornell University, USA.

[10] Ma, Wei, et al. "An Interpretable Produce Price Forecasting System for Small and Marginal Farmers in India using Collaborative Filtering and Adaptive Nearest Neighbors." Carnegie Mellon University, CoolCrop.

[11] Bhardwaj, Mayank Ratan, et al. "An Innovative Deep Learning Based Approach for Accurate Agricultural Crop Price Prediction."

[12] Klompenburg, Thomas van, Ayalew Kassahun, and Cagatay Catal. "Crop yield prediction using machine learning: A systematic literature review." Information Technology Group, Wageningen University & Research, Wageningen, the Netherlands; Department of Computer Engineering, Bahcesehir University, Istanbul, Turkey.

[13] Zhang, Dabin, et al. "Forecasting Agricultural Commodity Prices Using Model Selection Framework With Time Series Features and Forecast Horizons." College of Mathematics and Informatics, South China Agricultural University, Guangdong, China.

[14] Jin, Dong. "Forecasting of Vegetable Prices using STL-LSTM Method." Department of Computer Engineering, Sejong University, Seoul, South Korea.

[15] Madaan, Lovish, et al. "Price Forecasting & Anomaly Detection for Agricultural Commodities in India." IIT Delhi.