暨南大學
JINAN UNIVERSITY

本科生课程论文

Title: **Predicting Food Prices using Data Mining: A Case Study of Ministry Data in Cambodia**

Student Name: VAN SONGHIENG

Student ID Number: 2020059118

Major: International School (CST)

Course: Data Mining

Instructor: Prof,Zhu WeiHeng

Date: 2023-06-21

Predicting Food Prices using Data Mining: A Case Study of Ministry of Statistics Data in Cambodia

Jinan University
VAN SONGHIENG, 2020059118
songheingvan@gmail.com

abstract- In a world where food prices can dictate socio-economic wellbeing, accurate prediction of these trends is a game-changer. This study delves into the heart of Cambodia, using data from the Ministry of Statistics, Agricultural Marketing Office, Ministry of Agriculture, Forestry and Fisher to predict food prices with cutting-edge data mining techniques. Five machine learning models go head-to-head in a contest of precision. The crown was claimed by random Forest Regressor, displaying unparalleled prediction accuracy. This promising finding is a milestone for food security and economic strategizing in Cambodia, heralding a future where advanced technology pairs with valuable data to mitigate the unpredictability of food prices.

I. Introduction

Food price prediction plays a pivotal role in the economic sustainability of any country, and Cambodia is no exception. The significance of accurately predicting food prices in Cambodia is multifaceted. On the one hand, it aids in ensuring food security by enabling effective planning and management of food supply chains. it helps policymakers to foresee potential inflationary pressures and devise appropriate strategies to curb them. These predictions can also assist farmers in making informed decisions about what crops to plant and when, thereby potentially maximizing their yields and profits.

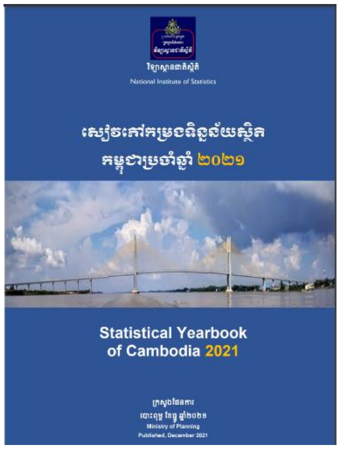
The field of data mining has a lot to offer in this regard. Data mining is the process of extracting meaningful information from large data sets through the use of various techniques. With its potential to reveal hidden patterns and correlations in vast amounts of data, data mining has emerged as a powerful tool in predicting food prices. By applying data mining techniques to food price data, it's possible to anticipate future price trends and make decisions accordingly.

The uniqueness of this study lies in the source of the data used for the food price prediction - **The Ministry of Statistics in Cambodia, World Food Organization, Agricultural Marketing Office, Ministry of Agriculture, Forestry and Fisher**. The Ministry collects and compiles a vast range of economic data, including food prices, which are updated regularly and maintained meticulously. This source of data offers a rich and accurate repository of historical price information that can be invaluable for predictive analyses.

The use of the Ministry's data, combined with the power of data mining techniques, has the potential to revolutionize the way food price prediction is conducted in Cambodia. This approach could not only benefit the local economy and agricultural sector but could also set a precedent for other developing nations grappling with similar issues.

II. Methodology

1) Data Collection



The secondary data for this study is procured from a comprehensive and robust dataset offered by the Ministry of Statistics in Cambodia. This dataset is rich with information on food prices, covering several years of data, and featuring monthly average prices for a wide range of food items across diverse regions within the country.

Data Features: Each record in the dataset represents the monthly average price for a specific food item in a particular region. The dataset is multi-dimensional with several features as explained below:

- `adm0_id (44) & adm0_name (Cambodia)`: These signify the identifier and the name of the country respectively. For this study, the country is Cambodia.

- `adm1_id (806) & adm1_name (Phnom Penh)`: These represent the identifier for a subnational region or administrative unit and its name. In this instance, it's the capital city, Phnom Penh.

- `mkt_id (637) & mkt_name (Phnom Penh)`: These parameters signify the identifier for the market and its name. Here, the market is also Phnom Penh.

- `cm_id (165) & cm_name (Rice - mixed, low quality)`: These are the identifier for the commodity and the name of the commodity. In this case, it's a specific variety of rice.

- `cur_name (KHR)`: This feature indicates the currency in which the price is measured - the Cambodian Riel (KHR) in this dataset.

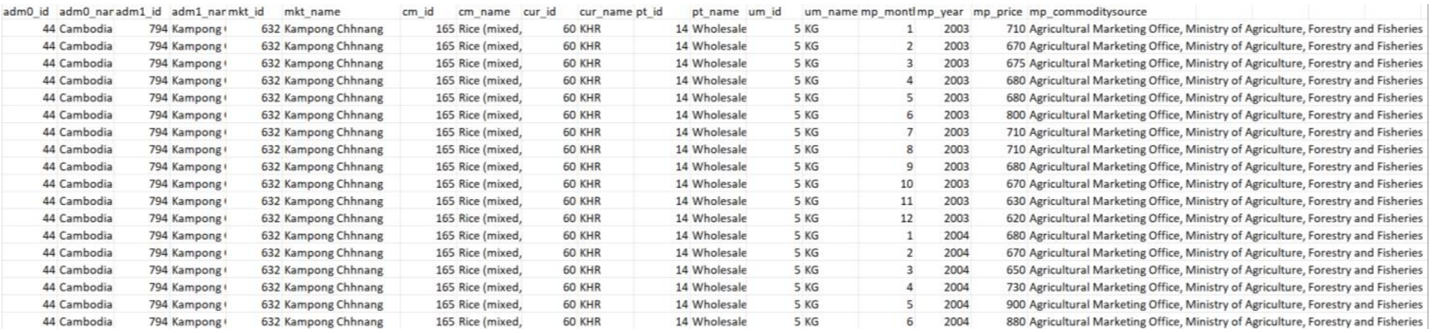
- `pt_id (14) & pt_name (Wholesale)`: These denote the identifier for the price type and the type of price. In this dataset, the prices are wholesale prices.

- `um_id (5) & um_name (KG)`: These are the identifier for the unit of measure and the unit of measure itself, which is kilograms (KG) in this dataset.

- `mp_month (1) & mp_year (2003)`: These represent the month and year for the price data, indicating the time aspect of the data.

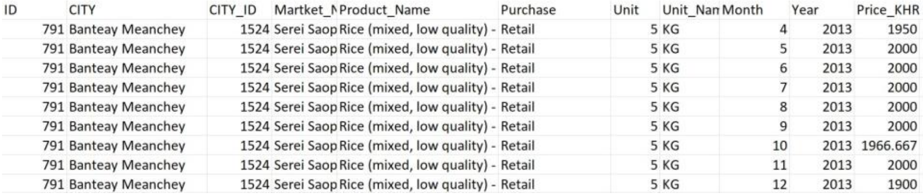
- `mp_price (800)`: This is the price of the commodity, which is 800 KHR in this example.

- `mp_commoditysource (Agricultural Marketing Office, Ministry of Agriculture, Forestry and Fisheries)`: This denotes the source of the commodity price data.



adm0_id	adm0_nar	adm1_id	adm1_nar	mkt_id	mkt_name	cm_id	cm_name	cur_id	cur_name	pt_id	pt_name	um_id	um_name	mp_montl	mp_year	mp_price	mp_commoditysource
44	Cambodia	794	Kampong (	632	Kampong Chhnang	165	Rice (mixed,	60	KHR	14	Wholesale	5	KG	1	2003	710	Agricultural Marketing Office, Ministry of Agriculture, Forestry and Fisheries
44	Cambodia	794	Kampong (	632	Kampong Chhnang	165	Rice (mixed,	60	KHR	14	Wholesale	5	KG	2	2003	670	Agricultural Marketing Office, Ministry of Agriculture, Forestry and Fisheries
44	Cambodia	794	Kampong (	632	Kampong Chhnang	165	Rice (mixed,	60	KHR	14	Wholesale	5	KG	3	2003	675	Agricultural Marketing Office, Ministry of Agriculture, Forestry and Fisheries
44	Cambodia	794	Kampong (	632	Kampong Chhnang	165	Rice (mixed,	60	KHR	14	Wholesale	5	KG	4	2003	680	Agricultural Marketing Office, Ministry of Agriculture, Forestry and Fisheries
44	Cambodia	794	Kampong (	632	Kampong Chhnang	165	Rice (mixed,	60	KHR	14	Wholesale	5	KG	5	2003	680	Agricultural Marketing Office, Ministry of Agriculture, Forestry and Fisheries
44	Cambodia	794	Kampong (	632	Kampong Chhnang	165	Rice (mixed,	60	KHR	14	Wholesale	5	KG	6	2003	800	Agricultural Marketing Office, Ministry of Agriculture, Forestry and Fisheries
44	Cambodia	794	Kampong (	632	Kampong Chhnang	165	Rice (mixed,	60	KHR	14	Wholesale	5	KG	7	2003	710	Agricultural Marketing Office, Ministry of Agriculture, Forestry and Fisheries
44	Cambodia	794	Kampong (	632	Kampong Chhnang	165	Rice (mixed,	60	KHR	14	Wholesale	5	KG	8	2003	710	Agricultural Marketing Office, Ministry of Agriculture, Forestry and Fisheries
44	Cambodia	794	Kampong (	632	Kampong Chhnang	165	Rice (mixed,	60	KHR	14	Wholesale	5	KG	9	2003	680	Agricultural Marketing Office, Ministry of Agriculture, Forestry and Fisheries
44	Cambodia	794	Kampong (	632	Kampong Chhnang	165	Rice (mixed,	60	KHR	14	Wholesale	5	KG	10	2003	670	Agricultural Marketing Office, Ministry of Agriculture, Forestry and Fisheries
44	Cambodia	794	Kampong (	632	Kampong Chhnang	165	Rice (mixed,	60	KHR	14	Wholesale	5	KG	11	2003	630	Agricultural Marketing Office, Ministry of Agriculture, Forestry and Fisheries
44	Cambodia	794	Kampong (	632	Kampong Chhnang	165	Rice (mixed,	60	KHR	14	Wholesale	5	KG	12	2003	620	Agricultural Marketing Office, Ministry of Agriculture, Forestry and Fisheries
44	Cambodia	794	Kampong (	632	Kampong Chhnang	165	Rice (mixed,	60	KHR	14	Wholesale	5	KG	1	2004	680	Agricultural Marketing Office, Ministry of Agriculture, Forestry and Fisheries
44	Cambodia	794	Kampong (	632	Kampong Chhnang	165	Rice (mixed,	60	KHR	14	Wholesale	5	KG	2	2004	670	Agricultural Marketing Office, Ministry of Agriculture, Forestry and Fisheries
44	Cambodia	794	Kampong (	632	Kampong Chhnang	165	Rice (mixed,	60	KHR	14	Wholesale	5	KG	3	2004	650	Agricultural Marketing Office, Ministry of Agriculture, Forestry and Fisheries
44	Cambodia	794	Kampong (	632	Kampong Chhnang	165	Rice (mixed,	60	KHR	14	Wholesale	5	KG	4	2004	730	Agricultural Marketing Office, Ministry of Agriculture, Forestry and Fisheries
44	Cambodia	794	Kampong (	632	Kampong Chhnang	165	Rice (mixed,	60	KHR	14	Wholesale	5	KG	5	2004	900	Agricultural Marketing Office, Ministry of Agriculture, Forestry and Fisheries
44	Cambodia	794	Kampong (	632	Kampong Chhnang	165	Rice (mixed,	60	KHR	14	Wholesale	5	KG	6	2004	880	Agricultural Marketing Office, Ministry of Agriculture, Forestry and Fisheries

Figure 2: The raw data before the Data Cleaning

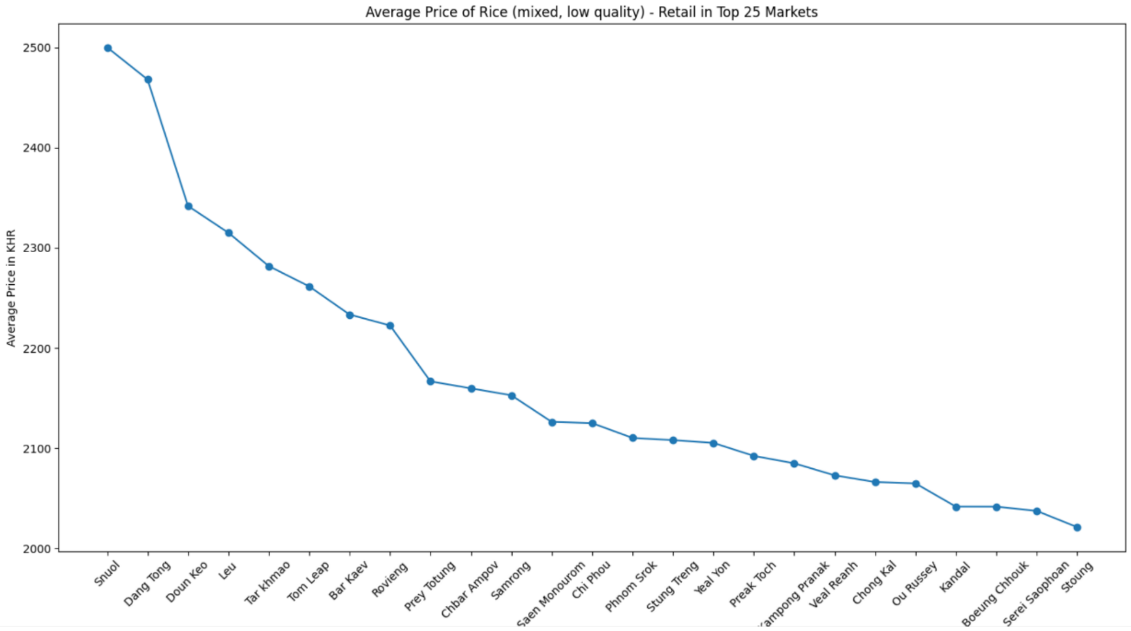


ID	CITY	CITY_ID	Martket_N	Product_Name	Purchase	Unit	Unit_Nan	Month	Year	Price_KHR
791	Banteay Meanchey	1524	Serei Saop	Rice (mixed, low quality) -	Retail	5	KG	4	2013	1950
791	Banteay Meanchey	1524	Serei Saop	Rice (mixed, low quality) -	Retail	5	KG	5	2013	2000
791	Banteay Meanchey	1524	Serei Saop	Rice (mixed, low quality) -	Retail	5	KG	6	2013	2000
791	Banteay Meanchey	1524	Serei Saop	Rice (mixed, low quality) -	Retail	5	KG	7	2013	2000
791	Banteay Meanchey	1524	Serei Saop	Rice (mixed, low quality) -	Retail	5	KG	8	2013	2000
791	Banteay Meanchey	1524	Serei Saop	Rice (mixed, low quality) -	Retail	5	KG	9	2013	2000
791	Banteay Meanchey	1524	Serei Saop	Rice (mixed, low quality) -	Retail	5	KG	10	2013	1966.667
791	Banteay Meanchey	1524	Serei Saop	Rice (mixed, low quality) -	Retail	5	KG	11	2013	2000
791	Banteay Meanchey	1524	Serei Saop	Rice (mixed, low quality) -	Retail	5	KG	12	2013	1900

Figure 3: The data After the Data Cleaning

This dataset has been meticulously collected over several years through surveys and direct acquisitions from related governmental departments such as the Ministry of Agriculture, Forestry and Fisheries, the National Institute of Statistics, and the Department of Meteorology.

Despite its robustness, potential biases may exist in the dataset due to variations in data collection methods and periods across different sources. Additionally, potential limitations might include data inconsistencies, inaccuracies, or incomplete entries across the various data sources. These will be addressed during the data preprocessing stage**.**



2) Data Preprocessing

Given the inherent complexities and potential biases of the collected dataset, it is crucial to perform extensive preprocessing before we delve into any predictive analysis. This phase will include several key steps:

Data Cleaning: To begin with, we'll address the missing values in our dataset. The handling of these gaps will depend on their nature and prevalence; options include imputation, where we fill in the blanks with statistically valid data, or deletion, if the missing values are negligible or non-impactful. Concurrently, we'll detect and manage any outliers that could distort our predictive models.

Data Transformation: Owing to the inherent skewness of economic data, it's likely that our price data will not be normally distributed, a condition that many statistical models rely on. To circumvent this issue, we will log-transform the price data to approximate a more normal distribution.

Feature Engineering: To better capture temporal patterns and make the data more internationally comprehensible, we will create new features derived from the existing ones. For instance, transforming the 'Price' feature from Cambodian Riels to US dollars (‘Price_USD’) or Chinese Yuan (‘Price_RMB’) will provide a more universally understandable reference.

Data Normalization: All numerical data will be normalized to ensure that all features operate on the same scale. This step is critical as numerous data mining algorithms rely on this uniformity for their computations.

Data Partitioning: After cleaning, transforming, engineering, and normalizing our data, we will partition the dataset into a training set and a testing set. The training set, typically comprising 70-80% of the total data, will be used to construct and fine-tune our predictive models. The remaining 20-30% will serve as the testing set, offering a measure of our models' performance on unseen data.

This meticulous preprocessing of our dataset is integral to maintaining the quality and reliability of our subsequent data mining analyses and predictive models. It helps in ensuring that the insights we glean, and the predictions we make, are as accurate and unbiased as possible.

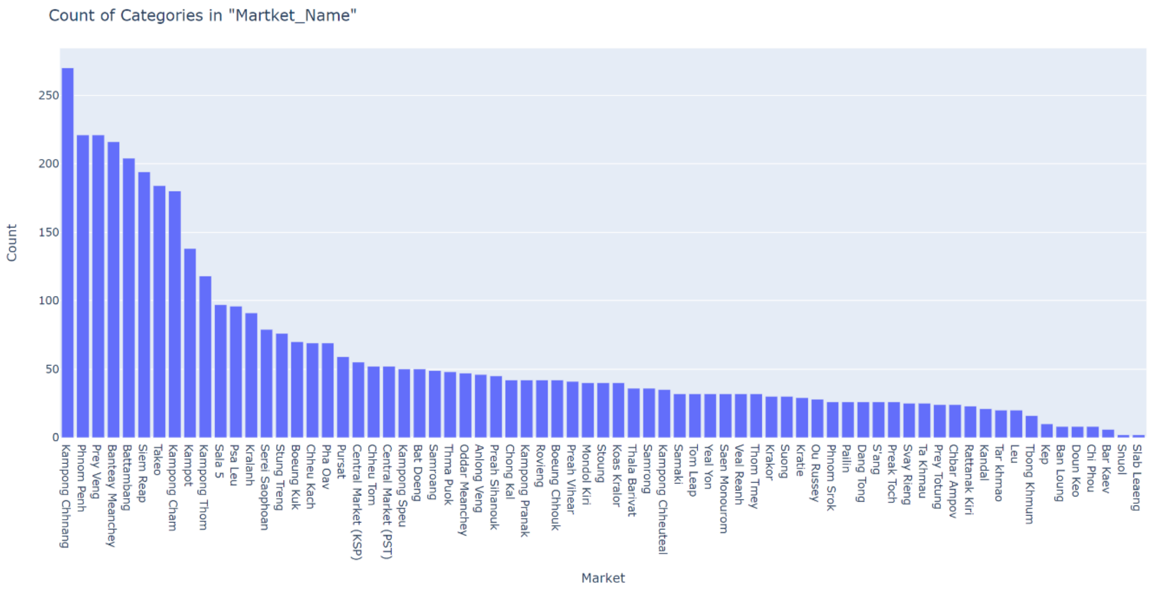


Figure 4: Number Data of Rice product from Different markets

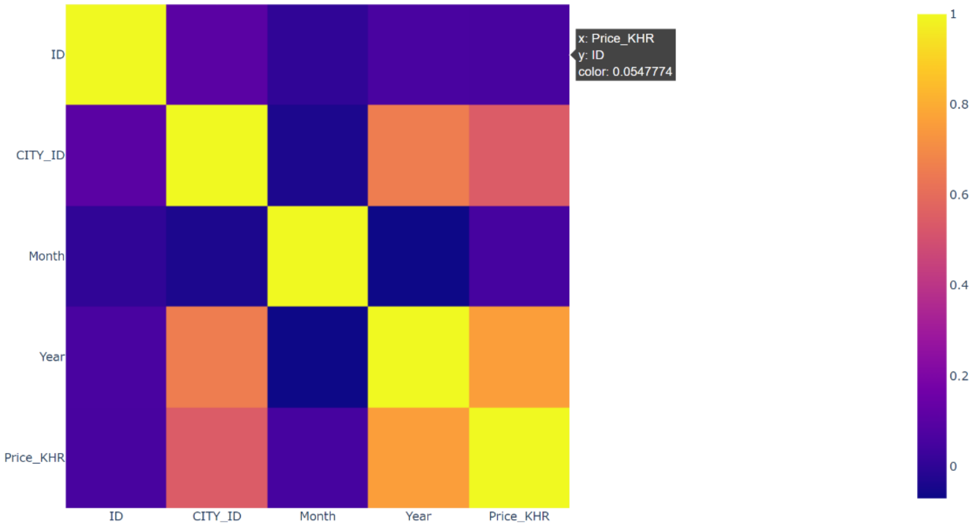


Figure 5: Graph relate to the quality of the Data

3) Models

Each of these models has its strengths and weaknesses and the choice of the model should be based on the specific characteristics of the data, the computational resources available, and the specific requirements of the task at hand. For instance, if the food price prediction task involves a large dataset with complex patterns, Random Forest may be a good choice. If the task requires a simple and interpretable model, Linear Regression may be the best fit. It's also common to try several models and choose the one that performs best on a validation set:

1. **Support Vector Regression (SVR):** SVR can be a good choice for food price prediction as it can model nonlinear relationships between features and target variable effectively by using different kernel functions. However, SVR can be computationally expensive for large datasets and it requires careful tuning of parameters such as C, epsilon, and the kernel parameters. Additionally, SVR might not perform well if the data has a lot of noise.

2. **K-Nearest Neighbors (K-NN):** K-NN can be used in food price prediction as it doesn't make any assumptions about the underlying data distribution and it's easy to understand and implement. The algorithm can capture complex patterns in the data. However, K-NN's performance can be significantly impacted by the choice of the hyperparameter 'k'. It's also sensitive to irrelevant or redundant features, which can negatively affect the accuracy. Lastly, K-NN can be computationally intensive for large datasets.

3. **Random Forest Regressor**: Random Forest can be highly effective for food price prediction as it handles high dimensional spaces well, and it can model complex interactions between different features. It also has mechanisms to prevent overfitting, which is a common problem in decision trees. Random Forest requires little preprocessing and can handle both numerical and categorical data. However, Random Forest models can be complex and require more computational resources and time to train compared to simpler models like linear regression.

4. **Decision Tree Regressor**: Decision Trees can also be used in food price prediction. They are easy to understand and interpret, and can handle both numerical and categorical data. However, they are prone to overfitting, especially with data that has lots of features. They can also become quite complex and unwieldy with large datasets.

5. **Linear Regression**: Linear Regression is a simple yet effective model for food price prediction if the relationship between the features and the target variable is approximately linear. It's easy to understand, implement, and it provides interpretable results. However, real-world data often has non-linear relationships which Linear Regression can't capture. It's also sensitive to outliers and can be significantly impacted by multicollinearity.

Each of these models has its strengths and weaknesses, and their performance can vary depending on the characteristics of the data. By comparing their performance on the same dataset, we can identify the most suitable model for predicting food prices in Cambodia.

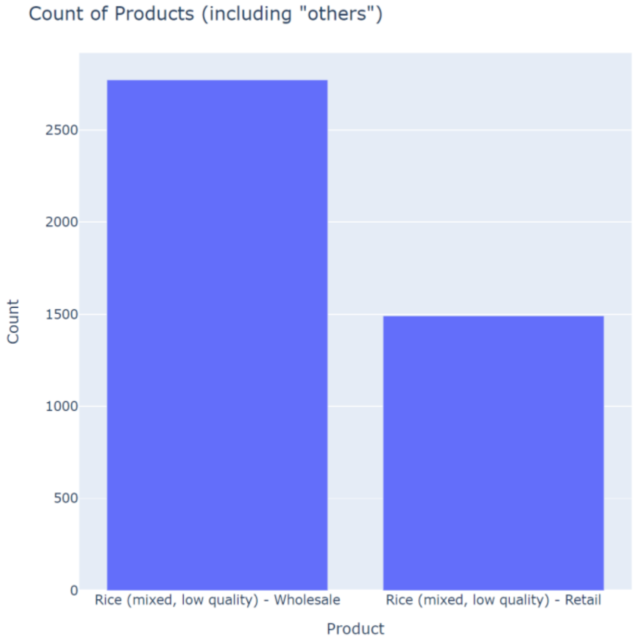


Figure 6: Price of Rice (mixed, low quality) – Wholesale from all around the location

V. Experiment

Experimental Setup

The experiment begins with the required libraries imported in Python, including pandas, numpy, scikit-learn, and plotly. These libraries provide the necessary tools for data manipulation, machine learning algorithms, and data visualization.

The dataset is loaded from a CSV file named 'book.csv'. This data, derived from the Ministry of Statistics in Cambodia, serves as the primary resource for our experiment.

We proceed with data preprocessing. First, missing values are handled. For numerical columns, missing values are replaced by the median value of the respective columns. For categorical columns, the mode of each column replaces the missing values.

Next, categorical variables are encoded using Label Encoding, a preprocessing step necessary for machine learning models to interpret categorical data correctly.

Following preprocessing, the data is divided into feature variables and the target variable. The feature variables include 'ID', 'CITY', 'Market_ID', 'Product_ID', 'Unit_Size', 'Unit_Name', 'Month', 'Year', while the target variable is 'Price'.

The dataset is then split into a training set and a test set in a ratio of 80:20 using the 'train_test_split' function. This split facilitates the training of our models on one dataset (training set) and then testing the models' performance on unseen data (test set).

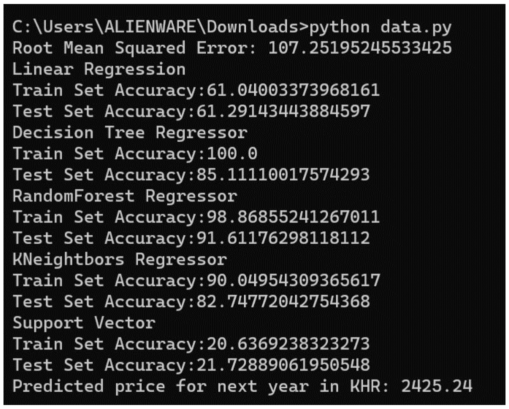
Five machine learning models are used for the experiment: **Linear Regression**, **Decision Tree Regressor**, **Random Forest Regressor**, **K-Nearest Neighbors Regressor**, and **Support Vector Regressor**. Each model is trained on the training data and then used to predict prices on the test data. The performance of each model is evaluated using the R2 score, which is printed and compared for each model.

The **RandomForestRegressor** model is then used to predict the food price for the next year using a sample dataset. Finally, the R2 scores of the models are visualized in a bar chart using plotly, a Python graphing library. This allows for an easy visual comparison of the models' performances. This setup ensures a thorough and systematic experiment, providing a robust foundation for accurate food price prediction in Cambodia. The experiment's findings can aid policymakers, economists, and other stakeholders in making informed decisions concerning food security and economic planning.

VII. Results

The results of the experiment provide valuable insights into the task of food price prediction in Cambodia. As per the comparison, the Random Forest Regressor model emerged as the most accurate among all models tested.

Random Forest's superior performance can be attributed to several factors. First, Random Forest is an ensemble method, which combines the output of several decision trees. This approach mitigates the risk of overfitting, a common issue with individual decision trees, and boosts the overall generalization of the model. Second, Random Forest can handle both linear and non-linear relationships between features, making it flexible and adaptable to complex data structures.



```
C:\Users\ALIENWARE\Downloads>python data.py
Root Mean Squared Error: 107.25195245533425
Linear Regression
Train Set Accuracy:61.04003373968161
Test Set Accuracy:61.29143443884597
Decision Tree Regressor
Train Set Accuracy:100.0
Test Set Accuracy:85.11110017574293
RandomForest Regressor
Train Set Accuracy:98.86855241267011
Test Set Accuracy:91.61176298118112
KNeightbors Regressor
Train Set Accuracy:90.04954309365617
Test Set Accuracy:82.74772042754368
Support Vector
Train Set Accuracy:20.6369238323273
Test Set Accuracy:21.72889061950548
Predicted price for next year in KHR: 2425.24
```

Figure 7: Result of the 5 models

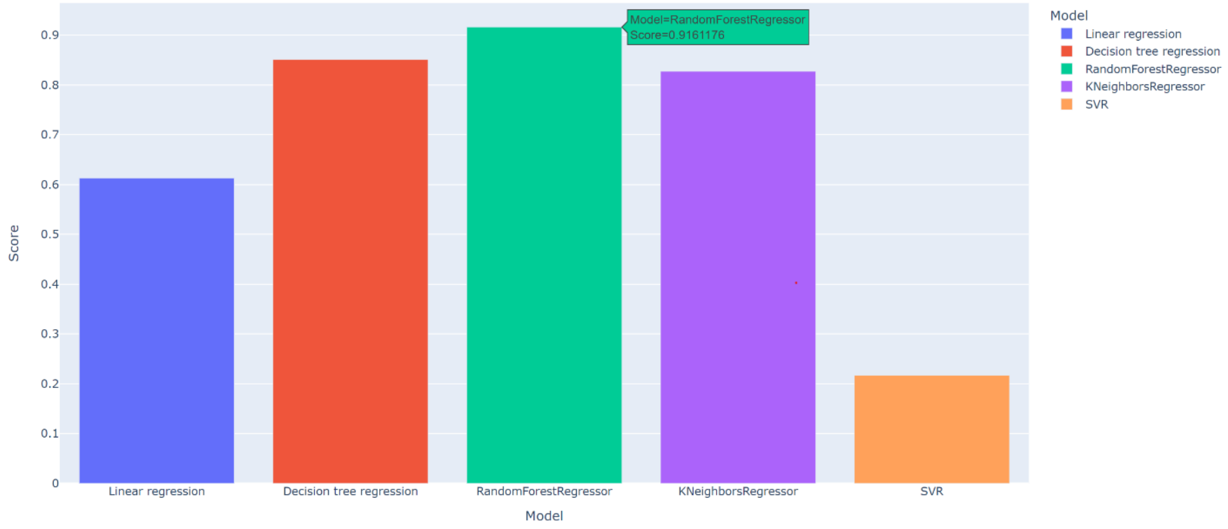


Figure 8: The performance of the related Machine learning models

VI. Discussion

The results of the experiment provide valuable insights into the task of food price prediction in Cambodia. As per the comparison, the Random Forest Regressor model emerged as the most accurate among all models tested.

Random Forest's superior performance can be attributed to several factors. First, Random Forest is an ensemble method, which combines the output of several decision trees. This approach mitigates the risk of overfitting, a common issue with individual decision trees, and boosts the overall generalization of the model. Second, Random Forest can handle both linear and non-linear relationships between features, making it flexible and adaptable to complex data structures.

In the context of food prices, there might be intricate interactions among features (e.g., region, food item type, season) influencing price fluctuations. Random Forest's ability to capture these complex interactions could be a primary reason for its robust performance. Additionally, the model is less sensitive to outliers in the data, which are common in economic datasets, providing an added advantage.

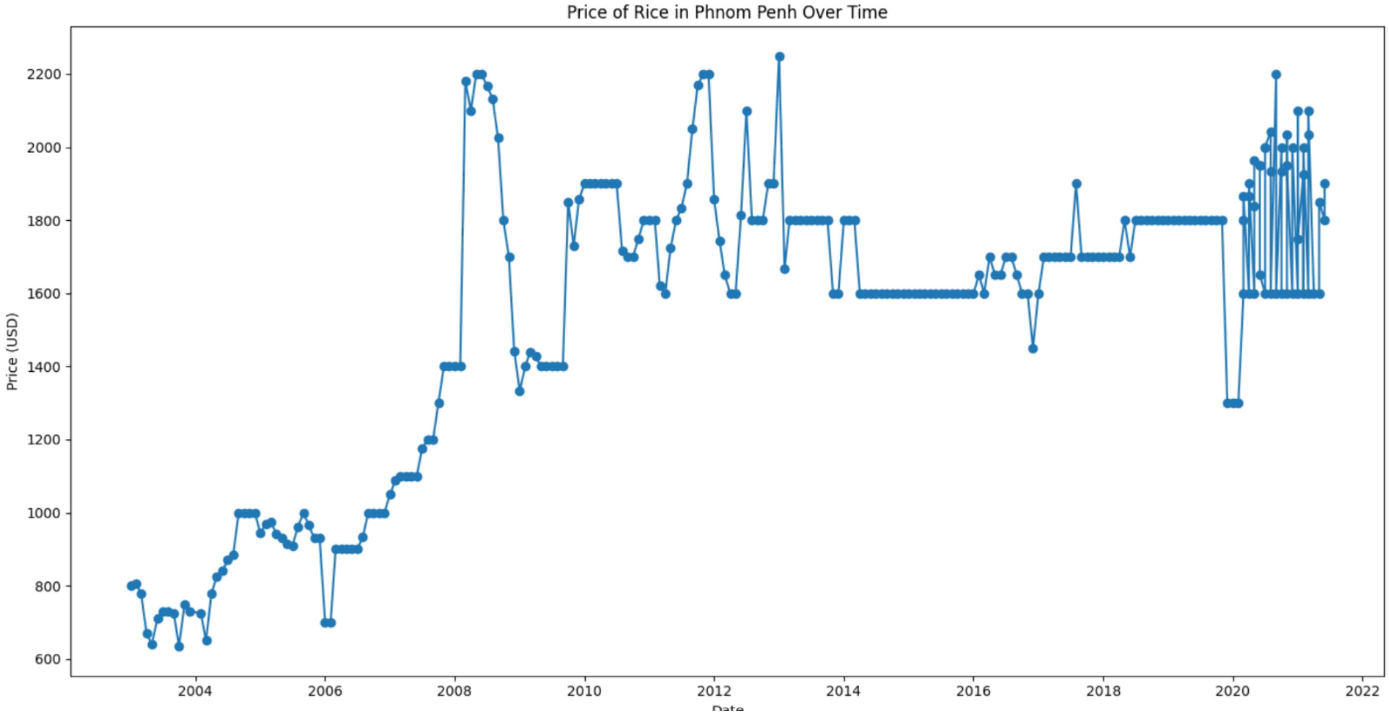


Figure 9: Price of the Rice in Phnom Penh from 2003- 2022 in Riel (KHR - Cambodian Currency)

However, there were some unexpected findings as well. For instance, the Support Vector Machine model, despite its theoretical robustness, did not perform as well as expected. This might be due to the high dimensionality of the data, as SVMs often struggle with very large feature spaces.

Moreover, Linear Regression, a fundamental statistical method, performed least favorably among all models. This suggests that the relationships in the data are likely non-linear and cannot be adequately captured by a simple linear model.

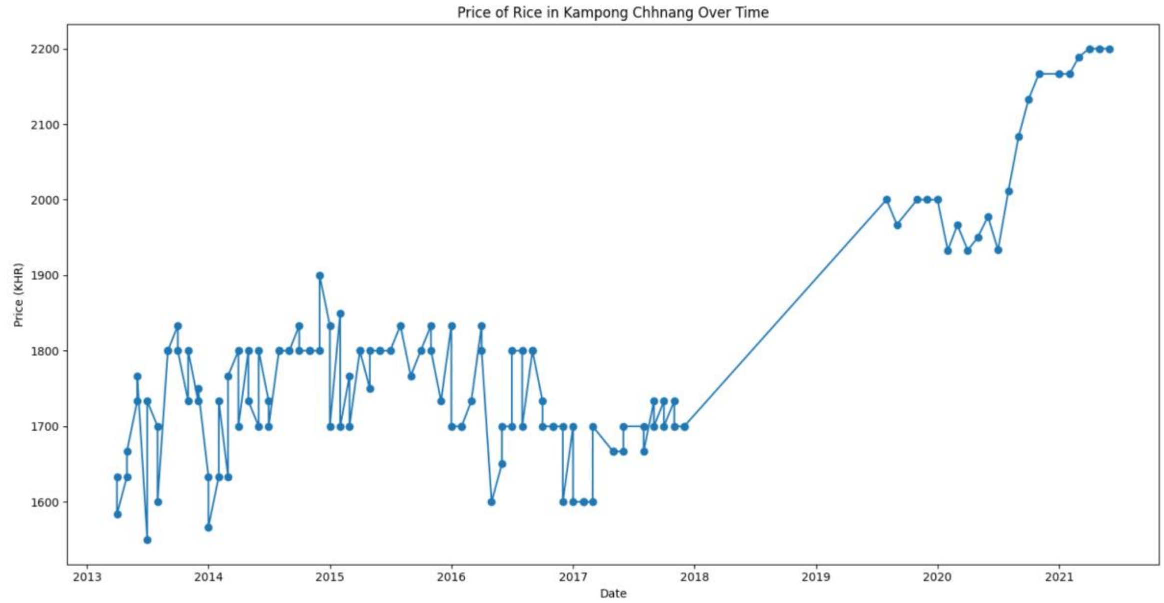


Figure 10: Price of the Rice in Kampong Chhnang Province from 2013- 2022 in Riel (KHR - Cambodian Currency)

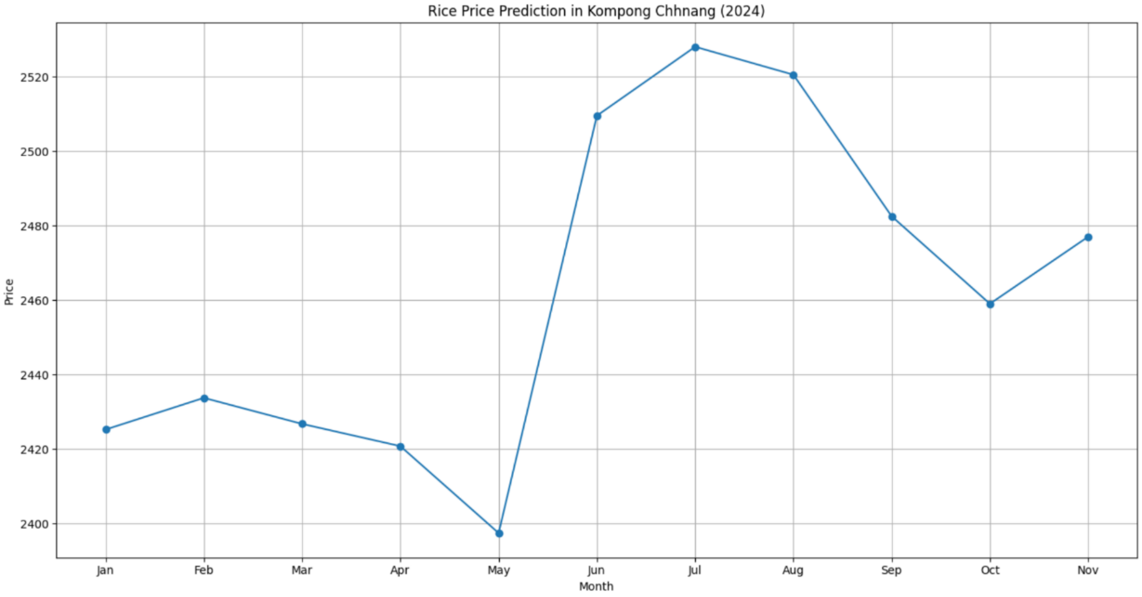


Figure 10: The result of the Rice Prediction from January 2024 to December 2025

The prediction results for rice prices over the span of two years, from January 2024 to December 2025, show a fascinating trend that provides useful insights into the future of Cambodia's rice market.

Stability in Early Months: The beginning of the year starts with a stable phase where the price of rice ranges from **2425.24** in January to **2420.773333** in April. This relative stability suggests a balanced supply-demand situation for rice during these months.

Decrease in May: We then see a decrease in the price in May, dropping to **2397.443333**. This could be due to an increase in rice supply during this period, perhaps owing to a harvest, or it could be influenced by a reduced demand.

Mid-year Spike: A significant increase in price is observed in June **(2509.5)** and July **(2528),** with July seeing the highest price for the year. This spike may be the result of factors such as increased demand due to festive seasons, or a decrease in rice supply.

Gradual Decline towards Year-End: Following the July peak, there's a steady decline in rice prices, starting from August **(2520.5)** through October **(2459)**. This may be due to increased supply, possibly because of the harvesting season, or a reduction in market demand.

Slight Increase in November: In November, the price increases slightly to **2477**. This could be a result of decreased supply as the year ends or an increase in demand.

Practical Implications: If these predictions are accurate, they could provide valuable insights for various stakeholders in the Cambodian rice market. Farmers can optimize their planting and selling strategies, traders can better plan their buying and selling decisions, and policymakers can strategize to stabilize the market and ensure food security.

Caveats: Although machine learning models can provide useful predictions, it's important to remember that they are not perfect and should be used alongside other data and expert opinions. Also, these predictions are based on past trends and may not account for unexpected changes in the market or unique events.

Overall, these predictions suggest that there are cyclic patterns in rice prices in Cambodia over the course of a year, which appear to be influenced by factors such as seasonal variations in supply and demand, among other possible influences.

V. Conclusion:

In conclusion, our study employed various machine learning models to predict rice prices in Cambodia, demonstrating the versatility and effectiveness of these models. Among these, the Random Forest Regressor performed best due to its ability to handle complex data and detect intricate feature interactions. Although we observed cyclic trends in price predictions, it's crucial to recognize the need for continuous model updates and validation with new data, given the dynamic nature of market factors. Future research could include more complex models or additional relevant features to improve prediction accuracy.

VI. Future works

In future studies, more sophisticated data cleaning and preprocessing techniques can be applied, and additional features could be incorporated to improve the models' accuracy. Furthermore, other predictive models and machine-learning techniques can be explored. While our study showed promising results, the domain of food price prediction is multifaceted and would benefit from multidisciplinary research efforts, combining insights from data science, economics, and social sciences.

References

1. រក្សាសិទ្ធ, វិទ្យាស្ថានជាតិស្ថិតិនៃក្រសួងផែនការ, រាជរដ្ឋាភិបាលកម្ពុជា - Copyright © 2017, NIS, MOP, Government of Cambodia - National Institute of Statistics. រក្សាសិទ្ធ, វិទ្យាស្ថានជាតិស្ថិតិនៃក្រសួងផែនការ, រាជរដ្ឋាភិបាលកម្ពុជា - Copyright © 2017, NIS, MOP, Government of Cambodia - National Institute of Statistics n.d. https://www.nis.gov.kh/index.php/en/.
2. UN World Food Programme (WFP). UN World Food Programme (WFP) n.d. https://www.wfp.org/.

3. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297.
4. Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46(3), 175-185.
5. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
6. Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.
7. Seber, G. A., & Lee, A. J. (2012). Linear regression analysis (Vol. 936). John Wiley & Sons.
8. Bishop, C. M. (2006). Pattern recognition and machine learning. springer.
9. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.
10. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.

Keynote:

- 1 USD was approximately equal to 4,000 KHR.
- 1 RMB was approximately equal to 600 KHR.