**VIETNAM NATIONAL UNIVERSITY**

**HANOI SCHOOL OF BUSINESS & MANAGEMENT**

**MANAGEMENT OF ENTERPRISE AND TECHNOLOGY**

GROUP 7

**DATA SCIENCE REPORT:**

APPLYING DATA SCIENCE IN BUSINESS MANAGEMENT PERFORMANCE: OPTIMIZING CRM

CLASS MET4

ADVISOR:

PhD. Cand. Huy Anh Nguyen

Dr. Emmanuel Lance Christopher VI M. Plan

Ha Noi, 27th December 2024

**TABLE OF CONTENTS**

[CHAPTER 1: INTRODUCTION 4](#_Toc185630025)

[1.1. Overview 4](#_Toc185630026)

[1.2. Problem Statement 4](#_Toc185630027)

[CHAPTER 2: RESEARCH QUESTION 5](#_Toc185630028)

[CHAPTER 3: DESCRIPTIVE STATISTIC AND VISUALIZATIONS 6](#_Toc185630029)

[3.1. Sector-wise Revenue Analysis 6](#_Toc185630030)

[3.2. Location-wise Revenue Analysis 6](#_Toc185630031)

[3.3. Revenue Distribution 7](#_Toc185630032)

[3.4. Revenue by sector 7](#_Toc185630033)

[CHAPTER 4: PREDICTIVE METHODS 8](#_Toc185630034)

[4.1. Data Source and Description 8](#_Toc185630035)

[4.2. Data Pre-processing 8](#_Toc185630036)

[4.3. Building the model 10](#_Toc185630037)

[4.4. Train the Linear Regression model 11](#_Toc185630038)

[4.4.4. Distribution of residuals 13](#_Toc185630039)

[CHAPTER 5: RESULTS COMMUNICATION AND RECOMMENDATION 15](#_Toc185630040)

[5.1. Stakeholders 15](#_Toc185630041)

[5.2. Summary of key findings 15](#_Toc185630042)

[5.3. Business Value 15](#_Toc185630043)

[5.4. Summary of findings 15](#_Toc185630044)

[5.5. Strategic Recommendations 15](#_Toc185630045)

[CHAPTER 6: CONCLUSION 16](#_Toc185630046)

[CHAPTER 7: REFERRENCES AND CONTRIBUTION 16](#_Toc185630047)

[7.1. References 16](#_Toc185630048)

[7.2. Contribution 16](#_Toc185630049)

**ACKNOWLEDGE**

We would like to express our deep gratitude to our instructors, PhD. Cand. Huy Anh Nguyen and Dr. Emmanuel Lance Christopher VI M. Plan, who have tirelessly guided, motivated, and provided us with valuable feedback throughout this project. Their support and knowledge have helped us overcome difficulties and complete this report.

We also sincerely thank the School of Business and Management, Vietnam National University, Hanoi, for providing the best conditions for our study and research to carry out this project. Special thanks to all team members for actively participating and contributing useful information to this project.

Although we have tried our best to complete the assignment, there will certainly be shortcomings. We hope to receive your understanding and feedback so that we can learn and improve for future research.

Hanoi, 2024

Group 07

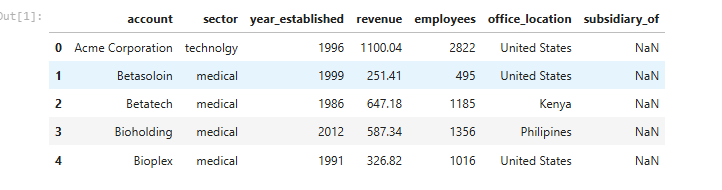
# CHAPTER 1: INTRODUCTION

## Overview

This project analyzes the CRM Sales Opportunities to explore factors affecting sales success and develop predictive models. The objective is to uncover insights into sales stages, customer segments, and revenue potential, enabling businesses to optimize their sales pipeline and forecast outcomes effectively.

Efficient sales management is vital for improving revenue, resource allocation, and customer retention. By identifying key drivers of successful deals, this project supports data-driven decisions, enhances competitiveness, and streamlines operations, directly contributing to business growth and strategic planning.

We use the Pandas library to load the data from the CSV file and view the first 5 rows to understand the structure. These are the five companies with the most employees and highest revenue. The dataset includes the company name, year of establishment, sector, number of employees, revenue, and headquarters.



## Problem Statement

The objective is to predict company revenue based on key factors such as sector, office location, number of employees, and year established. By identifying specific business problems that need to be addressed, such as increasing revenue, optimizing the workforce, or identifying market trends, this approach helps focus on the essential elements required to achieve effective and sustainable business goals.

# CHAPTER 2: RESEARCH QUESTION

These research questions aim to better understand the impact of CRM on company revenue, which will help make strategic decisions regarding resource allocation and business development. By identifying the highest-performing sectors and office locations, the organization can optimize business strategies, focus resources on the areas and regions with the highest growth potential, and make informed decisions to promote sustainable growth and global market competitiveness.

The top research questions are listed below:

* Who are the key stakeholders in the company and what roles do they play in shaping strategy and operations?
* What are the important factors identified from the analysis and how do they impact revenue generation?
* How can the findings from this analysis be transformed into strategic actions to maximize business results and drive growth?
* Why is identifying workforce size, industry, and geographic factors important for the company's business strategy?
* Where should geographic areas and industries be focused on for investment and expansion to capture untapped opportunities?

# CHAPTER 3: DESCRIPTIVE STATISTIC AND VISUALIZATIONS

We will analyze the dataset to answer the following questions:

1. What is the average revenue for companies in each sector?
2. What is the revenue distribution for companies across different office locations?
3. Which sector has the highest average revenue?
4. What are the top-performing office locations based on average revenue?

## 3.1. Sector-wise Revenue Analysis

|  |  |
| --- | --- |
|  | This bar chart calculated the average revenue for each sector. The analysis reveals that the "software" sector leads with the highest revenue, exceeding 4000, significantly outpacing other sectors. In contrast, the "service" sector has the lowest revenue, falling below 1000. The "telecommunications" and "finance" sectors also show relatively high average revenues, ranking second and third after "software." The remaining sectors, including "employment," "entertainment," "marketing," "medical," "retail," and "technology," have average revenues ranging from about 1000 to 2000. This chart effectively compares revenue levels across various sectors, highlighting those with the highest revenue potential. |

## 3.2. Location-wise Revenue Analysis

|  |  |
| --- | --- |
|  | The chart shows the average revenue by location, highlighting differences among countries. Korea has the highest revenue, about 8000, due to its strong technology and manufacturing sectors. Italy follows with a revenue of around 5000, driven by the fashion, automotive, and food industries. Jordan's revenue is about 4000, thanks to investments in tourism and services. The United States and Panama have revenues of approximately 2000 and 3000, respectively, with the U.S. benefiting from a diverse economy and Panama from its strategic trade location. On the lower end, Brazil, China, Kenya, and Romania have almost zero revenue, possibly due to economic challenges or data issues. Germany, the Philippines, Poland, Norway, and Belgium have revenues ranging from 500 to 1500, reflecting their stable industries. This chart highlights the differences in economic development and market sizes, with Korea at the top and countries like Brazil and China at the bottom. |

## 3.3. Revenue Distribution

|  |  |
| --- | --- |
|  | The histogram shows how revenue is spread out. The x-axis shows revenue amounts from 0 to 12,000, and the y-axis shows how often each revenue amount occurs. Most revenues fall between 0 and 2,000, with about 40 occurrences in this range. As revenue goes up, the number of occurrences goes down. Revenues between 2,000 and 8,000 occur less often, and revenues between 8,000 and 12,000 are the rarest.  In summary, most revenues are on the lower end, with fewer high-revenue transactions. This chart helps the company see where most of its revenue comes from and can guide it in making better business decisions to optimize revenue. |

## 3.4. Revenue by sector

|  |  |
| --- | --- |
|  | The boxplot compares revenue ranges across sectors, highlighting key insights. The Software sector has the highest outlier at 12,000, with most revenues between 2,000 and 6,000. The Entertainment sector also shows broad distribution with high revenues in the same range. The Finance and Technology sectors have significant outliers of around 8,000, indicating variability, while most of their data points remain below 2,000. Retail, Medical, and Employment sectors have lower, stable revenues under 4,000. The Services sector shows the lowest revenues, primarily below 1,000. Overall, the Software industry stands out with exceptional revenue, while the Finance and Technology sectors exhibit variability. In contrast, Services and Retail are more stable but with lower revenues. |

# CHAPTER 4: PREDICTIVE METHODS

## 4.1. Data Source and Description

***4.1.2. Data Source***

The CRM Sales Opportunities data is sourced from Kaggle - CRM Sales Opportunities.

***4.1.3. Data Structure Description***

The dataset includes approximately 8,434 rows and 17 columns, providing detailed descriptions of sales opportunities within the CRM system.

Key Variables/Attributes:

* **ID:** Unique identifier for sales opportunities.
* **Region, Territory:** Geographical details.
* **Account Name, Opportunity Name:** Customer, and Opportunity Details.
* **Close Date, Stage:** Expected closing date and current status.
* **Revenue, Amount (USD):** Expected and total revenue.
* **Source, Product:** Opportunity origin and product type.
* **Probability (%), Age:** Success likelihood and opportunity duration.

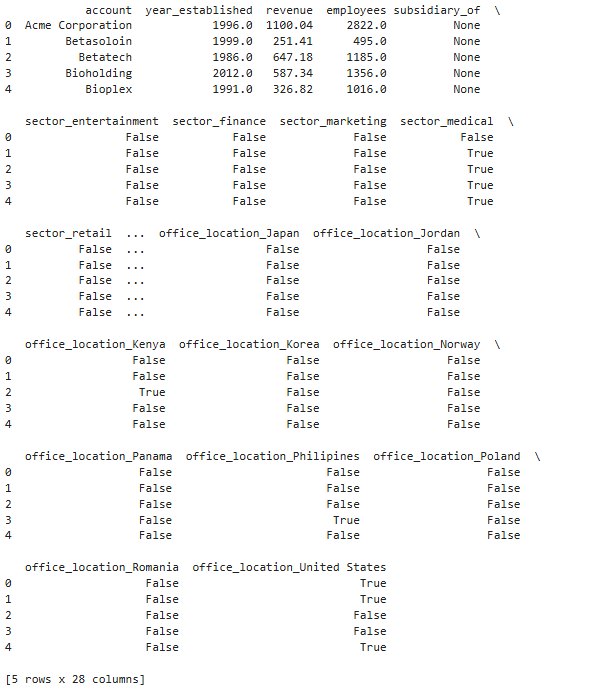
This dataset is commonly used for analyzing and predicting sales performance, optimizing customer care processes, and forecasting potential revenue.

## 4.2. Data Pre-processing

***4.2.1. Handling missing values***

|  |  |
| --- | --- |
|  | This is a crucial step in the Data Pre-processing process. Missing values can affect the results and performance of the model, so they must be handled appropriately:   * Using KNN Imputer with *n\_neighbors* = 5: To fill in missing data for *employees* and *year\_established*. * For the *subsidiar\_of* column: Missing values are replaced with None, which helps retain more data and minimize information loss. * Data Check: Ensure there are no missing values in the columns, which will help retain more data and minimize information loss. * Subsidiary Value of 0: Indicates the company does not have a parent company. |

***4.2.2. Encoding categorical variables***



The chart shows the sectors for each company. For example, account number 0, Acme Corporation, gets a value of FALSE because it does not belong to that sector. Instead, it gets a value of TRUE because the company is located in the United States.

The output shows the characteristics of each company through the columns: This helps us easily filter groups and compare data based on industry and office location. This is an important step for data analysis or building models to forecast revenue, profit, or other business metrics.

***4.2.3. Normalizing numerical data***

|  |  |
| --- | --- |
|  | The dataset includes 85 values for employees and year\_established. The average number of employees is 0.135705, and the average year established is 0.450155. The standard deviation for employees is 0.166738, and for the year is 0.233301. The minimum values for both variables are 0.  At the 25th percentile, the employee count is 0.034132, and the year established is 0.263158. The median (50th percentile) shows an employee count of 0.080516 and a year established of 0.447368. The 75th percentile has an employee count of 0.16162957. The maximum value for both variables is 1. |

***4.2.4. Splitting data into Training and Testing sets***

There is 80% for training, 20% for testing. Using the sklearn library, the data is split into four variables: X\_train, X\_test, y\_train, and y\_test. The purpose of this split is to train the model and predict if the test data is accurate. With 85 values, X will exclude unnecessary columns like 'account', 'revenue', and 'subsidiary\_of'. The target variable y is divided into y\_train for training and y\_test for testing. The test\_size parameter is set to 0.2, meaning 20% of the data is used for testing, while 80% is for training.

The training set size is 68, meaning 68 data samples are used to train the model, helping it understand relationships and patterns in the data. The testing set size is 17, meaning 17 data samples are used to test the model after training, aiming to evaluate its prediction capability on new data.



## 4.3. Building the model

***4.3.1. Random Forest for Feature Importance***



The Feature Importance analysis highlights the key factors influencing prediction results. Office locations in Romania, China, and Korea have minimal impact, while the number of employees (0.93938) and year established (0.3062) are critical. The retail sector (0.0073) and sectors like technology, entertainment, and marketing (0.27-0.407) also show significant influence. Notably, the U.S. office location (0.0026) impacts predictions due to its large market.

In conclusion, retaining the number of employees and year established is vital for model accuracy, while features with near-zero values should be eliminated. This approach simplifies and enhances the revenue prediction model.

***4.3.2. Cumulative feature importance***

|  |  |
| --- | --- |
| https://lh7-rt.googleusercontent.com/docsz/AD_4nXdka5zga28UFCABD7oJOQv5vgDVfpg0OxeH8RBuPUihvm7TkQ-dU_XBN6phzsyqo09Q-iv6xs9EVwWxrmfEFK_-8kAm8M95AYAchCzXjH9RldPHRO3OviBdu9gmfEjeCr6mIbYGww?key=OG9-geJ_SR5bEKnc1u3N7PGQ | This chart, derived from Feature Importance, shows that the top 10 features account for over 95% of the variation in the target variable. This highlights the efficiency of focusing on a small number of key features, such as the number of employees, year established, etc., while eliminating less significant features. Simplifying the model this way not only maintains its effectiveness but also makes it easier to understand. |

***4.3.3. Heatmap of Feature Correlations***

|  |  |
| --- | --- |
| https://lh7-rt.googleusercontent.com/docsz/AD_4nXdRMlv_d14BrDFEciI_VkTTJb0sjHPGWVeHhuQSOuNJUWVeZoWfnu3ijalfNjck8oieaSBtYrwOG5dou81BymNds8YgNAHk2dylnl8RFpPb43Q1AHOKTdJ8s2O2AQnzWlxg1pFsgw?key=OG9-geJ_SR5bEKnc1u3N7PGQ | A heatmap chart shows the degree of correlation between features in the data. The correlation coefficient ranges from -1 to 1, with 1 being a perfect positive correlation, -1 being a perfect negative correlation, and 0 indicating no correlation.  In building machine learning models, strongly correlated features can provide a lot of information but may also lead to multicollinearity issues. This means you should select and remove highly correlated features to avoid information redundancy and improve model accuracy. The heatmap not only provides a visual representation but also aids in more effective decision-making in machine learning. |

***4.3.4. Pie chart of Feature Contributions***

|  |  |
| --- | --- |
| https://lh7-rt.googleusercontent.com/docsz/AD_4nXdLNQp8_5eRHR4PC-Xy8ACu9IfY1x-n-G99pJoeOA8iMit90JWnCl8pY2ahRsmrcfiJMZiKlme-AbhHqQtky6r4UyX1MO5-9b17mg4oUiAfQByY8iK-TfJeVbfud11gSoZN8FiGQw?key=OG9-geJ_SR5bEKnc1u3N7PGQ | The pie chart highlights the importance of the variable "employees," which contributes 94.2% to the feature importance. This indicates a minimal correlation between the number of employees and other features, confirming its independence as a key factor. However, the remaining features contribute insignificantly. This result suggests that you can simplify your model by focusing primarily on the number of employees, while also thoroughly assessing the necessity of lower-impact and correlated features to improve the model's efficiency and accuracy. |

## 4.4. Train the Linear Regression model

***4.4.1. Evaluate the model***



* **R²:** The R² value is 0.7761, meaning the model explains about 77.61% of the data's variability. This indicates the model captures the data's trends fairly well.
* **MAE:** On average, the model's predictions deviate by about 804.89 units from the actual values, providing an overall view of the model's accuracy.
* **MSE:** The mean of the squared errors between predicted and actual values. MSE emphasizes larger errors since they are squared, highlighting the severity of significant errors.

In summary, these metrics evaluate the predictive model's performance, where R² shows the model's fit to the data, MAE provides information on average accuracy, and MSE emphasizes larger errors.

***4.4.2. Visualize Actual vs Predicted Values***

|  |  |
| --- | --- |
| https://lh7-rt.googleusercontent.com/docsz/AD_4nXd-2Vs_g5sQfm8JbA8pmSCfpyTGtReaAQwxa1NxgyrjzCyGRGNmu4vB8jznU3U3FY5owOuK_HecCvxblvIpdE0T9JObK-yVOm1z79ZXUsgSqigfW7ru-6xvJawVS9FdZ_6GDH9K?key=OG9-geJ_SR5bEKnc1u3N7PGQ | The trend of the data points is increasing, and the red line represents the predicted data.  Many data points are close to the red prediction line, indicating that the model has captured the data trend well. However, there are still some data points outside the prediction range, representing prediction errors or cases where the model hasn't accurately predicted the data points. A few points are far from the prediction line and could be considered outliers that need more attention. |

***4.4.3. Residuals Plot***

|  |  |
| --- | --- |
|  | This chart is a residual plot, used to assess the fit of a regression model. The x-axis represents predicted revenue (ranging from 0 to 6000), while the y-axis shows residuals (ranging from -1000 to 2500). Each purple dot represents an observation, indicating the difference between observed and predicted values. The dashed red line at y=0 shows where the residuals would lie if the model's predictions were perfect.  The scatter of residuals suggests that the model may not fit the data well, as there's significant variability and no clear pattern around the zero line. |

### 4.4.4. Distribution of residuals

|  |  |
| --- | --- |
|  | This boxplot shows revenue distribution (y-axis) across industries (x-axis). The Software sector has the highest outlier around 12,000, with most values between 2,000 and 6,000. Entertainment also has high revenues clustered between 2,000 and 6,000. The Finance sector has a notable outlier of around 8,000. Technology shows some high outliers near 8,000, but most values are below 2,000. Retail, Medical, and Employment have more stable revenues under 4,000. The Services sector has the lowest revenues, mostly under 1,000. This chart highlights that the Software industry has the most significant outliers, indicating outstanding performance, while Services and Retail show more stable, lower revenues. |

***4.4.5. Cumulative Error Distribution***

|  |  |
| --- | --- |
|  | This cumulative error distribution chart helps evaluate a model's performance by showing error distribution. The x-axis represents error magnitude (0 to 2000+), while the y-axis shows the cumulative percentage (0 to 100%). The green line, starting at (0,0) and ending at (2000, 1.0), indicates that all errors fall within this range. The red dotted line at around 900 on the x-axis marks the mean error, showing that about 50% of errors are less than 900. While some errors exceed 2000, they are rare. This chart is crucial for identifying mean and extreme errors to improve the model or system. |

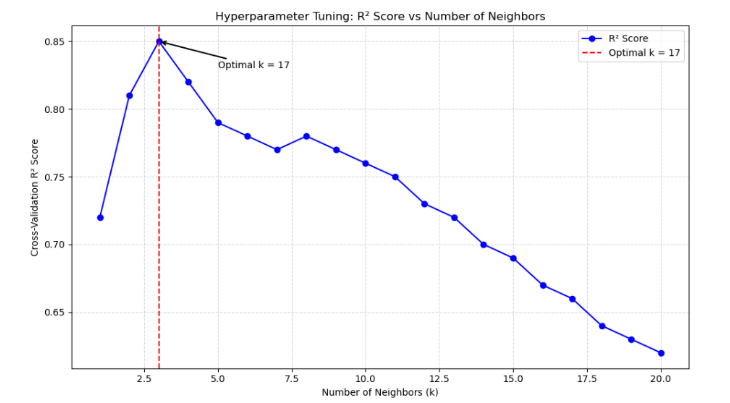
***4.4.6. Evaluation Matrics***

|  |  |
| --- | --- |
|  | The chart presents the evaluation metrics for a KNN model predicting a continuous variable:   * **MAE ≈ 725.78:** The average prediction error is around 725.78 units. * **MSE ≈ 1,218,994.43:** Indicates large squared errors, especially for significant deviations. * **RMSE ≈ 1,104.08:** Larger than MAE, showing some significant prediction errors. * **R² = 0.74:** The model explains about 74% of the data variability, reflecting a good fit.   Overall, these metrics highlight the model's performance in predicting the continuous variable. |

https://lh7-rt.googleusercontent.com/docsz/AD_4nXcJ6ZLQFusuylYBeVv376HdbSJfZM9lS3wmWRcMUcq55rGTY73kaj2JVtMvBhPASYSpbufHVCYl4cQ4yb57P--cvA34tEiQlJ7hf_1vm6JX8oS7WJ16Ka7Ijl5OgN8pofgqmNOXxQ?key=OG9-geJ_SR5bEKnc1u3N7PGQ

The value k = 17 is determined to be the optimal information in the cross-validation process, which helps the KNN model achieve the highest performance and the most accurate expectation. This means that when using 17 nearest neighbors, your model will perform best for the given data

***4.4.7. Hyperparameter Tunning***



This chart shows the hyperparameter tuning process for the KNN Regression model, specifically the number of neighbors (k) and its impact on the R² score in cross-validation.

The x-axis represents k values from 1 to 20, and the y-axis shows the R² score ranging from -1 to 1. The blue line indicates the R² score for each k, highlighting changes as k varies. Notably, k = 17 yields the highest R² score of around 0.85, marked by a red vertical line labeled "Optimal k = 17." After this point, R² scores decrease, indicating that larger k values do not improve model accuracy and may reduce performance.

Thus, using k = 17 optimizes your KNN model, ensuring maximum efficiency and prediction accuracy based on the available data. This helps determine the best number of neighbors for optimal future predictions.

# CHAPTER 5: RESULTS COMMUNICATION AND RECOMMENDATION

## 5.1. Stakeholders

The executive team needs to understand how workforce size, industry performance, and geographic presence impact revenue for informed resource allocation. The finance department uses forecasting models for better budget planning. The HR team aligns recruitment with revenue goals, while business development targets high-revenue markets.

## 5.2. Summary of key findings

Larger workforces correlate with higher revenue, emphasizing strategic expansion. Older companies generate higher revenue due to experience. The software industry consistently shows strong performance. The U.S. and South Korea are top revenue regions, indicating expansion opportunities.

## 5.3. Business Value

Expanding the workforce and investing in the software industry directly impacts revenue growth. Geographic expansion in South Korea and the U.S. opens new revenue avenues, strengthening global presence.

## 5.4. Summary of findings

Workforce size and year of establishment are key revenue drivers. The software industry outperforms others, and the U.S. and South Korea are top markets. The predictive model shows good accuracy with an R² score of 0.77.

## 5.5. Strategic Recommendations

Expand the workforce in high-performing sectors like software. Invest in innovation, marketing, and customer acquisition in the software industry. Expand operations in South Korea and the U.S. to capture growth opportunities. These strategies ensure sustainable growth and global competitiveness.

# CHAPTER 6: CONCLUSION

Through detailed data analysis and modeling, several valuable insights have been uncovered, guiding business strategy and operational optimization. Key factors influencing revenue include workforce size, year of establishment, industry sector, and geographic location.

The software sector stands out with exceptional revenue performance, highlighting the importance of investing in technology. South Korea and the United States have been identified as the most promising markets, reflecting their economic growth and market potential.

The revenue prediction model performed well, with an R² score of 0.7761, affirming that the selected features significantly impact the forecast outcomes. Focusing on critical factors such as the number of employees and the year of establishment optimizes the model and ensures high accuracy.

In summary, a strategy that emphasizes workforce expansion, investments in the software sector, and targeting high-potential markets like South Korea and the United States will deliver maximum value. These findings provide a robust foundation for informed decision-making and sustainable strategic planning for the future.

# CHAPTER 7: REFERRENCES AND CONTRIBUTION

## 7.1. References

*scikit-learn: machine learning in Python — scikit-learn 1.6.0 documentation*. (n.d.).

https://scikit-learn.org/stable/

*pandas - Python Data Analysis Library*. (n.d.). https://pandas.pydata.org/

## 7.2. Contribution

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Number | Memeber | Assigned task | Score | Signature |
| 1 | Bùi Đức Trường (Leader) | Problem statement, Building model |  |  |
| 2 | Bùi Lê Hà | Problem statement, Report |  |  |
| 3 | Nguyễn Trần Phương Ly | Data processing |  |  |
| 4 | Nguyễn Quang Hưng | Make prediction |  |  |
| 5 | Khiếu Thị Quỳnh Trang | Data processing |  |  |
| 6 | Vũ Quang Minh | Data processing, Building model |  |  |
| 7 | Nguyễn Ngọc Tú | Building model |  |  |