uOttawa
Faculté de génie
Faculty of Engineering

Phase 3: OLAP Queries, and BI Dashboard
Phase 4: Data Mining

CSI 4142- Introduction to Data Science

Winter 2023

School of Electrical Engineering and Computer Science

Group Members:

Abir Boutahri ,300074850
Elvis MAZIMPAKA, 300113276
Jusley Xavier AMANI MUTANGANA, 300094632

Submission due: April 11th 2023

TABLE OF CONTENTS

INTRODUCTION

This report outlines the deliverables for Phases 3 and 4, which involves conducting OLAP queries and BI dashboard creation in Phase 3, and data mining in Phase 4. The report provides detailed instructions for the team assignment, including the type of data mining techniques used. The report also provides relevant information on resources used for data preprocessing, handling missing values, and feature selection.

I. PHASE 4

A. PART A

In our code, the dataset "weatherAUS.csv" was read using the pandas library and stored in a DataFrame called "df". Since four columns ("Sunshine", "Evaporation", "Cloud9am", and "Cloud3pm") had most of their values missing, they were dropped from the DataFrame.

Next, empty rows were removed from columns with 50 or more missing values. The missing values in the remaining columns were then replaced with the median value of the respective columns using the function "replace_numerical". This function also added new columns for average temperature, wind, rainfall, humidity, pressure, and cloud by calculating the mean values from multiple columns.

Furthermore, object columns were filled with the mode value using the function "replace_object". Finally, duplicates were removed from the DataFrame and a surrogate key was added to identify each row.

The modified DataFrame was then saved to a new CSV file called "modified_weatherAUS.csv" using the pandas "to_csv" function with index=False to exclude the row numbers.

Overall, the preprocessing steps included dropping unnecessary columns, removing empty rows, filling missing values with appropriate measures, and adding a surrogate key to the DataFrame. These steps ensured the quality and completeness of the data for further analysis.

B. PART B

1. SECTION 2

Model	Accuracy
Decision Tree	0.7876682011331445
Gradient Boosting	0.8529567988668555
Random Forest	0.8560995042492918

Classification report for decision tree model:

Time taken to construct decision tree model: 1.1957478523254395 seconds

	Precision	Recall	F1-score	Support
No	0.87	0.86	0.86	17715
Yes	0.51	0.54	0.52	4877
Accuracy			0.79	22592
Macro avg	0.69	0.70	0.69	22592
Weighted avg	0.79	0.79	0.79	22592

Confusion matrix for decision tree model:

```
[[15148  2567][ 2230  2647]]
```

Classification report for gradient boosting model:

Time taken to construct gradient boosting model: 19.50440502166748 seconds

	Precision	Recall	F1-score	Support
No	0.87	0.95	0.91	17715
Yes	0.73	0.51	0.60	4877
Accuracy			0.85	22592
Macro avg	0.80	0.73	0.75	22592
Weighted avg	0.84	0.85	0.84	22592

Confusion matrix for gradient boosting model:

```
[[16798   917][ 2405  2472]]
```

<u>Classification report for random forest model:</u>
Time taken to construct random forest model: 13.532151937484741 seconds

	Precision	Recall	F1-score	Support
No	0.88	0.95	0.91	17715
Yes	0.74	0.52	0.61	4877
Accuracy			0.86	22592
Macro avg	0.81	0.73	0.76	22592
Weighted avg	0.85	0.86	0.85	22592

Confusion matrix for random forest model:
[[16825   890][2361 2516]]

2. *<u>SECTION 3</u>*

The team used three different algorithms (decision tree, gradient boosting, and random forest) to create models and predict whether it would rain tomorrow based on various weather attributes. The models were evaluated based on their precision, recall, F1-score, accuracy, and confusion matrix.

The decision tree model was the fastest to construct and had an accuracy of 79%. It had a higher precision and recall for predicting "No" rain, but performed poorly in predicting "Yes" rain.

The gradient boosting and random forest models had similar accuracy scores of 85% and 86% respectively. Both models had higher precision and recall for predicting "No" rain, but performed better in predicting "Yes" rain compared to the decision tree model.

From the confusion matrix, it was clear that all three models had more correct predictions for "No" rain than for "Yes" rain. This could be due to the imbalanced class distribution, where there were significantly more instances of "No" rain compared to "Yes" rain in the dataset.

In conclusion, the team found that the gradient boosting and random forest models outperformed the decision tree model in predicting whether it would rain tomorrow. However, further analysis may be needed to address the class imbalance issue and improve the models' performance in predicting "Yes" rain.

C. PART C

1. SECTION 2

After training the One-Class SVM model on the given dataset, we identified a total of 1130 outliers. These outliers were identified based on the feature set selected for the model, which included MinTemp, MaxTemp, Rainfall, WindGustSpeed, WindSpeed9am, WindSpeed3pm, Humidity9am, Humidity3pm, Pressure9am, Pressure3pm, Temp9am, Temp3pm, AvgTemp, AvgWind, AvgRainfall, AvgHumidity, and AvgPressure.

Based on these outliers, we can draw several insights from the dataset. For instance, it seems that extreme weather conditions such as high rainfall, wind gusts, and humidity can lead to outliers in the data. These outliers could also be a result of measurement errors or data entry errors. Removing these outliers may or may not help to improve the accuracy of our model and provide more meaningful insights. It is important to note that removing outliers should be done carefully as they may contain important information that can impact our analysis.

II. CONCLUSION

In conclusion, Phases 3 and 4 provided an opportunity for us to apply our knowledge in OLAP queries, BI dashboard creation, and data mining. The project requires the use of different tools and techniques to transform large amounts of data into actionable insights. The report provided detailed instructions for the team assignment, including the type of data mining techniques used. The report also provided relevant information on resources used for data preprocessing, handling missing values, and feature selection. Overall, the project was an excellent opportunity for us to develop our data science skills and to apply our knowledge in a real-world context.