**A logo of a university

Description automatically generated**

**machine learning report (classification project weather)**

**By:**

|  |  |
| --- | --- |
| Name | ID |
| Moustafa Karam | 42010428 |

**Under the Supervision of:**

Eng. Ahmed Nousir

Contents

[**1.** **Introduction** 3](#_Toc154864287)

[**2.** **The Problem understanding:** 3](#_Toc154864288)

[**2.1** **Data Description** 3](#_Toc154864289)

[**2.2** **Problem Statement** 3](#_Toc154864290)

[**3.** **Data Exploration** 3](#_Toc154864291)

[**3.1** **Descriptive Statistics** 3](#_Toc154864292)

[**3.2** **Correlation Analysis** 4](#_Toc154864293)

[**4.** **Data Preprocessing** 5](#_Toc154864294)

[**4.1** **Handling Missing Values** 5](#_Toc154864295)

[**4.2** **Encoding Categorical Variables** 5](#_Toc154864296)

[**4.3** **Feature Scaling** 5](#_Toc154864297)

[**5.** **Model Building** 5](#_Toc154864298)

[**5.1** **Splitting Data** 5](#_Toc154864299)

[**5.2** **Model Selection** 5](#_Toc154864300)

[**5.3** **Training the Model** 5](#_Toc154864301)

[**5.4** **Model Evaluation** 5](#_Toc154864302)

[**6.** **Conclusion** 6](#_Toc154864303)

[Figure 1 Statistics for data 4](file:///E:\Project%20Data%20mining%20and%20big%20data%20report%20for%20data%20preprocessing%20and%20big%20data.docx#_Toc154864308)

[Figure 2 Correlation 4](file:///E:\Project%20Data%20mining%20and%20big%20data%20report%20for%20data%20preprocessing%20and%20big%20data.docx#_Toc154864309)

[Figure 3 plots 4](file:///E:\Project%20Data%20mining%20and%20big%20data%20report%20for%20data%20preprocessing%20and%20big%20data.docx#_Toc154864310)

[Figure 5 Accuracy 6](file:///E:\Project%20Data%20mining%20and%20big%20data%20report%20for%20data%20preprocessing%20and%20big%20data.docx#_Toc154864311)

1. **Introduction**

Begin by providing a brief overview of the problem at hand – the challenges and implications associated with weather. Mention the importance of predicting and managing weather, and how the given dataset can contribute to this.

1. **The Problem understanding:**

The data set is about weather this data contain 21 columns and 25000 rows

* 1. **Data Description**

Explain the significance of each column in the dataset. Highlight the role of factors like rain today, wind, and others in contributing to rain tomorrow.

* 1. **Problem Statement**

Define the problem statement clearly. For example, you could state that the goal is to predict based on various factors in the dataset.

1. **Data Exploration**

This data from Kaggle is about 25000 rows and 21 columns in his part we need to get more information about data and another statistics.

* 1. **Descriptive Statistics**

Provide summary statistics for each column. This could include mean, median, standard deviation, and other relevant metrics. Identify any trends or patterns.

A screenshot of a computer

Description automatically generated

Figure 1 Statistics for data

* 1. **Correlation Analysis**

Explore the correlation between different factors and the target variable (Rain tomorrow). This will help understand which features are more influential.

A screenshot of a computer

Description automatically generated

Figure 2 Correlation

* 1. **Visualizations**

Create visualizations such as histograms, scatter plots, or box plots to better understand the distribution of data and relationships between variables.

A screenshot of a computer

Description automatically generated

Figure 3 plots

1. **Data Preprocessing**

This part for clean data from nulls and duplicates and check outliers to make data ready for deploy model.

* 1. **Handling Missing Values**

Address any missing values in the dataset through imputation or removal.

* 1. **Encoding Categorical Variables**

If there are categorical variables, encode them into numerical format for model compatibility.

* 1. **Feature Scaling**

Normalize or standardize numerical features if necessary to ensure fair treatment by the model.

1. **Model Building**
   1. **Splitting Data**

Divide the dataset into training and testing sets.

* 1. **Model Selection**

Choose a suitable machine learning model for the task (e.g., Random Forest, Logistic Regression).

* 1. **Training the Model**

Train the chosen model on the training dataset.

* 1. **Model Evaluation**

Evaluate the model's performance on the testing dataset. Mention metrics such as accuracy, precision, recall, and F1 score.

A screenshot of a computer

Description automatically generated

Figure 5 Accuracy

1. **Conclusion**

Summarize the key findings and insights from the analysis. Discuss the limitations and potential areas for improvement. Offer recommendations for further research or model enhancements.