Forecasting Local Weather in Exeter, UK using Historical Data to Predict Precipitation Probability

Student Sneha G G

1 Introduction

The most unforgettable aspect of Exeter is its pleasant weather, paired with the intriguing unpredictability of its climate—one moment it's sunny, and just five minutes later, it starts pouring, often reminding us that we've left our umbrella at home. The geo spatial aspect of the Exeter City with the rolling hills along the coastline contributes to the micro-climatic variations resulting in unpredictability. Even though there are technological advancements to accurately predict the local weather, predicting extreme, localized events like storms and sudden snowfall remains a significant challenge. Addressing this gap will help the people to prepare themselves for the unpredictable weather. Hence resulting in the research question which is addressed in this coursework: "**How can local weather, specifically the probability of precipitation in Exeter, UK, be forecasted through the analysis of historical data?**".

2 Dataset Overview

The data source for the research question is obtained from the famous data archive **https://www.data-is-plural.com/archive/** featuring the "**2024.09.25 edition**". This edition provides details about one of the Open-source weather APIs which is Open-meteo (**https://open-meteo.com/**). **Open-meteo**, an open source project built on the data from national weather services, offers a range of weather and climate APIs that are free for non-commercial use. They include weather forecasts (temperature, humidity, precipitation, wind speed, etc.), daily historical weather since 1940, climate change model outputs, marine wave forecasts, air quality assessments, and more. In order to source historical weather data , **Historical Weather API** has been used from the Open-meteo site to download hourly weather data starting from **January 1st, 2010** to **November 15th, 2024**. The dataset contains over 130,392 records of hourly weather data.

3 Overview of Data Wrangling

The data preprocessing is done by checking for missing values in the dataset. Then , the dataset is plotted into histograms to understand the distribution of each variable. In order to determine the precipitation category "**Rain**" or "**No rain**", precipitation (mm) variable is considered where if the value in mm > 0,it is assumed to rain otherwise it will not rain. In order to select the best combination of features, **Feature selection** and **Variance thresholding** methods are used along with Linear regression. The dataset is resampled using **SMOTE** method and test size is 0.2.

4 Machine Learning Technique

The ultimate goal of this research is to determine whether there will be any precipitation i.e.) finding out the probability of precipitation. Since the likelihood of precipitation falls under the categorical label values "**Rain**" and "**No Rain**" , "***Classification Machine Learning Technique***" has been used for this research.Since multiple features are involved in the modeling process, it is **Multi-variate classification task**.

Model	Precision		Recall		f1-score		Accuracy(%)
	No Rain	Rain	No Rain	Rain	No Rain	Rain	
Logistic Regression	0.81	0.78	0.77	0.82	0.79	0.8	79
Decision Tree	0.87	0.83	0.82	0.88	0.84	0.85	85
Random Forest	0.94	0.89	0.88	0.94	0.91	0.91	91
Gradient Boosting	0.86	0.84	0.84	0.86	0.85	0.85	85
k-Nearest Neighbours	0.97	0.83	0.8	0.97	0.87	0.9	89
Perceptron	0.97	0.58	0.29	0.99	0.45	0.74	64
SVM	0.97	0.56	0.21	0.99	0.35	0.72	61

Table 1: Overview of ML models with metrics

5 Overview of Machine Learning Models

For this research , classifier models like logistic regression, decision tree, random forest, gradient boosting , kNN, Perceptron, SVM are used to forecast the probability of precipitation. The accuracy is measured using confusion matrix for each model and corresponding classification report is generated. Apart from these classifier models, Polynomial regression techniques was also used to predict precipitation (mm) which is numerical value resulting in a r2 score of 0.31 with MSE 0.092.

6 Overview of Data Analysis and Outcomes

Initially, for **exploratory data analysis**, **PCA** is done on the dataset to have 2 principal components with the total explained variance ratio of **0.53**. Then, clustering techniques **DBScan and k-means** were used to find out the different categories of data. The Silhouette score for DBScan is **0.38** with 3 clusters whereas for k-means , it is **0.32** with 12 clusters. Further, the data visualization methods are applied to find the trends and patterns in 5 years of weather data. It is observed that Exeter's temperature peaks (80-90°F) in July and August, and drops (20-30°F) in December. September and October consistently saw the highest precipitation (0.15 mm). Humidity, dew, and cloud cover increased with precipitation . High wind speed is observed due to geographical location whereas the curved patterns suggest specific phenomena, like cyclones or jet streams. After training the models, **Random Forest Classifier** model topped the list with an accuracy of **91 percent** while the k-NN model came second with an accuracy of **90 percent**.

7 Limitations of Dataset and ML models

Micro climatic variations can be accurately predicted if the dataset has more geo-spatial information (elevation, terrain) along with radar and satellite data of the particular location. This also requires having more real-time sensors deployed to capture variations within less time. Scaling and sampling is required since real data may not be balanced in nature which decreases the efficiency of ML models. The subset of data is used for clustering since playing with the model parameters resulted in MemoryError (Personal computer)for large dataset. Hence, more computational power is required to analyze big data like weather data.

8 Conclusion

This research on historical weather data has provided valuable insights on the factors affecting the precipitation at a particular geographical location and importance of forecasting the same using machine learning as well as visualization techniques to avoid damages to life and property at large scale.

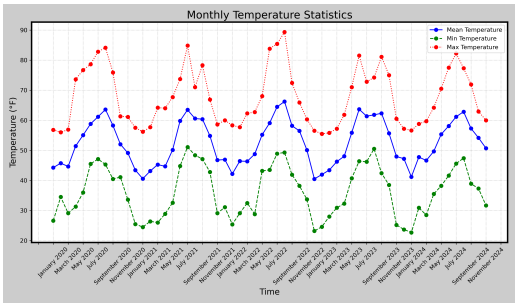

Figure 1: Temperature over Time

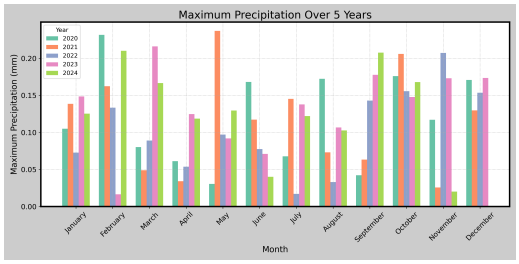

Figure 2: Maximum Rain over 5 years.

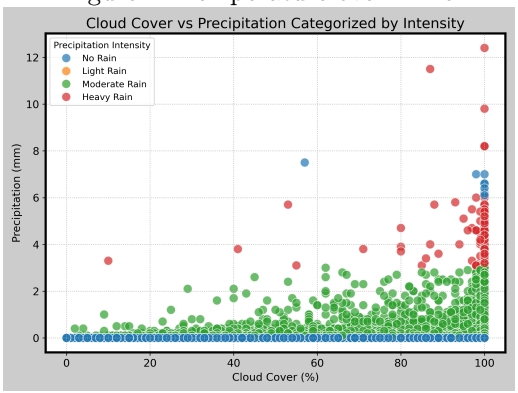

Figure 3: Cloud cover vs Precipitation

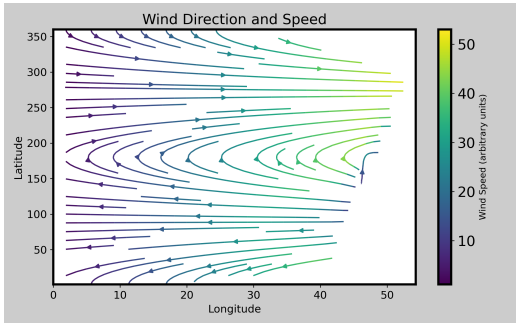

Figure 4: Wind Speed and Direction

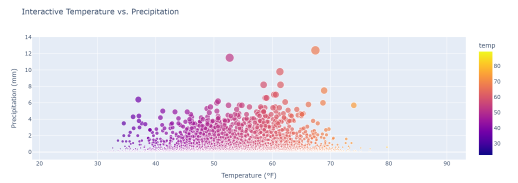

Figure 5: Temperature vs Precipitation

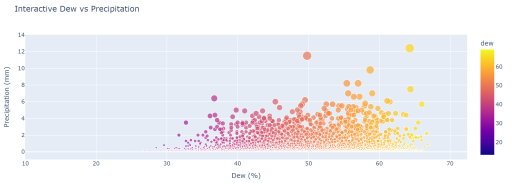

Figure 6: Dew vs Precipitation

References

[1] Gřon, Aurľien, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow concepts, tools, and techniques to build intelligent systems*, Book, Second Edition, 2019.

[2] Suhail Ashraf, *Weather Forecasting: Using Time Series Analysis Under Different Stations Of The UK*, `https://www.researchgate.net/publication/380150077_Weather_Forecasting_Using_Time_Series_Analysis_Under_Different_Stations_Of_The_UK`, April 2024.

[3] Zippenfenig, P., *Open-Meteo.com Weather API [Computer software]*, `https://open-meteo.com/en/docs/historical-weather-api`, 2023.

[4] Fiona M. Rust, Gavin R. Evans, *Improving the blend of multiple weather forecast sources by Reliability Calibration*, `https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.2142`, July 2023.

Generative AI Statement

AI-supported/AI-integrated use is permitted in this assessment. I acknowledge the following uses of GenAI tools in this assessment:

- (~~YES~~ / **NO**) I have used GenAI tools for developing ideas.
- (**YES** / ~~NO~~) I have used GenAI tools to assist with research or gathering information.
- (**YES** / ~~NO~~) I have used GenAI tools to help me understand key theories and concepts.
- (~~YES~~ / **NO**) I have used GenAI tools to identify trends and themes as part of my data analysis.
- (~~YES~~ / **NO**) I have used GenAI tools to suggest a plan or structure for my assessment.
- (~~YES~~ / **NO**) I have used GenAI tools to give me feedback on a draft.
- (**YES** / ~~NO~~) I have used GenAI tool to generate image, figures or diagrams.
- (~~YES~~ / **NO**) I have used GenAI tools to proofread and correct grammar or spelling errors.
- (~~YES~~ / **NO**) I have used GenAI tools to generate citations or references.
- (~~YES~~ / **NO**) Other: [please specify]
- I have not used any GenAI tools in preparing this assessment.

I declare that I have referenced use of GenAI outputs within my assessment in line with the University referencing guidelines.

Appendix

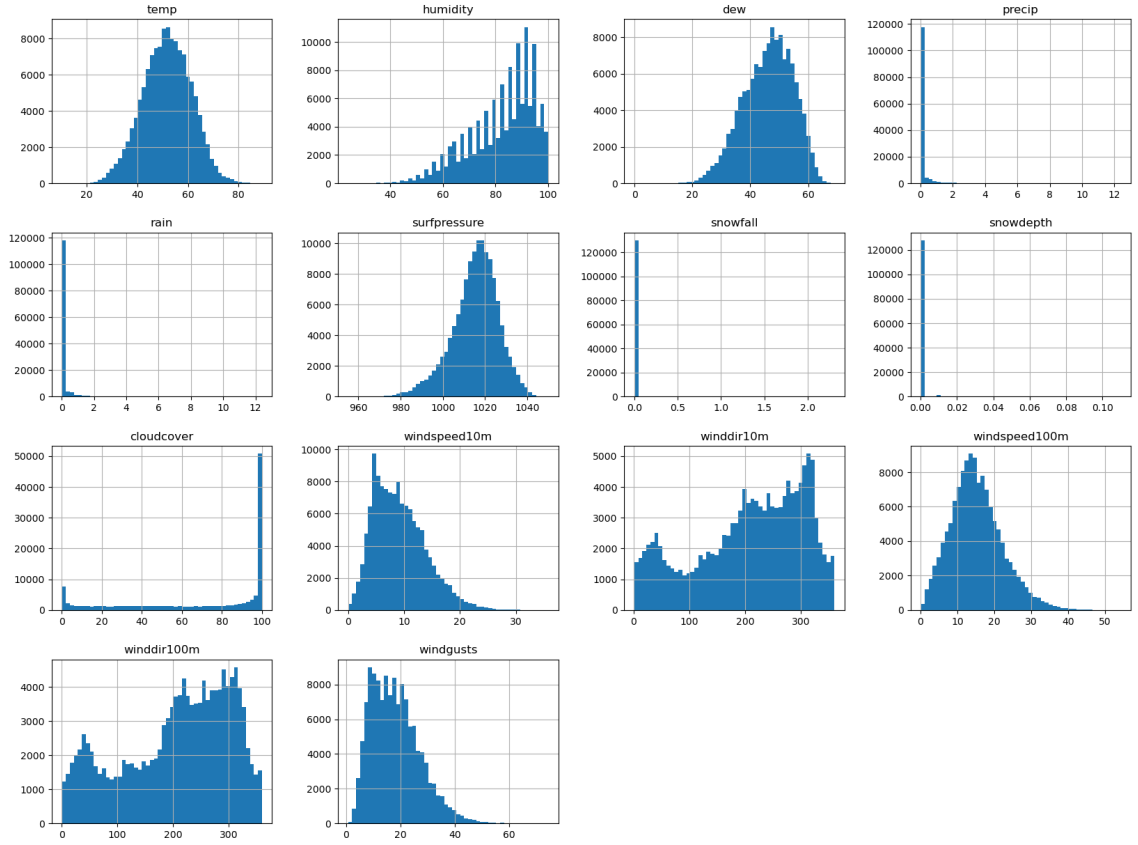

Figure 7: A histogram for each numerical attribute in the dataset. This shows how each attribute is distributed.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 130392 entries, 0 to 130391
Data columns (total 20 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   time           130392 non-null  datetime64[ns]
 1   temp           130392 non-null  float64
 2   humidity       130392 non-null  int64
 3   dew            130392 non-null  float64
 4   precip         130392 non-null  float64
 5   rain           130392 non-null  float64
 6   snowfall       130392 non-null  float64
 7   snowdepth      130392 non-null  float64
 8   wmocode        130392 non-null  int64
 9   surfpressure   130392 non-null  float64
 10  cloudcover     130392 non-null  int64
 11  windspeed10m   130392 non-null  float64
 12  windspeed100m  130392 non-null  float64
 13  winddir10m     130392 non-null  int64
 14  winddir100m    130392 non-null  int64
 15  windgusts      130392 non-null  float64
 16  date           130392 non-null  object
 17  month          130392 non-null  object
 18  year           130392 non-null  int32
 19  MM-YYYY        130392 non-null  object
dtypes: datetime64[ns](1), float64(10), int32(1), int64(5), object(3)
memory usage: 19.4+ MB
```

Figure 8: Structure of Weather Data in pandas dataframe

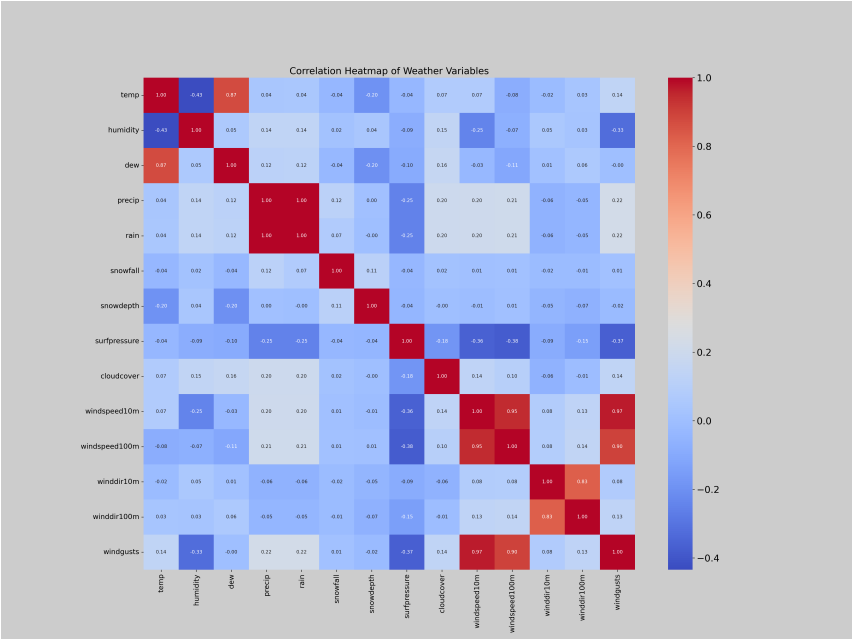

Figure 9: This figure illustrates Correlation between Weather Variables

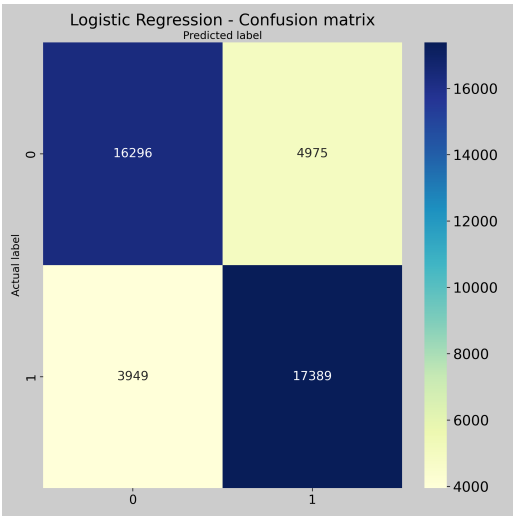

Figure 10: Logistic Regression

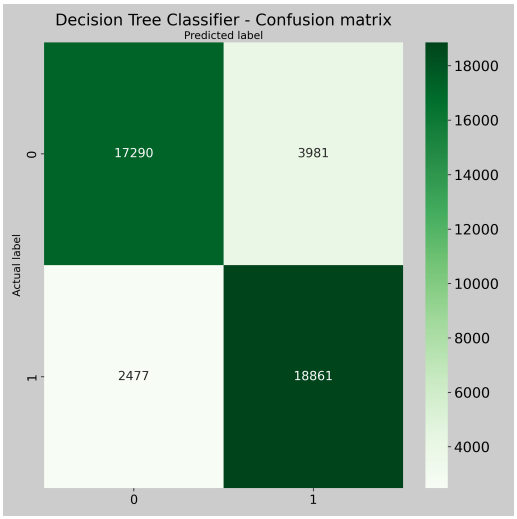

Figure 11: Decision Tree Classifier

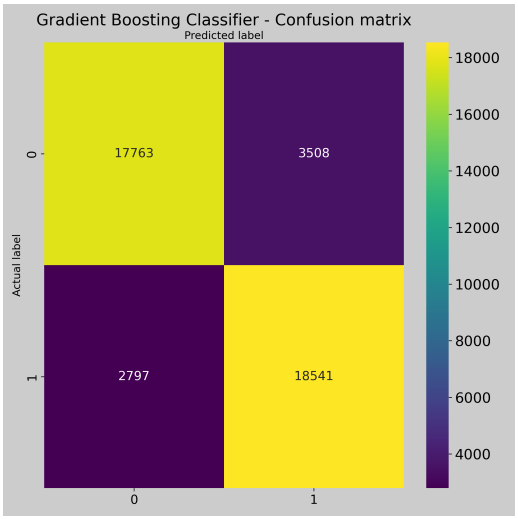

Figure 12: Gradient Boosting Classifier

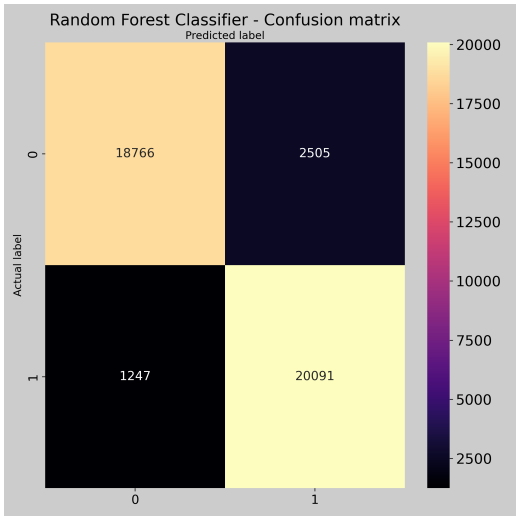

Figure 13: Random Forest Classifier

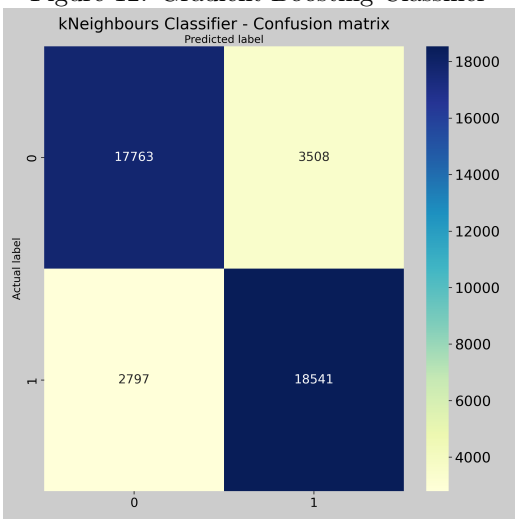

Figure 14: k-Nearest Neighbors Classifier

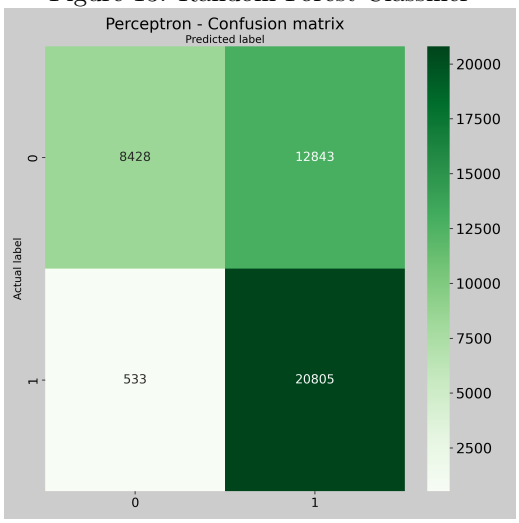

Figure 15: Perceptron

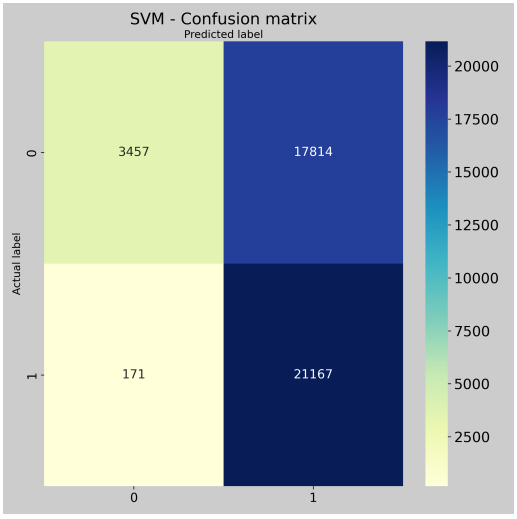

Figure 16: Support Vector Machine Model

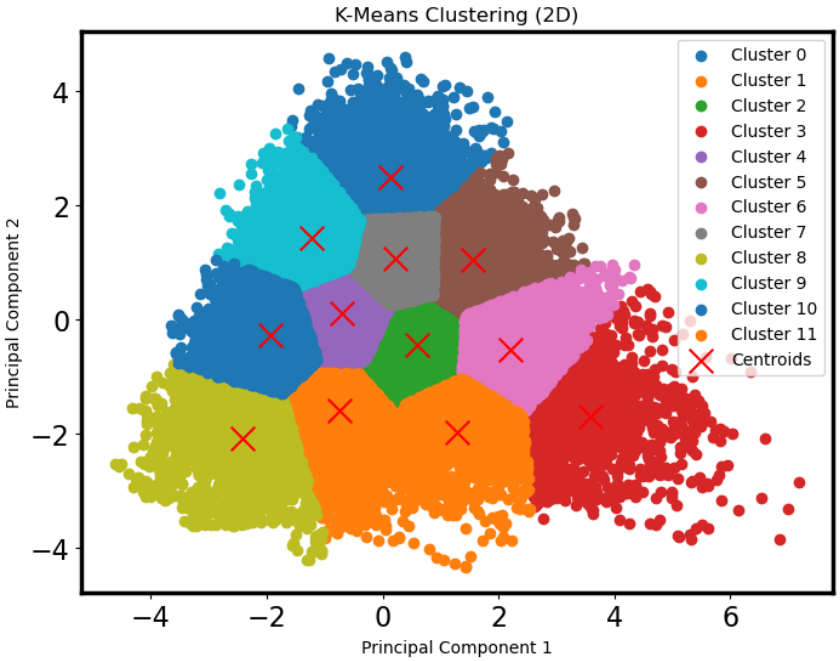

Figure 17: k-Means Clustering - Exploratory Data Analysis

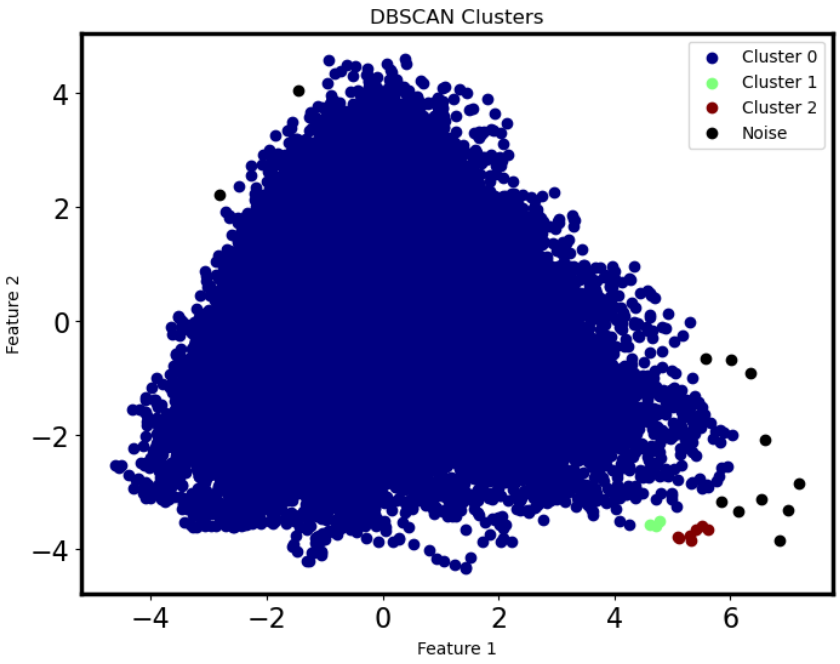

Figure 18: DBScan Clustering - Exploratory Data Analysis