University of Exeter

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

Coursework : Learning from Data

Student : Sneha G G

Student Id: 740003228

University of Exeter

INTRODUCTION:

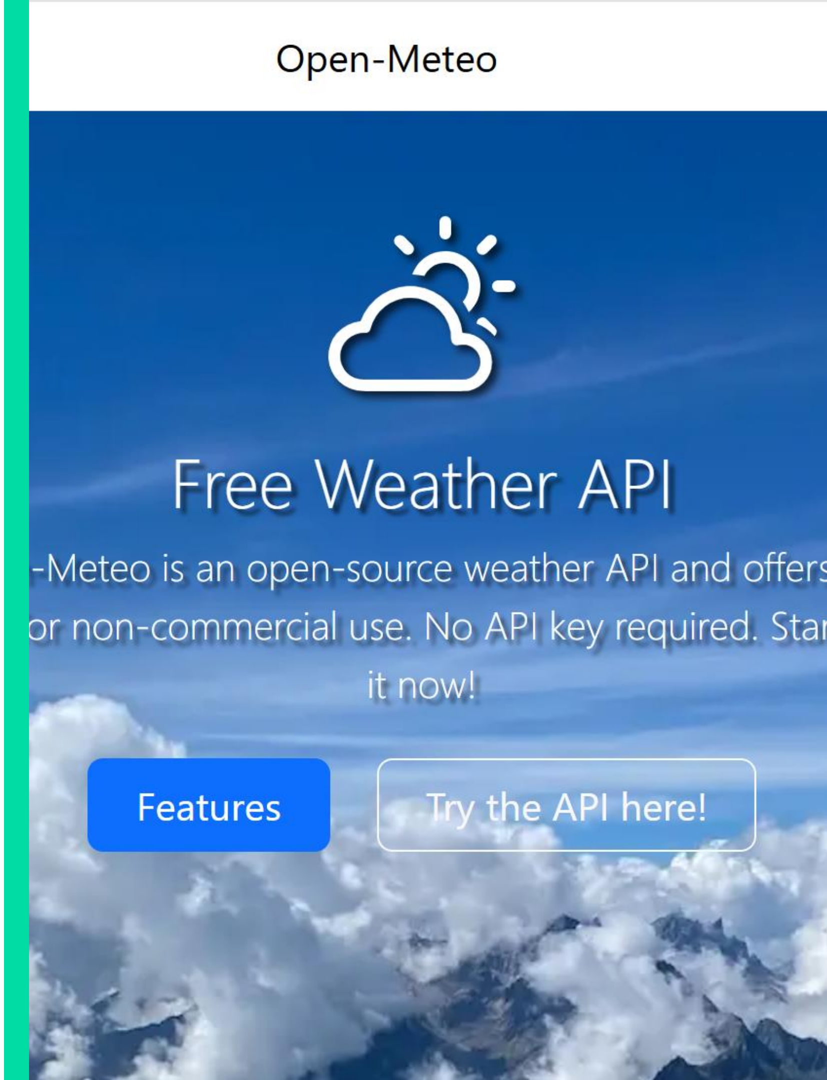
- The most unforgettable aspect of Exeter is its pleasant weather, paired with the intriguing unpredictability of its climate—one moment it’s sunny, and just five minutes later, it starts pouring, often reminding us that we’ve left our umbrella at home.
- The geo spatial aspect of the Exeter City with 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?"

University of Exeter

Dataset Overview:

- The data source is obtained from the 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 Open-meteo (https://open-meteo.com/).
- 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.



Open-Meteo

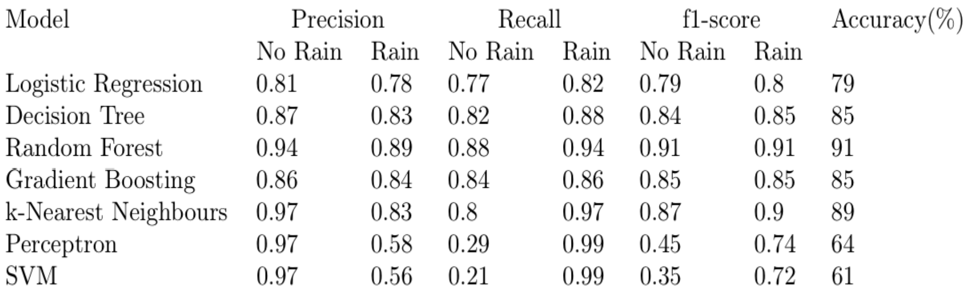
Free Weather API

-Meteo is an open-source weather API and offers
or non-commercial use. No API key required. Star
it now!

Machine Learning Technique Used:

- 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.
- 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.

Machine learning Models Used:



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

- 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, 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.

University of Exeter

Overview of Data Analysis and its 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 (0.15mm). Humidity, dew, and cloud cover increased with precipitation.
- High windspeed 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.

University of Exeter

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 as one of the parameters – temperature didn't have much variance.
- 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 analyse big data like weather data.

University of Exeter

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.
- It helped to study and classify the various trends and patterns in the dataset through the data visualization techniques and how the different variables are correlated to each other.
- On the other hand, the research has also its down side due to less time for analysis and complexities in handling a large dataset and various Machine learning models .

University of Exeter

Thank you