



ADS2001 -GROUP PROJECT

REPORT – MELBOURNE CLIMATE DATA

Callum Battilana: 33227667

Thi Nguyen: 33154791

Yijia Xue: 33051178

Date: 30/05/2023



Table of Contents

EXECUTIVE SUMMARY	2
INTRODUCTION	4
DATA QUALITY	5
MODEL DEVELOPMENT: FILTERING DATA	8
SIMPLE FORECASTING: USING NAIVE METHOD	12
ADVANCED FORECASTING	12
TIME SERIES FORECASTING	17
COMPARING FORECASTED VALUES TO ACTUAL VALUES	19
CONCLUSION	20
REFERENCES	21

Executive Summary

Purpose

This report evaluates the Melbourne Climate Data project, aiming to analyse and develop models for Melbourne's meteorological data from 2011 to 2022. The project's objective is predicting rainfall patterns using weather variables and their impact on rainfall.

Backgrounds

The project utilizes data obtained from the Bureau of Meteorology, recorded every 10 minutes, primarily from the Melbourne Airport automatic weather station.

The data source provides a total of 19 variables of climates, including temperature, humidity, wind speed, and rainfall.

Key Findings

Variables:

To begin with, we cleaned the dataset, removing missing values and addressing data type issues. We then conducted a descriptive analysis of the data, finding that the sampling time may have been irregular. In order to resolve this issue, we resampled the data and created separate tables for analyzing the data by month, day, hour, and half hour.

After assessing the raw data and cleaning it, it was necessary to simplify the model before proceeding.

- Through correlation analysis, we found that dew point temperature showed a strong linear correlation with atmospheric pressure (E), gamma, and calculated apparent temperature, and removed them from data.
- We use a machine learning model called Random Forest to analyze the most efficient variables about rainfall. Wind Direction, Humidity and MSLP are the most important variables.

Forecasting Model:

To account for variable influences and changing weather patterns, we used various modelling techniques, such as Random Forest, SVC, Adaboost, K-Means Clustering, and Logistic Regression, which were applied and compared for accuracy.

- The Random Forest model, along with its alternatives (Adaboost, K-Means Clustering and Logistic Regression), demonstrated favourable accuracy when using all available data variables, except for K-Means Clustering.
- For modelling that uses partial variables, the SVC model, considering MSLP and air temperature, exhibited good accuracy (70 %)for predictions.

Practice

In the time series forecasting, given the randomness of the data in the selected sample (May 18th -26th, 2016), We introduce a new model ARIMA (Autoregressive Integrated Moving Average) and use several models, including ARIMA, to forecast on the climate of May 26, 2016, and compare them with the real data.

- By comparing the predicted results with the actual data, the Random Forest Regression model using all the data variables for prediction has the highest accuracy.
- Secondly, the SVC model with only MSLP and Air temperature was used for prediction, which proved that the accuracy of two variables could be as good as that of multiple variables.
- The accuracy of the newly mapped ARIMA model is not high.

Conclusions

Based on the findings, the Random Forest Regression model, incorporating multiple variables, and the SVC model, focusing on MSLP and air temperature, are recommended for meteorological prediction. The Adaboost model also demonstrated accuracy and can be considered as an alternative approach.

Introduction

This dataset describes the various patterns of the weather at Melbourne airport from January 1st 2011 to the present day. The main weather features include the temperature, dewpoint, windspeed, humidity, and air pressure. The data was collected at ten minute intervals from the Australian Bureau of Meteorology.

To understand how this data is related to predicting rainfall, we need to understand the science of how rain forms.

Background Information

Rain is formed through a process called the "water cycle". Water vapour formed from oceans or plants from photosynthesis rises into the clouds. If the right conditions are present, these water molecules cool and condense causing the droplets to grow. When those water molecules are heavy enough, they fall to the ground again as rain.

Conditions for Rain: Low Relative Humidity

To have the right conditions to form rain, the air must have low relative humidity. Relative humidity is the percentage of water vapour air can hold at a particular temperature. If it is low, then it is easy to reach 100% humidity with a small amount of water vapour. At this point the water vapour condenses forming clouds. Relative humidity depends on the temperature. If the temperature is higher, then the relative humidity increases as well. It is also dependent on the air pressure, as pressure can be seen as an indicator of how much "stuff" (particles) is stored in the air. Therefore if pressure increases, relative humidity increases as well, as more water vapour can be stored. To measure relative humidity, we can calculate the difference between the drybolt temperature (the standard temperature of the air) and the dewpoint temperature (temperature to which air must be cooled in order to produce condensation (dew)). A larger difference represents a lower relative humidity.

Conditions for Rain: Wind Speed and Direction

The rain must be moved from the ocean to land. In order to do that there must be wind blowing towards it. This means that we need the wind direction to be pointing towards land (in this case the airport). To measure how much the wind

is blowing towards the airport, we calculate the air pressure difference (between the airport and at sea level). A greater difference means a stronger force pushing the rain towards the airport. This means it is more likely to move closer to the airport.

Challenges

Weather is notoriously difficult to predict accurately. The atmosphere is constantly changing and evolving. Indicators such as air pressure and temperature are constantly changing on a minute by minute basis in a chaotic fashion. Melbourne weather is especially difficult to predict, due to the vastly different climates of the Antarctic and the Australian desert interacting with one another.

Solution Methods

Due to the challenges in predicting rainfall, we will be only trying to predict rainfall for one day based on the rainfall of the previous day. We understand that the key to knowing when rainfall will occur is whether the relative humidity is low or not, as well as whether the wind is blowing towards the airport. We will therefore try and predict relative humidity and wind direction for the next day, based on our weather indicators as well as our prior knowledge of how they work. We will also only predict whether a particular day will rain or not, to simplify our model even further.

Data Quality

Data Manipulation and Cleaning

```

Year                object
Month              object
Day                float64
Hour               float64
Min                float64
Air Temp (degrees C) float64
Apparent Temp (degrees C) float64
Dew Pt Temp (degrees C) float64
Humidity (%)        float64
Wind Direction      object
Wind Speed (km/h)   object
Wind Gust (km/h)    float64
MSLP (hPa)          float64
Rainfall since 9 am (mm) object
gamma              object
Calculated Dew Pt Temp (degrees C) object
E (hPa)             float64
Calculated Apparent Temp (degrees C) object
dtype: object

```

At the beginning of data cleaning, after we use `df.dropna` remove all the empty blanks from the data frame, we use `df.dtypes` to check all columns' data types.

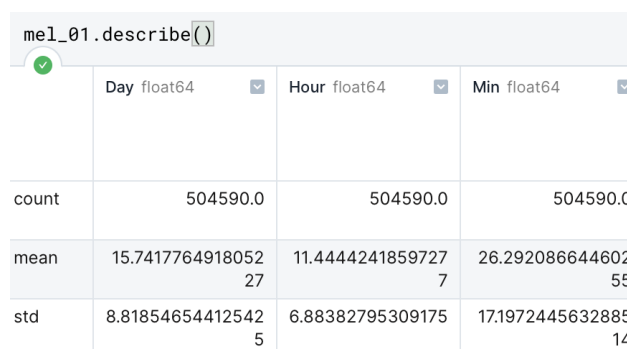
According to the result in figure 1, there are potential quality issues with the data sets based on the problem of data types.

Figure 1: Data types

For time units such as years and months, they should be in integer format. Given that most variables should record numbers, not words, and that objects, not floats, are used as data styles in the series, means that there are several errors in the raw data that need to be removed or fixed.

After we solved the data types of problems of the data, `df.describe` was used to analyse the data set and derive descriptive statistics to find problems in numerical aspects.

As you can see from figure 2, the data in the descriptive analysis is not ideal for the fixed hours and minutes of the day. In addition to the fact that we delete data to format the data correctly, it is also possible that the data sampling time is irregular.



	Day float64	Hour float64	Min float64
count	504590.0	504590.0	504590.0
mean	15.741776491805227	11.44442418597277	26.29208664460255
std	8.818546544125425	6.88382795309175	17.197244563288514

Figure 2: Data description

To ensure the accuracy of subsequent analysis and data modelling, we will re-sample the data and independently build tables according to the month, day, hour, and half hour to carry out different analyses.

In addition, as shown in the picture above, the wind direction's data type is an object because Data is represented using abbreviations for cardinal directions rather than angles. This prevents the code from being automatically parsed and should be converted to the corresponding Angle before analysis.

After resampling the collected data into daily and monthly statistics, we compared air temperature and rainfall monthly data with the data from the Bureau of Meteorology website to ensure the data used is valid.

It is noticeable in figure 3, that the temperature data are similar between the collected data and the data from the website.

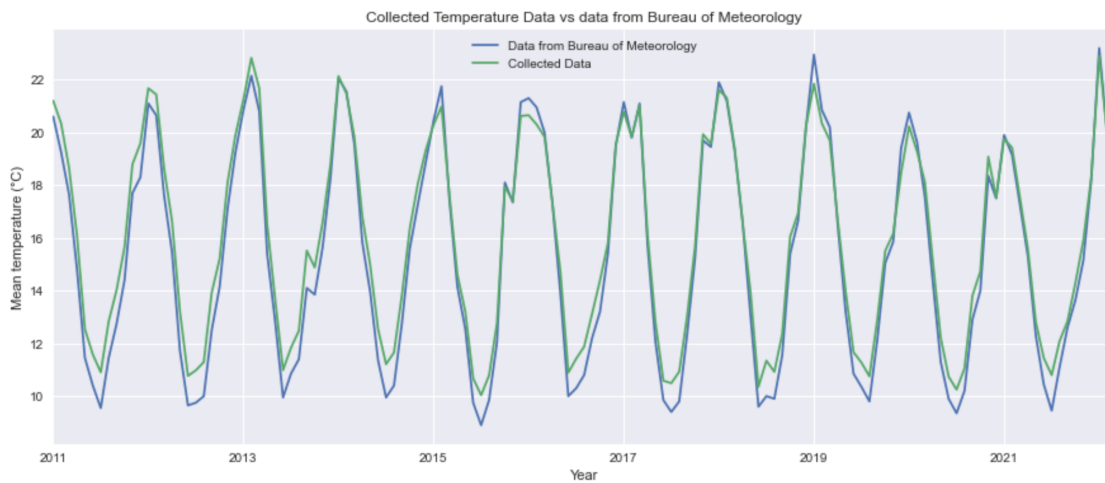


Figure 3: Collected temperature data vs data from the website

For the rainfall data in figure 4, there is a difference as some data relating to the amount of rainfall are missing in the collected data. But the trend pattern of both data is still pretty similar, therefore, the data can still be used for developing forecasting models later on.

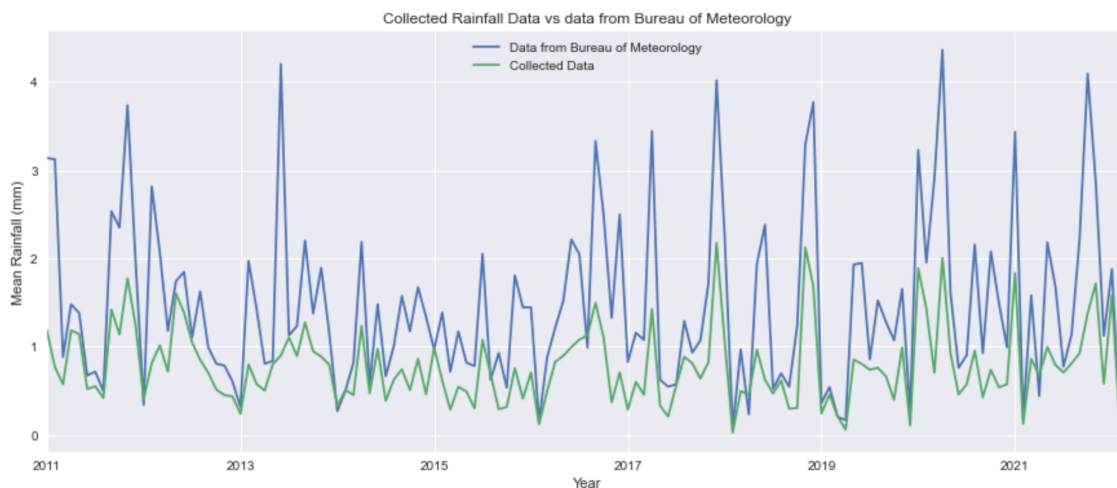


Figure 4: Collected rainfall data vs data from the website

What we finally checked was whether there were significantly more days that rained compared to those that did not. We want to make sure that our data is not biased to days with rainy or non-rainy conditions.

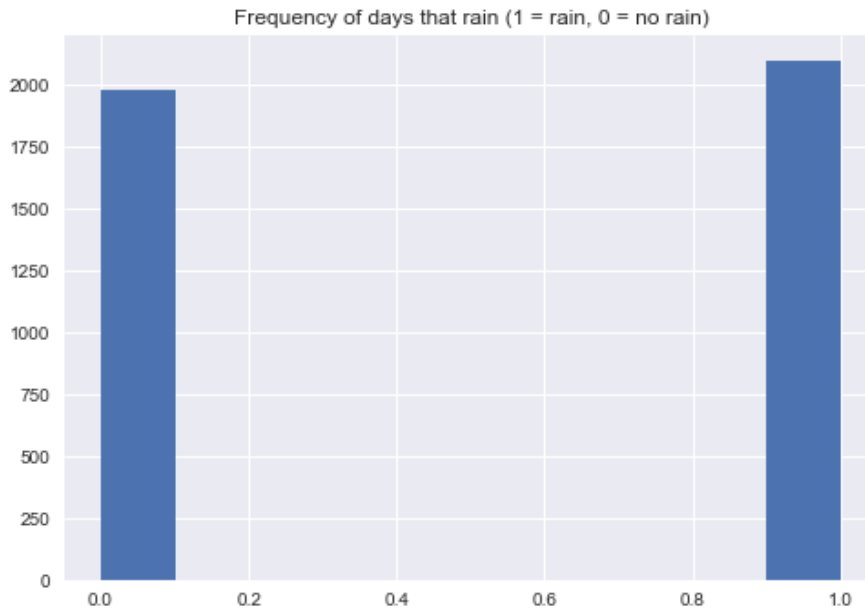


Figure 5: Frequency of days that rain

We found that we have the same amount of days that rain as those that did not (figure 5). This means that we do not need to split the data into even sections for rainy and non-rainy days.

Model Development: Filtering Data

Before we do any analysis, we need to simplify our data as much as possible and only keep the “important” variables. This means that we will not have any redundant variables that do not help at all in predicting rainfall. Another reason is that the weather changes constantly, and as such we need to limit the variables we have to predict for forecasting purposes to a minimum.

Finding Multicollinearity

Multicollinearity is where one feature variable has a very high correlation with another. Multicollinearity removes our ability to determine the actual importance of feature variables to the target. This is because our feature importance assumes that one variable is affecting the target completely independently from the other ones. This of course is not true if there is high correlation between the variables.

The threshold for removing variables with high multicollinearity is subjective, but we will use the common benchmark of when r (correlation coefficient) is greater than 0.9 or -0.9. To put things into perspective, an r -value of 1 or -1,

means that you can perfectly predict one variable exactly based on the other. 0.9 or -0.9 is very close to that level.

We plotted the r-value between each of our features with just how well they correlate to each other on a straight line (Linear Regression) (figure 6).

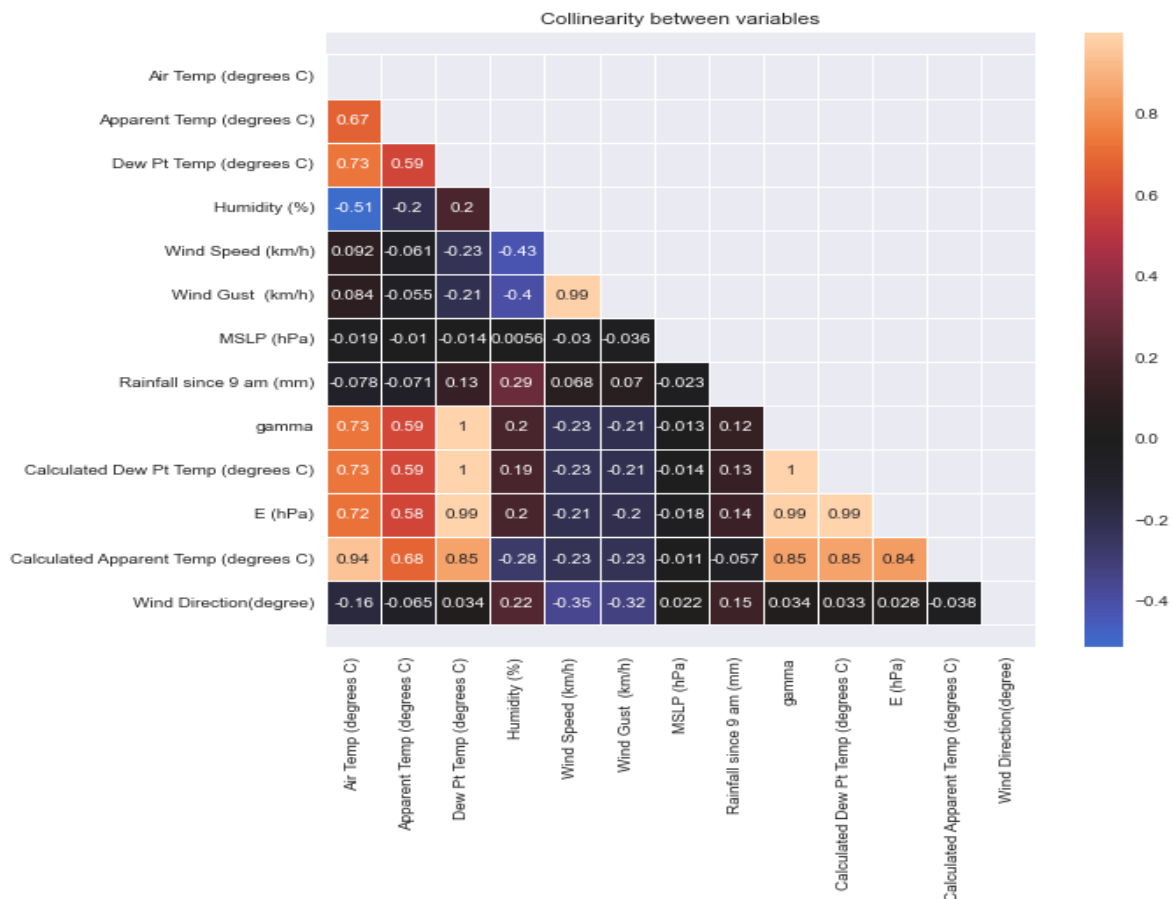


Figure 6: Collinearity between variables

We can see that dew point temperature is perfectly linearly correlated with E (atmospheric pressure), gamma, and calculated apparent temperature.

We realize that E, gamma, and calculated apparent temperature are all just calculated from the Magnus Formula which is shown here.

https://en.wikipedia.org/wiki/Dew_point

Where T is the apparent temperature, and RH is the relative humidity. Gamma we see is just a variable that is part of the Magnus Formula to simplify calculations, and the calculated temperature and atmospheric pressure are calculated based on this formula. We can therefore remove these values from

our data when classifying whether a day rains or not. They have perfect correlation to each other since they are derived from a formula.

Finding Feature Importance

We will finally use a machine learning model called Random Forest to analyse which variables most affect whether a day rains or not.

Random Forests is an ensemble of decision trees. A decision tree predicts what target characteristics a dataset will have (for example if the day is raining or not), by looking at the rest of the data (humidity, temperature ect). It then sorts the data into categories. For example, it may compare days in which humidity was greater than 50%, and those less than 50%, creating branches to put those days into. It will keep doing this until either the category is pure (only the data points with one target feature are in them) or until the depth of branching categories is bigger than we allow. When it creates branches, it will make sure that they are as pure as possible. We keep creating decision trees for many random selections of our data, and average the criterion for them to be put in certain categories.

We find that the accuracy, precision, and recall of the model is about the same. They are high enough (about 76%) that this model can generate useful insights.

We will now use this model to predict feature importance. This is determined by seeing how the model's accuracy changes with the removal of each variable.

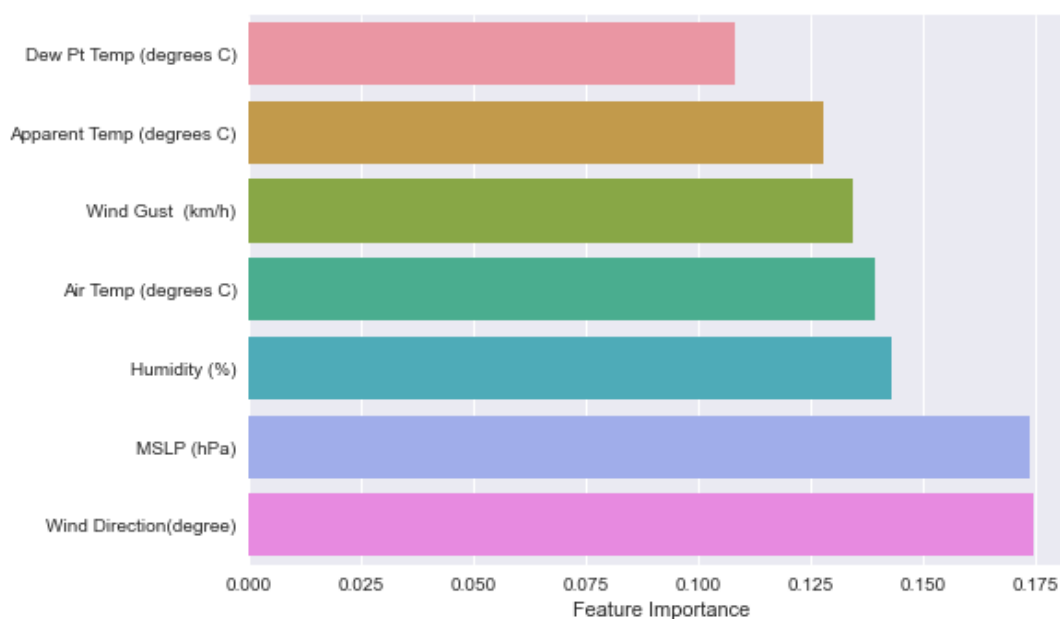


Figure 7: Feature importance

We can see from figure 7 that wind direction, humidity, and MSLP (mean sea level pressure) are the most important variables. This is what we predicted in our hypothesis.

MSLP is the atmospheric pressure recorded at the weather station (E) extrapolated to mean sea level. This means that it is always reduced from "standard" atmospheric pressure. MSLP is important for knowing the direction of wind. The lower the pressure, the clouds will be moved more strongly from the ocean to Melbourne Airport.

We predicted that the difference between the dew point temperature and the drybolt temperature (Apparent Temperature) is also an important indicator of our rainfall. Our model predicts that it is the second most important feature in our data. Thus our hypothesis was proven correct.

We therefore removed our Dew Point and Apparent temp from our model entirely and replaced them with their difference.

This was the final result of our feature importance of our model (figure 8).

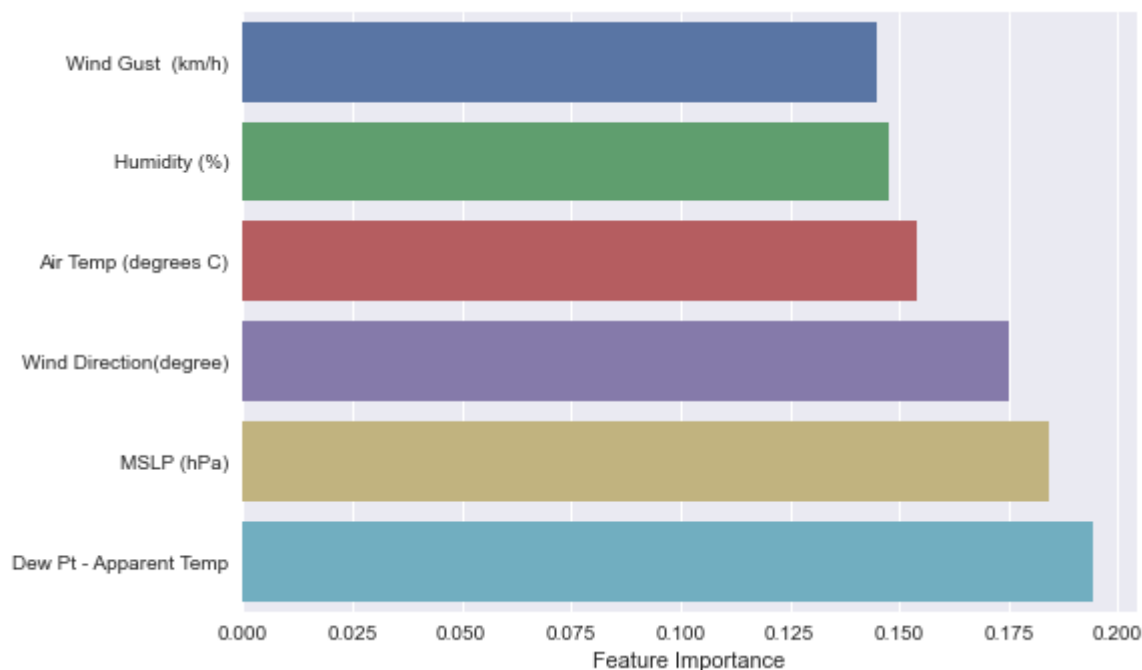


Figure 8: Feature importance

We have now simplified our model so that we only include important variables without any multicollinearity. We will now use only these variables to try and predict whether a particular day rains based on the day before.

Simple Forecasting: Using Naive Method

What we will firstly do, is use a "naive" method of prediction. This means that we use the current day's average conditions (Wind Gust (km/h), Humidity ect) and see whether we can classify accurately whether the next day will rain or not based on the current data.

When we try to predict the day before based on the current day using Random Forest Classification, we get an accuracy of only 52%. This is as accurate as flipping a coin to predict the weather. This is expected, since the weather drastically changes and as such the conditions of one day are drastically different from the next. The reason why we used this method was just as a benchmark to measure other forecasting methods against.

Advanced Forecasting

We now need to develop more advanced forecasting. We know that the weather conditions on the previous day are not the same as the day before. We will therefore try to predict the next days' conditions, by using some sort of regression model on each of our feature data. We will then forecast the mean of each feature variable for the next day. We will use half-hourly data from a week before to predict the next day. This is because we were given advice that this will increase the accuracy of our model rather than just using the previous day. Before we embark on this however, we will try to look at more classification models to predict whether a day will rain or not besides Random Forest Classification. Perhaps they might give us better accuracy.

Classification Model: SVC

SVC (Support Vector Classification) works by creating a "boundary" for data that rains and does not rain when comparing different features.

The reason why SVC will be considered, is that it has the potential to simplify our model even more. If we have good clear boundaries between two variables, then we only need to use two variables instead of six to predict whether a day will rain or not. If those two variables are also predicted well by our regression model, then we can safely just use those instead of the other more inaccurate feature variables.

Firstly, we created a scatterplot of all the feature variables. We saw that there were still a few "odd" values in our data, such as wind gusts and MSLP being less than zero for some of the data (which is impossible). What we did was

remove those values and all other outliers in our data (where the value is less or greater than 1.5 times the interquartile range).

We now saw that all the data is clustered in relatively the same area.

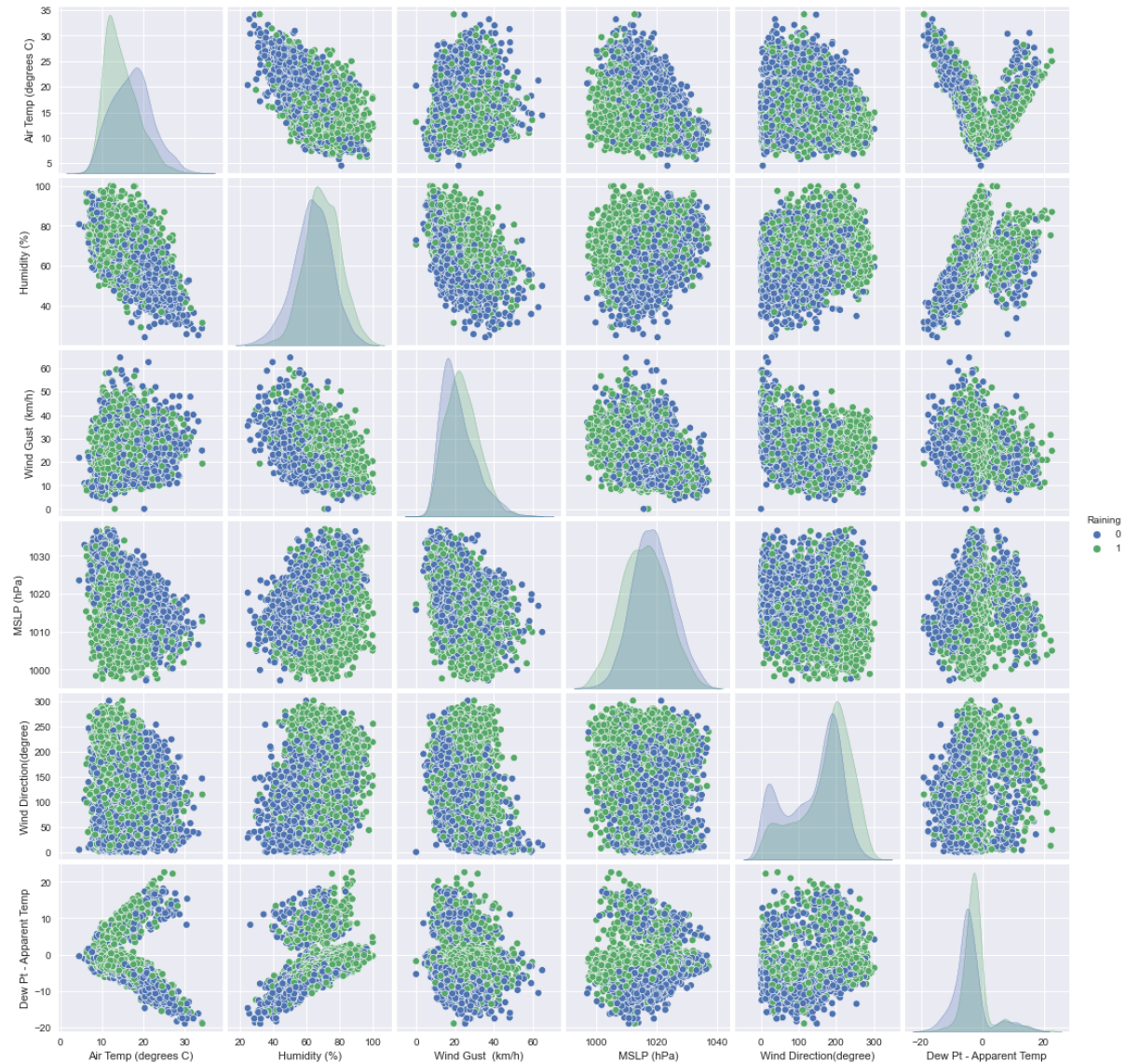


Figure 9: Scatterplots

Analysis

Looking at these scatterplots in figure 9, there seems to be some interesting features. We see a clear cluster of rainy days and non-rainy days when we compare windgust to humidity and MSLP to air temperature. This means that we could try doing some SVC to find the best boundary of those clusters.

We will now draw the best lineary boundary comparing MSLP and air temperature. We will use soft boundaries as it takes too long to create otherwise (with such a large amount of data).

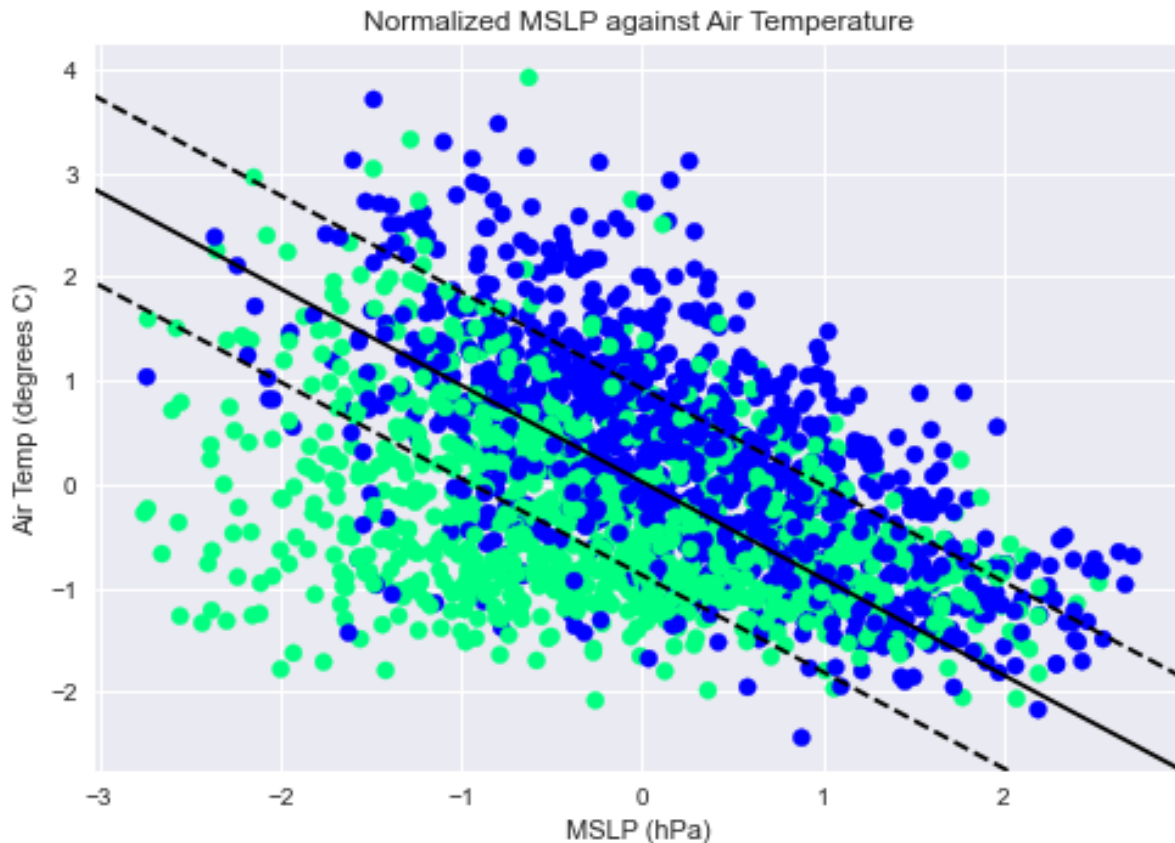


Figure 10: Normalized MSLP Against Air Temperature

We can see that in figure 10, the green data (representing the days with rain) is located below the black line, and the blue data (days with rain) is above the line. The accuracy for this model is 70%. Similar tests with wind gust and humidity show an accuracy of 67% which is decent as well. We see that these models can certainly be used for our forecasting analysis later.

After looking at the obviously good boundaries, we looked at the rest of the scatterplots and performed other SVC boundary methods. There are also rbf and polynomial boundaries that we can use on our data. We found that all other SVC models cannot be used. There was substantial overfitting. In other words, the models fitted too exactly with any outliers present. There was therefore probably no strong barrier between the variables. This is shown in figure 11.

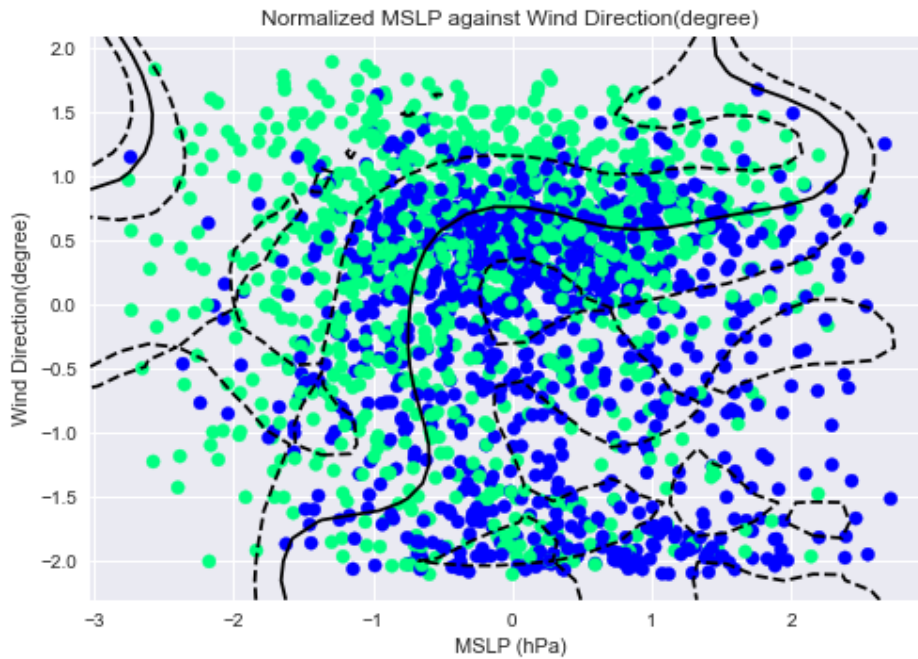


Figure 11: Normalized MSLP against Wind Direction

Classification Models: Adaboost, K-Means Clustering, Logistic Regression

All of the other models we tried were simply alternatives to Random Forest. They use all of the “important” variables, to try and classify whether a particular day will rain or not. They use different approaches however.

Adaboost

The iterative process for AdaBoost is based on associating weights with each observation. Then after each iteration the weights are adjusted to place more emphasis on classifications that the model incorrectly predicts. These weighted target values are then used to create an updated model. Then the process is repeated. The rate at which the weights are updated at each iteration is governed by a parameter called the learning rate, and choosing the optimal learning rate is critical to optimizing AdaBoost. If the learning parameter is too large you get overshoot, if it is too small the convergence is very slow.

To compare with Random Forest, we checked the accuracy of using the data with our Random Forest classification on Adaboost. We saw that it was 72% without any parameter tuning. We therefore tuned the parameters to find the best accuracy. This is done by checking manually different values of the learning rate.

The optimum accuracy is 74% for sixty-five iterations and a learning rate of 0.8. We see it is lower than Random Forest Regression, but not by much.

K-Means Clustering

kMeans clustering works by investigating the features and assuming that clusters in that data that lie close to each other in feature space will have the same classification. Seeing that our clusters for rainy and non-rainy data are very close to each other in our scatterplots from earlier, this probably is not a suitable model for us to use. We see that the accuracy of this model is 16%, which proves this fact. We will not use this model.

Logistic Regression

Logistic Regression is the most simplest model we are exploring. It predicts the probability that our datapoint is classified as rainy or not-rainy depending on previous features. We find that the accuracy of this model is 74%. This is just as good as our Adaboost model.

Time Series Forecasting

We can now finally try and forecast our feature variables based on half hourly data up to a week before. What we did was try and forecast one specific day, and use the models we forecasted to predict every other day in our data.

Forecasting May 26th 2016

For all features we started by graphing them over the previous week (May 18th-25th). We will look at the Air Temperature over time as an example (figure 12).

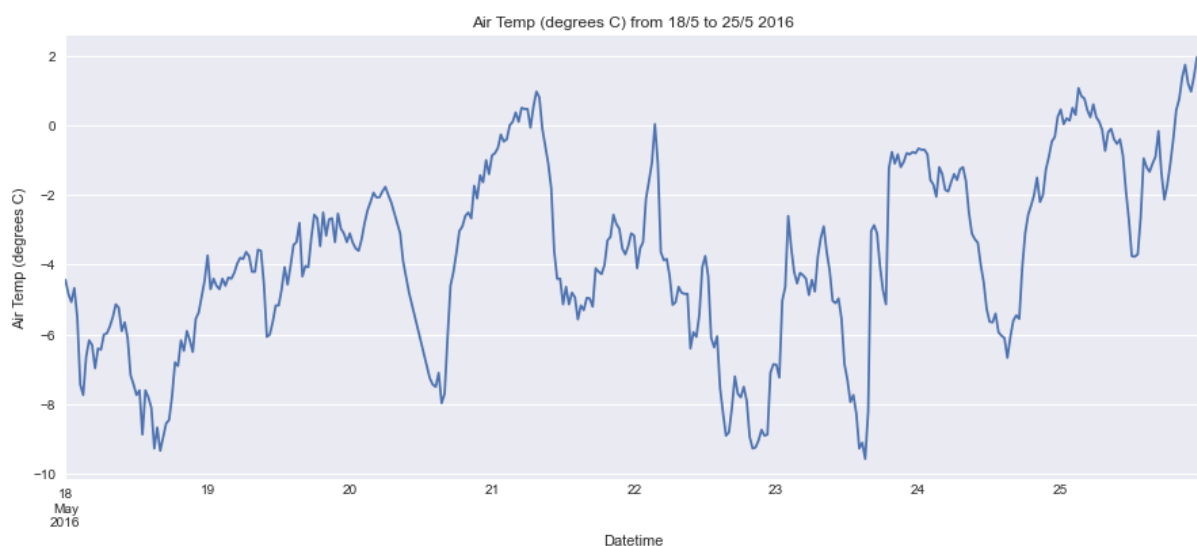


Figure 12: Air Temperature from 18th of May to 25th of May 2016

All the feature variables had a graph like this. We can see no real patterns in the data at all at first glance. It seems almost random. Any pattern we can use to

create a forecast is not obvious using the human eye. Therefore, simple regression models such as linear regression, or any polynomial or logarithmic regression cannot be used here.

ARIMA modelling

ARIMA (Autoregressive Integrated Moving Average) models can tell us if there are patterns in our data. It can tell us whether the previous data influences the present (Autoregressive), and if data is correlated with data a certain period of time behind, in this case every previous day (Differencing, which is opposite to Integration). Our ARIMA model did difference the data by one. This means that our data is not random according to our model. This was the case for all of our features, not just Air Temperature.

Indeed, our ARIMA model can also be used for forecasting. This is our forecast for Air Temperature on May the 26th 2016 (figure 13).

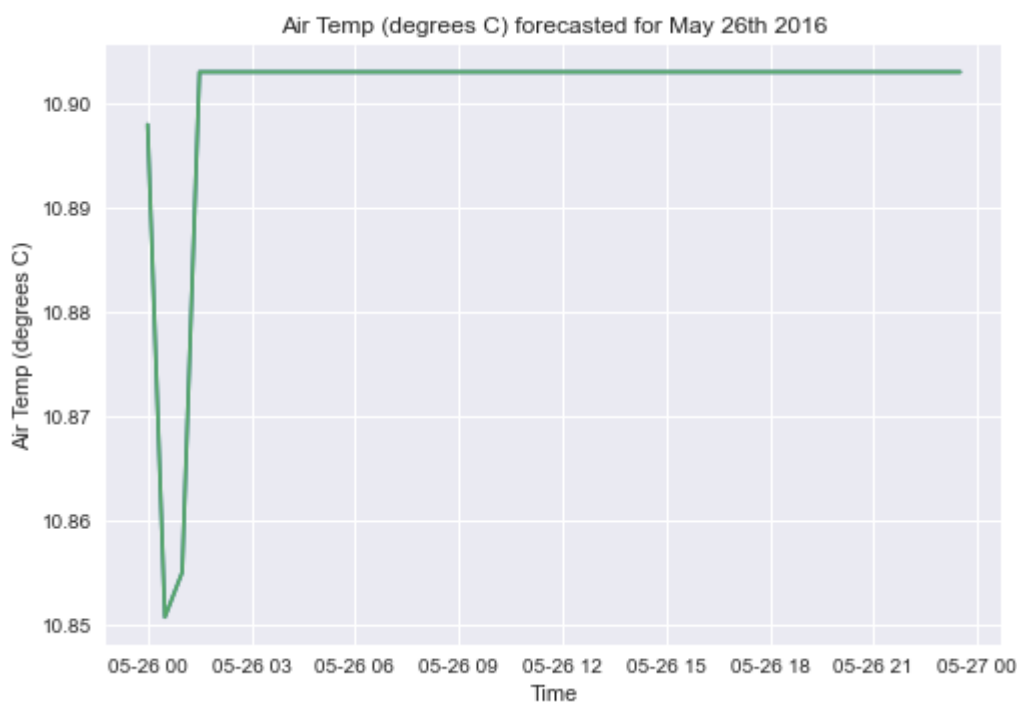


Figure 13: Air Temperature forecasted for May 26th 2016

As you can see, it is not very good. It is mostly a straight line. This is because for all our features the model did not forecast based on the previous data (Autoregressive part was 0). The data was too complicated to see obvious consistent patterns for the previous data. ARIMA models are no better than using the naive method for forecasting.

For all of our data however, Random Forest Regression improved on ARIMA for forecasting. It was able to use the data before for a longer period of time. We therefore will use Random Forest Regression to forecast whether the day will rain or not at midday.

Accuracy of Forecast

For May 26th 2016, the accuracy of our results was surprisingly accurate. Here is the accuracy of some of our variables in figure 14.

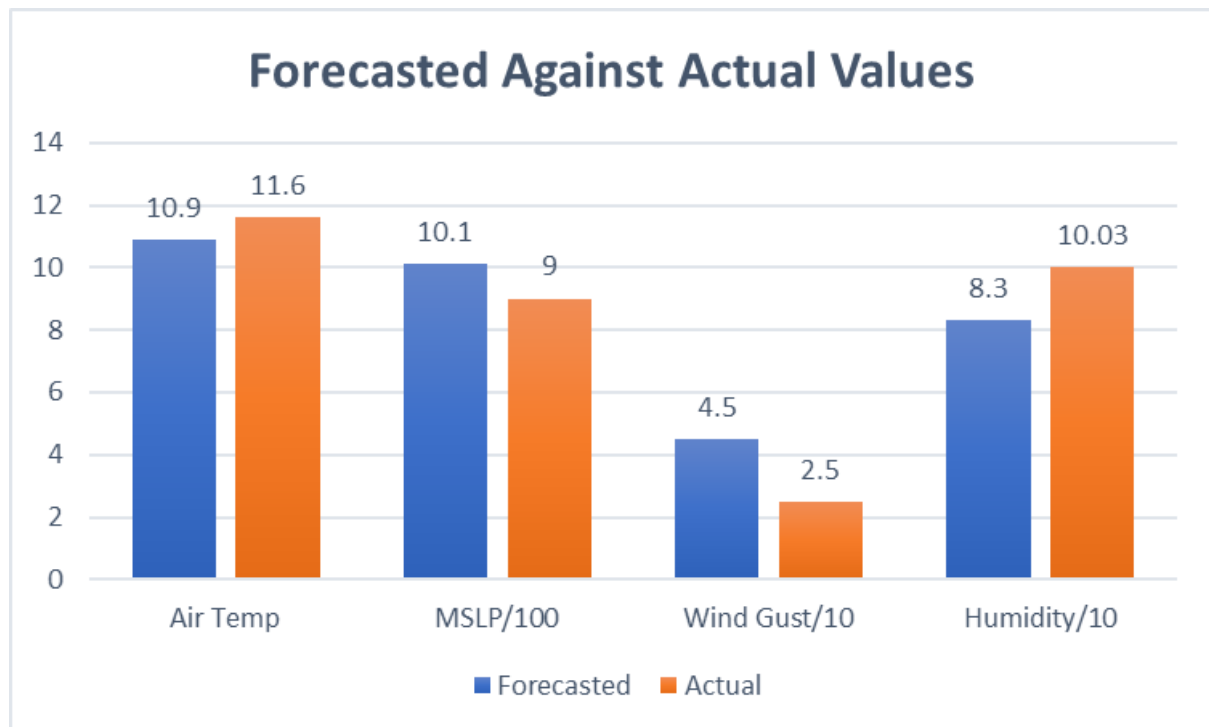


Figure 14: Predicted Feature Variable to Actual Feature Variable

When we look at the variables, we can see that our SVC model comparing Air Temperature to MSLP should give us better results than our other SVC model comparing Wind Gust to Humidity.

Comparing Forecasted Values to Actual Values

We performed the same method with predicting all of our days in our data based on the previous seven days using Random Forest Regression. Firstly, we found the accuracy of our forecasts using the various classification models we described earlier (figure 15).

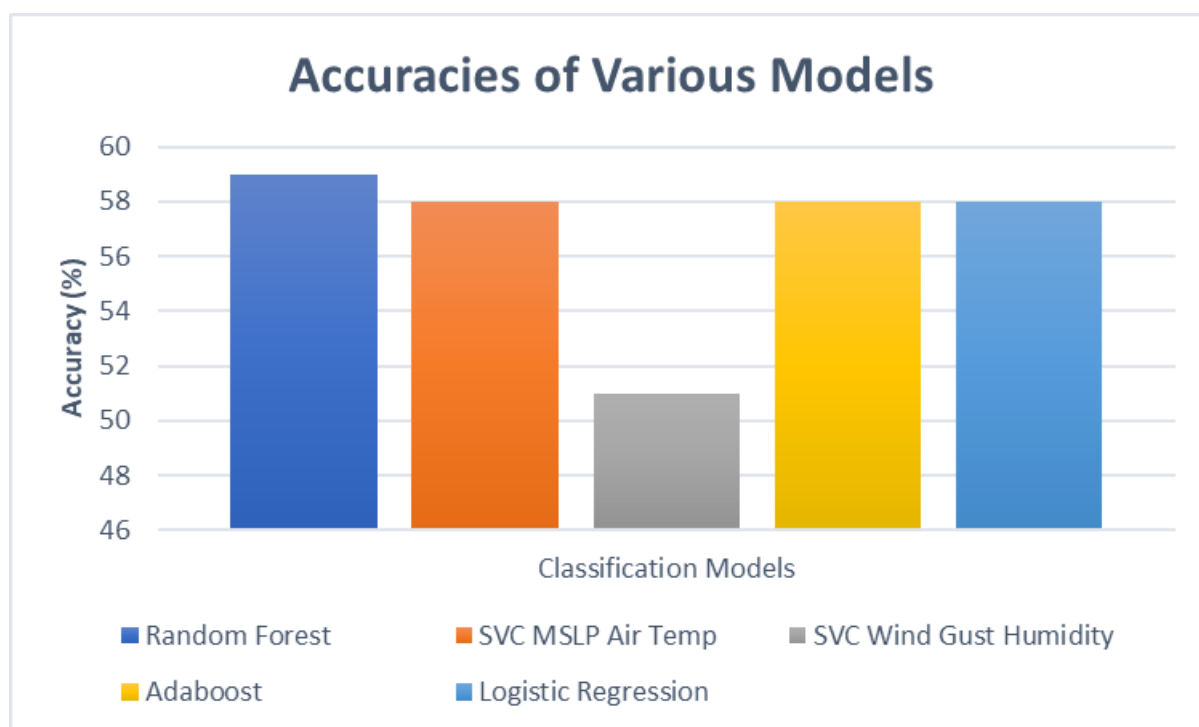


Figure 15 Accuracies of Various Models

We can see that for most of our models, the accuracy is about the same as each other at about 60%. Random Forest is however the most accurate.

We get some powerful insights when looking at this chart. Firstly, we know that we have done better than the naive method. Secondly, we can see that our SVC model for MSLP and Air Temperature is just as accurate as using all the features for our other models. This means that only using two variables can be as accurate for forecasting as using six. Thirdly, Wind Gust and humidity are indeed harder to predict than MSLP and Air Temperature, since the accuracy is lower. This was what we found in our analysis earlier.

Conclusions

In this analysis, we examined how various weather variables were related to rainfall in Melbourne. Using machine learning, we predicted whether a particular day would rain based on the weather conditions the previous day. Furthermore, we addressed data quality issues by checking data types and conducting descriptive analysis then identified important variables. A number of key conclusions can be drawn from the model results and outputs regarding the prediction of rainfall in Melbourne.

As a result of the feature importance analysis using the Random Forest model, the most influential factors predicting rainfall were wind direction, humidity,

and mean sea level pressure (MSLP). This indicates that these variables play a crucial role in determining whether rainfall will occur. Moreover, it has been demonstrated that the difference between dry bulb temperature and dew point temperature is a significant predictor of rainfall. This corresponds to meteorological principles since a large difference in temperature indicates a greater probability of precipitation (NOAA's National Weather Service, n.d.).

Based only on the current day's average conditions, using Random Forest Classification method to predict rainfall for the next day resulted in only a 52% accuracy rate, which illustrates the difficulty in predicting weather conditions from current data alone. To improve the prediction accuracy, regression models were used to forecast features for midday the following day. These models aimed to capture more dynamic patterns in weather conditions using half-hourly data.

In spite of the use of advanced machine learning techniques, rain predictions remain challenging because of the complex and dynamic atmospheric environment (Ongoma, n.d.). The best-performing model amongst models used is Random Forest achieving an accuracy of approximately 60%, compared with SVC, Logistic Regression and Adaboost. This indicates that more research is needed in order to improve the model.

Optimising the parameters of the models can further enhance prediction accuracy. Changing parameters such as the learning rate in Adaboost or tuning hyperparameters in SVC can help optimize the models for the purpose of predicting rainfall. In addition, the quality of data and the incorporation of additional data sources can improve forecasts. It is possible to improve forecast accuracy by ensuring the accuracy and completeness of input data as well as by adding relevant data from other sources, such as weather forecasts and satellite imagery.

Overall, valuable insights into predicting rainfall in Melbourne can be gained from the models and analyses conducted in this project. Moreover, by optimizing parameters, improving data quality, leveraging machine learning and exploring different models, we can enhance our forecasting capabilities and provide more accurate future rainfall predictions.

References

Met Office - UK Weather. (2014, April 16). *How does rain form and what is the water cycle?* [Video].

YouTube. <https://www.youtube.com/watch?v=zBnKgwnn7i4>

Nambiar, K. (2022). Why Is It So Hard To Predict The Weather? *Science ABC*.

<https://www.scienceabc.com/eyeopeners/why-is-it-so-hard-to-predict-the-weather.html>

NOAA's National Weather Service. (n.d.). *ADAS Glossary*.

https://www.weather.gov/mlb/adas_glossary

NOAA's National Weather Service. (n.d.-b). *Pressure Definitions*.

https://www.weather.gov/bou/pressure_definitions#:~:text=MEAN%20SEA%20LEVEL%20PRESSURE%3A%20This,of%20elevation%20from%20pressure%20readings

Ongoma, V. (n.d.). *The science of weather forecasting: what it takes and why it's so hard to get right*.

The Conversation.

<https://theconversation.com/the-science-of-weather-forecasting-what-it-takes-and-why-its-so-hard-to-get-right-175740>

Sam Burt, University of Melbourne & Dr Linden Ashcroft Lecturer in climate science and science communication, School of Earth Sciences, Faculty of Science, University of Melbourne.

(2023). Explaining Melbourne's crazy but predictable weather. *Pursuit*.

<https://pursuit.unimelb.edu.au/articles/explaining-melbourne-s-crazy-but-predictable-weather>

Weather Melbourne Australia: Four Seasons in One Day. (n.d.).

<https://www.we-love-melbourne.net/weather-melbourne.html#:~:text=The%20unpredictable%20weather%20Melbourne%20is,dry%20and%20hot%20desert%20winds>

What Makes It Rain? (n.d.). NOAA SciJinks – All About Weather. <https://scijinks.gov/rain/>

Wikipedia contributors. (2023). Dew point. *Wikipedia*. https://en.wikipedia.org/wiki/Dew_point