Data: Feature extraction and visualization

Australian weather

Group members

Anton Mosquera Storgaard - s214611

Jacob Borregaard - s181487

Mathias Correll Damsgaard - s214647

Group 230

Date: 4th october 2022

Academy: Technical University of Denmark - DTU

Subject: Introduction to Machine Learning and Data Mining - 02450

Assignment: Project

Contents

1 Description of The Data Set

The dataset is found at this location.

1.1 Description

Our data set is a report of weather measurements in Australia containing different information about rain, sunshine, temperature, wind, air humidity and atmospheric pressure. We want to be able to predict the likelihood of rain in the specific city Sydney. We want to classify for the **RainToday** attribute. It is a simple binary class transformation with either "Yes" or "No" as its labels, where yes becomes 1 and no becomes 0. We want to do this with the **MinTemp**, **MaxTemp**, **Evaporation**, **Sunshine**, **WindGustSpeed**, **Humidity9am**, **Pressure9am**, **Cloud9am**, **Temp9am** attributes. We also want to do the regression based on the same nine attributes, but instead we want to find the value of the continuous **Rainfall** attribute. So in summary our goal is to be able to estimate whether or not it will rain today and how much it will rain.

1.2 Summary of previous analysis

This previous analysis wanted to make a reliable weather prediction machine learning software based on the given data. They started out by checking the data and correcting any missing values by taking the average of the column and one-hot encoding for example locations. They also discarded rows from the test data set that did not have information in the prediction column. They then selected the best features based on their correlation with the prediction column. They divide their training data into a ratio of 80:20 - training and test data to then train six different classifiers. They compare their accuracy for predicting the true weather for the following day, and optimize the best model to then finally use it to predict the weather on the given test-data.

2 Explanation of the attributes

2.1 Selection of location and attributes

Our data set is very large with a shape of (99516,23) which means that we have 99.5 thousand rows, with 23 different attributes. We have decided to lower our amount of data by focusing on 11 of the attributes as stated in the previous section. The data set has a lot of missing values. One particular reason for this, is because the data is collected from different locations in Australia. This means that since the various values are measured differently, the types of data that is being recorded in different locations is simply not the same. An example of this is the data point **Sunshine** where 48% of the data points are missing. If you look at the csv-file you will quickly realize that the missing data is correlated with the locations, which means that a lot of the locations don't measure this specific data. One way to deal with this problem, and lowering the amount of data at the same time, is to select a specific location. We have selected to focus on the data from Sydney, since it narrows our scope and Sydney has data from all the attributes that we are interested in.

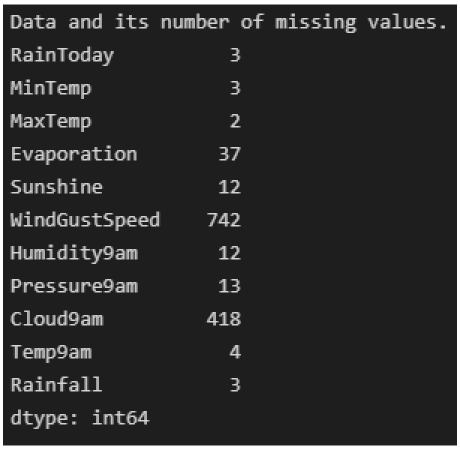
Here we have categorized our attributes into their different categories:

- **Discrete data:**
  - First we have **RainToday**, which is a normial data type, since it belongs to one of two categories.
  - **Cloud9am** is also a discrete variable but with ordinal data type, since it is ranked between 0 and 8 for how many cloud there is in the sky. The same characterization applies to the attribute **Humidity9am** which is ranked between 0 and 100.
- **Continuous data:**
  - Where the **MinTemp**, **MaxTemp** and **Temp9am** is an interval data type, since it is in degree Celsius.
  - And the the rest being **Evaporation**, **Sunshine**, **Rainfall**, **WindGustSpeed** and **Pressure9am** are a ratio data type, since a value of zero for these variables have a general meaning.

3 Data visualization

3.1 Data preprocessing

The subsection of the data set that we are analysing consists of 2361 rows of data and of course 11 columns of attributes. To get an understanding of our data we started by looking for any missing values.



```
Data and its number of missing values.
RainToday          3
MinTemp            3
MaxTemp            2
Evaporation       37
Sunshine          12
WindGustSpeed    742
Humidity9am       12
Pressure9am       13
Cloud9am         418
Temp9am            4
Rainfall           3
dtype: int64
```

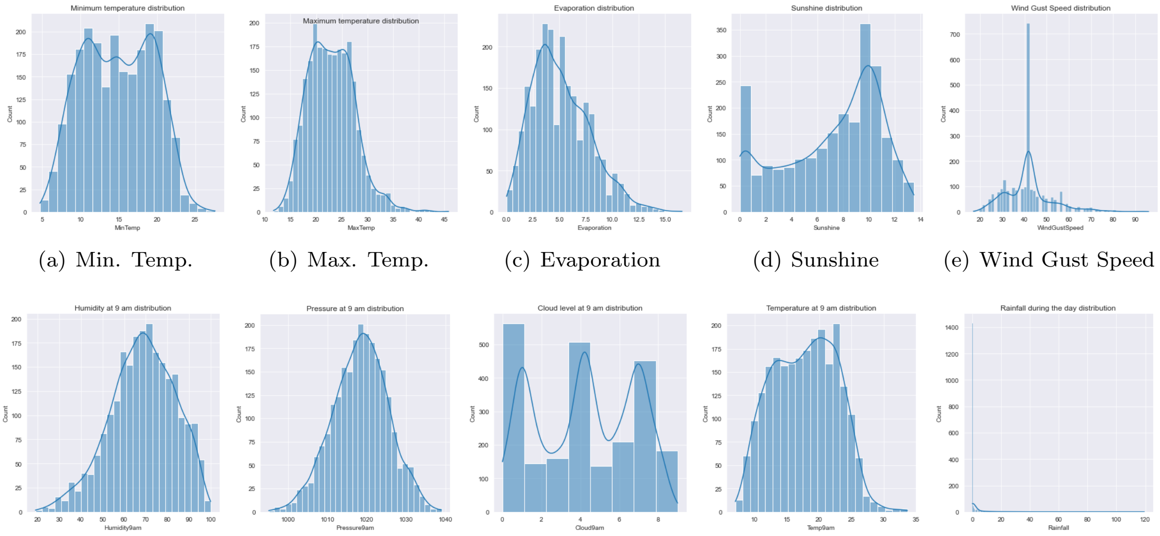
Figure 1: Sum of missing values for each attribute

Here we see that both **RainToday** and **Rainfall** have three missing values. It would be impossible for us to do use these data rows for our machine learning algorithms later, since we would not know the true label or value. Thus we decided to remove these rows entirely.

For the rest of the missing values we computed their mean value and inserted that as a replacement. We also binarized **RainToday** so that 1 corresponds to ”Yes” and 0 to ”No”. We have not one-out-of-K encoded the attribute, since it will not engage in the machine learning algorithms but only be used as a true label to train and test the classifier.

3.2 Normal Distributions

We have then plotted the attribute to see their distributions. These can be seen below in Figure 2.



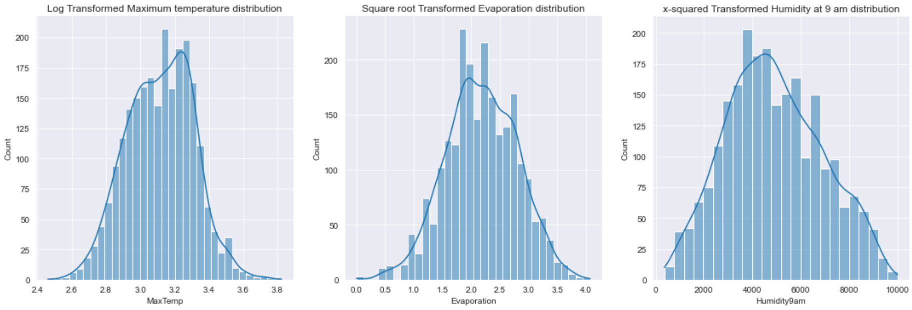
(a) Min. Temp. (b) Max. Temp. (c) Evaporation (d) Sunshine (e) Wind Gust Speed

(f) Humidity 9am (g) Pressure 9am (h) Cloud 9am (i) Temp. 9am (j) Rainfall

Figure 2: Distribution of each attribute

Plot **(a)** appears to be a bimodal distribution. This could be caused by the fact that the summer temperature follows a normal distribution and the winter temperature follows a normal distribution which added together gives this result. The same can be seen to a lesser degree in plot **(b)** which is also right skewed. In plot **(c)** we also see a right skewed distribution, this could be caused by the fact that the **Evaporation** attribute does not go below zero. The same applies to plot **(d)** as there cannot be negative sunlight, this also creates the non normal distribution which is to be expected. Plot **(e)** appears to have some extreme values and also many equal values around the expected value. The reason for this is the amount of missing data points associated with this attribute. These missing values where given the value of the mean, and when there were over 700 missing it produces this distribution. For this reason this attribute will unlikely be useful in our analysis. Plot **(f)** appears normal but skewed to the left. Plot **(g)** appears normal. Plot **(h)** has three peaks. Plot **i** appears somewhat normal distributed. Plot **j** has many data points with value 0, which is of course days with zero rainfall and some extreme values, this distribution was to be expected. There is no issue with outliers in the data.

We want to transform some of the data so PCA and other statistical tools will have better performance. The transformed data looks like this:



(a) Log Max. Temp. (b) Squared Evaporation (c) Squared Humidity

Figure 3: Distribution of three transformed attributes

We see that the distributions appear more normal distributed.

3.3 Correlations

The correlation between two attributes shows how they are intertwined. The heat-map in Figure 4. shows the correlation between the individual attributes, and are colored to make it easier to gather an overview of the results. Looking at correlations can help you see links between different attributes of the data set. In our data set most of the correlations make sense, for example **MaxTemp** appear to correlate with **MinTemp**, **Evaporation** and **Temp9am** as well as having a negative correlation with **Cloud9am**, which all makes sense. Another observation is that **WindGustSpeed** doesn't appear to correlate very well with any of the other attributes, which means that it might not be the most important attribute to include if you are trying to make a classifier for any of the other data.

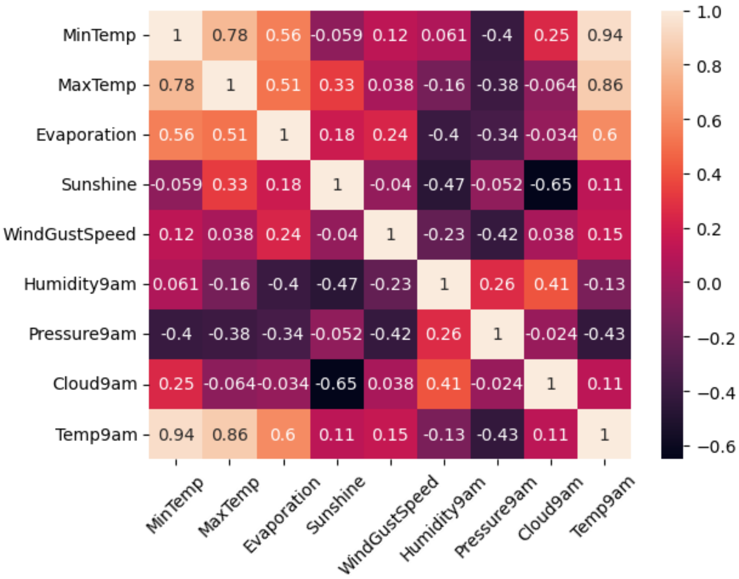


Figure 4: Correlation heatmap

3.4 Principal Component analysis

3.4.1 Explained Variance

The explained variance is a metric that captures how much of the information in the data that can be explained by a feature or a group of features in form of more principal components. We have plotted the variance explained by the first few principal components, as well as the cumulative explained variance. We see that the first principal component explains close to 60% of the variance and that the first 3 explain 90% of the variance as well as 4 of the PC's explain over 95% of the variance, which means that you could lower the amount of computation by a lot, without losing too much information by only using the first 5 or 6 PC's.

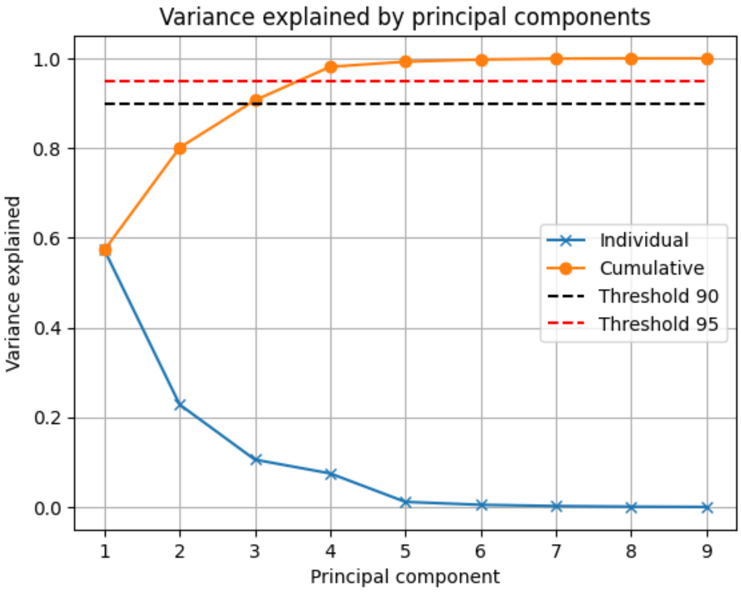


Figure 5: Explained variance plot

3.4.2 Principal directions

In the plot below we can see, how each of the nine attributes is described in a 4D space consisting of our four first principal components (PCs). We have chosen this as the threshold, since four PCs describe 95% of the variance in the data.

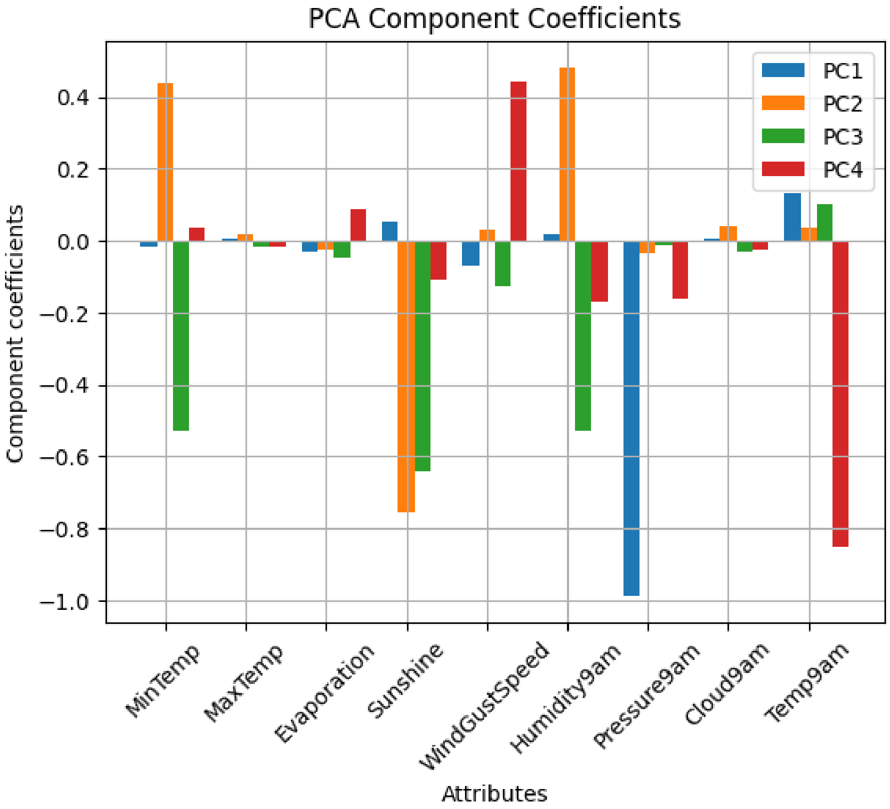


Figure 6: Plot of first four PC coefficients

From this figure we can see that the direction of the first PC is dominated by the value of the attribute **Pressure9am**, which also means that this attribute is the one that explain most of the variance in the data. The second and third PCs are mostly directed by **MinTemp**, **Sunshine** and **Humidity9am** in different compositions. Lastly **WindGustSpeed** and **Temp9am** directs the fourth PC. From this information, which is also pretty visible in the plot, we can infer that the attributes **MaxTemp**, **Evaporation** and **Cloud9am** describes very little of the explained variance with the first four PCs.

3.4.3 Projected Data

The projected data plot in Figure 7, shows the data projected into a two-dimensional subspace, containing two principal components at a time. We have projected onto pairs of the first four PC's. What you can see in the plot is that there are no clear tendencies in the distribution, which can be seen by the way the data is scattered in the plots. This means that the four first principle components capture different features in the data set.

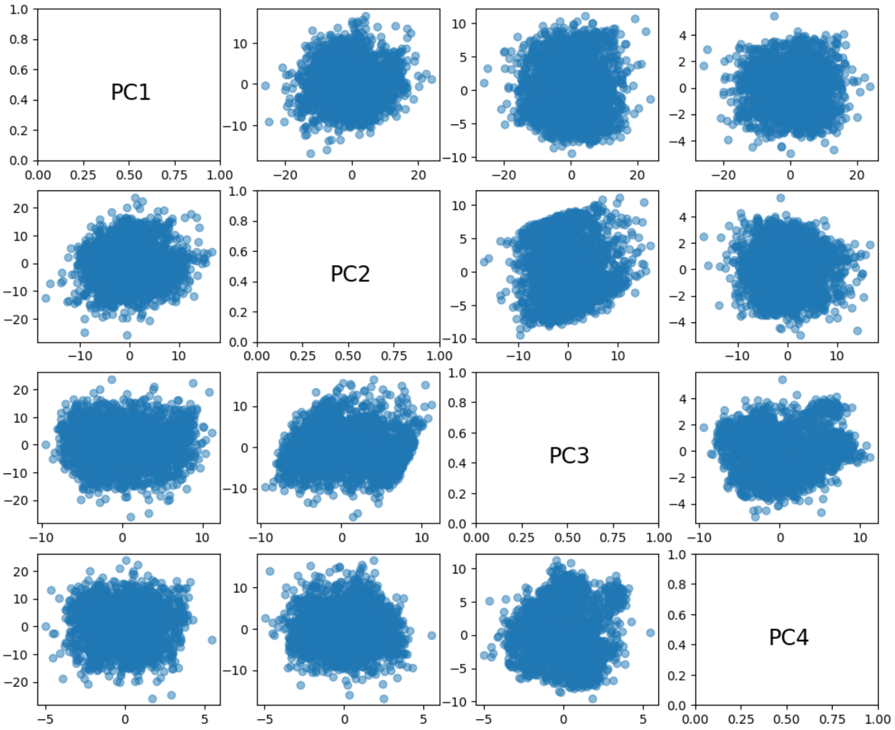


Figure 7: Plot of data on two of the first four PCs

4 Discussion

We started out by looking through our data set and quickly realized that we needed to specify how we wanted to handle the the many attributes and rows of information to meet our needs for both this project and the coming. Therefore, we chose eleven attributes that we found suiting to work with. Of these eleven two of them is for our classification and regression algorithms in the next project. We also chose to lessen the amount of data rows by only looking at the weather data from a single city being Sydney that the entire of Australia. Hence we wanted to look closer at the nine attributes that would be used for these tasks.

By analysing our data further through visualization, we learned that the data for each of our attributes generally can come close to follow a normal distribution. Especially **Pressure9am** has a good bell shape. We did also see an effect of our earlier data preprocessing in **WindGustSpeed**, where there was a large spike at the mean value. Some of the attributes had a skewed distribution so we tried to transform those attributes which resulted in a better normal distribution for those attributes. We then investigated the correlation between our attributes. It made sense that we saw a high correlation between the three temperature attributes, who all also correlate well with **Evaporation**. **Sunshine** and **Cloud9am** also had a negative correlation with each other, which intuitively again made sense. We could lastly also conclude that **Pressure9am** had a decent correlation with most of the other attributes, where **WindGustSpeed** had close to no correlation with the other attributes. This information is helpful in understanding how our information scales between attributes.

After having completed a principal component analysis we looked closer at the explained variance by each principal component. By calculating the cumulative explained variance we saw that four PCs would explain over 95% of the variance in the data, hence we could reduce the dimensionality of the data set from nine without losing much information. We also took at look at how each of the nine attributes would be projected on the first four principal components coefficients. Here we could see how the value of an attribute would determine the direction in this four-dimension space. This we lastly visualized by plotting the projected data in every combination of the first four PCs. From the plots it is hard to truly understand the data as it clumps together, but it was possible to see some differences in the shapes of the clumps indicating a bit of how the data would flow in that two-dimension space.

To end of we can say that we learned a lot about of data and its attributes and we believe it will be possible to complete our machine learning aims on this data based on our visualizations and current understanding of the data and goals of the next project, but should be carried out without the use of the **WindGustSpeed** attribute.

5 Appendix

5.1 Responsibility of each student

Here is shown the contribution for each section for each student in approximate percentage:

Contribution %	Part 1	Part 2	Part 3	Part 4	Exam Questions
Anton	40	20	55	15	10
Jacob	20	60	15	15	60
Mathias	40	20	30	70	30

5.2 Exam questions

I Question 1.

Answer is D.

The time of day is listed as a number from 1 to 27, in which the distance can be measured with subtraction/addition, so it must be an interval. And traffic lights and running over has a universal zero, so they must be ratio. Also the congestion level increasing with a higher number means it is ordinal. Hence the answer is **D**.

II Question 2.

Answer is A.

When p goes toward infinity the max-norm distance can be calculated as the maximum absolute difference between each pair of observations. That means we have the biggest difference between the first pair, having a value of 7, while the rest are 0 expect for one being 2. Hence the right answer is **A**.

III Question 3.

Answer is A.

The amount of cumulative explained variance can be calculated the following way:

$$\frac{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2}{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2} = 0.867$$

which means that the first four principle components counts for 86.7% of the explained variance, which isn't surprising since the principle components count for less and less of the explained variance, so the first four out of five, must be more that 80%. The answer is **A**.

IV Question 4.

The answer is D.

The method in this question was just trial and error, in option D, we got the equation

$$low * 0.58 + high * 0.23 - high * 0.01 + high * 0.33$$

Summing these numbers up and assuming that the high numbers are about equally high, then the projection on pc2 will typically have a positive value.

V Question 5.

The answer is A.

The Jaccard similarity is calculated by following calculation:

$$\frac{f11}{M - f00} = \frac{2}{20000 - 19987} = 0.153846$$

which is the answer **A**.

VI Question 6.

The answer is D.

In order to find the probability of x2=0 and y=2, we chose to sum over the events given these two probabilities by summing the two occurrences we get the following equation:

$$0.81 * 0.23 + 0.03 * 0.23 = 0.193$$

which gives us the answer to be **D**.