RAIN IN AUSTRALIA

PREDICT RAIN TOMORROW IN AUSTRALIA

MACHINE LEARNING FINAL PROJECT

Prepared By:

Erencan Çabuk
130403008

2019

Contents

FIGURE LIST

TABLE LIST

1. ABSTRACT

Nowadays, large amount of data is available everywhere. Therefore, it is very important to analyze this data in order to extract some useful information and to develop an algorithm based on this analysis. This can be achieved through data mining and machine learning. Machine learning is an integral part of artificial intelligence, which is used to design algorithms based on the data trends and historical relationships between data. Machine learning is used in various fields such as bioinformatics, intrusion detection, Information retrieval, game playing, marketing, malware detection, image deconvolution and so on. This paper presents the work done by various authors in the field of machine learning in various application areas

Thanks to weather forecasts, we make some plans in our daily lives. We even make weekly plans so we can arrange ourselves. Weather forecasts are made by computer systems. But history was also made by observations of nature. Nowadays, when the experience and technology are confronted, the weather forecasts are usually made without errors.

Getting to know the weather, getting dressed, taking precautions, getting off the road, planning the transportation alternatives and getting things done during the day are very good. For this reason, people do not miss the weather forecasts immediately after the news. Weather forecasts also refer to data such as wind and humidity with day and night temperature differences. Besides, if we don't know if it doesn't rain, we can be caught off guard.

The weather forecast; The prediction of meteorological events that can be seen in a given country, region or center in a time frame by using subjective or objective methods based on observations and analyzes is called as weather forecast.

Weather Forecast Three Stages:

1. Observations
2. Analysis
3. Guess

The data I use is daily weather observations from the weather stations of Australia. The process to be performed here is to analyze the data and then make an estimate. In other words, according to the data I use, I have to guess what really happened. I'm using the 'WeatherAus' data set and this data set contains moisture, temperature, precipitation and pressure. Using this data, we look to see if the weather is rainy the next day.

The target variable 'Rain Tomorrow' in the data grid means.

Did it rain the next day? Yes or no, there are observations. It is wise to use the classification method to obtain estimates.

In summary, we predict whether the next day is rainy. Here, our 'observation' section is in the data. We are analyzing this data with the help of computer. As a result, we 'predict' that the air of the next day is rainy or not rainy.

2. INTRODUCTION

Weather forecasts are one of the most important studies made with machine learning recently. In fact, the weather forecast consists of 3 main headings: observation, analysis and prediction. In today's conditions, machine learning operations are as follows; In the observation part, daily, monthly, weekly and annual data are used and data are prepared. The next step is to analyze the data and some unnecessary data should be removed from the data. The last stage is to try to obtain the target data from the data we analyze.

To achieve successful results in such classification methods, we need very large data.

One of the problems here is that it is hard to classify very large data on normal computers. So, you can take a long time to make the forecast as a result of the received data. Therefore, it is important that the data is selected well during data analysis.

The content of the data I use is related to the weather in Australia. The ‘Rain Tomorrow’ target variable in the data generated from weather observations in Australia means the following. Did it rain the next day? Yes or no contains observations. It is wise to use the method of classification to obtain the predictions. Finding yes or no information in ‘Rain Tomorrow’ target section shows that it will be more advantageous to solve us with binary classification method. In other words, if the method used here is not raining with binary classification = 0, it is defined as rain = 1.

I have had several difficulties in the preprocessing part of the data when using some of the available methods. Each data has its own pre-processing methods. So, it is important to select and investigate the techniques to be used according to the data.

3. Methods

3.1. Logistic Regression

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).

In logistic regression, the dependent variable is binary or dichotomous, i.e. it only contains data coded as 1 (TRUE, success, pregnant, etc.) or 0 (FALSE, failure, non-pregnant, etc.).

The goal of logistic regression is to find the best fitting (yet biologically reasonable) model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of presence of the characteristic of interest:

Linear Regression: $y = b_0 + b_1X$

$$logit(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k$$

where p is the probability of presence of the characteristic of interest. The logit transformation is defined as the logged odds:

$$odds = \frac{p}{1-p} = \frac{probability\ of\ presence\ of\ characteristic}{probability\ of\ absence\ of\ characteristic}$$

Rather than choosing parameters that minimize the sum of squared errors (like in ordinary regression), estimation in logistic regression chooses parameters that maximize the likelihood of observing the sample values.

In logistic regression, b0 moves the slope to the right and left, while b1 defines the slope of the curve. The logistic regression equation can be written with the probability ratio (logit (p)).

$$logit(p) = \ln\left(\frac{p}{1-p}\right) \qquad ln\left(\frac{p}{1-p}\right) = b_0 + b_1x$$

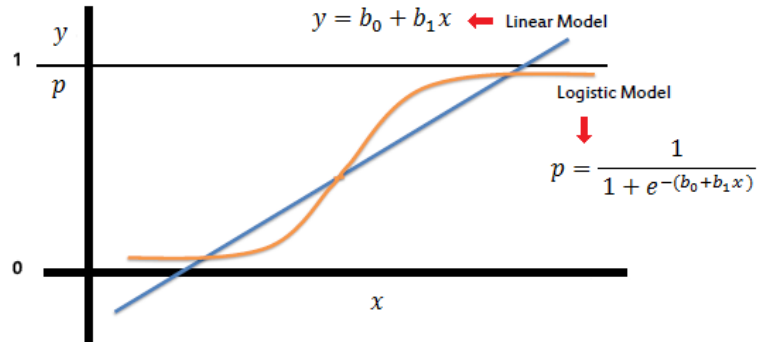

Figure 1 Logistic Model

3.2. SVM (Support Vector Machine)

It is one of the most effective and simple methods used in classification. For classification, it is possible to separate two groups by drawing a border between two groups in a plane. The place where this limit will be drawn should be the most distant place for the members of both groups. Here SVM determines how this limit is drawn.

In order to do this, two boundary lines are drawn close to each other and parallel to each other and these boundary lines are brought together and common boundary line is produced. For

example, consider the two groups in the figure below:

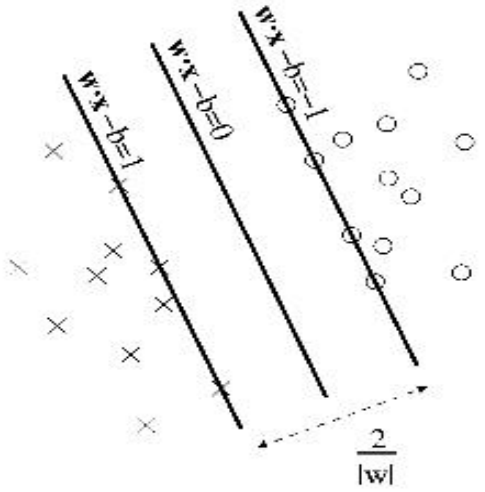

Figure 2 Two groups are shown on a two-dimensional plane

In this way, two groups are shown on a two-dimensional plane. It is possible to think of this plane and dimensions as a feature. In other words, a feature extraction is made in each input (input) in the simplest way, resulting in a different point indicating each input in this two-dimensional plane. The classification of these points means the classification of the inputs according to the characteristics. Above is the offset between the two classes. The definition of each point in this plane can be made by the following representation:

$$\mathcal{D} = \{(\mathbf{x}_i, c_i) | \mathbf{x}_i \in \mathbb{R}^p, c_i \in \{-1, 1\}\}_{i=1}^n$$

It is possible to read the picture above. For each x, c is binary, X is a point in our vector space, and c is the value that indicates that this point is -1 or 1. This point set moves from i = 1 ton.

In other words, this representation refers to points in the previous figure.

Consider this representation on the hyper plane. It is possible to express each point in this note by equation (wx - b = 0). W overload is the changing vector of the vector x and the normal vector is perpendicular to the vertical and b is the scroll ratio. It is possible to liken this equation to the correct equation of calcic 'ax + b'.

Again, according to the equation above b / || w || The value gives us the difference in distance between the two groups. This distance difference before the offset (offset) had given the name. In order to increase the distance to the highest value according to this distance difference equation, in the equation giving 0, -1 and +1 values shown in the first figure above, 3 / || w || The formula is used. In other words, the distance between the lines is determined as 2 units.

The two equations obtained according to this equation:

(wx - b = -1) and (wx + b = 1). In fact, these equations are the result of the process of finding the highest values obtained by shifting the lines. It is also accepted that the problem is linearly separable with these equations.

As can be expected, it is not possible that the overload (hyperplane) between the two groups is unidirectional. Here is an example of this situation:

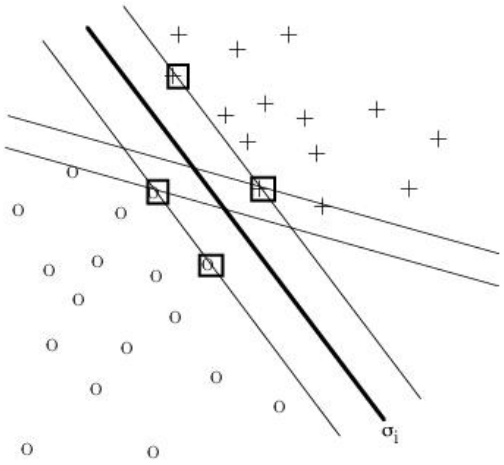

Figure 3 Hyperplane

In the above figure, although there are two different types of hyperplane, in the 'SVM' method, those with the greatest tolerance are taken

3.3. KNN Classification

A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor.

Distance functions

Euclidean $$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Manhattan $$\sum_{i=1}^{k}|x_i - y_i|$$

Minkowski $$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$$

It should also be noted that all three distance measures are only valid for continuous variables. In the instance of categorical variables, the Hamming distance must be used. It also brings up the issue of standardization of the numerical variables between 0 and 1 when there is a mixture of numerical and categorical variables in the dataset.

Hamming Distance

$$D_H = \sum_{i=1}^{k}|x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

Choosing the optimal value for K is best done by first inspecting the data. In general, a large K value is more precise as it reduces the overall noise but there is no guarantee. Cross-validation is another way to retrospectively determine a good K value by using an independent dataset to validate the K value. Historically, the optimal K for most datasets has been between 3-10. That produces much better results than 1NN. If k is too small means that overfitting algorithm performs too good on the training set, compared to its true performance on unseen test data
If small k = less stable, influenced by noise
If larger k = less precise, higher bias

4. RESULTS AND DISCUSSION

By using WeatherAus ’dataset, 'Rain Tomorrow’ which is the target variable in data, has been classified. The results obtained in 6 different classification methods are as follows.

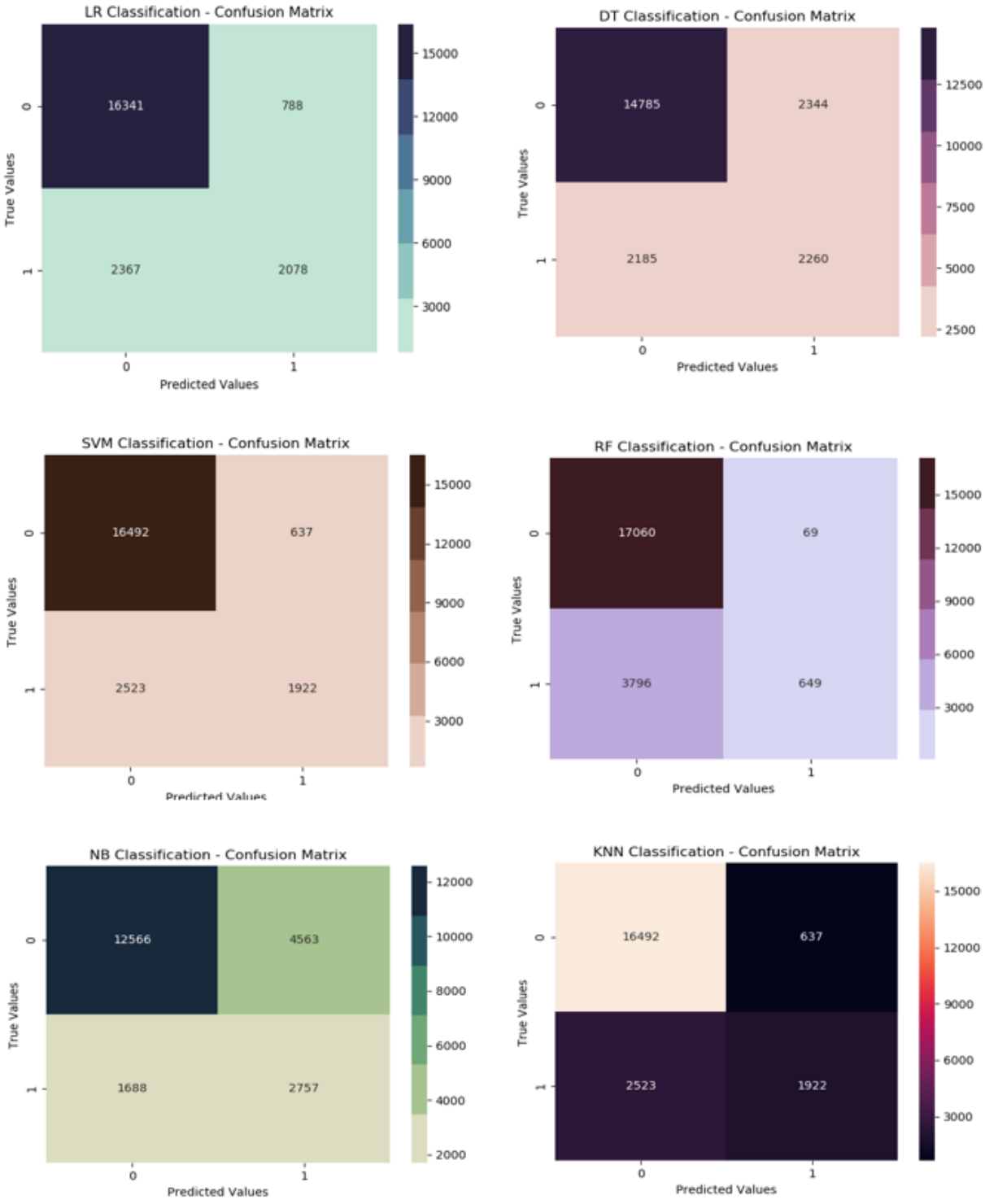

Figure 4 Confusion Matrix

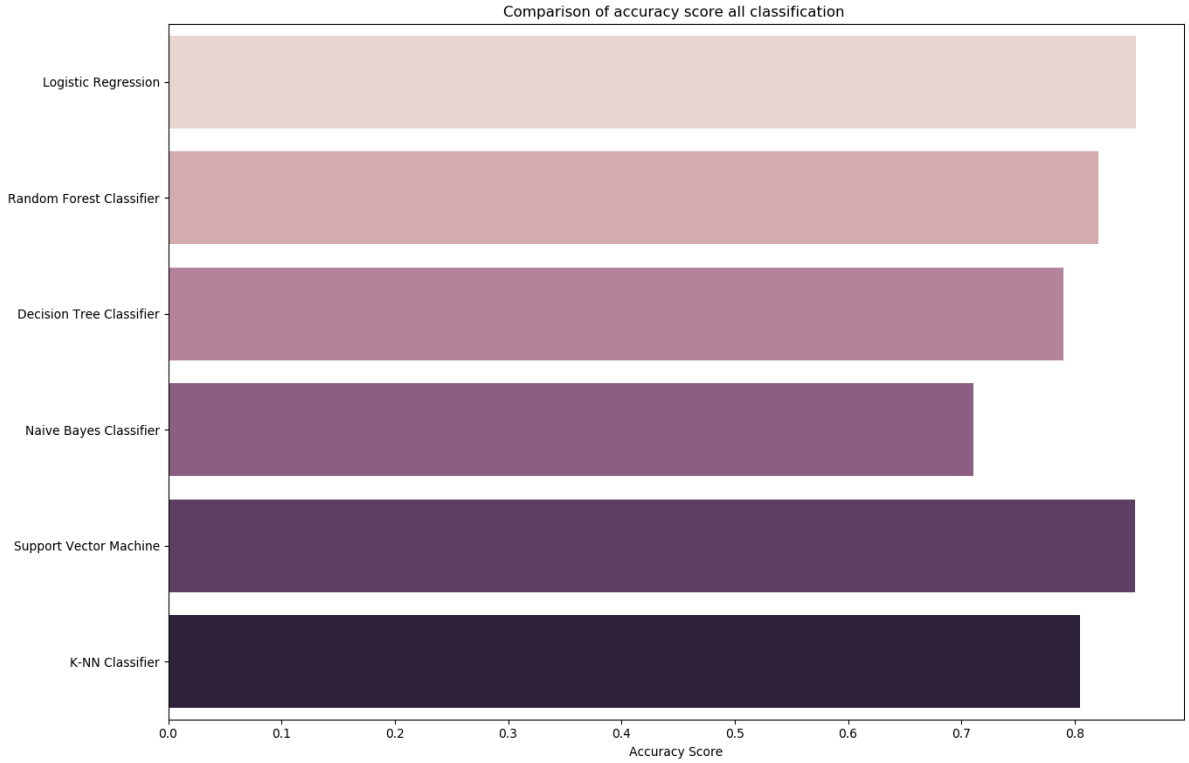

Figure 5 Comparison of accuracy score all classification

The above bar plot shows the accuracy scoring of classification methods.

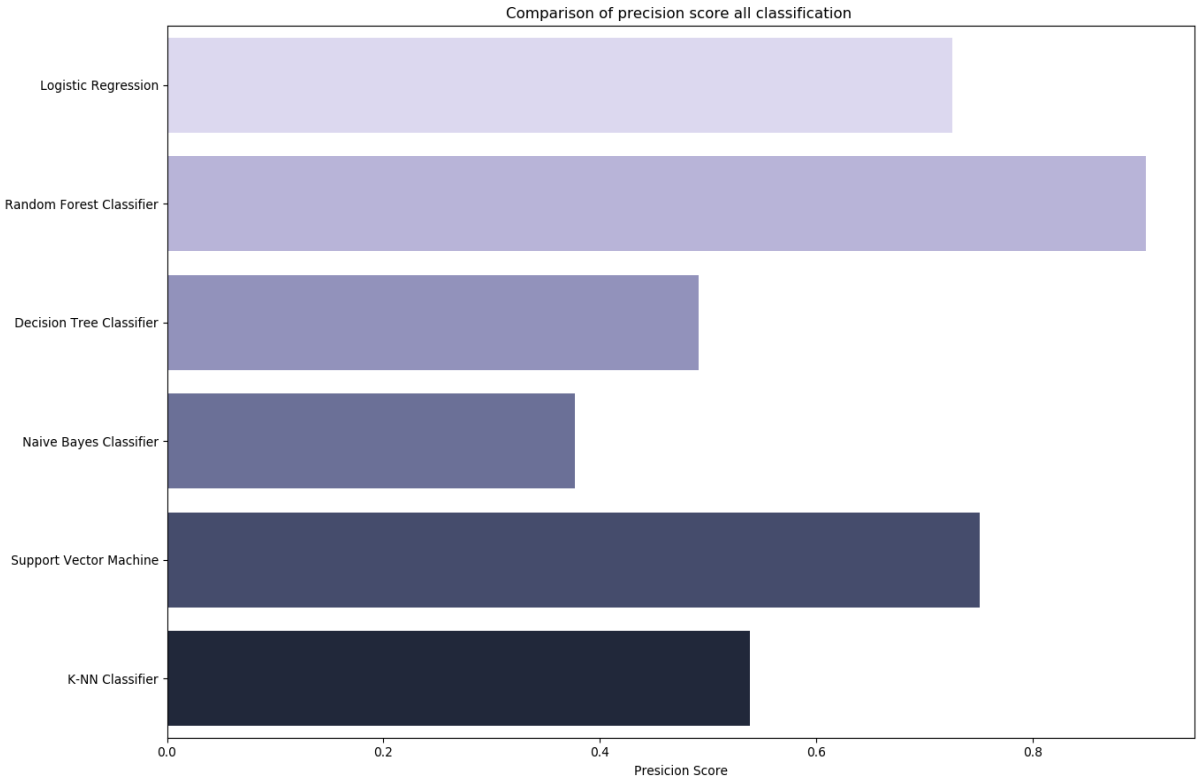

Figure 6 Comparison of precision score all classification

The above bar plot shows the precision scoring of classification methods.

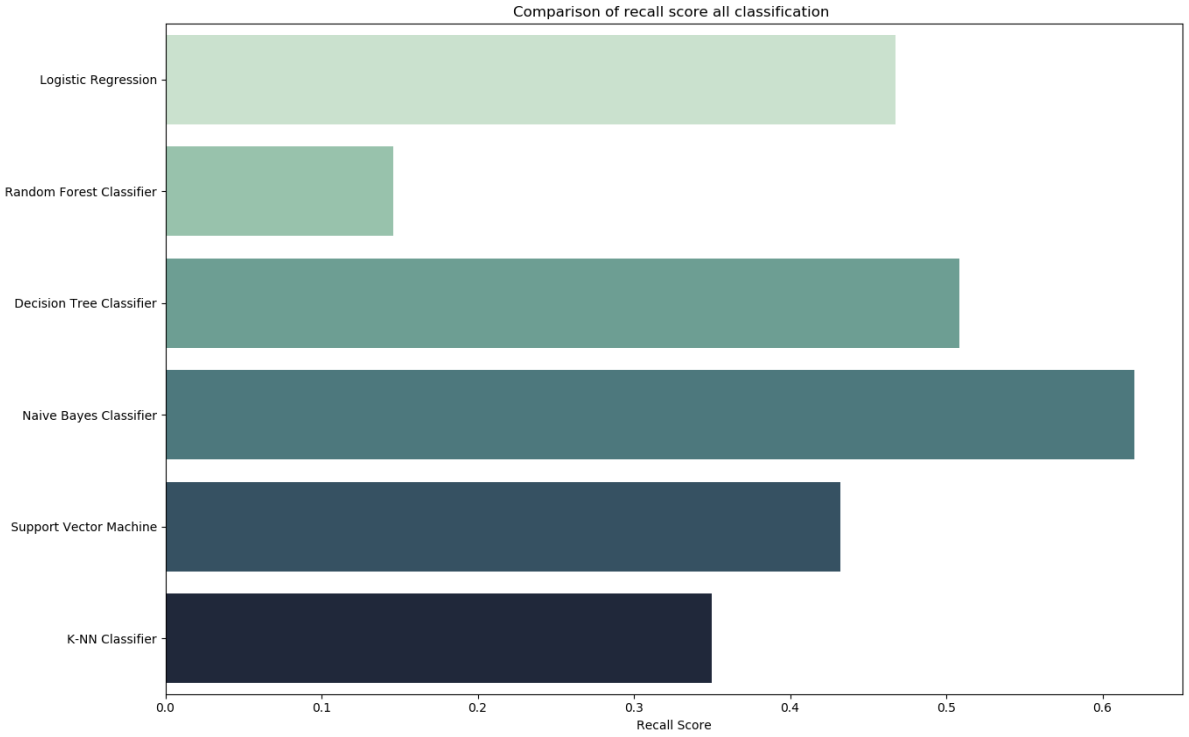

Figure 7 Comparison of recall score all classification

The above bar plot shows the recall scoring of classification methods.

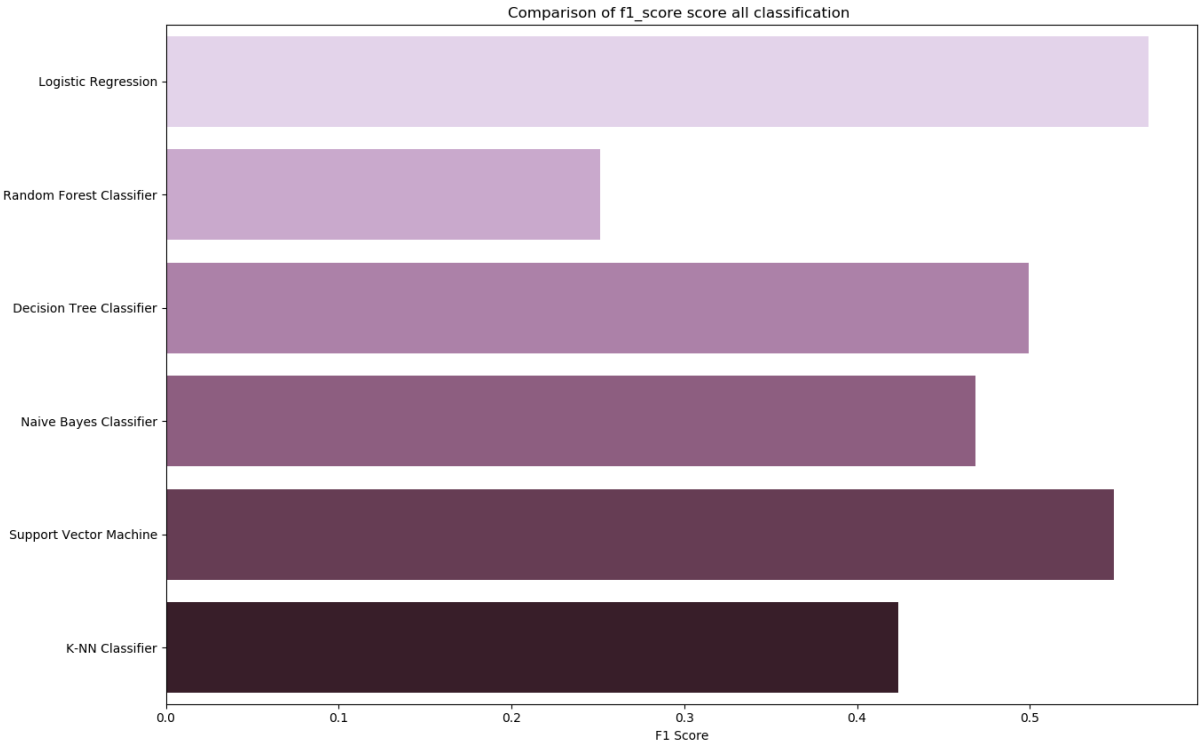

Figure 8 Comparison of f1 score all classification

The above bar plot shows the f1 scoring of classification methods.

Table 1 Classification score table

	Logistic Regression	Random Forest Classifier	Decision Tree Classifier	Naïve Bayes Classifier	SVM Classifier	K-NN Classifier (k=3)
Accuracy Score	0.8536	0.8208	0.7901	0.7103	0.8535	0.8043
Precision Score	0.7251	0.9039	0.4909	0.3766	0.7511	0.5388
Recall Score	0.4675	0.1460	0.5084	0.6203	0.4324	0.3496
F1 Score	0.568	0.2514	0.4995	0.4687	0.5488	0.4241

The scores of accuracy, precision, recall and f1 are shown in the table above.

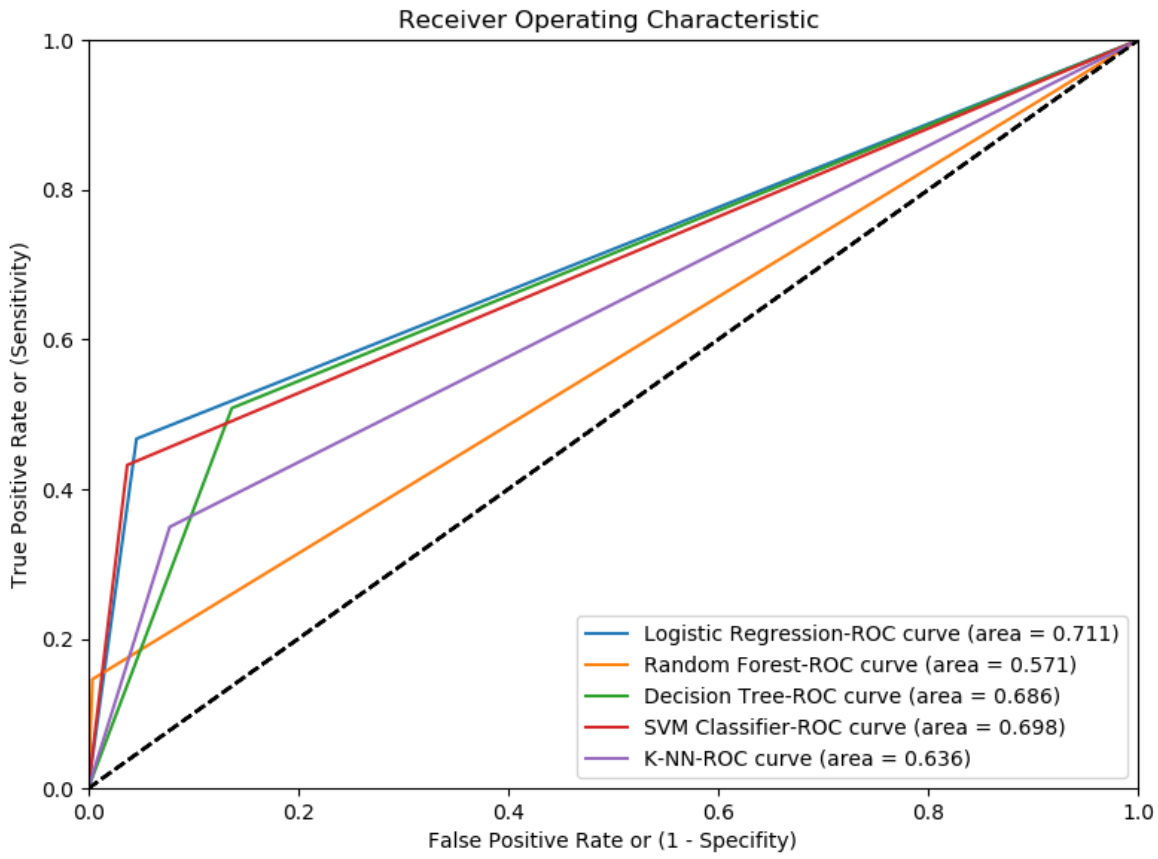

Figure 9 ROC Curves for classification

The ROC curves of applied classification methods are shown in the figure above.

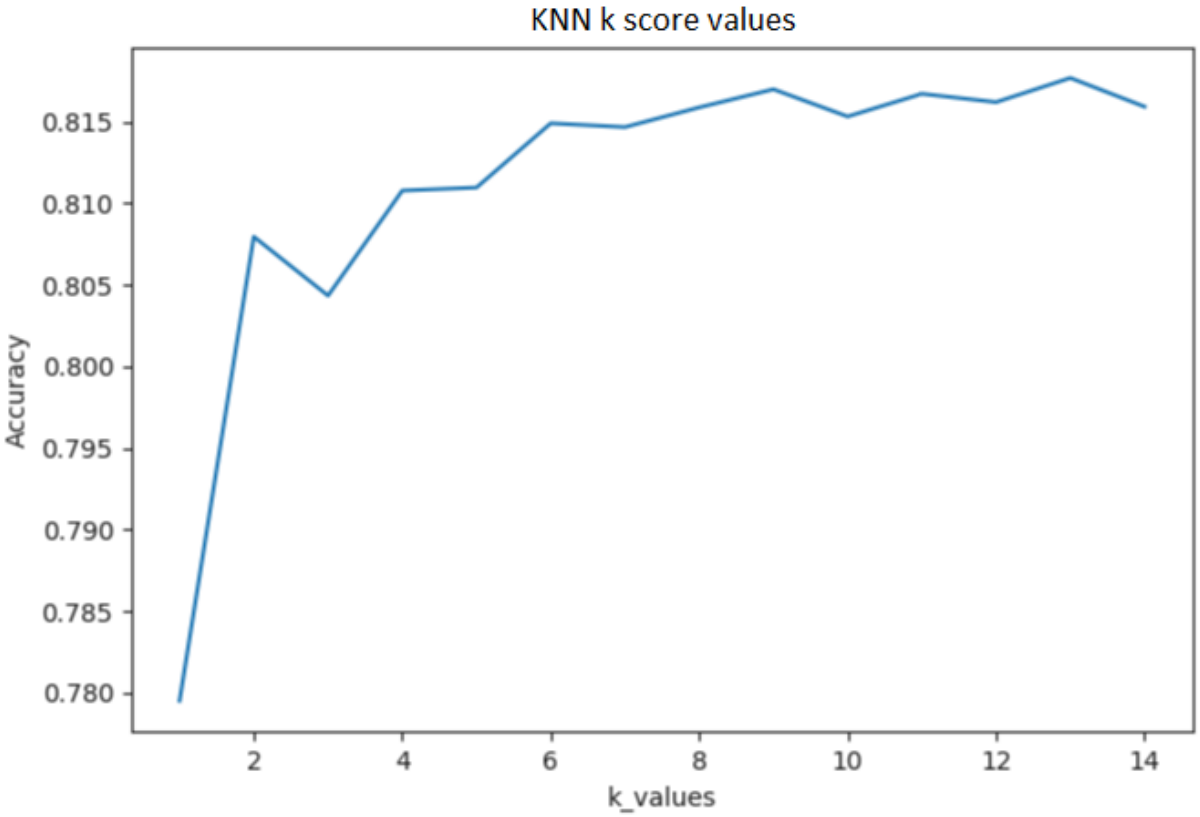

Figure 10 KNN k score values

Accuracy scores according to the k values of the applied knn classification method are shown in the figure above.

Table 2 Classification scores by k values

	K score = 3	K score = 5	K score =7	K score = 9
Accuracy Score	0.8043	0.8110	0.8147	0.817
Precision Score	0.5388	0.5760	0.6059	0.629
Recall Score	0.3496	0.3130	0.2878	0.2719
F1 Score	0.4241	0.4055	0.3902	0.3798

Accuracy, precision, recall, f1 scores according to the k values of the applied Knn classification method are shown in the table above.

Choosing the optimal value for K is best done by first inspecting the data. In general, a large K value is more precise as it reduces the overall noise but there is no guarantee. Cross-validation is another way to retrospectively determine a good K value by using an independent dataset to validate the K value. Historically, the optimal K for most datasets has been between 3-10. That produces much better results than 1NN. If k is too small means that overfitting algorithm performs too good on the training set, compared to its true performance on unseen test data.

5. CONCLUSION

As a result, data such as temperature, pressure, humidity, rainfall, wind direction, wind speed in 'weatherAus' data obtained from Australian daily weather observations were used. It was predicted that the data given in the 'RainTomorrow' would rain or not rain the next day.

There are 142k rows x 24 columns in our data. Some unnecessary columns were discarded from the data and the missing values were cleared. We also used the Z-score to detect and remove outliers from our data.

In this application, because our goal is binary classification we used. (not raining = 0, raining = 1) We used 6 classification methods. These; Logistic Regression, Random Forest Classifier, Decision Tree Classifier, Naïve Bayes Classifier, Support Vector Machine Classifier and KNN Classifier. Among the classifier methods applied Logistic Regression and Support Vector Classifier have the highest accuracy values. We also obtained accuracy score, precision score, recall score and f1 score for all classifications.

In addition, we have obtained ROC curve from all classification methods with specificity and sensitivity. In the K-NN classification method, a score of k score was drawn and accuracy score change was shown according to the value of k.

Future plans;

using different classification methods and larger data will be to increase accuracy scores. Also, it is planned to analyze the most appropriate columns related to the part to be estimated by using different feature selection methods.