**Rain Prediction Comparative Analysis Through Multiple Model Building**

**Submitted by**

Urmila Singh 18BCE1259

Vivek Gorania 18BCE1310

**J Component - Report**

**CSE3506 – Essentials of Data Analytics**

**BACHELOR OF TECHNOLOGY**

in

**COMPUTER SCIENCE AND ENGINEERING**



**May 2021**

**TABLE OF CONTENTS**

|  |  |  |  |
| --- | --- | --- | --- |
| **Chapter** | **Title Page** | |  |
|  |  | |  |
| **1** | **ABSTRACT 3** | |  |
|  |  | |  |
| **2** | **SECTION I - Data Set Description 4-5**    2.1 Will it Rain Tomorrow 5 | |  |
|  |  |  | |
|  |  |  | |
|  |  | |  |
| **3** | **SECTION II - Basic Models 6-9** | |  |
|  | 3.1 Linear Regression 6 | |  |
|  | 3.2 Logistic Regression 6  3.3 Decision Tree 7  3.4 Random Forest 7  3.5 Bagging 8  3.6 Boosting 9 | |  |
|  |  | |  |
| **4** | **SECTION III – Alternate/Advanced Models 10-11**  4.1 Neural Network 10  4.2 Support Vector Machine 11 | |  |
|  |  |  | |
|  |  | |  |
| **5** | **SECTION IV – Comparison & Conclusion 12** | |  |
|  |  |  | |
|  |  |  | |

**ABSTRACT**

In this project, we have mainly tried to observe various machine learning algorithms and we also tried to develop a rain prediction model for various locations in Australia using the daily weather data set supplied by the Australian Bureau of Meteorology. Mainly, it predicts whether there will be a rainfall tomorrow or not – in which case it is labelled **Yes**, otherwise **No**. For getting fair prediction accuracy we have used various basic and advanced models. In this project we have used *logistic regression, Decision Tree, Random Forest, Neural Network* & *SVM*.

We also tried to explore a simple model like linear regression but due to the non-linear structure of data set & it has produced fair prediction accuracy and brings out the linear structure of data set.

**Data Set Description**

1. **Will it Rain Tomorrow?**

This dataset was taken from kaggle.com and includes 142,193 daily weather observations from 49 weather stations across Australia with 24 variables describing various factors. Variables include information such as wind speed, humidity, temperature, and cloud cover. Most variables are missing a number of values, so we had to sanitize the data before proceeding.

The objective of this project was to predict rain based upon historical weather data so that it will help farmers in better yielding of crops and this project will also help various organizations like Australian Cricket Board to predict rain beforehand and thus they will be able to conduct matches without any disturbances.

The following are 24 original variables in this dataset:

* **Date**-The date of observation,
* **Location**-The common name of the location of the weather station, 1
* **MinTemp**-The minimum temperature in degrees celsius,
* **MaxTemp**-The maximum temperature in degrees celsius,
* **Rainfall**-The amount of rainfall recorded for the day in mm,
* **Evaporation**-The so-called Class A pan evaporation (mm) in the 24 hours to 9am,
* **Sunshine**-The number of hours of bright sunshine in the day,
* **WindGustDir**-The direction of the strongest wind gust in the 24 hours to midnight,
* **WindGustSpeed**-The speed (km/h) of the strongest wind gust in the 24 hours to midnight,
* **WindDir9am**-Direction of the wind at 9am,
* **WindDir3pm**-Direction of the wind at 3pm,
* **WindSpeed9am**-Wind speed (km/hr) averaged over 10 minutes prior to 9am,
* **WindSpeed3pm**-Wind speed (km/hr) averaged over 10 minutes prior to 3pm,
* **Humidity9am**-Humidity (percent) at 9am,
* **Humidity3pm**-Humidity (percent) at 3pm,
* **Pressure9am**-Atmospheric pressure (hpa) reduced to mean sea level at 9am,
* **Pressure3pm**-Atmospheric pressure (hpa) reduced to mean sea level at 3pm’
* **Cloud9am**-Fraction of sky obscured by cloud at 9am. This is measured in "oktas", which are a unit of eigths. It records how many eigths of the sky are obscured by cloud. A 0 measure indicates completely clear sky whilst an 8 indicates that it is completely overcast,
* **Cloud3pm**-Fraction of sky obscured by cloud (in "oktas": eighths) at 3pm.
* **Temp9am**-Temperature (degrees C) at 9am,
* **Temp3pm**-Temperature (degrees C) at 3pm,
* **RainToday**-Boolean: 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0,
* **RISK\_MM**-The amount of rain. A kind of measure of the "risk"
* **RainTomorrow**-The target variable. Did it rain tomorrow

**Basic Models**

1. **Linear Regression**

Linear regression is one of the most well-known and well understood model in machine learning.

Linear regression is a linear model, a model that assumes a linear relationship between the input variables and the single output variable.

The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.

Linear regression attempts to model the relationship between two models by fitting a linear equation to observed data.

As we did not get good prediction accuracy, we concluded that the data is non linear and moved on for non linear models.

1. **Multiple Linear Regression**

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the [linear relationship](https://www.investopedia.com/terms/l/linearrelationship.asp) between the explanatory (independent) variables and response (dependent) variable.

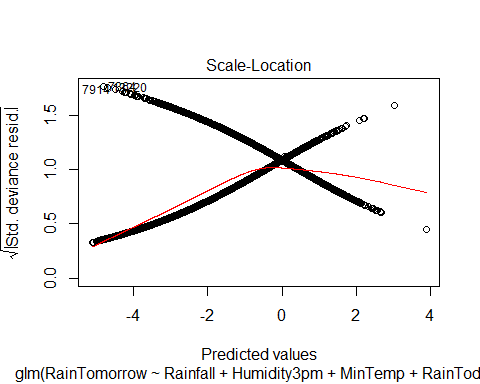
In essence, multiple regression is the extension of ordinary least-squares (OLS) [regression](https://www.investopedia.com/terms/r/regression.asp) that involves more than one explanatory variable.

1. **Logistic Regression**

Logistic regression is a classification and not a regression algorithm and it models the probabilities for classification problems with two possible outcomes. It's an extension of the linear regression model for classification problems.

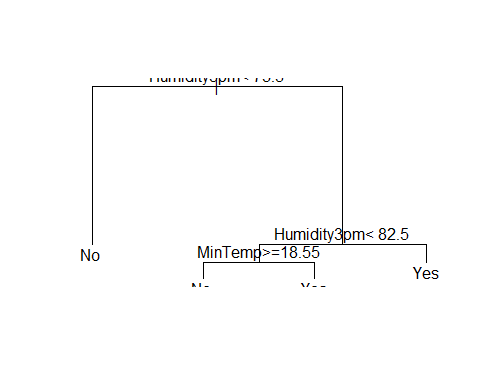
Logistic regression can be seen as a special case of the generalized linear model and thus analogous to linear regression. The model of logistic regression, however, is based on quite different assumptions (about the relationship between the dependent and independent variables) from those of linear regression.

Here, in our project we have treated **RainTomorrow** as the dependent variable and **Rainfall, Humidity3pm, MinTemp** & **RainToday** as independent variable and tried to predict results using logistic regression model.



1. **Decision Tree**

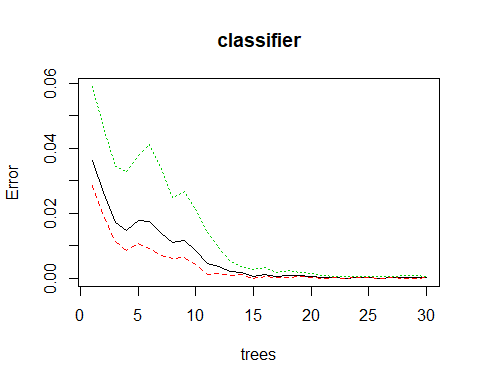
It is a type of supervised learning algorithm that is mostly used for classification problems. Surprisingly, it works for both categorical and continuous dependent variables. In this algorithm, we split the population into two or more homogeneous sets. This is done based on most significant attributes/ independent variables to make as distinct groups as possible.



In the above tree, we have tried to predict rainfall on the basis of **humidity** at various times and **MinTemp** and we achieved accuracy of about 82.3% by using this basic model.

1. **Random Forest**

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.



**Alternate/Advanced Models**

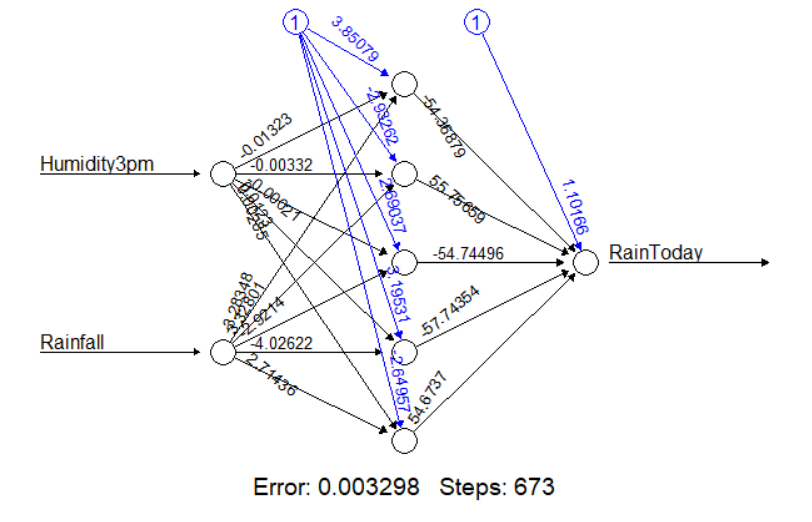
1. **Neural Network**

A neural network is a series of algorithms that endeavours to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. In this sense, neural networks refer to systems of neurons, either organic or artificial in nature. Neural networks can adapt to changing input; so the network generates the best possible result without needing to redesign the output criteria.

Basically, they interpret sensory data through a kind of machine perception, labelling or clustering raw input. The patterns they recognize are numerical, contained in vectors, into which all real-world data, be it images, sound, text or time series, must be translated.

In our project, we have used **Humidity3pm, Rainfall & RainToday**, for training our neural net model.

The following neiural network is obtained when we use a neural net with 3 hidden layers. This imporves the prediction accuracy



1. **Support Vector Machine (SVM)**

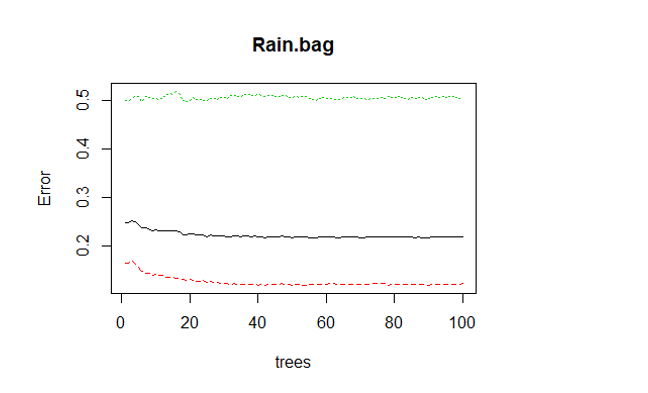
Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well. The SVM classifier is a frontier which best segregates the two classes

Pros:

1. It works really well with a clear margin of separation.
2. It is effective in high dimensional spaces.

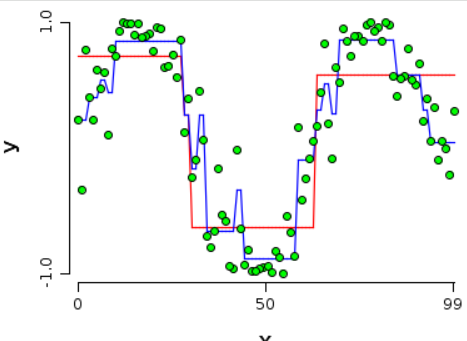
Cons:

1. It will doesn’t perform well when we take large data set because the required training time is higher
2. SVM doesn’t directly provide probability estimates, these are calculated using five-fold cross-validation.
3. **Bagging**

The decision trees suﬀer from high variance. This means that if we split the training data into two parts at random, and ﬁt a decision tree to both halves, the results that we get could be quite diﬀerent. In contrast, a procedure with low variance will yield similar results if applied repeatedly to distinct data sets; linear regression tends to have low variance. A natural way to reduce the variance and hence increase the prediction accuracy of a statistical learning method is to take many training sets from the population, build a separate prediction model using each training set, and average the resulting predictions. This is called bagging.

1. **Boosting**

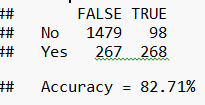
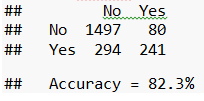
Unlike bagging, boosting does not involve bootstrap sampling; instead each tree is ﬁt on a modiﬁed version of the original data set. Unlike ﬁtting a single large decision tree to the data, which amounts to ﬁtting the data hard and potentially overﬁtting, the boosting approach instead learns slowly. Given the current model, we ﬁt a decision tree to the residuals from the model. That is, we ﬁt a tree using the current residuals, rather than the outcome Y, as the response. We then add this new decision tree into the ﬁtted function in order to update the residuals. Each of these trees can be rather small, with just a few terminal nodes.



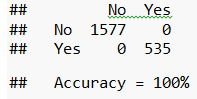
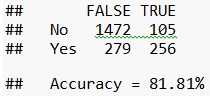
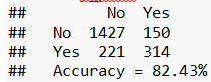
**Comparison & Conclusion**

When we applied linear regression on the dataset, we didn’t get good accuracy for the results, which led us to conclude that that data is non linear and thus we have used various advanced models for achieving a higher prediction accuracy.

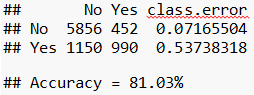
 





* For Rain Prediction Analysis, Neural Net and Random Forest models are recommended.

|  |  |
| --- | --- |
| **Method** | **Accuracy** |
| Logistic Regression | 53.8% |
| Neural Net | 100% |
| Decision Tree Classifier | 82.3% |
| Random Forest | 100% |
| Support Vector Machine | 82.43% |
| Boosting | 81.81% |
| Bagging | 81.03% |

Though Neural Net and Random Forest have the same accuracy, **Random Forest is preferred over Neural Nets because it has a lower time complexity**. Moreover, a random forest can reduce the high variance from a flexible model like a decision tree by combining many trees into one ensemble model.