Evaluation of forecasting algorithms in team building games.

Research Question:

How is **Logistic Regression** better than
Naive Bayes Classifier at predicting ***precision*** and ***accuracy***
in team building games?

Word: 3947

Session: May 2023

Candidate Number: kny549

Table of contents

Figures:

Introduction

Predictions are made in our daily lives among our friends, families and a lot of times at school especially during any competitions. Machine learning is an algorithmic study under Computer Science that is allowing programmers to make the most accurate predictions using statistical measures and algorithms. Modern technology has made machine learning a buzzword, and it is expanding quickly. Without even realizing it, we

are running daily on tools and techniques of machine learning. To reach a particular destination we usually make a prediction which lane will be faster or likely be jammed. But now we use Google Maps to navigate when we want to travel anywhere since it displays the best route possible displaying various modes to reach the same. It forecasts traffic patterns for example whether it's okay to move, slowing down, or heavily trafficked using real-time location and average travel time.

The main aim of this essay is to examine two most widely used machine learning algorithms in prediction of a variety of applications like facial recognition, health care or email spam detection. Forecasting the outcomes of a game which involves multiple team players on general accuracy based parameters and giving a judgment on a specific scenario with data is the whole purpose of this study.

This study will be used in basic understanding of machine learning algorithms along with a specialized tool that helps in applying the learnt concepts. Prediction in the form of legal or illegal betting in areas of sports or stock markets have been witnessed widely on a global horizontal. Data is key in every application where it needs to be collected, preprocessed, validated before it is used for results. To automate this process and be able to find more less biased based responses or outcomes, machine learning algorithms fit the goal perfectly. Comparative study of which algorithm proves better than the other will be presented with the general parameters utilized to judge accuracy and even a personal opinion of which one would be easier to understand and execute holds importance.

This research may be used in a variety of team prediction games. Predictions for team building games have improved dramatically in recent years and continue to do so. Naive Bayes is the most fundamental and widely used prediction method. It is one of the simplest and most successful categorization methods for building quick machine learning models competent enough to make agile estimations. A thought occurred to me: are there any other algorithms that are superior to Naive Bayes? If there is, how is it superior to it and in what ways?

Background Theory

Before we dive into the execution of the algorithms there was major time dedicated to understanding from root to top the concepts underlying prediction based algorithms.

Predictive based algorithms use statistics and mathematical operations to calculate the result by using assistance from previous data available as well as tools and techniques under machine learning. In conjunction with past data ongoing facts and figures also are required to produce the correct outputs.

Prediction Based Model building or analysis involves a general structure as represented below in stepwise manner:

1. Identification: Decide on your data to be utilized as well as the algorithm depending on your goal or problem you are trying to figure out.
2. Data Collection: This is a crucial step which involves options like either data is readily available or data needs to be found out from distinct sources or even sometimes in cases of complex problems or newer ones you need to build your own data. Your results will depend on your data and the algorithm you choose. So figuring out the right dataset is a difficult role.
3. Analysis: The data that you discover or gather is not what could be the requirement of the next stages; it can have a lot of disturbances and look unstructured. To make it useful for analysis purposes it needs to undergo cleansing and alterations so that the noise can be avoided and appropriate good information can be discovered to give better results.
4. Statistical Conclusions: Conclusions have to be supported with calculations and numbers and not just only results. To judge the accuracy and comparative analysis statistical measures have to be imparted on the data to figure out whether the execution was fruitful or not.

Stated below are algorithms studied and evaluated under predictive analytics.

Linear Regression

A linear regression technique calculates the rate of one thing based on the value of something else. The variable you are using to locate or forecast the other is referred to by the term predictor variables, and the amount you seek to forecast is referred to as the predictor variables.

This form of reasoning calculates the coefficient of a continuous equation with inclusion of one or maybe more explanatory variables and perfectly forecasts the dependent variable's value. The differences between projected and real target value were reduced by matching a perfect line or curves using linear regression. (“About Linear Regression”)

The linear regression which is in its simplest form comprises authoring tools that utilize the "least squares" approach to choose the ideal line for a bunch of coupled data. Value of another variable Y is adapted to calculate the value of X.
Formula used in calculating linear regression:

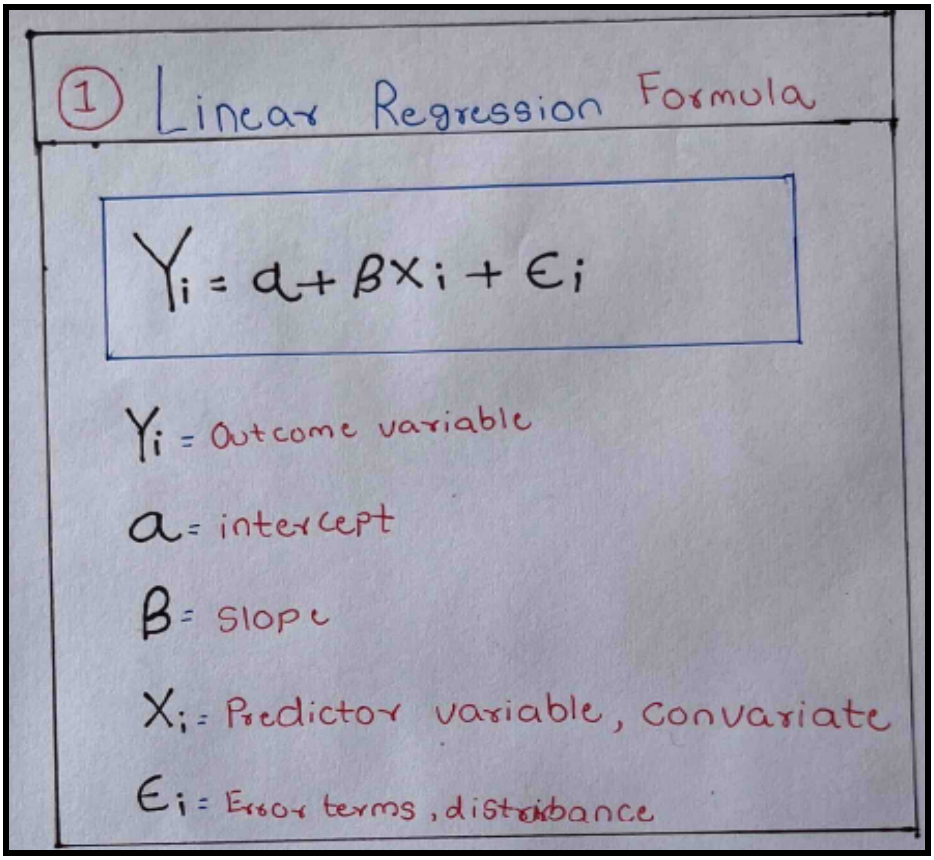


Figure 1: Formula of linear regression

Logistic Regression

Logistic regression is a popular AI approach which falls to the Guided Learning class. It's utilized to make predictions of the classified response variable depending on the group of predictor factors. (“Logistic Regression in Machine Learning”)

A logit model is used to anticipate the result of a class predictor variable. As just a consequence, the perception or reasoning has to have classes or definitions. You can expect boolean values in the state of 0 or 1 or true or false or yes or no. To portray a good amount of accuracy usually it does not show binary values but displays the probable values that occur between 0 and 1. (“Logistic Regression in Machine Learning”)

Although the methods in which regression is applied or utilized differ in their approaches they are still very similar in a lot of other aspects. If your problems are related to classifying something then logistic regression suits otherwise for nonlinear problems one can make use of linear regression.

To compute the values under Logistic Regression we can make use of Linear Regression which can be depicted in Figure 2, 3 and 4. A simple method can be shown for the same.

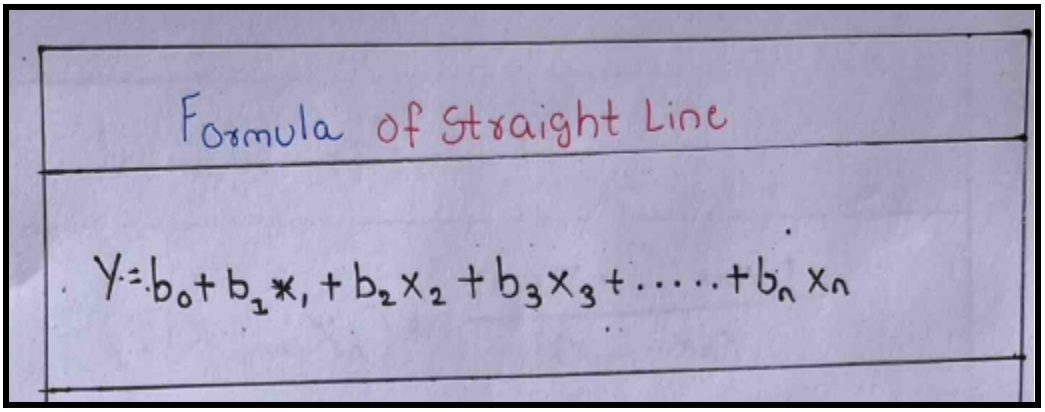


Figure 2: Straight Line

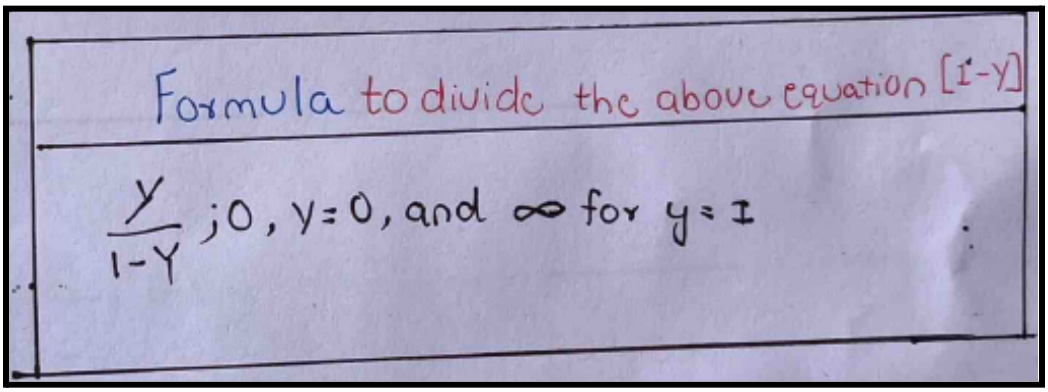


Figure 3: Divide the equation in Figure 1

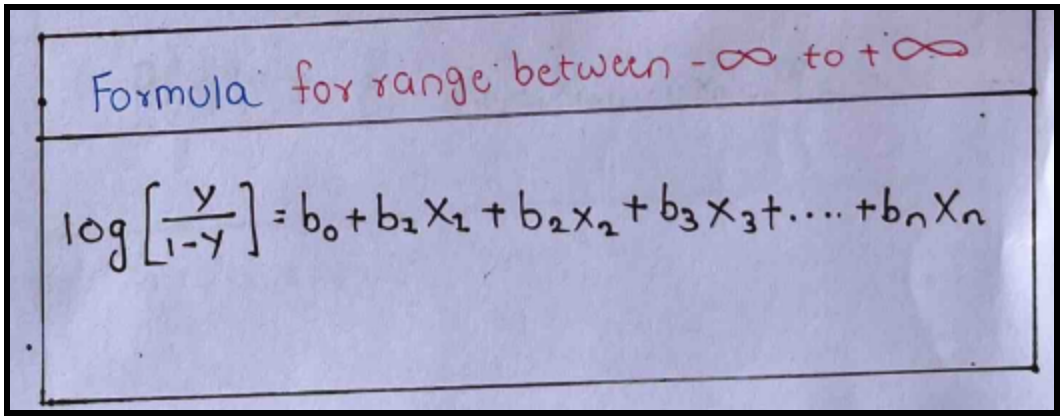


Figure 4: Range of values

Neural Networks

To comprehend the notion of artificially intelligent neural networks design, we must first grasp what such a neural network is. As a result of the process of creating a neural network, the overwhelming bulk of interconnected neurons, referred to here as a unit, are organized into a sequence of layers. Below listed and explained are the levels that can be found within an artificially formed neural network. (“What are Neural Networks? - India”)

There are three layers of artificial neural network explained:

1. **Input layer**:As the name suggests, it accepts input in a number of formats defined either by programmer.

2. **Hidden layers**: This concealed layer may be noticed between the intakes and focused on extracting. It performs every one of the math required to reveal hidden traits and patterns.

3. **Output layer**: The input is subjected to a series of adjustments with the hidden units, yielding outcomes which can be conveyed with some of this layer.

The input is subjected to something like a sequence of changes with both the hidden units, producing results that may be communicated through a portion of this layer.

It generates a balanced total, that's subsequently fed into such an input signal to obtain the outcome. The possibility that what a node will fire is determined by activation functions. People that get fired are the only ones who contribute towards the output units. There are various notable ones available that may be utilized based on the sort of task we are doing. (“What are Neural Networks? - India”)

Formula used in neural is:

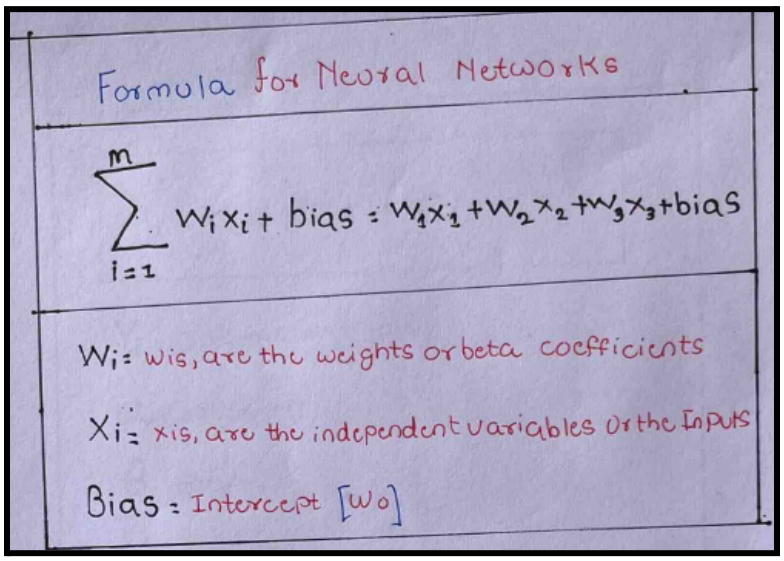


Figure 5: Formula of neural network

Assume that each node represents a distinct linear regression model, complete with input information, weights, a bias, and also an output.

Decision Tree

Regression as well as classification issues can be handled very well by Decision Tree Algorithm which is a supervised model. Mainly it is utilized for solving classification problems. In this form of forest classification, a tree would comprise nodes that depict attributes of a dataset, while branches show the rules made in a decision and each ending leaf node would display and give the result.

Every Decision Tree consists of two important nodes - Decision Node and Child Node. Judgment nodes are engaged to think critically with several branches, whereas Child nodes are the outcomes of such choices that don't have any additional branches located.

Use of decision trees:

While there are several algorithms throughout computer vision, the most essential feature to consider when constructing a predictive model for machine learning is to determine the proper approach for such data collected and situation. The two arguments for utilizing the Tree structure are as follows:

1. Decision trees are designed to mirror human decision-making abilities, making them simple to grasp.
2. Since this decision tree does have a forest structure, the logic behind it is easily understood.

Naive Bayes

Naive Bayes has a set of methods under the category of learning based algorithms. They make use of the Probability concepts of Bayes theorem under mathematics and statistics. Given a class variable y and an independent different feature x_1 through x_n, the Bayes theorem establishes the following relationship:

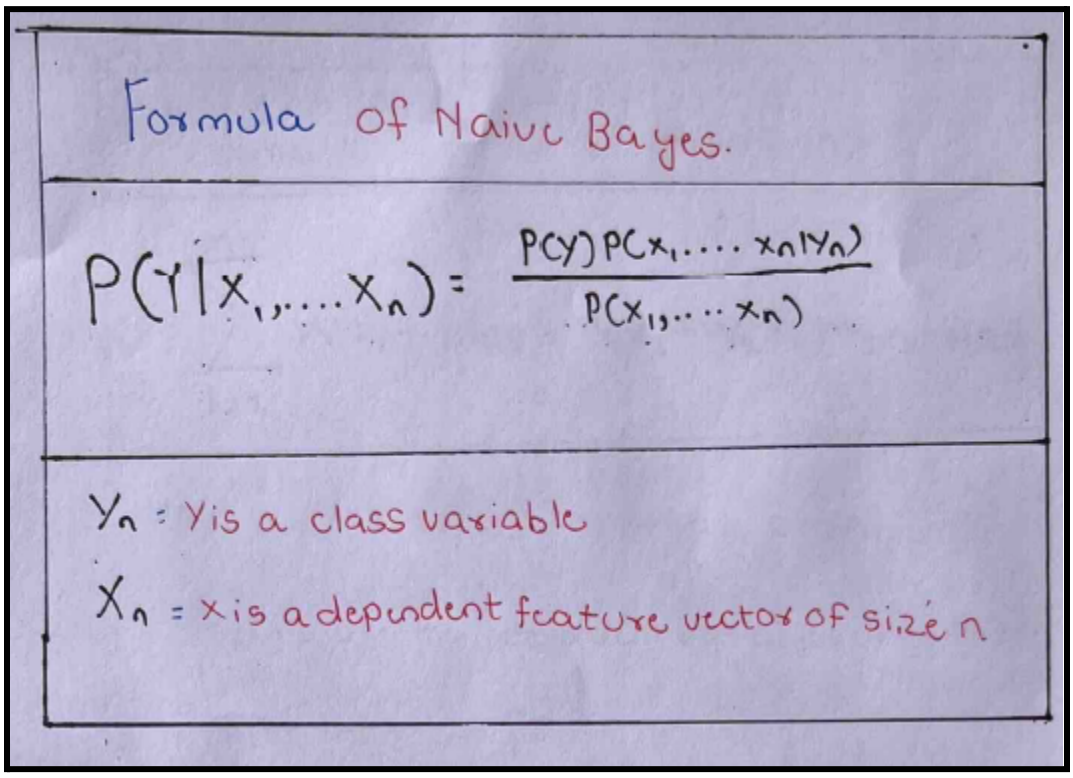


Figure 6: Formula of naive bayes

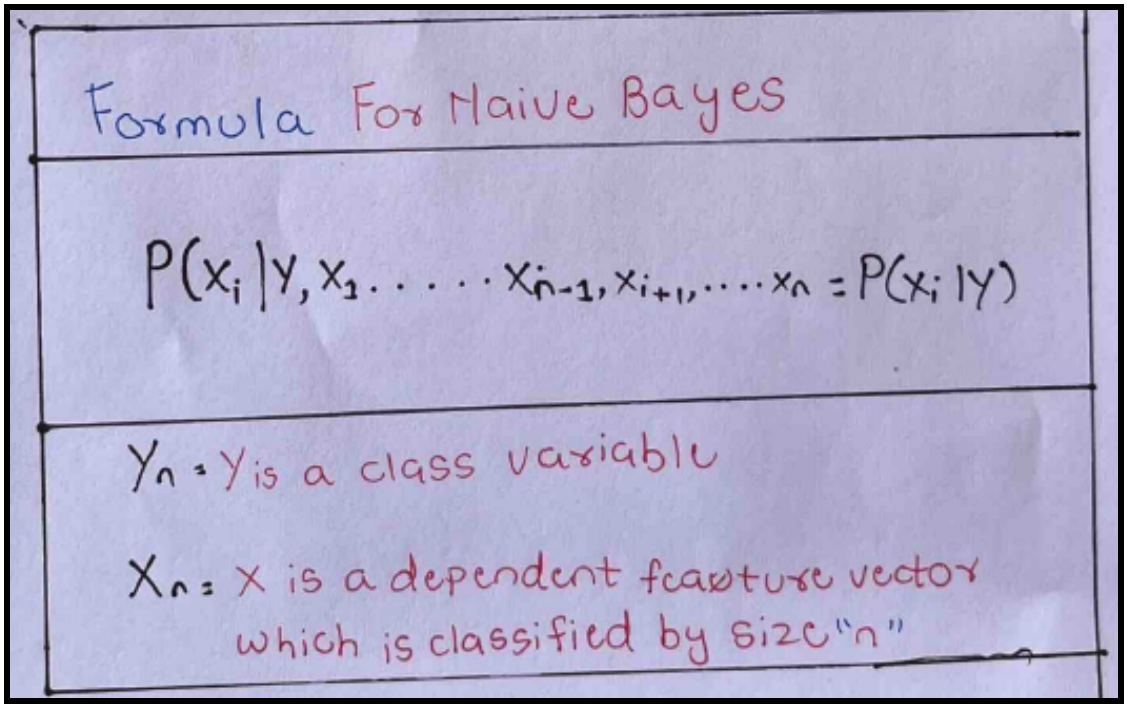


Figure 7: Second formula of naive bayes

Despite their apparent oversimplification of premises, naive Bayes classifiers have a positive effect in a wide range of real-world applications, emphasizing more on spam filtering and classification. For forecasting significant attributes some amount of training data is needed..

Naive Bayes Classification is known for its speed in execution when it is differentiated with the other more complicated methods. There is an independent evaluation of the distribution curve as there is a decoupling evident in the conditional class attribute distribution. To give feedback it mitigates the issues arising from the dimension curve.

Naive Bayes does an excellent job at classification but it is poor at estimation. In conclusion the results arising from probable values should be really looked into.

Below listed are the types of naive bayes algorithms to solve various problems.

1. Gaussian Naive bayes: This is a purely classification based machine learning problem. Using a randomized technique using a Gaussian distribution.

2. Multinomial Naive Bayes : This algorithmic approach is commonly used for classifying texts utilizing statistical evaluation on their contents.

3. Complement Naïve bayes: Instead of calculating the likelihood of an item falling to a particular category, we compute the chance of the item corresponding to any and all classes in complement Naïve Bayes classifier.

4. Bernoulli Naive bayes:It really is built upon that Bernoulli Distribution and is unable to take binary values that are 0 or 1.

5. Categorical Naive bayes:The categorical classifier based on a naive Bayes seems appropriate for categories distribution of identified.

Execution of both the algorithms were executed on Weka tool. As it was found difficult and time consuming to code the actual algorithms in any programming language and then post it compute the results. The Weka tool was researched and studied in detail. Below highlighting the information read and studied over the course of time about the tool.

Weka Tool

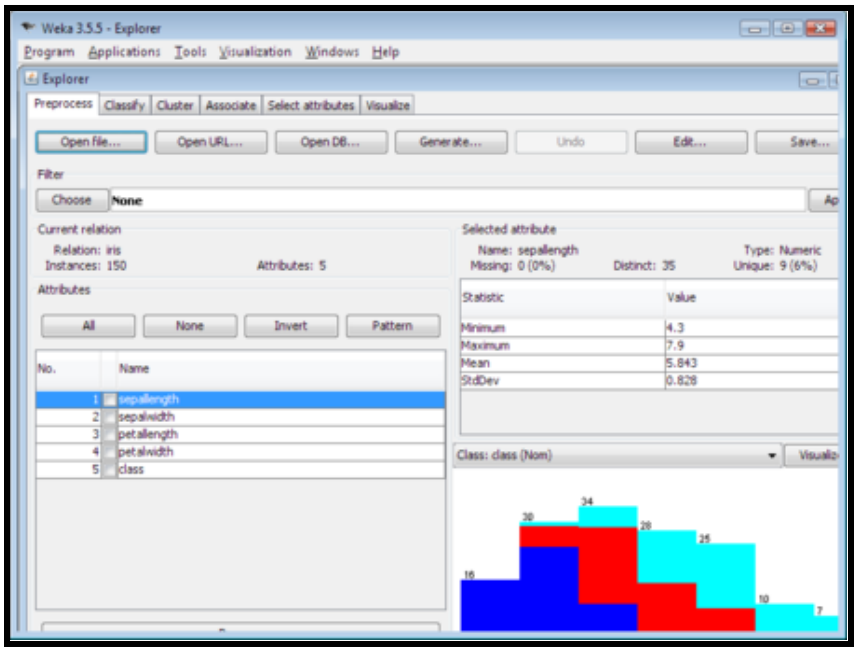


Figure 8: Snapshot of WEKA tool

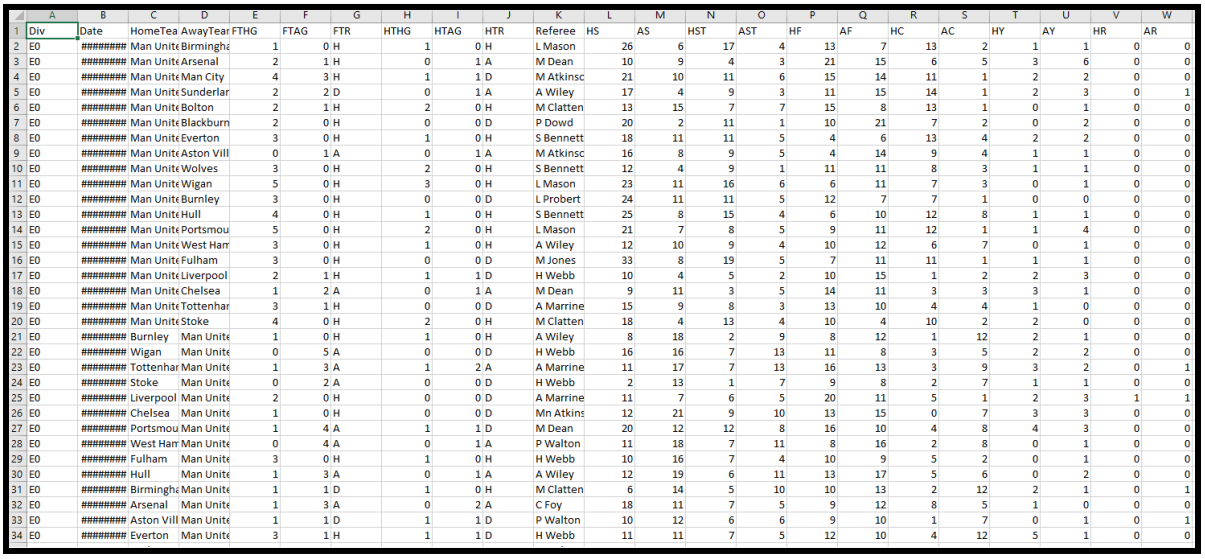
Weka is an open source platform that comprises a variety of machine learning methods which are used with your data to solve various real complex problems by utilizing pre installed techniques. (“Weka (machine learning)”)

The general method in the Weka tool involves taking in data, running various preprocessor checks to clean the data with any null or missing values. Collection of normal or data that is available as it is kept or stored. The stored data will have too many problems that can be resolved with pre- processing. Post this you make use of this new form of clean data and apply whichever algorithm you wish to it can be a clustering or classification. You can view the accuracy and other related parameters on successful completion of your algorithm on your dataset with detailed and intricate conclusions. (“Weka (machine learning)”)

Dataset

Football:

The Barclays Premier League seems to be the greatest in the world. It includes 20 teams who have qualified for the championship. Between all these 20 teams, five have been crowned champions in the previous 12 seasons: Manchester City, Liverpool, Manchester United, Chelsea, and Leicester, with two outsiders Arsenal and Tottenham. Every set of data comprises the last 12 seasons of the English Premier League and is in csv format. There are several Attributes in these data sets:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	Div	Date	HomeTea	AwayTear	FTHG	FTAG	FTR	HTHG	HTAG	HTR	Referee	HS	AS	HST	AST	HF	AF	HC	AC	HY	AY	HR	AR
2	E0	########	Man Unite	Birmingha	1	0	H	1	0	H	L Mason	26	6	17	4	13	7	13	2	1	1	0	0
3	E0	########	Man Unite	Arsenal	2	1	H	0	1	A	M Dean	10	9	4	3	21	15	6	5	3	6	0	0
4	E0	########	Man Unite	Man City	4	3	H	1	1	D	M Atkinsc	21	10	11	6	15	14	11	1	2	2	0	0
5	E0	########	Man Unite	Sunderlar	2	2	D	0	1	A	A Wiley	17	4	9	3	11	15	14	1	2	3	0	1
6	E0	########	Man Unite	Bolton	2	1	H	2	0	H	M Clatten	13	15	7	7	15	8	13	1	0	1	0	0
7	E0	########	Man Unite	Blackburn	2	0	H	0	0	D	P Dowd	20	2	11	1	10	21	7	2	0	2	0	0
8	E0	########	Man Unite	Everton	3	0	H	1	0	H	S Bennett	18	11	11	5	4	6	13	4	2	2	0	0
9	E0	########	Man Unite	Aston Vill	0	1	A	0	1	A	M Atkinsc	16	8	9	5	4	14	9	4	1	1	0	0
10	E0	########	Man Unite	Wolves	3	0	H	2	0	H	S Bennett	12	4	9	1	11	11	8	3	1	1	0	0
11	E0	########	Man Unite	Wigan	5	0	H	3	0	H	L Mason	23	11	16	6	6	11	7	3	0	1	0	0
12	E0	########	Man Unite	Burnley	3	0	H	0	0	D	L Probert	24	11	11	5	12	7	7	1	0	0	0	0
13	E0	########	Man Unite	Hull	4	0	H	1	0	H	S Bennett	25	8	15	4	6	10	12	8	1	1	0	0
14	E0	########	Man Unite	Portsmou	5	0	H	2	0	H	L Mason	21	7	8	5	9	11	12	1	1	4	0	0
15	E0	########	Man Unite	West Ham	3	0	H	1	0	H	A Wiley	12	10	9	4	10	12	6	7	0	1	0	0
16	E0	########	Man Unite	Fulham	3	0	H	0	0	D	M Jones	33	8	19	5	7	11	11	1	1	1	0	0
17	E0	########	Man Unite	Liverpool	2	1	H	1	1	D	H Webb	10	4	5	2	10	15	1	2	2	3	0	0
18	E0	########	Man Unite	Chelsea	1	2	A	0	1	A	M Dean	9	11	3	5	14	11	3	3	3	1	0	0
19	E0	########	Man Unite	Tottenhar	3	1	H	0	0	D	A Marrine	15	9	8	3	13	10	4	4	1	0	0	0
20	E0	########	Man Unite	Stoke	4	0	H	2	0	H	M Clatten	18	4	13	4	10	4	10	2	2	0	0	0
21	E0	########	Burnley	Man Unite	1	0	H	1	0	H	A Wiley	8	18	2	9	8	12	1	12	2	1	0	0
22	E0	########	Wigan	Man Unite	0	5	A	0	0	D	H Webb	16	16	7	13	11	8	3	5	2	2	0	0
23	E0	########	Tottenhar	Man Unite	1	3	A	1	2	A	A Marrine	11	17	7	13	16	13	3	9	3	2	0	1
24	E0	########	Stoke	Man Unite	0	2	A	0	0	D	H Webb	2	13	1	7	9	8	2	7	1	1	0	0
25	E0	########	Liverpool	Man Unite	2	0	H	0	0	D	A Marrine	11	7	6	5	20	11	5	1	2	3	1	1
26	E0	########	Chelsea	Man Unite	1	0	H	0	0	D	Mn Atkins	12	21	9	10	13	15	0	7	3	3	0	0
27	E0	########	Portsmou	Man Unite	1	4	A	1	1	D	M Dean	20	12	12	8	16	10	4	8	4	3	0	0
28	E0	########	West Ham	Man Unite	0	4	A	0	1	A	P Walton	11	18	7	11	8	16	2	8	0	1	0	0
29	E0	########	Fulham	Man Unite	3	0	H	1	0	H	H Webb	10	16	7	4	10	9	5	2	0	1	0	0
30	E0	########	Hull	Man Unite	1	3	A	0	1	A	A Wiley	12	19	6	11	13	17	5	6	0	2	0	0
31	E0	########	Birmingha	Man Unite	1	1	D	1	0	H	M Clatten	6	14	5	10	10	13	2	12	2	1	0	1
32	E0	########	Arsenal	Man Unite	1	3	A	0	2	A	C Foy	18	11	7	5	9	12	8	5	1	0	0	0
33	E0	########	Aston Vill	Man Unite	1	1	D	1	1	D	P Walton	10	12	6	6	9	10	1	7	0	1	0	1
34	E0	########	Everton	Man Unite	3	1	H	1	1	D	H Webb	11	11	7	5	12	10	4	12	5	1	0	0

Figure 9:This is the image of data set

Preprocessing

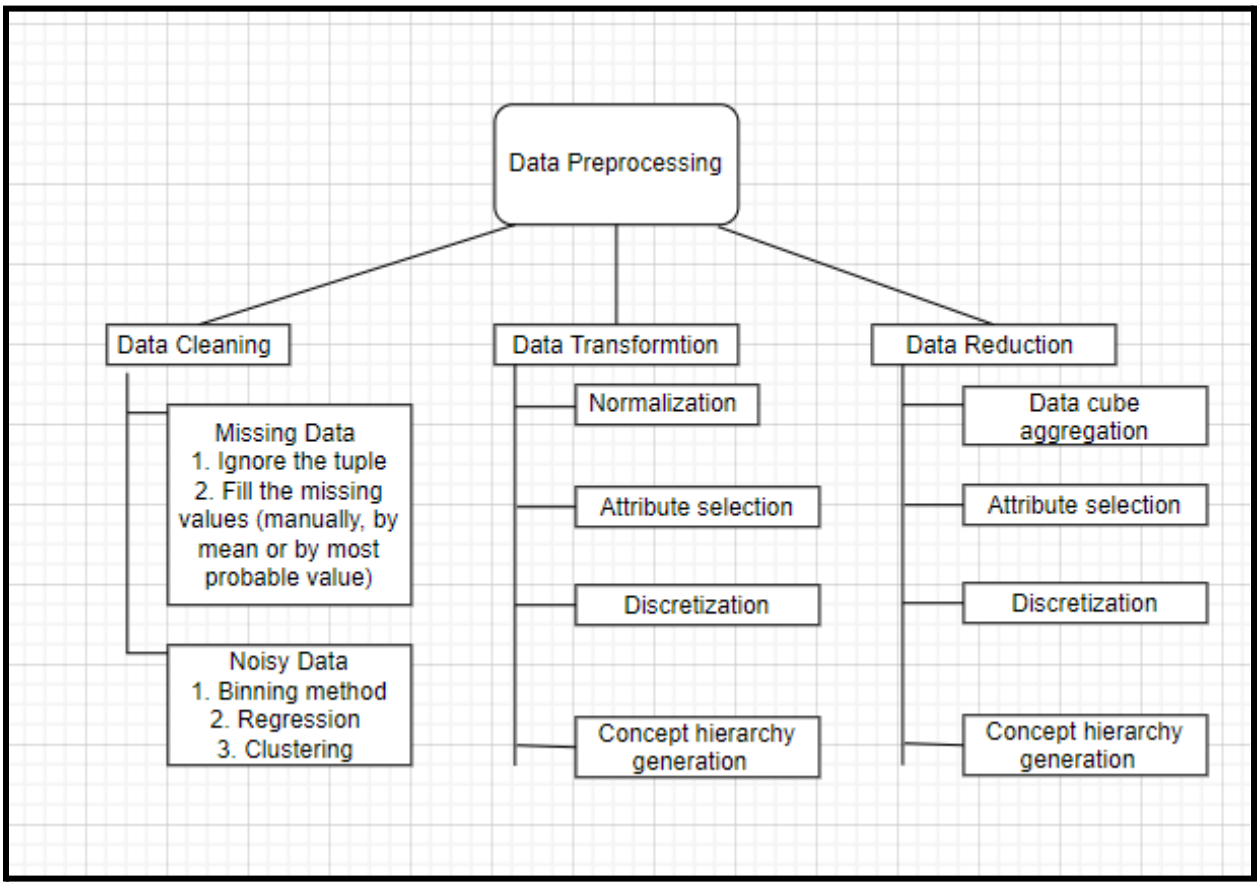


Figure 10: Preprocessing

Data preprocessing is a technique in data mining used to turn raw information into a usable and accessible manner**.** Weka's preprocessing capabilities are known as "Filters." Considerations pulls information from a file, SQL database, or URL . Weka's preprocessing tools may be employed to categorize the data. (“Data Preprocessing in Data Mining”)

Implementation

Step 1: I had chosen Weka as the primary application because it is very user friendly, interactive and also flexible in many ways. For example, it supports many formats for

the files When you launch the Weka programme, it appears shown in Figure 11. At the left side of the screen there we are going to choose Explorer. The other options are the following:

1. **Experimenter:** The Weka Experimenter lets you create your own experiments by running programs on databases, then execute them and examine the outcomes.
2. **knowledgeFlow:** Knowledge flows are movements of knowledge through people, organizations, locations, and time that show changes, shifts, and applications.
3. **Work bench:**The Classification Workbench is a graphical user interface (GUI) environment that unifies all of the User flow into a single interface. It's handy if you often switch between a number of different platforms, such as the Navigator and the Experimental Platform.
4. **Simple CLI:** Simple CLI is a graphical command-line interface that enables you to execute Weka functions and operations.

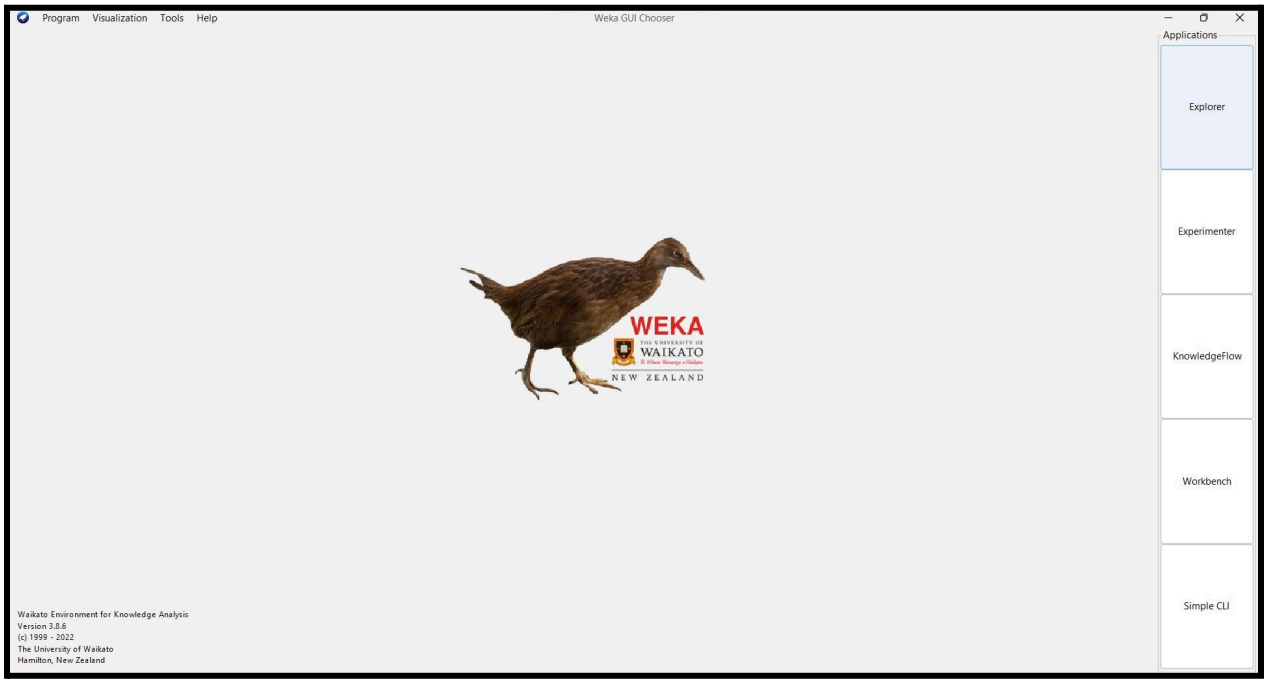


Figure 11: Home screen of weka

Step 2: This is the user interface after pressing explorer. There are several options like open file, open URL, open DB, generate and etc however out of which I am going to select Open file which will lead to the interface as shown in the next image. There are several functions of those buttons such as open URLs that would help in accessing the file if they are online on websites, open DB has a function to include a graphical tool for inspecting the data. The different models can be utilized on the same dataset. You may then analyze the results of several models and choose the best one for your needs. As a result, using WEKA speeds up the building of machine learning models in general. General is used if the resulting code is inherited from an identical class hierarchy, it could be used as a standard classification algorithm within Weka.

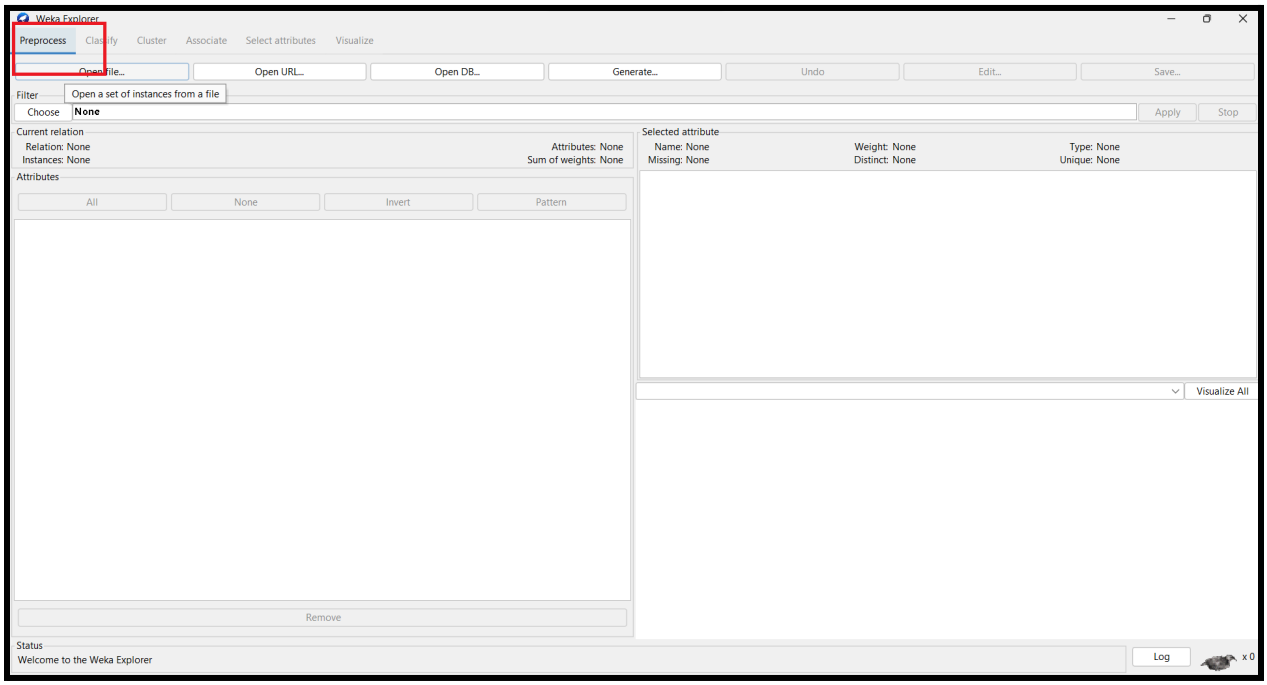


Figure 12: First Stage Preprocessing

Step 3: As we can see in Figure 13, once we press on an open file a different screen pops up which shows all the files that are located in the laptop, however the default format of Weka tool is ARFF which needs to be changed to csv which is an extension for the Excel file. There are various other format files available.

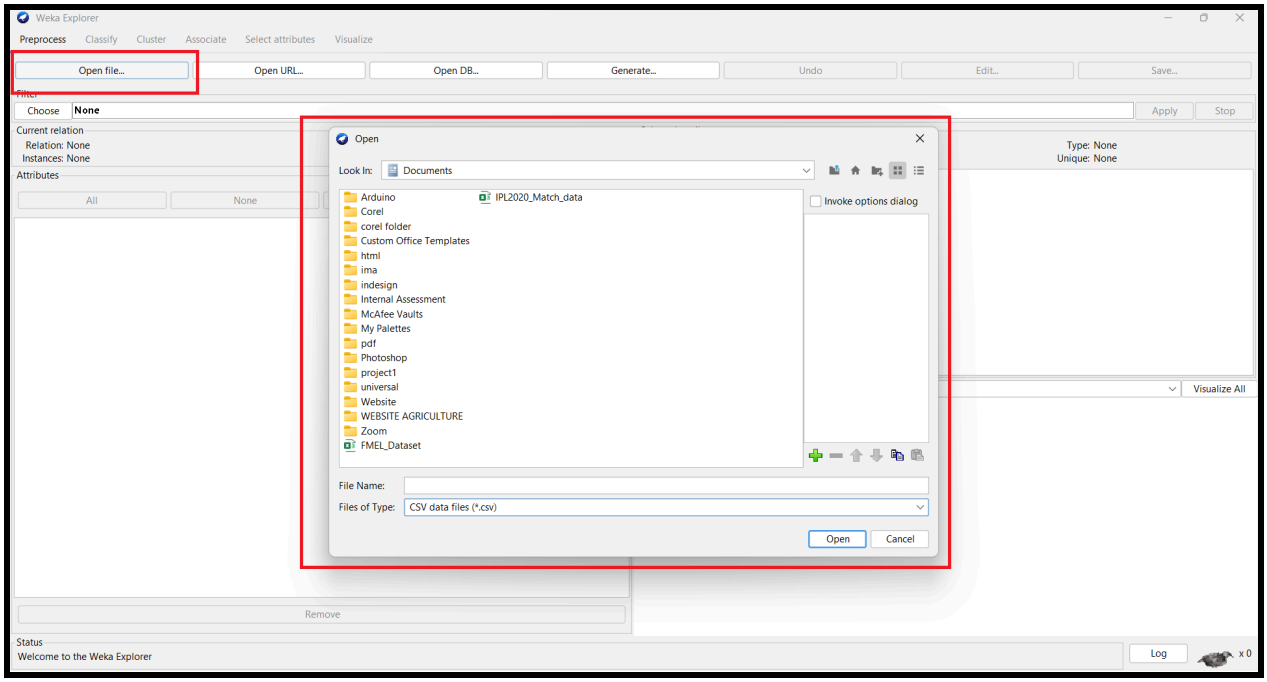


Figure 13: File location

Step 4:

The interface displays the prediction, with a graph at the bottom right of the screen visualizing which team will win, and visualization for the chosen qualities shown along the left side of the screen. At the top of the interface there are several options such as Preprocess, classify, cluster, associate, selection attributes and visualize.

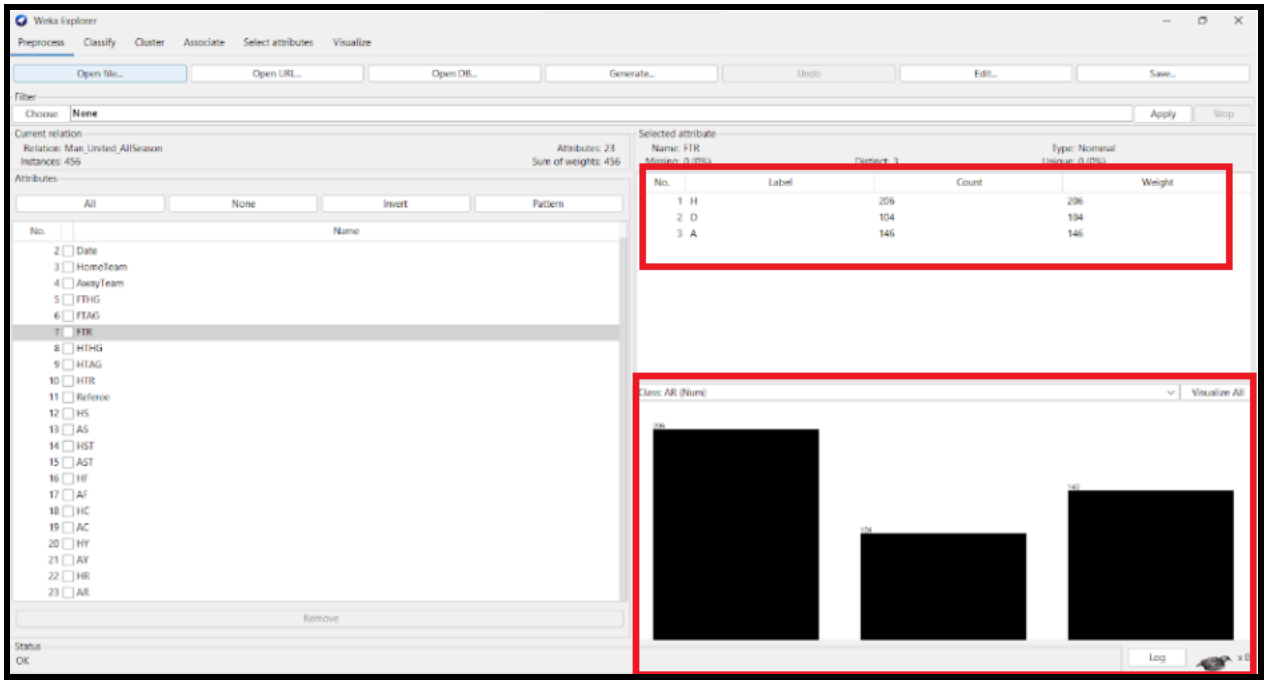


Figure 14: Visualization of dataset

Step 5: At the top of the screen after selecting classify I pressed the option choose within which there were several categories of prediction based algorithms i choose naive bayes and logistic regression as my primary algorithms.

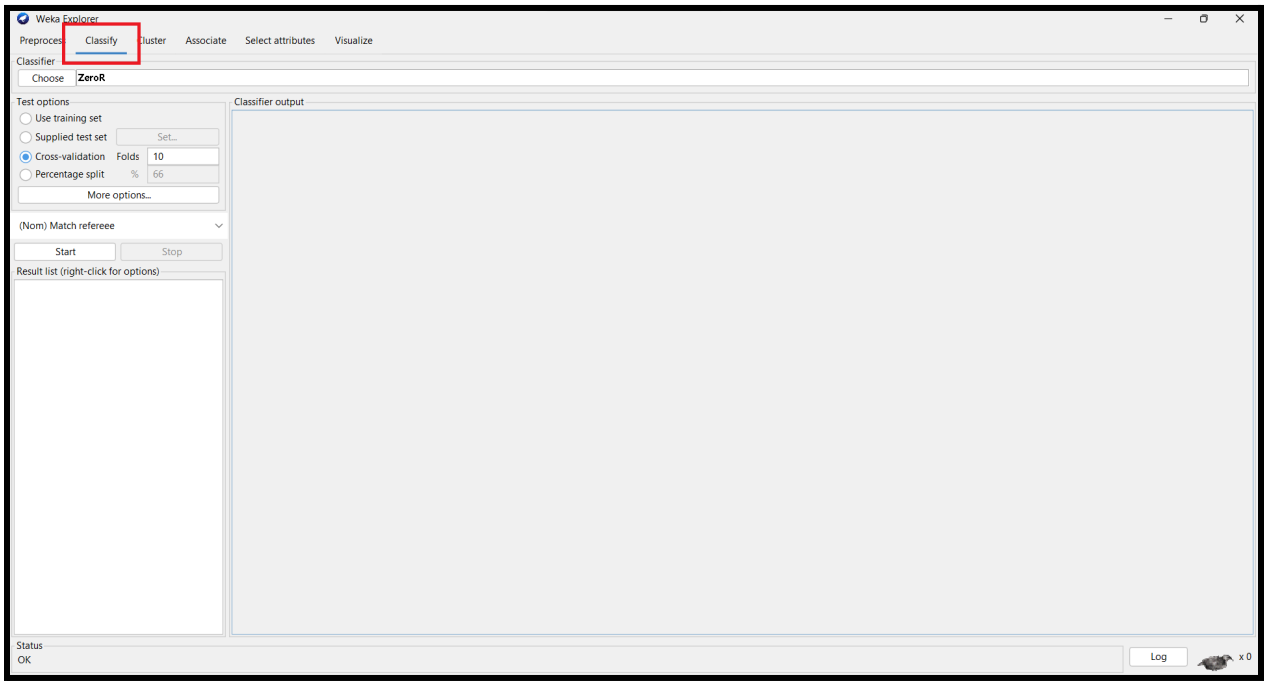


Figure 15: Classification

Step 6: Then, when I've chosen one of numerous possibilities, I'll see naive bayes in the classifiers area. When we choose naive bayes, the Weka tool displays the definition of naive bayes as well as the formulae utilized to create these algorithms. The similar procedure was used for logistic regression, except the spot to look for logistic regression is inside rules.

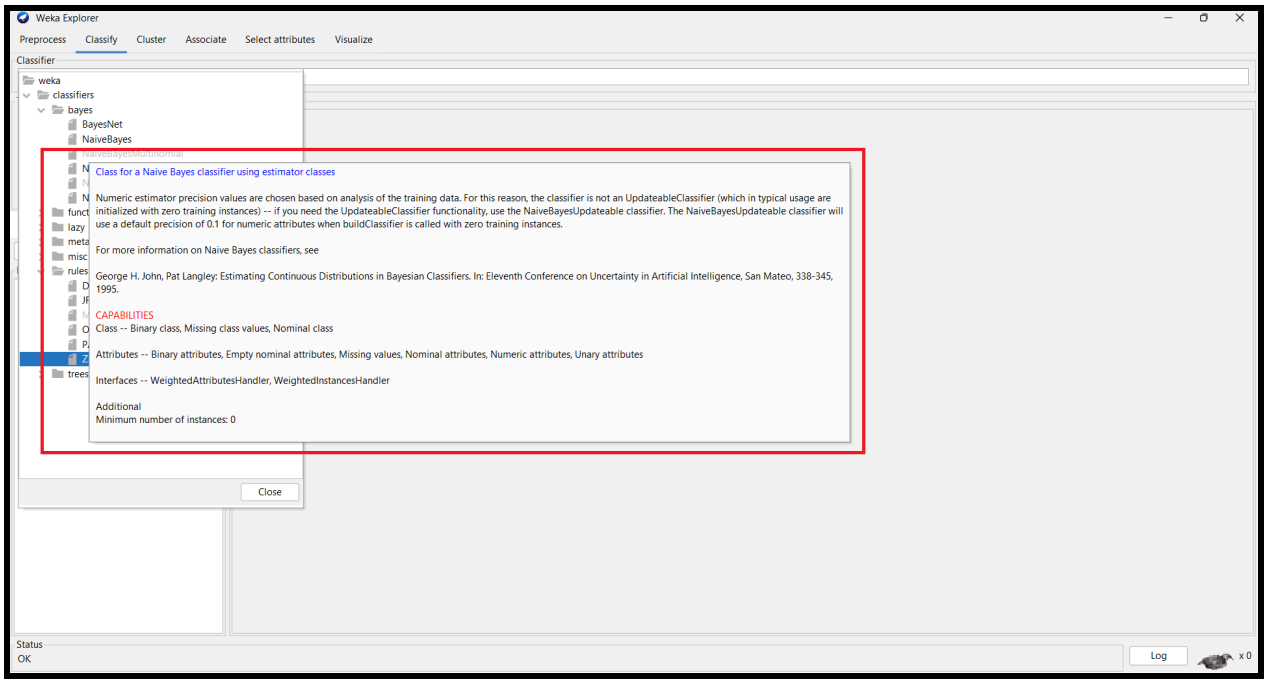


Figure 16: Description of Naive Bayes

Image of logistic regression:

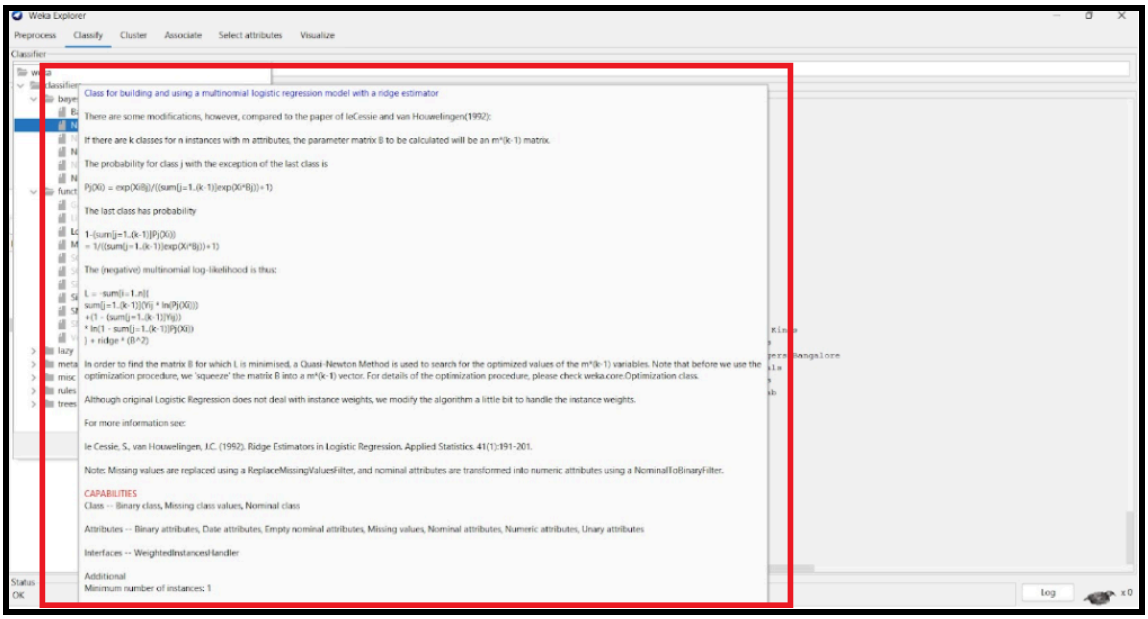


Figure 17: Description of Logistic Regression

Step 6: The last step displays the naive bayes and logistic regression results. The naive bayes were computed using cross-validation inside the Test options, and the data was given in a tabular and matrix style. The logistic regression was calculated in the same manner.

FTR: Full time result for man city: (naive bayes)

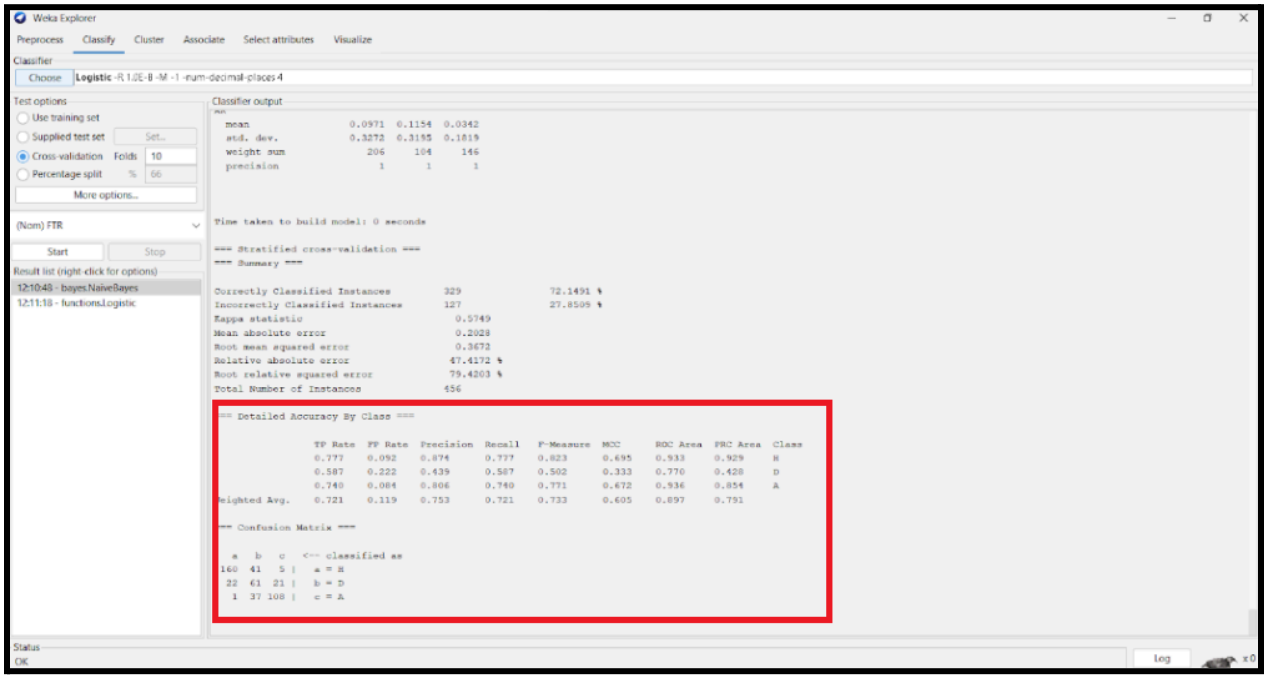


Figure 18: Result analysis of Naive Bayes

FTR: Full time result for man city: (Logistic Regression)

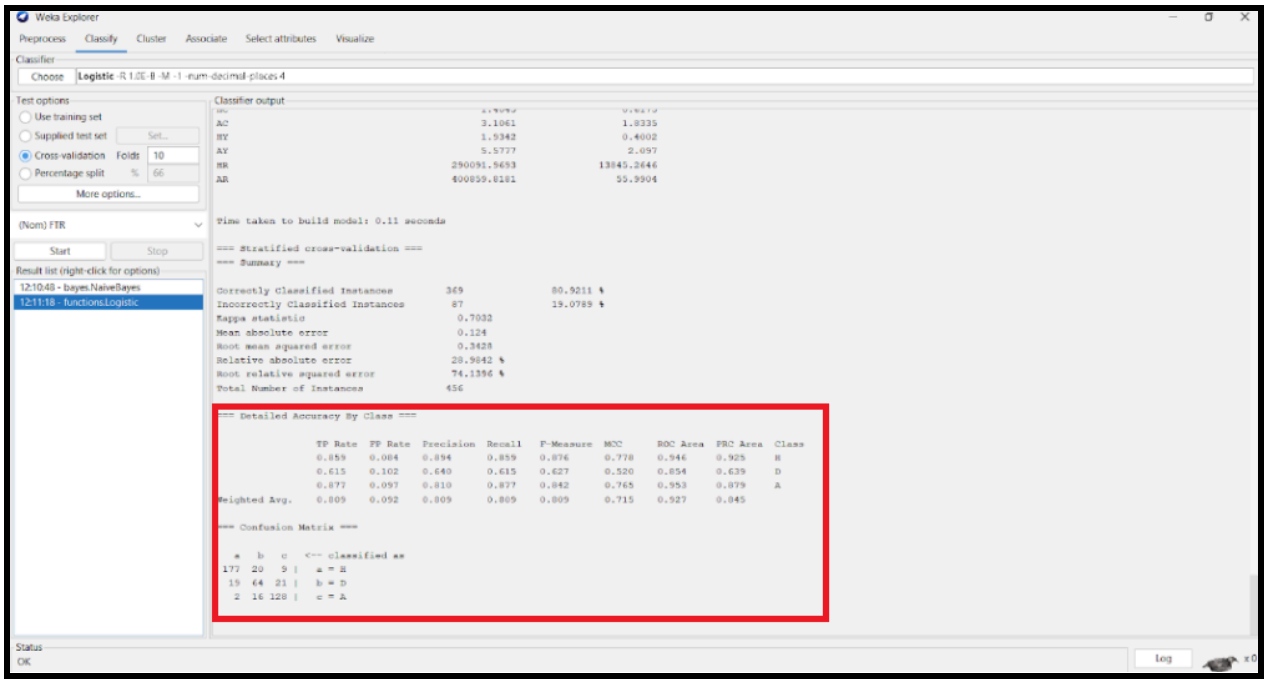


Figure 19: Result Analysis of Logistic Regression

Procedure to perform the algorithms:

Step 1: Exploration and discovery of data sets on kaggle.com, followed by download.

Step 2: Load the data sets into the Weka programme, but make sure that the file format is permitted by Weka.

step 3: Select naive bayes from the classify option and then run the algorithm. Same steps can be used to perform logistic regression.

Step 4: The results will be displayed at the right side of the interface as shown on above images.

Step 5: The format of the results are shown in the format as shown below:

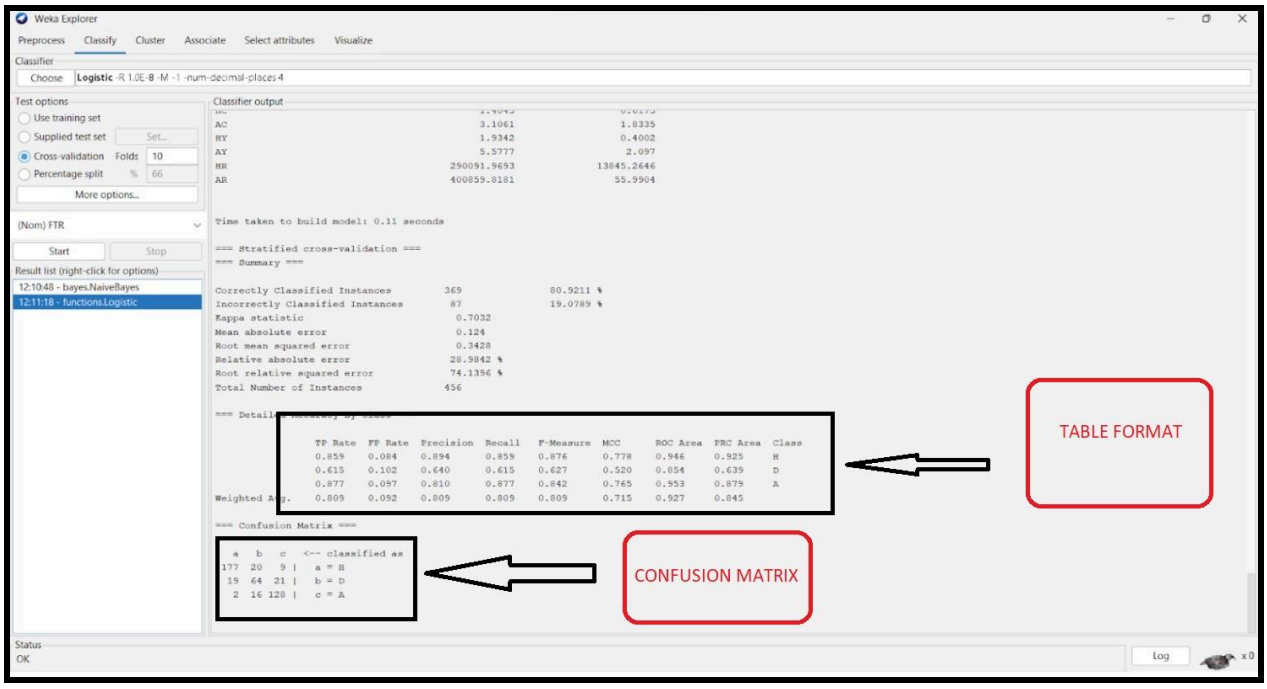


Figure 20: Confusion Matrix

Results

After running multiple trials for these algorithms, the results were displayed in two configurations on the right side of the screen: table and matrix. (Awadallah #)

I had referred several papers, each and every paper had different views however my results were different since my datasets were completely different, in there paper they had supported naive bayes as the accurate classifier however logistic regression was more accurate for my research (Awadallah #)

TP rate: represents the proportion of data that is properly categorized.

FP rate: amount of false positives. This is estimated as FP/FP+TN, where FP is the figure of negative positives and TN is the figure of true negative things (the statistic of negatives becoming FP+TN). It is the likelihood of an untrue alarm being constructed: that a positive result will be rebounded when the real negative value is returned.

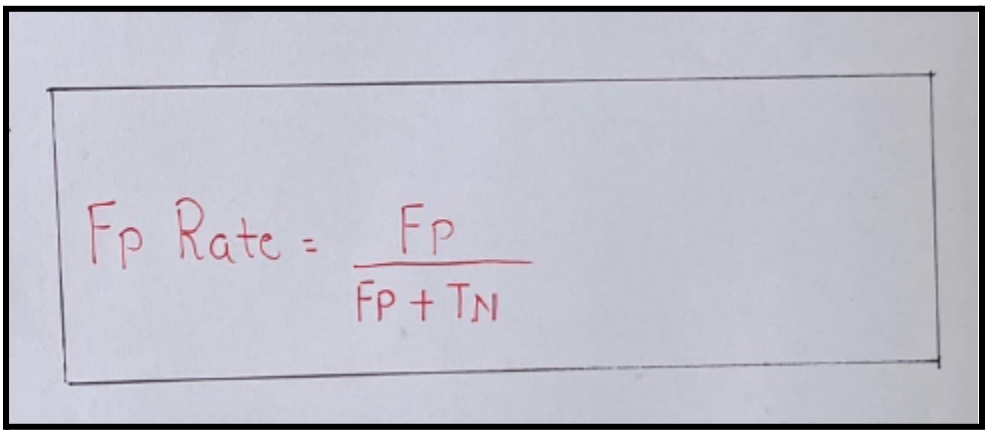


Figure 21: Formula for FP rate

Precision: Precision is measured by dividing the total amount of correctly predicted (TP + FP) by the number of correct positive predictions (TP). The maximal precision is 1.0, while the minimal is 0.0.

Formula for Precision:

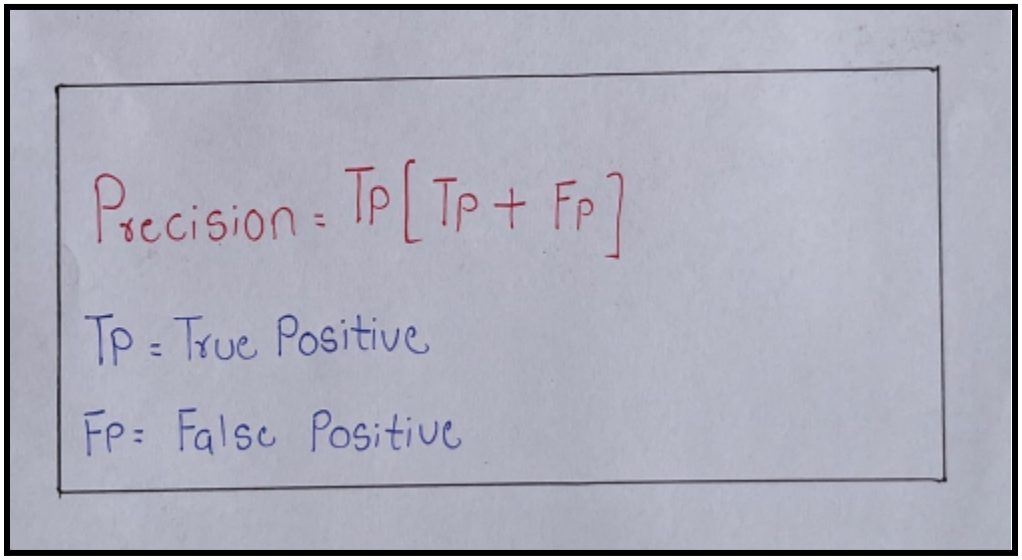


Figure 22:Formula for precision

Recall: This is calculated by correctly finding out positive samples from the given whole set of samples. It is used as a metric especially in the case of displaying accuracy

F-measure: The harmonic average of a system's accuracy and recall values is used to get an F-score. It is determined by the formula:

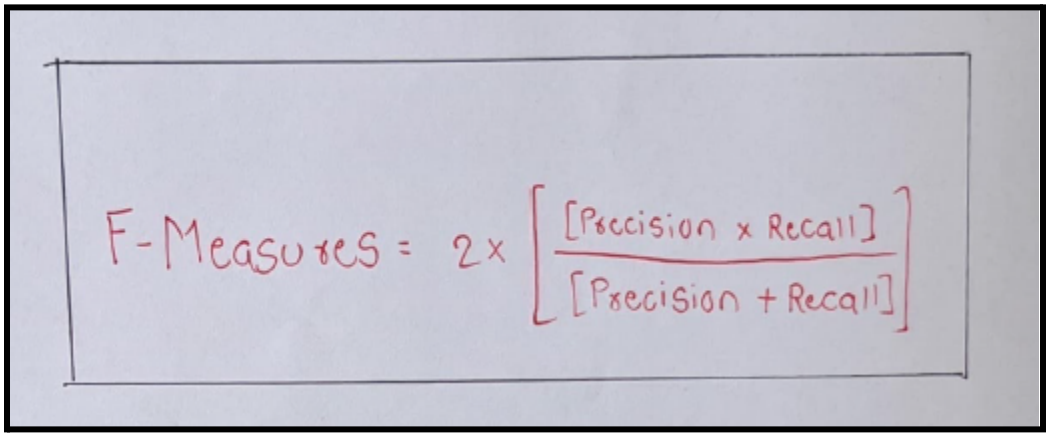


Figure 23: Formula for F-measures

MCC: MCC It's a relationship between projected and actual classes. It may be determined using the confusion matrix numbers.

ROC: For geographical reasons, the curve is called "Receiver - operating characteristic Characteristic." The region underneath the ROC curve is printed by Weka.

Kappa statistic: The Kappa statistic compares a Reported Accuracy to an Estimated Accuracy.

Mean absolute error: A measurement of how similar forecasts or projections are to the actual events.

Root mean squared error: computed as the Mean absolute percentage error reduced by that of the ZeroR classification error.

Relative absolute error: computed as the mean absolute error reduced by the error of something like the ZeroR classification. (“High Relative absolute error and Root relative squared error in classification”)

Root relative squared error: computed by dividing the Average absolute error through the ZeroR classifier error. (“High Relative absolute error and Root relative squared error in classification”)

Confusion matrix:

A confusion matrix is utilized to thoroughly investigate the potential for a classification purpose. In the matrix, we can see diagonality wherein the correctly classified outcomes in element form are placed. The ones that are not classified correctly are kept outside.

A confusion matrix that comprises fundamentals only in the diagonal area and the remaining are all zero is considered to be the best. Post the classification process the actual and predicted elements are seen.

RESULT ANALYSIS

Naive Bayes:

Correctly Classified Instances	329	72.1491%
Incorrectly Classified Instances	127	27.8509%
Kappa statistic	0.5749	

Mean absolute error	0.2028	
Root mean squared error	0.3672	
Relative absolute error	47.4172%	
Root relative squared error	79.4203%	

Table 2: Naive Bayes accuracy based results

TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Area**	**PRC Area**	**Class**
0.777	0.092	0.874	0.777	0.823	0.695	0.933	0.929	**H (Home)**
0.587	0.222	0.439	0.587	0.502	0.333	0.770	0.428	**D(Draw)**
0.740	0.084	0.806	0.740	0.771	0.672	0.936	0.854	**A(Away)**
0.721	**0.119**	**0.753**	**0.721**	**0.733**	**0.605**	**0.897**	**0.791**	**(Average)**

Table 3: Parameters of Recall & Precision for Naive Bayes

Logistic Regression:

Correctly Classified Instances	369	80.9211%
Incorrectly Classified Instances	87	19.0789%
Kappa statistic	0.7032	
Mean absolute error	0.124	

Root mean squared error	0.3428	
Relative absolute error	28.9842%	
Root relative squared error	74.1396%	

Table 4: logistic Regression accuracy based on results

TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Area**	**PRC Area**	**Class**
0.859	0.084	0.894	0.859	0.876	0.778	0.946	0.925	**H (Home)**
0.615	0.102	0.640	0.615	0.627	0.520	0.854	0.639	**D(Draw)**
0.877	0.097	0.810	0.877	0.842	0.765	0.953	0.879	**A(Away)**
0.809	0.092	0.809	0.809	0.809	0.715	0.927	0.845	**(Average)**

Table 5: Parameters of Recall & Precision for Logistic regression

We can see that logistic regression is obviously more accurate than naïve bayes in many ways. Below listed are the reasons logistic regression seems to be more accurate than naive bayes:

1. **Correctly Classified Instances**

Correctly Classified Instances	329	**Naive bayes**
Correctly Classified Instances	369	**logistic regression**
Total	**40**	

Table 6: Final Comparison of Naive Bayes and Logistic Regression

This shows that 40 more correct instances were presented, in logistic regression than naive bayes.

The overall percentage of naive bayes is 72.1491% whereas logistic regression is 80.9211 %.

2. **Precision:**
   Logistic regression indicates that the weight average of precision is 0.809 whereas naïve bayes has 0.753, indicating that logistic regression is more accurate since the closer the value is to one, the more accurate the predictions.

3. **Recall:**
   The recall weighted average for logistic regression is 0.809, whereas the average for naive bayes is 0.721. The logistic regression is more accurate since it evaluates its ability to recognise positive data. The greater the notification, the more positive samples discovered.
4. **F-measure:**
   The weighted average for f-measures in logistic regression is 0.809, whereas the average for naive bayes is 0.733. We may conclude that logistic regression is more accurate since the f-measure is a mix of recall and precision measurements. F-measures are calculated using the formula 2 x (Precision x Recall / Precision + Recall).

Challenges:

Below listed are the challenges post undertaking the research:

1. Machine Learning is an attractive field of computer science and my inclination towards sports made me opt for a topic that could keep me interested throughout the year.But it was tough to understand the basics and then learn the mathematical and coding aspects of so many different algorithms. Finally it was a

difficult task to code both the algorithms with such a huge dataset .Even to discover a tool like WEKA it took a good amount of time and a lot of features and techniques had to be studied.

2. Finding data sets was a huge problem because the large datasets for football did not fulfill my needs; nevertheless, scouring the internet and discovering kaggle. Kaggle has practically every sporting dataset, as well as the occurrences and projections for the datasets. However, because Weka only supports a restricted number of extensions, locating the data set with the proper extension proved difficult.

Future Scope

1. The data set I employed may be old and out of date, which might also impair the accuracy of the comparison; nevertheless, in the near future, I may use newer, recent data sets or develop my own data set.

2. Another possibility for the future scope is to develop my individual algorithm that is tailored and customized, making the comparisons between classified and easier to contrast among the two or more classifiers. I can write the algorithm below in Python or Java.

Conclusion:

I learned about many types of classifiers from the background theory, such as Naive Bayes, logistic regression, decision trees, linear regression, and neural networks. Following the completion of the study, I was able to separate the classifiers and determine which classifiers to utilize. I chose naives bayes and logistic regression since

they both had comparable qualities for comparison, which aided and simplified the analysis and comparison.

After picking the classifier, I needed to choose an algorithm that could run both of the classifiers as well as the algorithm, about which I used the weka tool. I chose weka since it's more versatile and simple to use. I utilized a data set from kaggle, which featured many sorts of data sets, and the data set was quite precise and accurate.

The experiment tried to determine whether the classifier is more accurate in forecasting a football team's odds of winning. At the completion of the trial and data analysis procedure, it is possible to infer that naive bayes are less accurate than logistic regression for the data sets utilized in the experiment. The findings may differ for others, however logistic regression is considerably more accurate than naive bayes in practically every area in my instance. Despite the investigation's limitations, this result is trustworthy. This inquiry is just meant to show the tremendous potentials of prediction utilizing the Weka tool as machine learning algorithms.

Works Cited

Weka Error Measurements When the class value is nominal, the kappa statistic is given. When the class value is numeric, the cor, https://katie.mtech.edu/classes/csci347/Resources/Weka_error_measurements.pdf. Accessed 4 November 2022.

Data Analytics, Kappa statistic, https://faculty.kutztown.edu/parson/fall2019/Fall2019Kappa.html. Accessed 4 November 2022.

"About Linear Regression." *IBM*, https://www.ibm.com/in-en/topics/linear-regression. Accessed 16 July 2022.

Awadallah, Ahmed Amr. "Football Match Prediction using Deep Learning (Recurrent Neural Network)." *CS230*, p. 9.

"Data Preprocessing in Data Mining." *GeeksforGeeks*, 29 June 2021, https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/. Accessed 3 November 2022.

"Decision Trees for Classification: A Machine Learning Algorithm." *Xoriant*, https://www.xoriant.com/blog/decision-trees-for-classification-a-machine-learning-algorithm. Accessed 16 July 2022.

"False Positive Rate | Split Glossary." *Split Software*, https://www.split.io/glossary/false-positive-rate/. Accessed 4 November 2022.

Gad, Ahmed Fawzy. "Accuracy, Precision, and Recall in Deep Learning." *Paperspace Blog*, https://blog.paperspace.com/deep-learning-metrics-precision-recall-accuracy/. Accessed 2 March 2023.

"High Relative absolute error and Root relative squared error in classification." *Stack Overflow*, 26 April 2016, https://stackoverflow.com/questions/36876749/high-relative-absolute-error-and-root-relative-squared-error-in-classification. Accessed 4 November 2022.

"Logistic Regression in Machine Learning." *Javatpoint*, https://www.javatpoint.com/logistic-regression-in-machine-learning. Accessed 2 March 2023.

PASSERAT-PALMBACH, Dr. Jonathan. "Predicting Football Results Using Machine Learning Techniques." June 20, 2018, p. 73.

"prathameshtari/Predicting-Football-Match-Outcome-using-Machine-Learning: Football Match prediction using machine learning algorithms in jupyter notebook." *GitHub*, 6 November 2018, https://github.com/prathameshtari/Predicting-Football-Match-Outcome-using-Machine-Learning. Accessed 4 November 2022.

"Predicting the Winner of a Tennis Match Using Machine Learning Techniques." *NORMA@NCI Library*, 20 December 2018, https://norma.ncirl.ie/4299/1/akshayasekar.pdf. Accessed 4 November 2022.

"Weka Data Mining Tutorial for First Time & Beginner Users." *YouTube*, 21 March 2012, https://youtu.be/m7kpIBGEdkI. Accessed 4 November 2022.

"Weka (machine learning)." *Wikipedia*, https://en.wikipedia.org/wiki/Weka_(machine_learning)#Extension_packages. Accessed 29 September 2022.

"What are Neural Networks? - India." *IBM*, 17 August 2020, https://www.ibm.com/in-en/cloud/learn/neural-networks. Accessed 16 July 2022.

"what is f-measure for each class in weka." *Stack Overflow*, 24 January 2014, https://stackoverflow.com/questions/21342449/what-is-f-measure-for-each-class-in-weka. Accessed 4 November 2022.

"- YouTube." - *YouTube*, 4 May 2022, https://www.kaggle.com/datasets/kaushiksuresh147/ipl-2020-data. Accessed 3 November 2022.