Football Prediction System

Akshay Gujjari (001372912)
Computer Science Department
University at Albany, SUNY
Albany, NY
agujjari@albany.edu

Abstract— **Predictive systems have been employed to predict events and results in virtually all walks of life. Football results prediction has gained popularity in recent years. In this paper, we develop a project that predicts the trends in the football, the tactics, strategies used by the teams and players in different situations like winning losing. We have visualized the data of all teams for the past 10 years. The visualization explains how a certain team plays and their preference to pass or run in any moment of the match. We can predict which players will have the highest impact on an individual game or play. This project predicts if the play will be a run or pass, who will be the receiver, and how many yards the play will be before the next pass or run. Overall, we have performed a complete analysis on the data of the football game that we collected from Kaggle website to predict, describe, and recommend the strategies, techniques.**

Keywords— **datamining, sports, football, visualization, analysis, prediction, winning, losing, tied, national football league**

I. INTRODUCTION

As we are aware that these days so much information is available that the challenge has become deciphering what is relevant. The Internet is updated frequently with information. Information about the stock market, databases of student results, telecommunication information, and weather records are a click away. The entertainment industry is no exception; for instance, football statistics flood the Internet every now and then. The English Premier League, in particular, produces a great deal of data because it is so popular. Fixed odds betting markets and researchers make use of these data to analyze and predict football match results. Different statistical techniques have been used to develop models for football match result prediction. Although some of these predictions have reasonable levels of accuracy, limitations remain and including the fact that some features affecting matches are not considered due to their complexity. Despite several studies in this area, there is room for improvement in terms of developing a system with higher prediction accuracy without much complexity.

Football is a team sport played by two teams of eleven players on a rectangular field with goalposts at each end. The offense, which is the team controlling the oval-shaped football, attempts to advance down the field by running with or passing the ball, while the defence, which is the team without control of the ball, aims to stop the offense's advance and aims to take control of the ball for themselves. The offense must advance at least ten yards in four downs or plays, and otherwise, they turn over the football to the defence; if the offense succeeds in advancing ten yards or more, they are given a new set of four downs. Points are primarily scored by advancing the ball into the opposing team's end zone for a touchdown or kicking the ball through the opponent's goalposts for a field goal. The team with the most points at the end of a game wins.

American football evolved in the United States, originating from the sports of association football and rugby football. American football is the most popular sport in the United States. The most popular forms of the game are professional and college football, with the other major levels being high school and youth football. As of 2012, nearly 1.1 million high school athletes and 70,000 college athletes play the sport in the United States annually, almost all of them men, with a few exceptions. The National Football League, the most popular American football league, has the highest average attendance of any professional sports league in the world; its championship game, the Super Bowl, ranks among the most-watched club sporting events in the world, and the league has an annual revenue of around US$10 billion. As we see that many organizations invest billions of dollars and this, in turn, provide huge employment and revenue, there is a high need for accurate prediction and analysis for teams and players.

There are 32 teams that compete each other to win the football tournament organised by the NFL. We ca see the 32 teams in the below pic with the symbols that represent their team.

We collected the dataset that is made available on Kaggle containing all the regular season plays from the 2009-2018 NFL seasons. The dataset has 356,768 rows and 100 columns. Each play is broken down into detail containing information on the game situation, players involved, results, and advanced metrics such as expected point and win probability values. After collecting the required data, we removed the unnecessary columns from the CSV file and performed the classification of the data into team wise and based on the situation in the game like losing, winning, and tied. We have visualized the data to understand the correlation between yardage and play type. We have predicted the frequent patterns in the playing style of the team, such as common Quarterback and receiver combinations using association rule mining. We also predicted a play will be pass or run if it is passed then which player is a receiver in any given situation of the match. Performing such analysis on the data and getting an understanding from the trends in the data helps the teams, players, organizations to grade performance of teams and players. These kinds of analytics help the teams to predict the playing style of the opponents and to develop some new techniques to surprise other teams.

II. LITERATURE REVIEW

Past research focusing on some of the prediction algorithms to predict football results for different results has been conducted. Kou-Yuan Huang and Wen-Lung Chang used neural for prediction of 2006 world cup football game. They developed a project that used Bayesian Network with other machine learning algorithms like decision tree learner and K-nearest neighbor to predict the results of the games. In this paper, they tried to predict the accuracy of the teams that wins the 2006 world cup. They also came up with 3 outcomes for each match they predicted like whether a team will win, lose or draw a match. In this paper, they proved that the Bayesian Network algorithm has superiority when compared to other algorithms in terms of accuracy in predicting the outcomes of an event.

J. Hucaljuk and A. Rakipovic have developed a project to predict football scores using machine learning algorithms. They used Bayesian Network along with KNN, Random Forest, Artificial Neural Network to predict the results of Europe Champions League. Instead of simply developing a project they thought creatively and came up with a software solution to solve the problem of predicting the results of the games. We know that the algorithm that we choose for classification and feature selection. As feature selection describes about the various important features like current form of teams shown on the basis of results achieved in the last six games, the outcome of the previous meeting of the teams that play the game, the current position in the rankings, number of injured players from the first team, the average number of scored and received goals per game, which affect a football match are been chosen. After choosing the correct features, among the ones mentioned above is selected. During the development of the system, a number of tests have been carried out in order to determine the optimal combination of features and classifiers.

A neural network method is adopted to predict the football game's winning rate of two teams according to their previous stage's official statistical data of 2006 World Cup Football Game. The input data is transformed into the relative ratios between two teams of each game. New training samples are added to the training samples at the previous stages. The adopted prediction model is based on multi-layer perceptron with back propagation learning rule. They input the average data of each team into the well-trained MLP, and then they compare the output value to determine the relationship and defeat. with bigger output value, which means the more ability to win the game, is the winner. This paper applies the Artificial Neural Network to the official statistical data of the 2006 World Cup and based on the adopted MLP prediction method.

There are various other papers that use computing techniques like Fuzzy Logic, Artificial Neural Network, and Genetic Programming to perform the predicting the outcomes of an event.

III. METHOD

When we decided to do the project on the football prediction system, one of the main challenges we faced is acquiring the raw data. We did a lot of research to acquire the data the exactly satisfies the motto of our project. So finally, we found the data on the Kaggle website that closely matches our requirements. The dataset which we collected has three hundred and fifty thousand rows of data with a hundred columns. The dataset was in the comma-separated values format.

Firstly, after collecting the raw data we have classified the data into team wise. To classify the data, we have understood the format of the data and removed the redundant columns. We have performed data pre-processing and extracted the data into folders each name with the name of the team. In the folder of the team, there are 3 more folders that contain the data in JSON format. These three folders contain data of the teams where each team is losing, winning and tied.

After performing the data pre-processing, we have visualized the data of the teams in situations like losing, winning, tied. Visualization helps us to get a clear picture of the trends in the data. In our project, visualization helped to understand the techniques like pass or runs used by the teams in different scenarios like winning, losing, and tied. This helped us to understand the preference of the teams to make a pass or run with the ball.
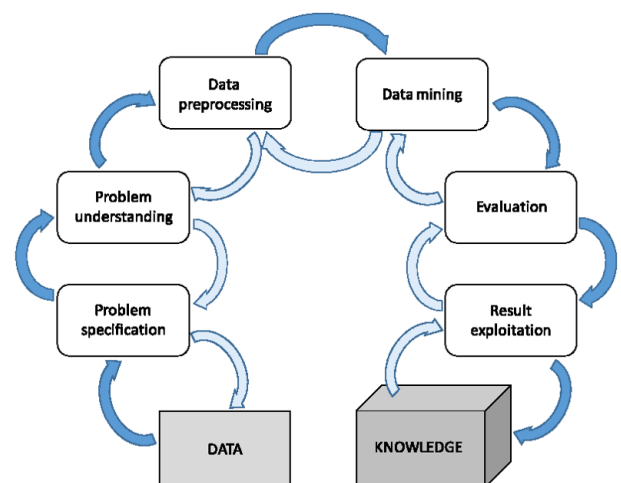
Association rule prediction is a concept that comes after visualization in our project. We have used the association rule to predict the frequent patterns in the data. When we talk about the frequent patterns regarding football it can say as Quarterback and receiver combination. This combination is the key for any team, and this helps the teams to score more points. Using this prediction, we found all the Quarterback and receiver combination of all the teams.

Markov model is used to predict if a certain play will be a pass or run. If this is a run, then who will be the runner if it is a pass then who will be the receiver. Analysing all the combinations helps the teams to know how the other teams' player in any moment of the game. Hidden Markov Model is used to find a play will be pass or run and how many yards a certain play will be played. These are the methods that we used to determine and understand the playing style of players and teams. This analysis helped us to predict the future outcomes of the events.

IV. IMPLEMENTATION

The implementation of the project involves a series of steps that we have followed from data collection to prediction of event outcomes. Once we have classified the data into different categories, we made sure that we need to follow certain steps to achieve a successful prediction of the data. The series of steps that we followed are:

1. Data Collection
2. Classification
3. Visualization
4. Association Rule Prediction
5. Markov Model Prediction
6. Hidden Markov Model Prediction

Data Collection: When we are developing a project that deals with the prediction of the outcomes of an event, there are a lot of parameters that need to be taken into consideration. So for this project, we concerned with the accurate acquisition of data. We as a team believed that selection of appropriate data collection tools and instruments, which may be existing, modified or totally new, and with clearly defined instructions for their proper use, reduces the chances of errors occurring during collection. So we could able to collect data from Kaggle where a group of Carnegie Mellon University statistical researchers including Maksim Horowitz, Ron Yurko, and Sam Ventura, built and released clean datasets at the individual play, player, game, and season levels. The trio went on to develop reproducible methods for building expected point and win probability models for the NFL. We have used this dataset as we this closely matched our requirements and we could able to derive the expected output.

Classification: Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. Data classification is the process of sorting and categorizing data into various types, forms or any other distinct class. Data classification enables the separation and classification of data according to data set requirements for various business or personal objectives. It is mainly a data management process. For this project, we have classified the data into team wise folders and also depending on the situation of the game which is losing, winning, tied.


Visualization: Visualization is the use of computer graphics to create visual images which aid in the understanding of complex, often massive representations of data. Visual Data Mining is the process of discovering implicit but useful knowledge from large data sets using visualization techniques. In particular, visualization may be used for outlier detection, which highlights surprises in the data, i.e. data instances that do not comply with the general behavior or model of the data. Through the visualization of the original data, the user can browse to get a feel for the properties of that data. We have visualized the number of passes and runs a certain team has made depending on the number of yards to the goal post, in every match. From this visualization, we able to understand the playing and preferences of the teams like making pass or runs depending on the yards.

Association Rule Prediction: Association rule prediction involves the use of machine learning models to analyze data for patterns, or co-occurrence, in datasets. Association rules are if-then statements that help to show the probability of relationships between data items within large data sets in various types of databases. Association rule prediction has a number of applications and is widely used to help discover correlations in the data sets. We used Association rule prediction to predict whom a passer will throw the ball. So we found the passer, receiver correlations in different

matches a team has played. We have also taken a lot of other features into consideration to get more accurate results. We have also predicted what side of the field a QB is most likely to throw to based on other features.
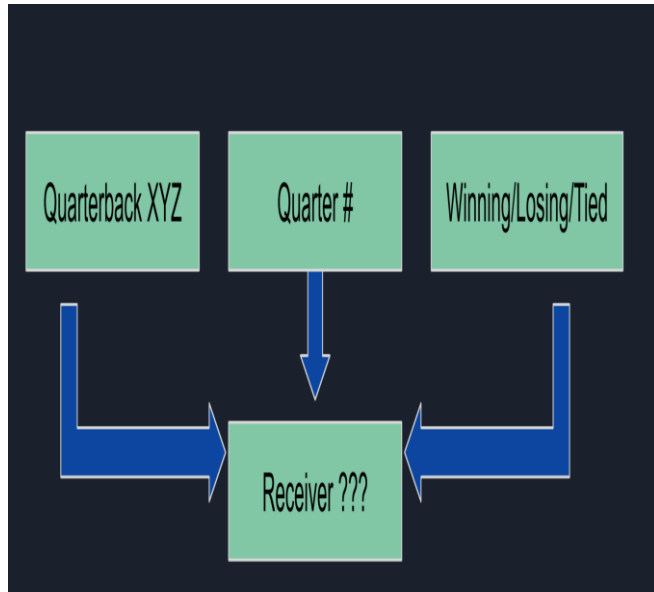
Markov Model Prediction: The Markov Model is a statistical model that can be used in predictive analytics that relies heavily on probability theory. In probability theory, a Markov model is a stochastic model used to model randomly changing systems.It is assumed that future states depend only on the current state, not on the events that occurred before that event. Generally, this assumption enables reasoning and computation with the model that would otherwise be intractable. For this reason, in the fields of predictive modelling and probabilistic forecasting, it is desirable for a given model to exhibit the Markov property. Markov models are particularly useful to predict a wide variety of behavior such as outcomes of an event, network formation. In this project we used Markov Model to predict who the ball will be thrown to based on a prior sequence of plays.

Written as a formula, the Markov Assumption looks like this:

$$P\left(event_t | event_{t-1}, event_2, \ldots, event_n\right) \approx P\left(event_t | event_{t-1}\right)$$
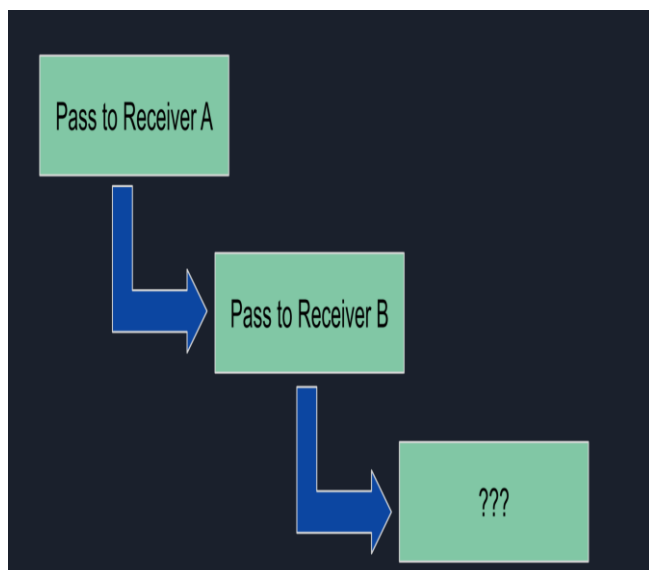
A second order Markov prediction includes just the last two events that happen in sequence. From the equation just given, the following widely used equation can also be derived:

$$\left(event_1, event_2, \ldots, event_n\right) = \prod_n^2 P\left(event_t | event_{t-1}\right)$$

Hidden Markov Model Prediction: Hidden Markov Model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved states. Hidden Markov models are especially known for their application in reinforcement learning and prediction. The hidden Markov model can be represented as the simplest dynamic Bayesian network. A hidden Markov model can be considered a generalization of a mixture model where the hidden variables, which control the mixture component to be selected for each observation, are related through a Markov process rather than independent of each other. We predict who the ball will be thrown to based on a prior sequence of plays as well as the length of the play. This prediction helps the teams to know how the opponents strategy the player who is most probable receiver.
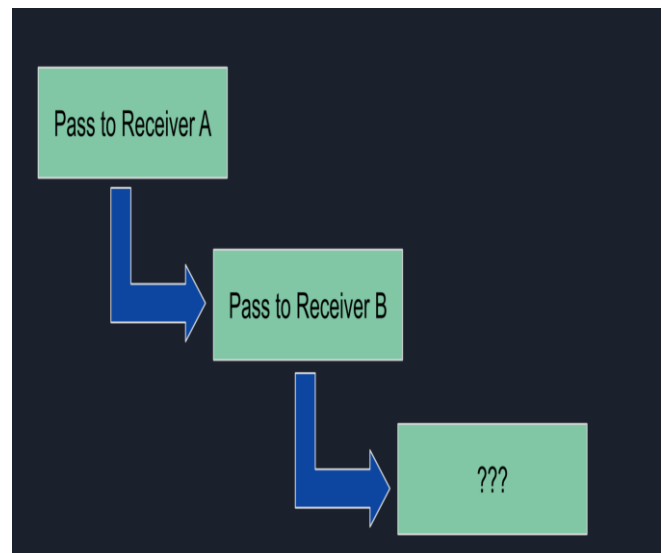
In the Hidden Markov Model, the probability of observing a sequence is explained as:

$$Y = y(0), y(1), \ldots, y(L-1)$$

of length L is given by

$$P(Y) = \sum_X P(Y \mid X) P(X),$$

where the sum runs over all possible hidden-node sequences

$$X = x(0), x(1), \ldots, x(L-1).$$

We have followed a step by step procedure to implement this project. We as a team decided motto of our project and decided the methodology, implementation which made us come with successful results.

V. RESULTS AND ANALYSIS

Data Preprocessing: The data which we collected to implement this project contains has 356,768 rows and 100 columns. Each play is broken down into detail containing information on the game situation, players involved, results, and advanced metrics such as expected point and win probability values. In the raw data which we collected has a lot of redundant data, so we had to remove the columns which are not required for our project. The below pic shows how the raw data is stored in the file.

We then implemented a scrapper python file to perform the pre-processing on the raw data and transform into JSON format and stored this data in according to teams, quarters, situation of the match with respect to that team.
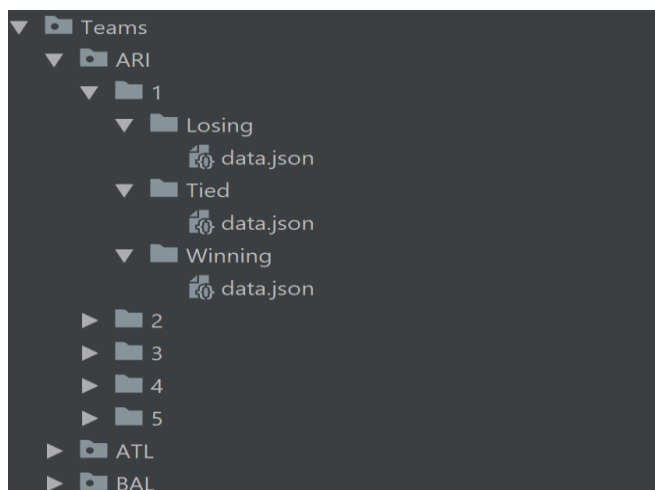
Visualization: After performing the data preprocessing, we then implemented a Data Visualization python which visualizes the data of all the teams like the number of passes, runs in a certain quarter.

After performing on the data of the team Green Bay Packers, ee can see in the below plot that the team Green Bay Packers have heavily rely on the passes when they were

losing in the fourth quarter as they are approaching the Red Zone of the game.
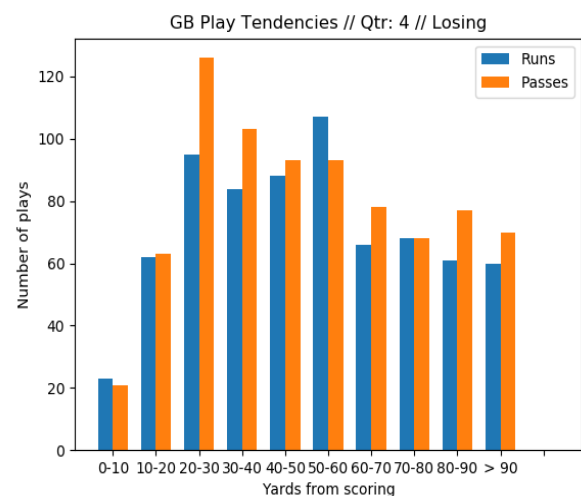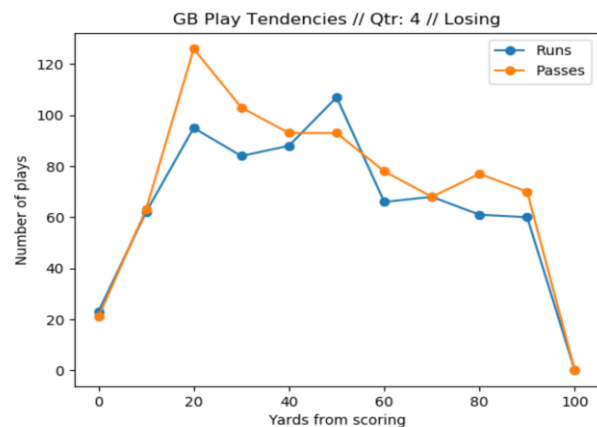
Lets us see the visualized data of another team Baltimore. Based on data from quarter 4 when Baltimore is tied, we can see Baltimore is a run-heavy team in their own half of the field, and then become a pass heavy team on the opposing side of the field.
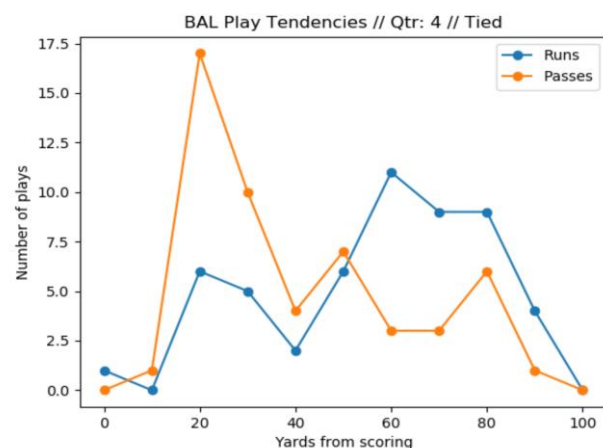
Association Rule Prediction: With the help of association rule mining, we can see that for the team Jacksonville Jaguars in the 1st Quarter when losing (left), Blake Bortles

has a different set of receivers he prefers to throw to than when the game is in the 2nd Quarter when losing (right).

```
Supp  Conf   Rule

 0.2   1.0   M.Lee -> B.Bortles

 0.2   0.2   B.Bortles -> M.Lee

 0.1   0.1   B.Bortles -> D.Moncrief

 0.1   0.9   D.Moncrief -> B.Bortles
```

```
Supp  Conf   Rule

 0.1   0.9   D.Westbrook -> B.Bortles

 0.1   0.1   B.Bortles -> D.Westbrook

 0.1   1.0   T.Yeldon -> B.Bortles

 0.1   0.1   B.Bortles -> T.Yeldon

 0.1   1.0   M.Lee -> B.Bortles

 0.1   0.1   B.Bortles -> M.Lee
```

Markov Model Prediction: If we consider the data of team Arizona Cardinals, the based on the data from the previous 3 plays (Run, Pass to L. Fitzgerald, Passto L.Fitzgerald), the next three plays we predict will be a Run, a pass to L.Fitzgerald, and a Run. We can see in the below pictures that based on the input given to model it will display all the player names involved with that play and based on the sequence that we give to the model, it will predict which player will play the next three moves.

The below picture says markov model prediction for team Indianapolis Colts.

Hidden Markov Model Prediction: Using the Hidden Markov Model, we can predict that based on the previous three plays who will be the next pass and how much will the pass and if it's a run how many yards will the run be in any moment of the game.

In the below pic, for the team Chicago Bears, based on the previous three plays that is a gain of 10 yards and loss of 2 yards and then again a gain of 4 yards, the model has predicted that the next will be a pass to Cohen and then it will a run and a pass to R. Williams.
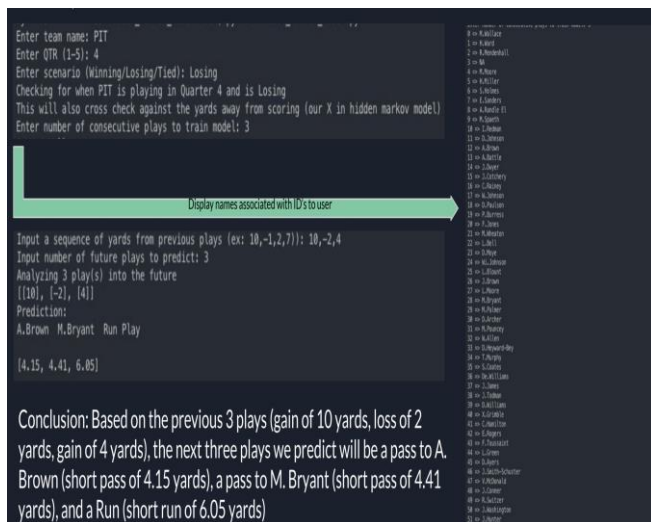

6

VI. CONTRIBUTION

When we decided that we need to project we had the planned to do a project that deals with a huge amount of data. We also wanted to do a creative, interesting, and challenging project. So we as a team decided to contribute individually with the same kind of mindset so that er all have the same ideas for the development of the project. We divided the project into multiple tasks so that each and everyone put in the same amount of efforts towards the project. We wanted to gain a good knowledge of the Data Mining Concepts.

I have contributed to the data collection, data preprocessing. After we are done with data preprocessing we have I have dealt with some statistical analysis to our data which helps us for the successful implementation of the project. I have handled the total visualization for our project. I explained my team members like why visualization is important, how it is helpful and how we can get insights from the data using visualization. Then we decided on how exactly we need to used visualization for our project and then finally we came up with a good solution. We have also created a GitHub repository for our project to track our task and also get updates from our team members. I am very much happy with what I contributed and also with the contribution of my team members. I have gained hands-on skills in python programming and got in-depth knowledge of Data Mining Concepts.

VII. CONCLUSION

The improved football result Prediction System explores the use of Data Mining techniques in the framework of Knowledge Discovery in Database. This research is driven by the overwhelming increase in the pool of available sports data in the National Football League. The datasets collected was successfully implemented using data mining technique in different aspects of the work.

This project is very much helpful to the teams, companies, players. Since Football has a lot of scope in America and many invest huge amounts of money in Football, there needs an accurate analysis of all the trends in football. Using the Data Mining we can know some surprising results, playing style and many other features.

In many instances, predicting the results of sporting procedures has always been a challenging and rewarding venture, therefore forecasting problem provides a growing need to conduct research in this area. Sports outcomes predictive techniques arise and this motivates the need to find more valuable datasets to improve the prediction accuracy and make precise decisions at key. Past comprehensive statistical data has been kept to Football games and other sporting events. Both players and teams' present varying forms of these statistical facts kept as data season in and off-season. As the dataset set grows with the Football games, it has become the preferred test platform. This pool of information will keep motivating different groups, ranging from public, statisticians and sports enthusiasts to discover embedded knowledge in it. This project also provides an ideal data mining environment, data mart containing comprehensive Football game information, which can be reused by future research. Finally, this project can be considered as a successful exploration of using Knowledge Discovery in Database for sports result prediction and it provides a good mainstay for future research works. We can predict the future outcomes of the games using the current data. We can see into the future using the current data available.

VIII. ACKNOWLEDGEMENT

IX. REFERENCES

[1] Kou-Yuan Huang and Wen-Lung Chang. A neural network method for prediction of 2006 world cup football game. In The 2010 International Joint Conference on Neural Networks (IJCNN), pages 1 –8, july 2010.

[2] J. Hucaljuk and A. Rakipovic. Predicting football scores using machine learning techniques. In MIPRO, 2011 Proceedings of the 34th International Convention, pages 1623 –1627, may 2011.

[3] Kou-Yuan Huang and Wen-Lung4 Chang. A neural network method for prediction of 2006 world cup football game. In The 2010 International Joint Conference on Neural Networks (IJCNN), pages 1 –8, july 2010

[4] Gangal, Anurag & Talnikar, Abhishek & Dalvi, Aneesh & Zope, Vidya & Kulkarni, Aadesh. (2015). Analysis and Prediction of Football Statistics using Data Mining Techniques. International Journal of Computer Applications. 132. 8-11. 10.5120/ijca2015907263.

[5] En.wikipedia.org. (2019). Hidden Markov model. [online] Available at: https://en.wikipedia.org/wiki/Hidden_Markov_model [Accessed 17 May 2019].

[6] Pdfs.semanticscholar.org. (2019). [online] Available at: https://pdfs.semanticscholar.org/0cb4/75365178811cc63cff286a52efade065d51f.pdf [Accessed 17 May 2019].

[7] Kaggle.com. (2019). Detailed NFL Play-by-Play Data 2009-2018. [online] Available at: https://www.kaggle.com/maxhorowitz/nflplaybyplay2009to2016 [Accessed 17 May 2019].

[8] En.wikipedia.org. (2019). Markov model. [online] Available at: https://en.wikipedia.org/wiki/Markov_model [Accessed 17 May 2019