

|  |
| --- |
| **Election Prediction System** |
| **By** |
| **Arsh Imtiyaz Assadi** |
| **(4NM19CS032)**  **Chris Xavier Mathais**  **(4NM19CS046)** |
| **Bachelor of Engineering** |
| **in** |
| **COMPUTER SCIENCE AND ENGINEERING** |
| **Under the guidance of** |
| **Dr. Aravinda C. V** |
| **Associate Professor** |
| ***Mini Project – Report Submitted to NMAM Institute of Technology, Nitte an Autonomous Institution affiliated to VTU Belagavi*** |
| **DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING** |



ISO 9001:2015 Certified

Accredited with ‘A’ Grade by NAAC

*Department of* ***Computer Science & Engineering***

CERTIFICATE

*Certified that the project work entitled* ***Election Prediction System*** *carried out by* ***Arsh Imtiyaz Assadi.,*** *USN* ***4NM19CS032*** *and* ***Chris Xavier Mathias*** *USN* ***4NM19CS046 bonafide*** *students of* ***NMAM Institute of Technology****,* ***Nitte*** *has been carried out satisfactorily. The Mini Project report has been prepared as per the prescribed format.*

**Name & Signature of Guide(s) Name & Signature of HOD**

### Project – Part I Evaluation

**Name of the Examiners Signature with Date**

*1.*

*2.*

# ACKNOWLEDGEMENT

The success and final outcome of this project required a lot of guidance and assistance from many people, and I am extremely fortunate to have got their support all along with the completion of our project.

I take this opportunity to express my profound gratitude and deep regards to my Project Guide **Dr. Aravinda C V.,** Associate Professor, Department of Computer Science and Engineering, for his exemplary guidance, monitoring, and constant encouragement throughout the course of this project.

I also express a deep sense of gratitude to **Dr. Jyothi Shetty.,** Head of the Department, of Computer Science and Engineering, for her cordial support, valuable information, and guidance, which helped me in completing this project through various stages.

Lastly, I thank almighty, my parents and friends for their constant encouragement.

- Arsh Imtiyaz Assadi (4NM19CS032)

Chris Xavier Mathias (4NM19CS046)

# ABSTRACT

**Motivation:** Election forecasting appeals to a basic human urge to peek into the future. Ever since elections were invented to choose leaders, humans have been tempted to find ways that would tell them with some degree of certainty who would win an election. The highly quantitative nature of elections aids them in such an endeavor. Elections are highly chaotic and volatile and it’s important to find the factors of the victor or the defeated.

**Goal:** To determine certain factors and predict if the person will win or lose with the data given and find the analysis of how the factors affect the results of the election.

**Results:** To predict the results of the election we have to determine certain factors that would affect the result of the election. The determined factors are belonging party, age of the candidate, the number of criminal cases of the past, gender of the candidate and the educational qualification. The following project aims to analyze the given factors and produce the results on form of effect and the quantity of each factor’s dependency on the prediction.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Chapter No. Title** | | **Page**  **No.** |
|  | **ABSTRACT** |  |
| **1** | **INTRODUCTION** | **1** |
| **2** | **LITERATURE SURVEY** | **2-3** |
| **3** | **METHODOLOGY** | **4-7** |
| **4** | **RESULTS AND DISCUSSION** | **8-13** |
|  | **REFERENCES** |  |
|  |  |  |

1. **INTRODUCTION**

An election is a formal [group decision-making](https://en.wikipedia.org/wiki/Group_decision-making) process by which a population chooses an individual or multiple individuals to hold [public office](https://en.wikipedia.org/wiki/Public_administration). Elections have been the usual mechanism by which modern [representative democracy](https://en.wikipedia.org/wiki/Representative_democracy) has operated since the 17th century. Elections may fill offices in the [legislature](https://en.wikipedia.org/wiki/Legislature), sometimes in the [executive](https://en.wikipedia.org/wiki/Executive_(government)) and [judiciary](https://en.wikipedia.org/wiki/Judiciary), and for [regional and local government](https://en.wikipedia.org/wiki/Local_government). This process is also used in many other private and [business](https://en.wikipedia.org/wiki/Business) organisations, from clubs to [voluntary associations](https://en.wikipedia.org/wiki/Voluntary_association) and [corporations](https://en.wikipedia.org/wiki/Corporation).

When we talk about Election prediction, it’s a subset of political forecasting. Political forecasting aims at [forecasting](https://en.wikipedia.org/wiki/Forecasting) the outcomes of political events. Political events can be a number of events such as diplomatic decisions, actions by political leaders and other areas relating to politicians and political institutions. The area of political forecasting concerning elections is highly popular, especially amongst mass market audiences. Political forecasting methodology makes frequent use of mathematics, statistics and data science. Political forecasting as it pertains to elections is related to [psephology](https://en.wikipedia.org/wiki/Psephology).

In this project, the data-set of the goa elections have been used to analyse the factors pertaining to a person’s eligibility of winning an election. We have considered multiple factors which could reason out to be the attributes which can lead to a candidate assuring a victory in the present election or a defeat. These factors are chosen based on the general attribute a candidate has for example his education, gender etc.

This project will help us give a general idea on the attributes which could lead to defeat or victory of a can.

**2.LITERATURE SURVEY**

Several Machine Learning techniques used in Election Forecasting have been published in literatures. It is obvious from our study that in the bid to apply machine learning to solve the Election Prediction problem, different learning algorithms are proposed each time thereby adding to the ever-expanding pool of machine learning algorithms for predicting the results of an election.

Some studies have been devoted to evaluating the effectiveness of prediction under similar situations using certain benchmarks. Such studies help us to identify the techniques that have superior performance under some circumstances.

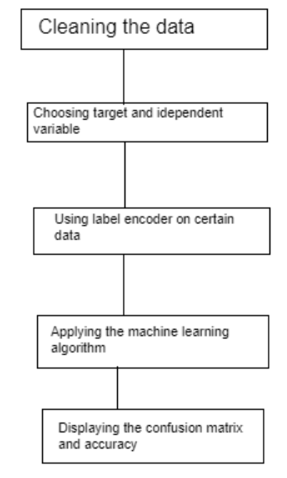
Random Forest is an ensemble learning method used for classification and regression. Developed by [Breiman (2001)](https://www.tandfonline.com/doi/full/10.1080/21642583.2014.956265), the method combines Breiman's bagging sampling approach (1996a), and the random selection of features, introduced independently by [Ho (1995)](https://www.tandfonline.com/doi/full/10.1080/21642583.2014.956265); [Ho (1998)](https://www.tandfonline.com/doi/full/10.1080/21642583.2014.956265) and [Amit and Geman (1997)](https://www.tandfonline.com/doi/full/10.1080/21642583.2014.956265), in order to construct a collection of decision trees with controlled variation. Using bagging, each decision tree in the ensemble is constructed using a sample with replacement from the training data. Statistically, the sample is likely to have about 64% of instances appearing at least once in the sample. Instances in the sample are referred to as in-bag instances, and the remaining instances (about 36%) are referred to as out-of-bag instances. Each tree in the ensemble acts as a base classifier to determine the class label of an unlabeled instance. This is done via majority voting where each classifier casts one vote for its predicted class label, then the class label with the most votes is used to classify the instance.

Qing-yun Dai, Chun-ping Zhang and Hao Wu researched that Decision tree algorithm is one of the most important classification measures in data mining. Decision tree classifier as one type of classifier is a flow- chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node represents a class.

The method that a decision tree model is used to classify a record is to find a path that from root to leaf by measuring the attributes test, and the attribute on the leaf is classification result.

K-Nearest Neighbour classifier or popularly known as KNN classifier is used to classify unlabeled observations by assigning them to the class of the most similar labeled examples. Characteristics of observations are collected for both training and test dataset. For example, fruit, vegetable and grain can be distinguished by their crunchiness and sweetness. For the purpose of displaying them on a two-dimension plot, only two characteristics are employed. In reality, there can be any number of predictors, and the example can be extended to incorporate any number of characteristics. In general, fruits are sweeter than vegetables. Grains are neither crunchy nor sweet. Our work is to determine which category does the sweet potato belong to. In this example we choose four nearest kinds of food, they are apple, green bean, lettuce, and corn. Because the vegetable wins the most votes, sweet potato is assigned to the class of vegetable. You can see that the key concept of KNN classifier is easy to understand.

**3. METHODOLOGY**



**Fig 3.1**. Block diagram of Proposed method

The following figure involves the following steps:

1.The first step to the any project is forming the problem statement.

2. Here, we collected our data from the Goa Election through the Election Commission of India website. We visualized the data using various graphing methods in order to find the relationship between various variables.

3. Choosing the right factors that could be responsible for the prediction of the election which is a quintessential process while performing prediction. That is choosing the target and the independent variables.

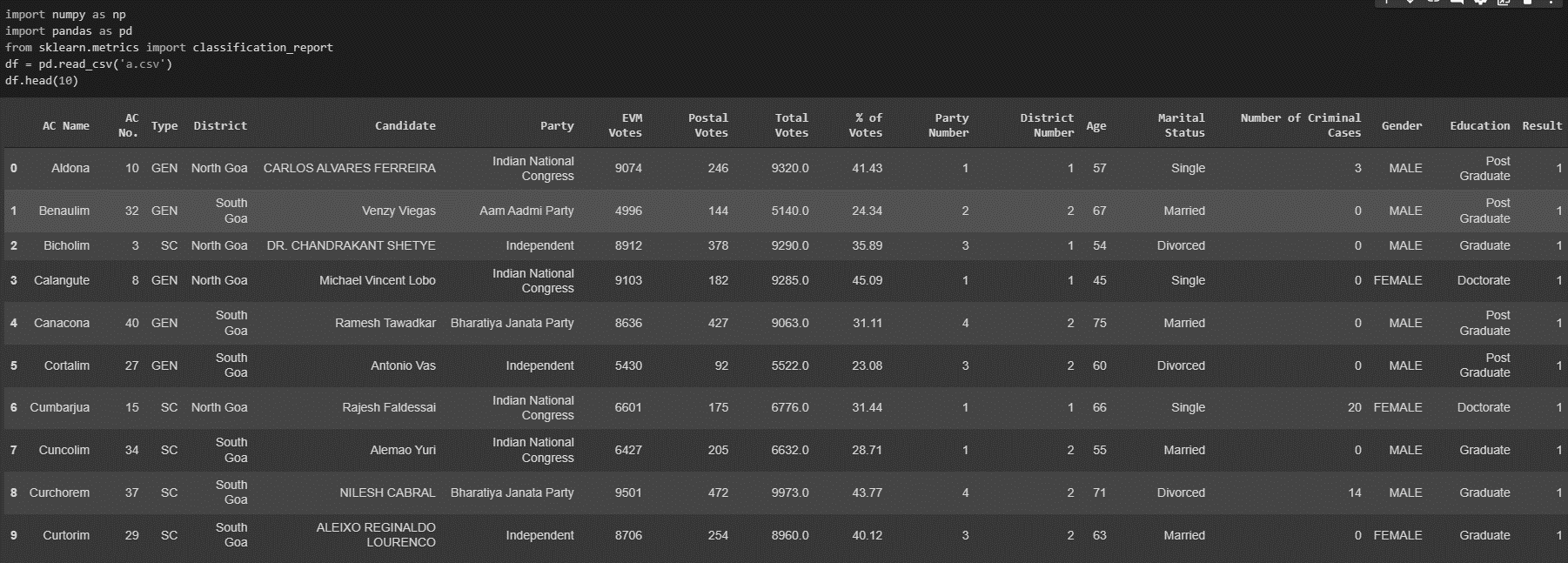
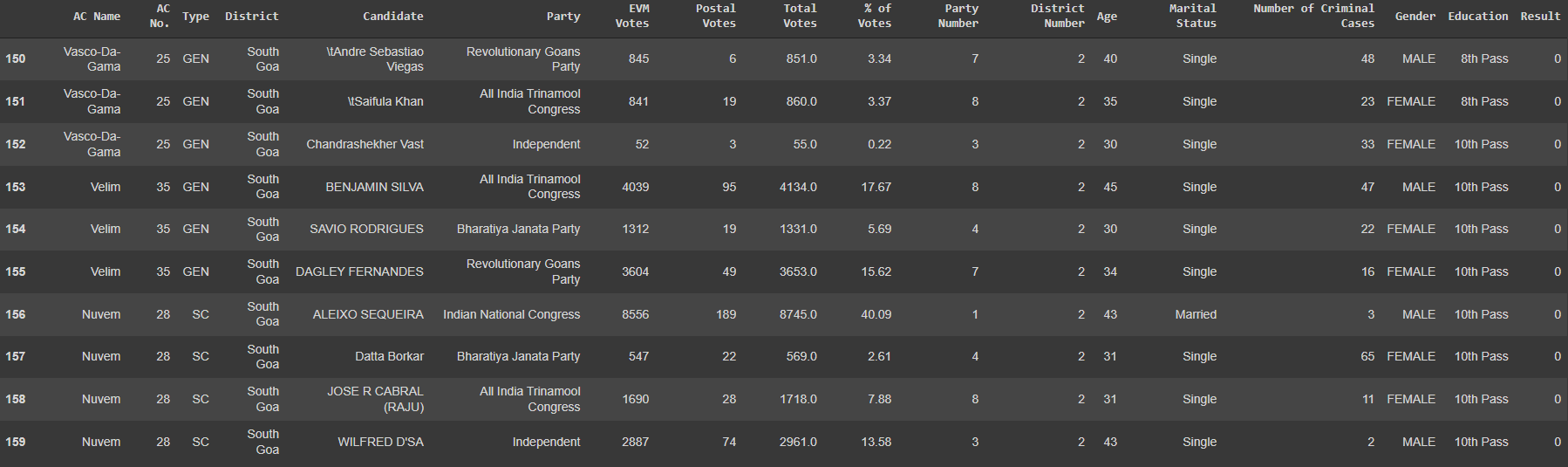
4. Since some data are categorical values, we apply One Hot Encoding. One hot encoding is a crucial part of feature engineering for machine learning.

5. Applying three machine learning algorithms that is Random Forrest classifier, Decision Tree Classifier and K-Nearest Neighbhour Classifier for prediction.

6. Comparing accuracy and the confusion matrix for each machine learning model.

## Dataset Collection

The dataset of 159 values were collected from the election commission website. The value that were taken down are party, number of votes, result of that constituency, type of candidate, candidate name, percentage of votes, district and we arbitrarily added certain data to ease the process of the project. The arbitrary data added was Marital Status, Gender, Education level and Number of Criminal Cases.



## Machine Learning Algorithms

1. Random Forest Classifier: Random Forest is a comprehensive machine learning (ML) classification approach. An ensemble of decision trees enables RF. Each tree separately anticipates a classification and “votes” for the necessary parameters, with the majority of the votes determining the entire RF predictions. Use of multiple trees reduce the risk of overfitting.
2. Decision Tree (ID3): ID3 stands for Iterative Dichotomiser 3 and is named such because the algorithm iteratively (repeatedly) dichotomizes(divides) features into two or more groups at each step. Invented by [Ross Quinlan](https://en.wikipedia.org/wiki/Ross_Quinlan), ID3 uses a **top-down greedy** approach to build a decision tree. In simple words, the **top-down** approach means that we start building the tree from the top and the **greedy**approach means that at each iteration we select the best feature at the present moment to create a node. Most generally ID3 is only used for classification problems with nominal features only.
3. K-Nearest Neighbors (KNN): The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It’s easy to implement and understand, but has a major drawback of becoming significantly slows as the size of that data in use grows. KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression).

## Hardware and Software used

All the Machine Learning models were parallelly trained and tested on Windows 10 and Ubuntu operating system with an 8th generation Intel i5 processor and 8GB of RAM. The models were developed on Jupyter Notebook using Python 3.7.9 including other libraries like Numpy, Pandas, Matplotlib, Sklearn, etc.

### Libraries used:

**NumPy:** It stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed.

**Pandas:** Itis a software library written for the Python programming language for data manipulation and analysis.

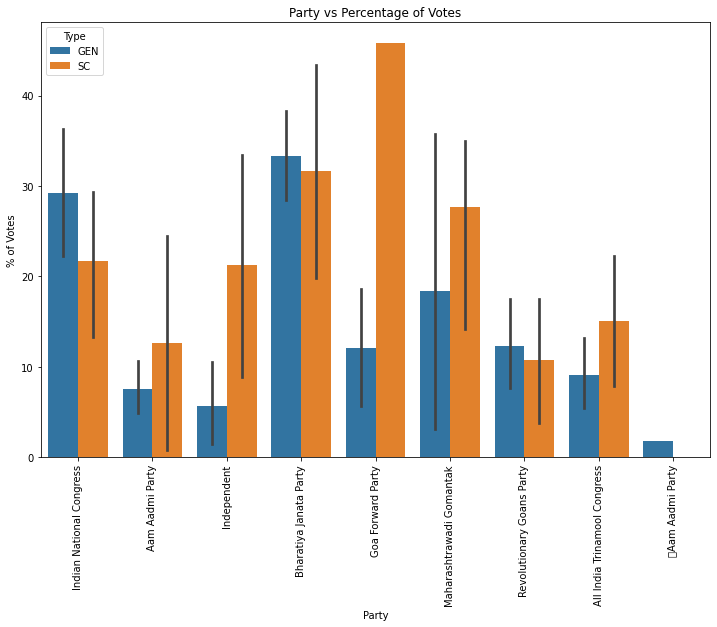
**Seaborn:** Itis a library mostly used for statistical plotting in Python. It is built on top of Matplotlib and provides beautiful default styles and color palettes to make statistical plots more attractive.

**Plotly:** Itis python graphing library makes interactive, publication-quality graphs.

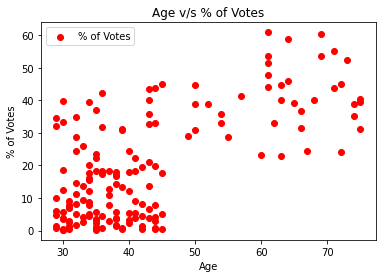
**Scikit-learn (Sklearn):** Itis the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

# RESULTS AND DISCUSSION

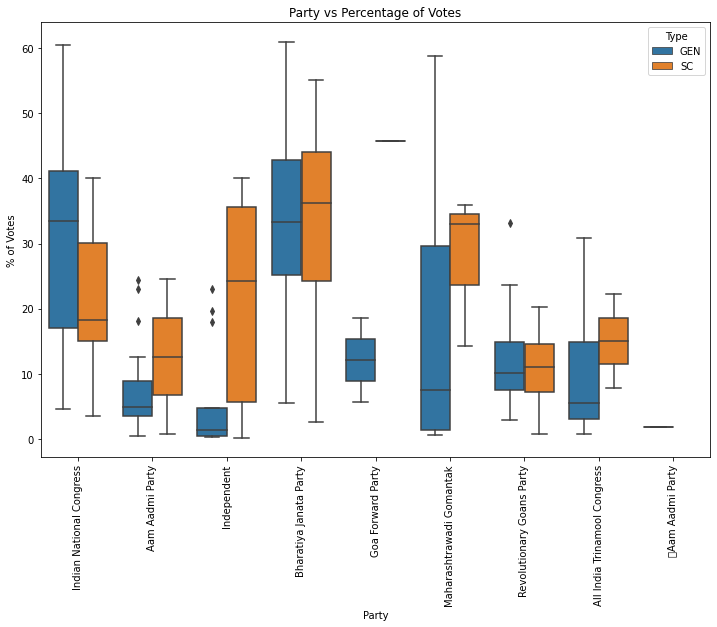
## Data Analytics



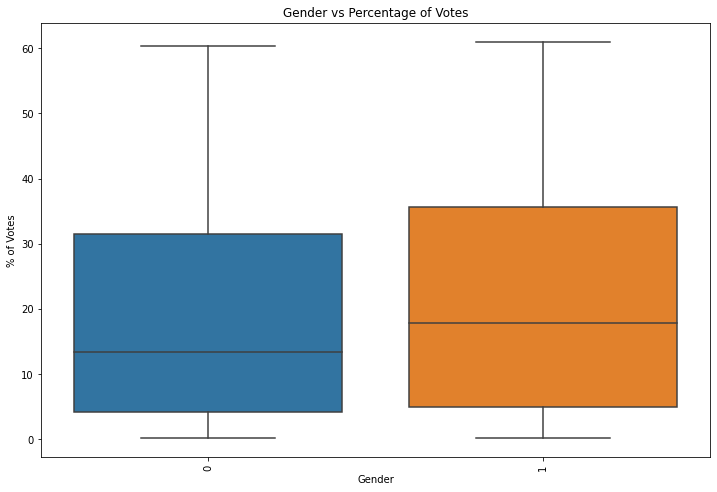
The above graph shows the percentage of votes between General Category and Scheduled Caste(SC) category. The graph is also for party vs the percentage of votes.



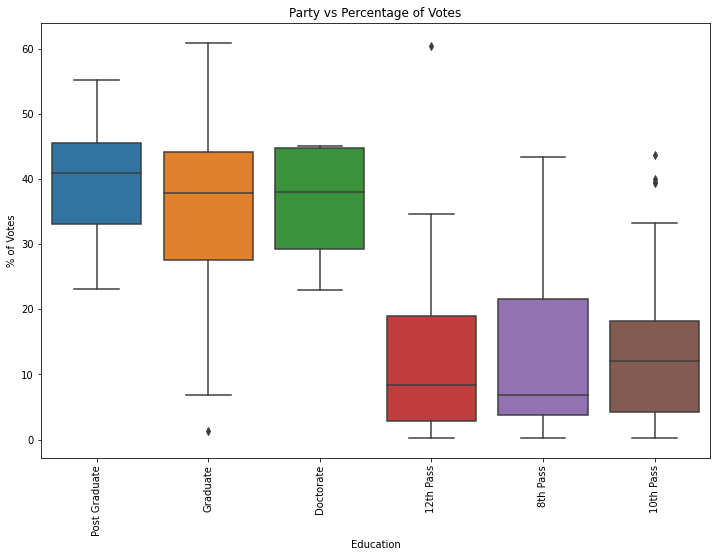
The above graph shows the age vs the percentage of votes. We can infer from the above graph that people of higher age get a greater number of votes and vice versa.



The above graph shows party vs percentage of votes for General Category and Schedule Caste candidates. The mid-line shows the mean of the votes.



The above graph shows the Gender vs Percentage of votes where mid-line is for the mean of the votes where 0 correlates to females and 1 to males. From this graph, we can infer that male get more amount of votes.



The above graph is for party vs percentage of votes for each education category. We can infer that higher the education a greater number of votes.

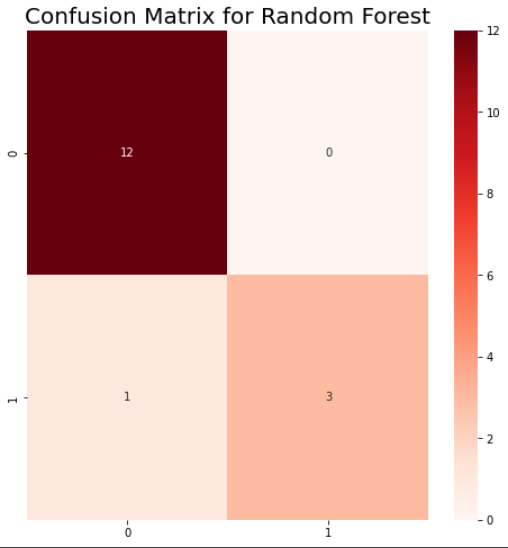
## Classification Report:

## Random Forest Classifier

## The Following shows the testing and training accuracy of the random forest classifier.

## 

## The following shows the confusion matrix of the random forest classifier:



## Decision Tree Classifier

## The Following shows the testing and training accuracy of the Decision Tree Classifier

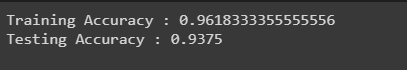
## 

## The following shows the confusion matrix of the Decision Tree Classifier:

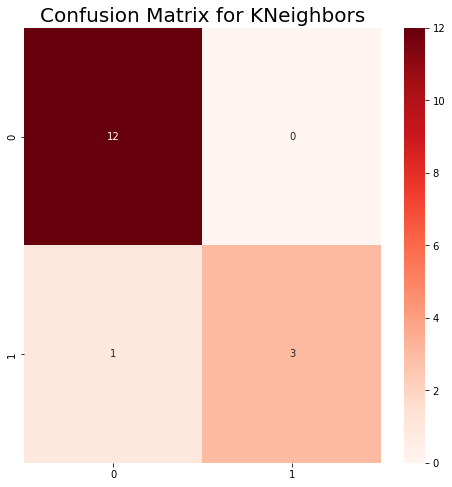
## 

## K-Nearest Neighbour Classifier

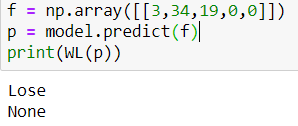
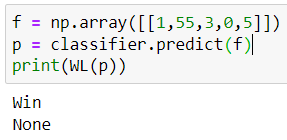
## The Following shows the testing and training accuracy of the Decision Tree Classifier



## The following shows the confusion matrix of the Decision Tree Classifier:



Clearly, using KNN Classifier or Decision Tree will give us a more prompt result depending on the Training accuracy. On passing the parameters to the model, the model can give either of the two results win or lose.

# REFERENCES

1. https://www.oxfordbibliographies.com/view/document/obo-9780199756223/obo-9780199756223-0023.xml
2. https://www.tandfonline.com/doi/full/10.1080/21642583.2014.956265#:~:text=4.-,Random%20forest,(2001).
3. https://towardsdatascience.com/getting-acquainted-with-k-nearest-neighbors-ba0a9ecf354f
4. An Introduction to Kernel and Nearest Neighbor Nonparametric Regression by N.S. Altman
5. <https://scikit-learn.org/stable/>
6. <https://matplotlib.org/stable/users/index>
7. “Election Forecasting, Scientific Approaches”, by [Michael Lewis-Beck](https://www.researchgate.net/profile/Michael-Lewis-Beck), Mary Stegmaier