**ANALYSIS AND PREDICTIVE MODELLING OF KERALA ASSEMBLY ELECTIONS DATASET, USING PYTHON**

*A project report submitted to ICT Academy of Kerala*

*in partial fulfillment of the requirements*

*for the certification of*

**CERTIFIED SPECIALIST**

**IN**

**DATA SCIENCE & ANALYTICS**

submitted by

**Athira K**

**Hari Krishna S**

**Rakhi Raj R**

**Sethulakshmi V A**

**Sofia John**



**ICT ACADEMY OF KERALA**

**THIRUVANANTHAPURAM, KERALA, INDIA**

**Oct 2022**

**List of Figures**

##### Fig No Figure Name

7.1 The null values present in the dataset

7.2 Number of participation of candidates on the basis of gender

7.3 Number of candidates of each parties

7.4 The education of the candidates

7.5 Age Distribution

7.6 Criminal cases of parties

7.7 The gender base visualization number of criminal cases

7.8 The number of criminal cases of candidates of each district.

7.9 The criminal cases on the basis of education of the candidates

7.10 The winning percentage of first eight parties

7.11 The participation of first 8 parties in the election

8.1 Box plot of the dataset

8.2 Correlation between the features.

**List of Abbreviations**

ML Machine learning

np numpy

pd pandas

plt matplotlib.pyplot

sns seaborn

**Table of Contents**

**1. Abstract**

**2. Problem Definition**

**2.1 Overview**

**2.2 Problem Statement**

**3. Introduction**

**4. Literature Survey**

**5. Project explanation**

**6. Dataset - Basic Info**

**7. Data Visualization**

**8. Data Preprocessing**

**8.1 Data Cleaning**

**8.1.1 Treating Null Values**

**8.2**  **Data Transformation**

**8.2.1 Changing Variable Class**

**8.2.2 Detecting and resolving data value concepts**

**8.2.3 Treating Outliers**

**8.3** **Data Reduction**

**8.3.1** **Dimensionality reduction**

**9. Feature Engineering**

**9.1** **Encoding**

**9.2** **Splitting Dataset**

**9.3 Feature Scaling**

**10. Predictive Modeling**

**11. Model Fine Tuning**

**12. Website Hosting**

**13. Result**

**14. Conclusion**

**15. Reference**

1. **Abstract**

The Kerala state (province of India) legislative assembly election was held in April 2021.This database contains detailed candidate-level data for Kerala Legislative election 2021. In the study we could predict whether the candidate would win or lose the election based on the information given in the feature columns.

Attribute Information:

* Year - Year of the election
* District - name of the district
* Ac.no - Constituency number
* Ac name - Constituency Name
* Candidates- Name of the candidates
* Sex - Sex of the candidates
* Age - Age of the candidates
* Category- Category of the candidates
* Party - Party of the candidates
* Symbol - Symbol of the candidates
* Valid Votes Polled in General -Number of valid votes
* Valid Votes Polled in Postal - Number of valid postal votes
* Total Valid Votes - Total number of valid votes
* % Votes Polled - Percentage of votes polled
* Total Electors -Total number of electors
* Criminal Cases - Number of criminal cases of the candidates
* Education - Education qualification of the candidates
* Total Assets - Assets of the candidates
* Liabilities - Liabilities of the candidate
* Winner - Winner of the given constituency

**2. Problem Definition**

**2.1 Overview**

The Kerala state (province of India) legislative assembly election was held in April 2021, and the left-led LDF won a second consecutive term in office with a historical lead. The data released by the Indian Election Commission reveals information about the ground reality in each of Kerala's 140 electoral constituencies. The data is obtained from the Election commission of India and is publicly available. The results of the election were declared on 2 May, 2021.The data could shed light on some voting patterns in specific districts, and the nature of leads in different regions of Kerala.

**2.2 Problem Statement**

To make a machine learning algorithm to predict whether the candidate will win or lose the election based on the information given in the feature columns.

##### ➔ Objective of the project

The dataset is a listing of all the details of the candidates in each constituency of Kerala during the 2021 Legislative election,containing data about their age,party, gender,education, criminal cases,assets,liabilities and vote got in the election.Considering how the winning candidate, by analyzing this dataset to learn how the data given is balanced against each other and to potentially identify the best insight from data using exploratory data analysis with Python.

**3. Introduction**

Kerala has a unicameral house of legislation, Niyamasabha, consisting of 140 members elected for a period of five years, unless dissolved earlier, and one nominated member for the [Anglo-Indian community](https://en.wikipedia.org/wiki/Anglo-Indian). The dataset contains the information of all the candidates in 140 contituencies.it includes district wise constituency list of all candidates, their basic details like age,sex,religion, education etc and voting percentages. In this project, analysis of the dataset is done using python. This includes looking at missing values and getting some simple calculations from it, visualizing some results using Matplotlib, Seaborn and Tableau and concluding the venture with some machine learning algorithm to try and predict the winner of the future years.

The basic steps involved in the data analysis process are as follows: first, select an interesting real-world dataset. Second step is data preparation & data cleaning, which includes loading the dataset into a dataframe using Pandas and exploring the number of rows & columns, ranges of values etc. Next step is handling missing, incorrect and invalid data. Perform any additional steps if required (parsing dates, creating additional columns, merging multiple dataset etc.). Then Perform exploratory analysis & visualization, Compute the mean, sum, range and other interesting statistics for numeric columns. Explore distributions of numeric columns using histograms etc. Explore relationships between columns using scatter plots, bar charts etc. Make a note of interesting insights from the exploratory analysis. As a fourth step Ask & answer questions about the data. Answer the questions either by computing the results using Numpy/Pandas or by plotting graphs using Matplotlib/Seaborn/Tableau. Create new columns, merge multiple dataset and perform grouping/aggregation wherever necessary. Summarize the inferences & write a conclusion.

**4. Literature Survey**

Predictive modeling is a statistical technique using machine learning and data mining to predict and forecast likely future outcomes with the aid of historical and existing data. It works by analyzing current and historical data and projecting what it learns on a model generated to forecast likely outcomes. One of the most common predictive analytics models are classification models. These models work by categorizing information based on historical data. Classification models are used in different industries because they can be easily retrained with new data and can provide a broad analysis for answering questions. Classification models can be used in different industries like finance and retail, which explains why they are so common compared to other models.This paper presents different yields with different separation utilized as a part of calculation and may know the reaction of the classifier for the coveted application.

Eager learners construct a classification model based on the given training data before receiving data for classification. It must be able to commit to a single hypothesis that covers the entire instance space. Due to the model construction, eager learners take a long time to train and less time to predict.The Classification method gauges the absolute and forecast models to foresee nonstop esteemed capacities. For the most part, grouping is the way toward arranging information into classifications for its most powerful and able utilization. The information grouping technique makes fundamental information that is anything but difficult to discover and recover. In this paper the execution of three Bayes classifiers calculations in particular Naive Bayes, Bayes Net also, Naive Bayes Multinomialare broke down.

**5. Project explanation**

The project programming has been done on Jupyter notebook. As an initial step of programming we have imported necessary libraries in the python environment. Imported libraries are:

➔ Pandas

➔ numpy

➔ matplotlib

➔ seaborn

➔ sklearn

Since the data are in the form of a dataframe, by using the pandas library the dataset is read to the python environment.

**6. Dataset - Basic Info**

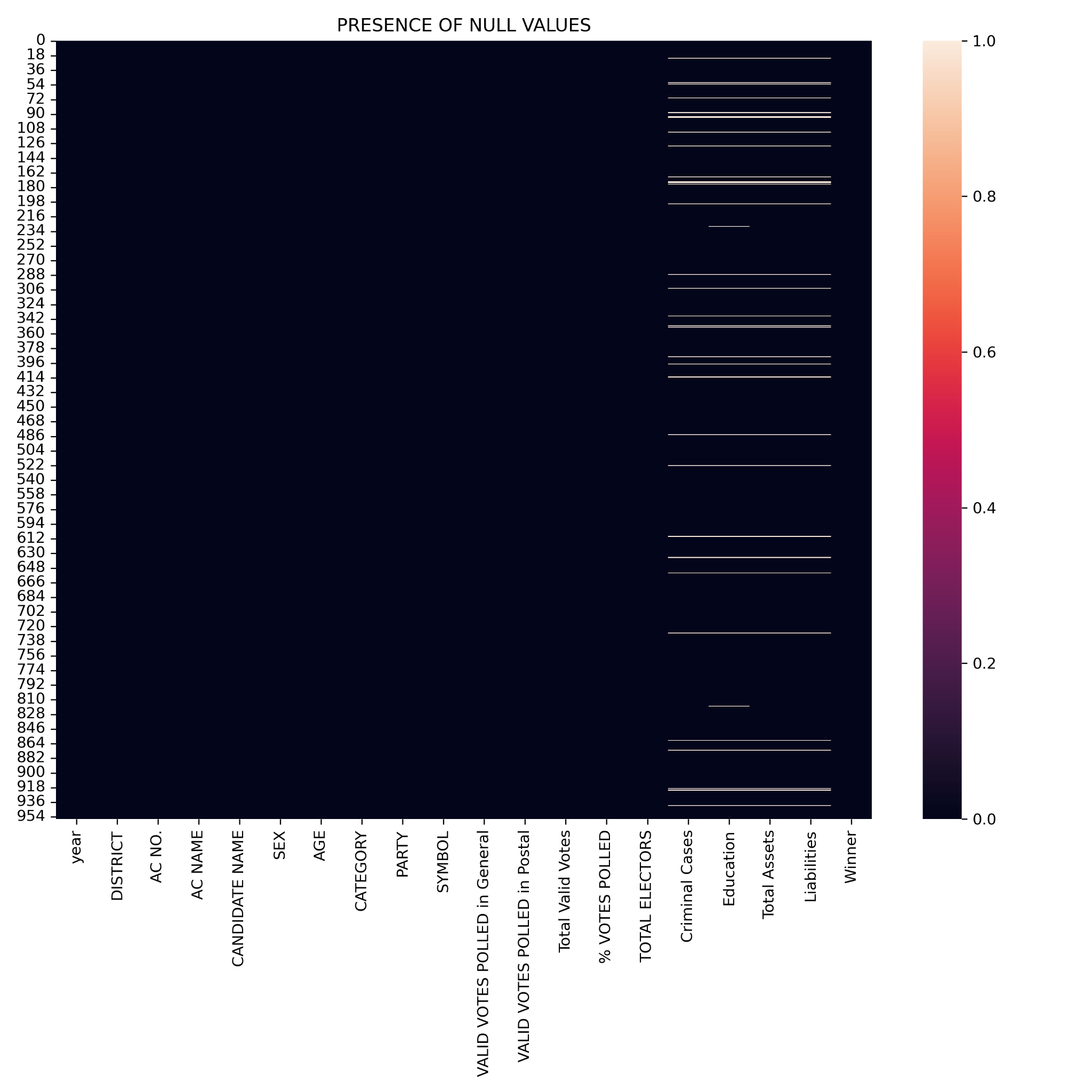
Checking the basic information of the dataset helps to understand the data very well.

Some of the insights are:

* The dataset consists of 957 observation and 20 attributes
* There are 2 floats, 8 integers, and 10 categorical columns respectively from this it's clear some of the columns have null values and most of the attributes need to be encoded.
* The observations are district wise categorized.
* Therefore the dataset can be corrected by implementing necessary preprocessing steps

**7. Data Visualizations**

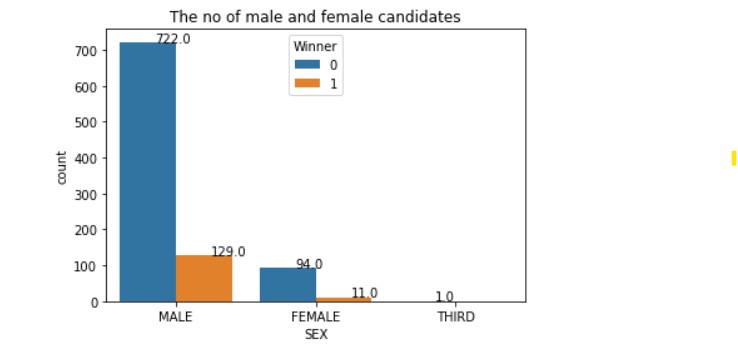
We'll start by taking a look at the null values present in the dataset. The following figure shows the null values of the dataset. 4 features such as criminal cases,education total assets and liabilities contain null values.



**7.1 The null values present in the dataset**

Second visualization examines the gender wise participation of candidates in the election. The visualization of the gender wise participation of candidates is shown in figure. The following visualization helps us to analyze the number of male, female and third gender candidates and a few interesting results standout:

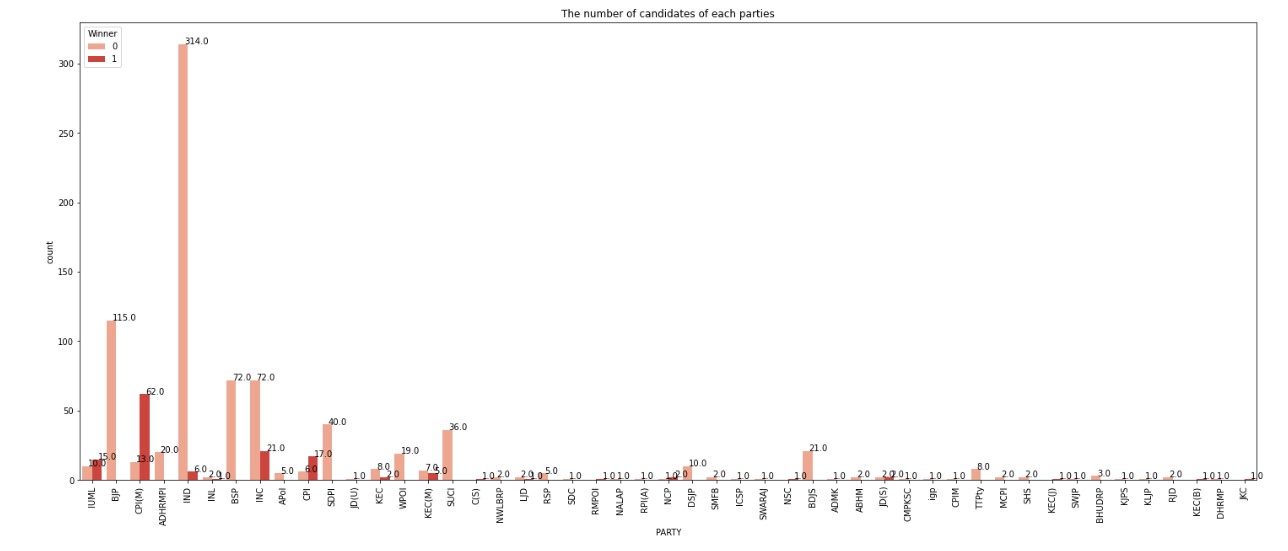
* There is one candidate in the third gender, who contested in Vengara constituency in Malappuram District.
* The participation of male candidates is greater than female candidates.
* 15% of male candidates and 10% of female candidates have won the elections.

****

**7.2 Number of participation of candidates on the basis of gender**

The below figure shows the candidates list of each party and the winning candidates. There are 48 different parties in the election and some candidates independently participated in the election.

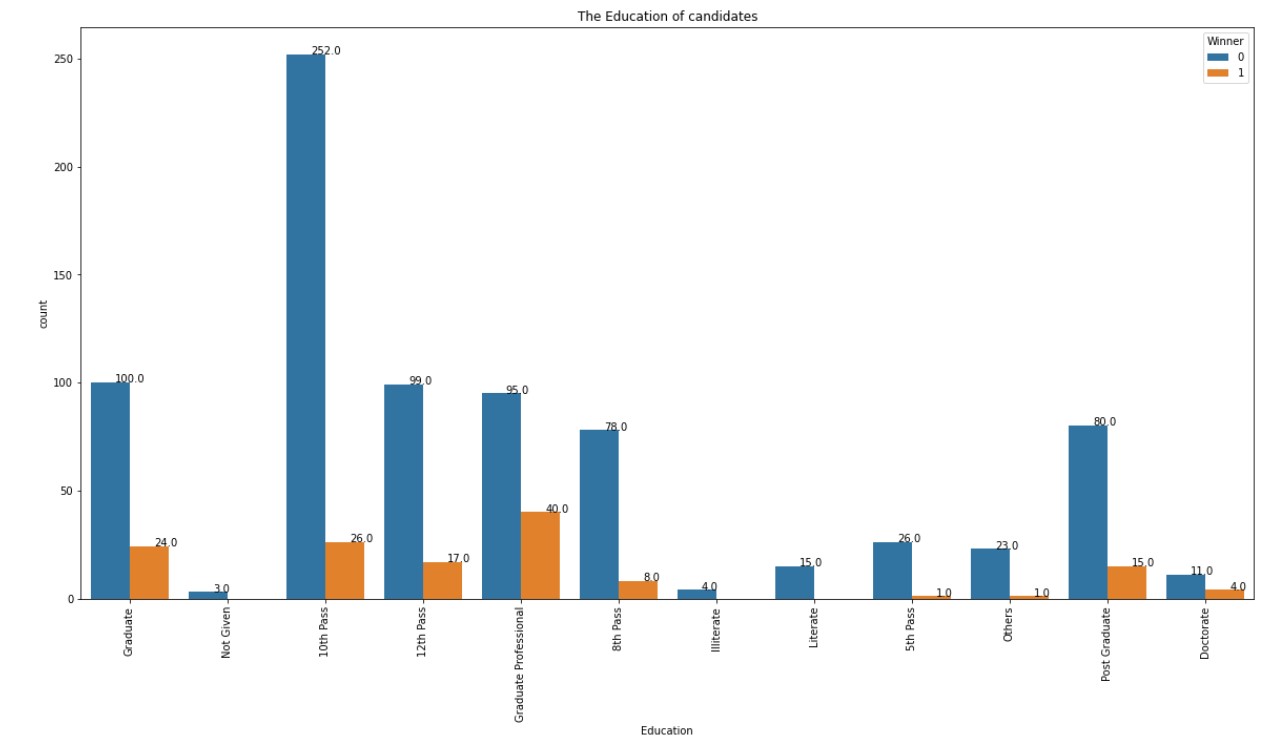
* Most of the candidates have independently contested in the election.
* The parties are combined to form alliances like LDF,UDF,NDA etc. Alliance with the most number of winning candidates will be the governing political alliance in Kerala.

****

**7.3 Number of candidates of each parties**

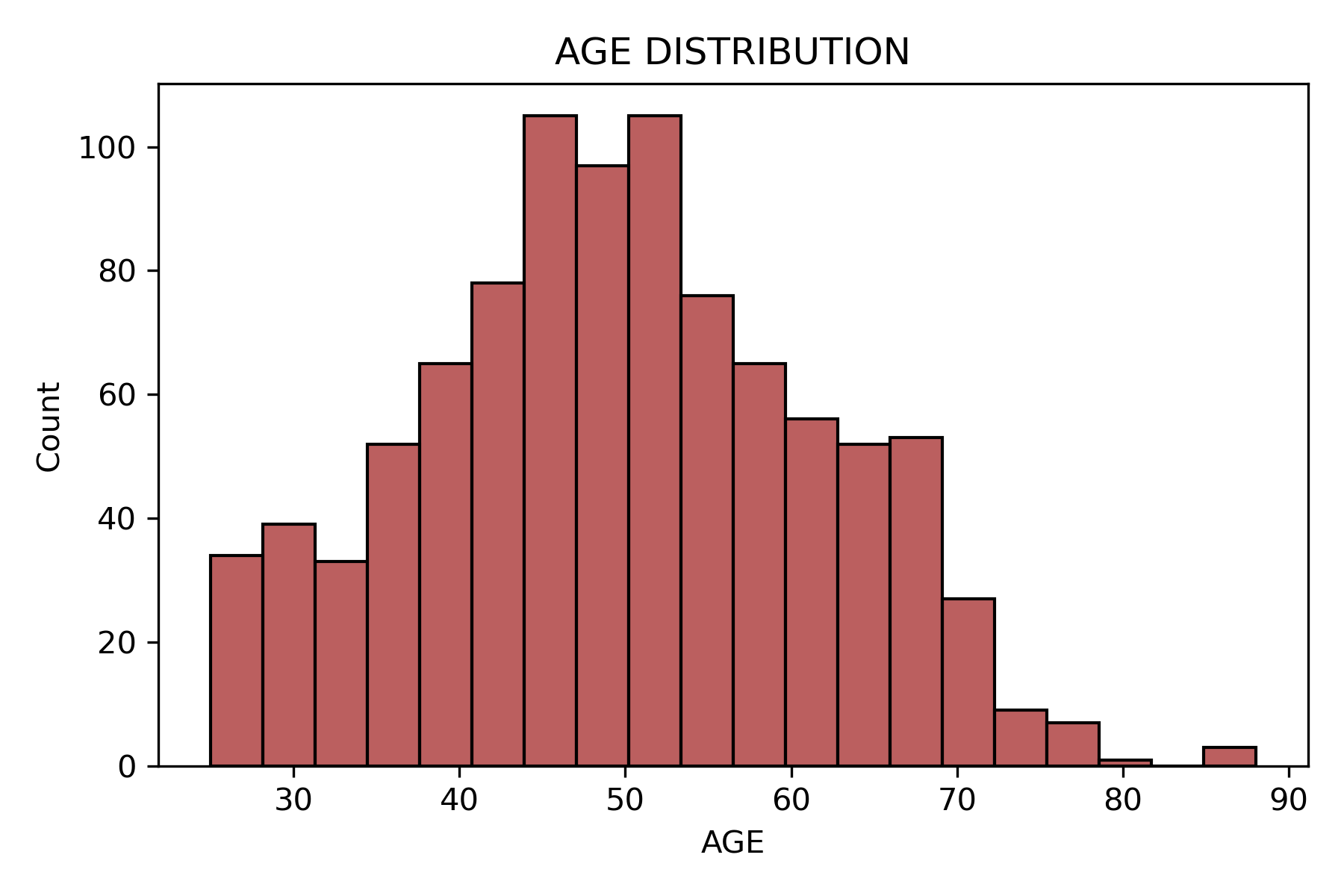
The next visualization is based on the education of the candidates. The highly educated candidate as well as illiterate candidates have participated in the election. By examining the below figure some interesting inferences are found:

* Most of the candidates are 10th pass.
* There are a very low number of illiterate candidates.
* Out of the winners we have 40 Graduate professionals, 24 Graduates,15 post Graduates and 4 candidates with Doctorate.

****

**7.4 The education of the candidates**

The age of the candidates are visualized below .

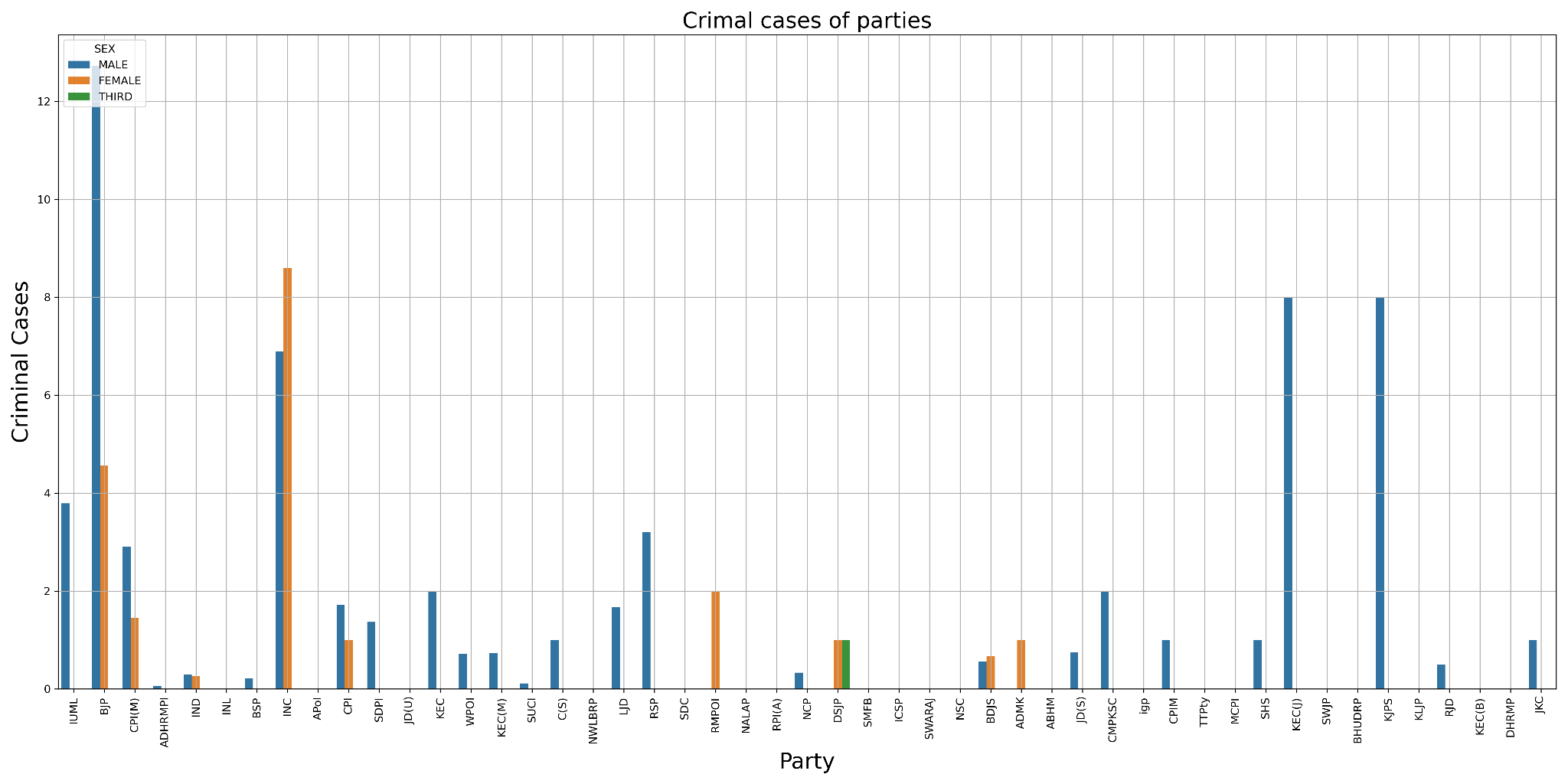
****

**7.5 Age Distribution**

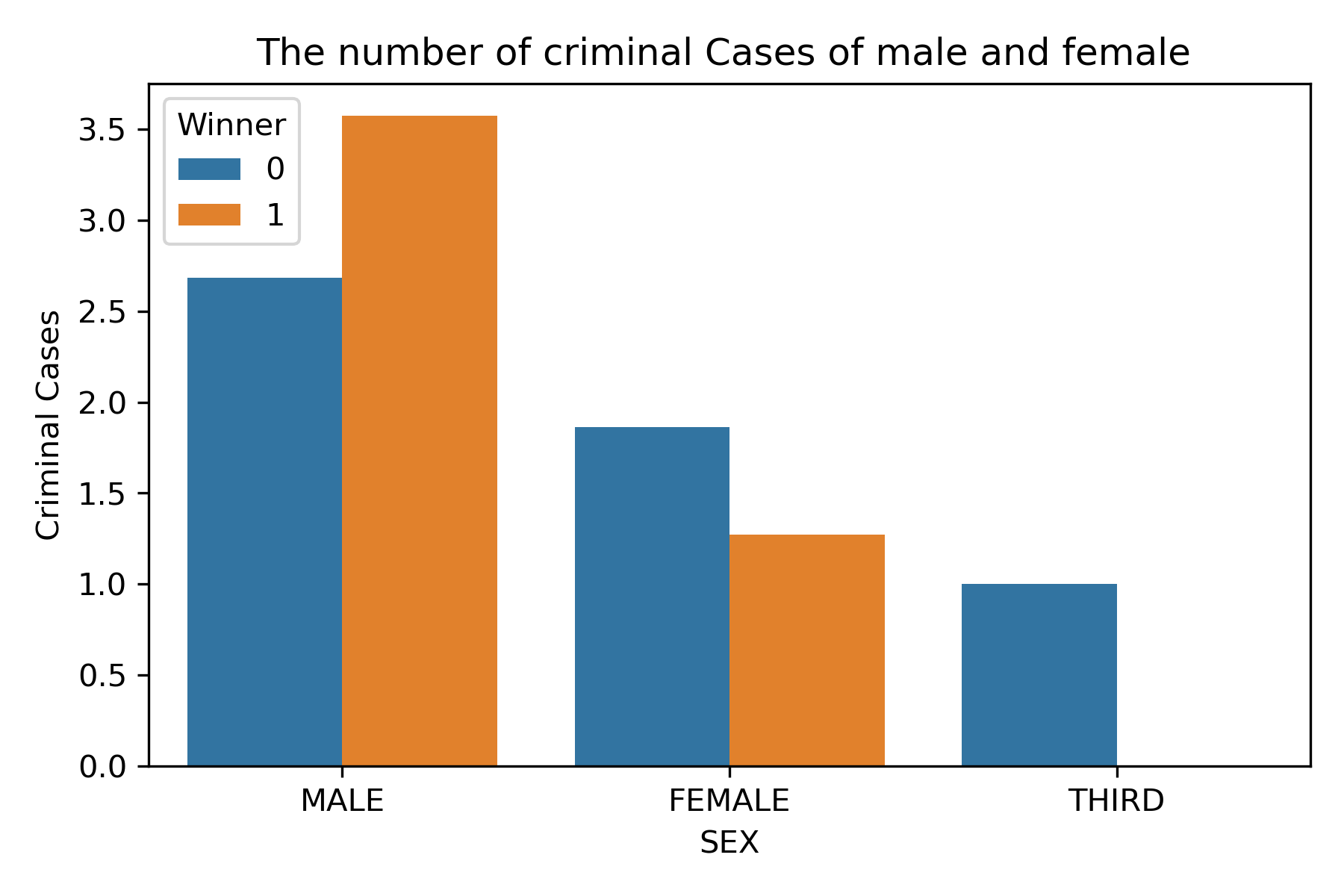
As shown in figure the ages are normally distributed and most of the candidates age between 45 to 55.

Next visualization is based on the criminal cases of parties. The below figure shows the criminal cases of both male and female candidates of all parties. By examining the below figure some inferences are found:

* Most of the parties have criminal cases.
* Criminal cases are large in BJP.
* In INC the criminal cases of Female candidates larger than that of male candidates



**7.6 Criminal cases of parties**

****

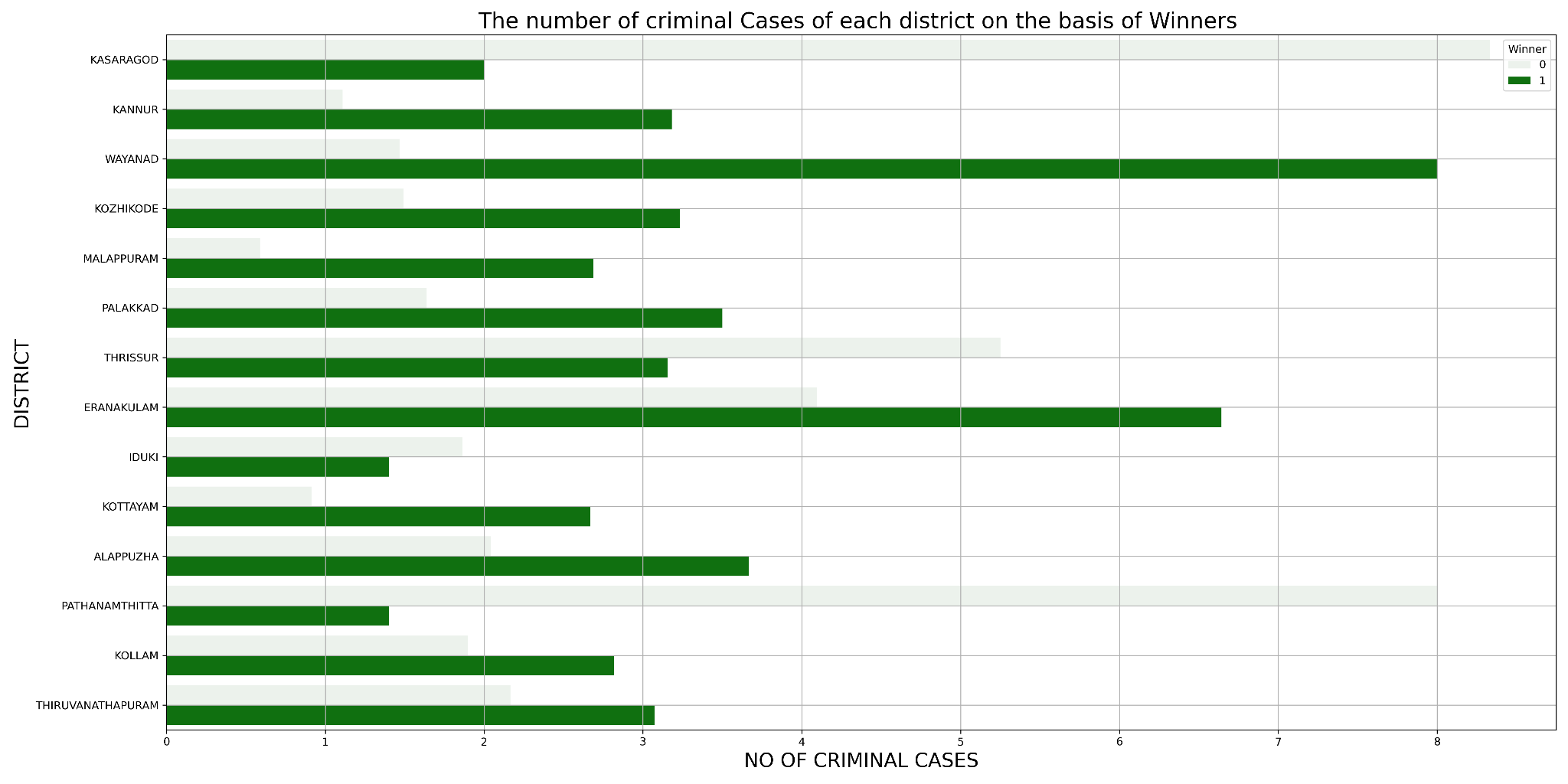
**7.7 The gender base visualization number of criminal cases**

The above diagram visualizes the number of criminal cases of male,female and third gender categories. The interesting fact is that

* The number of winning male candidates with criminal background is greater than the number of failing male candidates with criminal background.
* There is only one candidate in the third gender who has a criminal background.
* Female candidate with a criminal background also won the election.

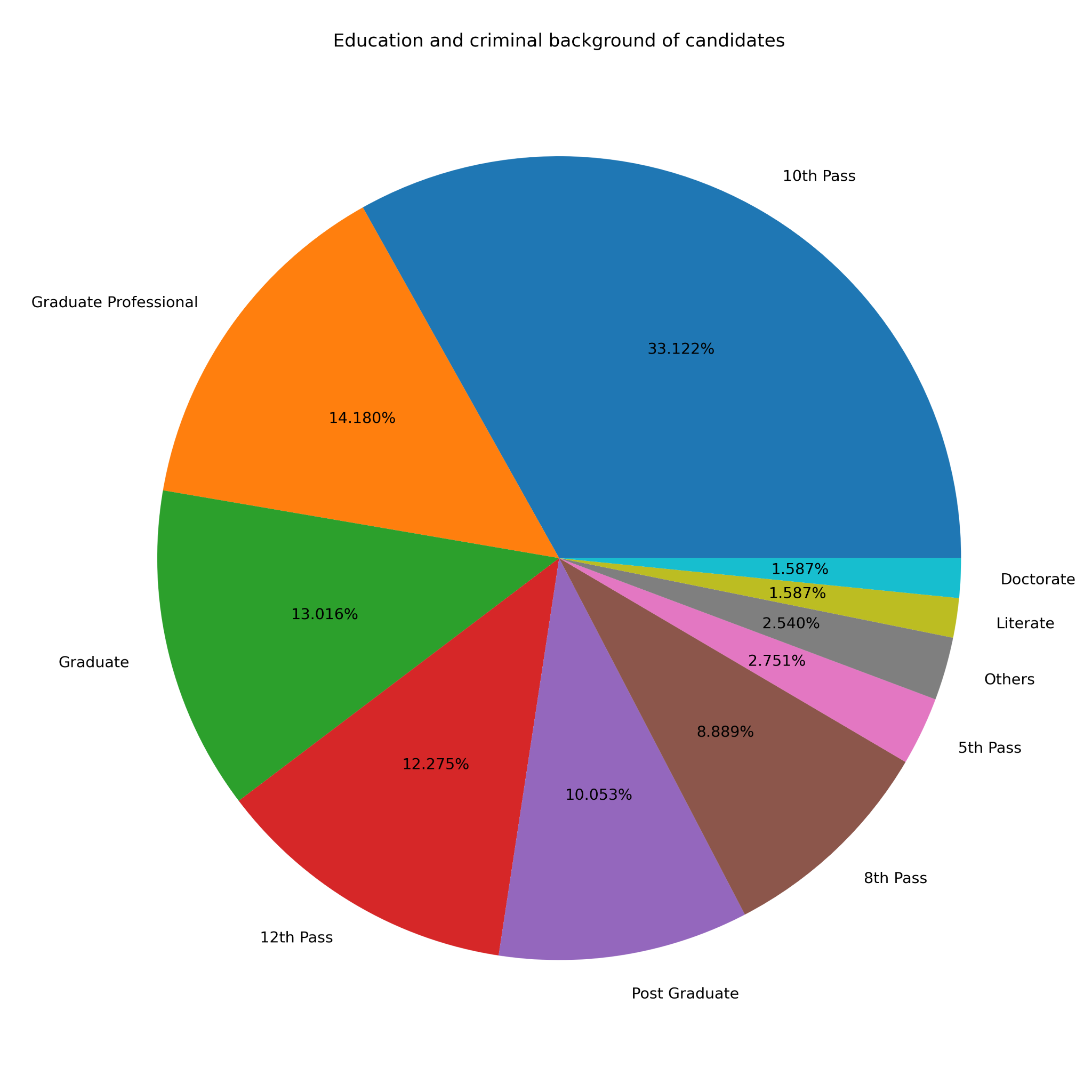
The below figure shows the criminal background of candidates in all districts. The figure shows that

* Most of the district the number of winning candidates with criminal background is greater than the number of winning candidates without criminal background.
* Wayanad has a large number of winning candidates with criminal backgrounds.
* Only 4 districts have the larger number of winning candidates without criminal background.
* Kasargod is the first district in the above category.



**7.8 The number of criminal cases of candidates of each district.**

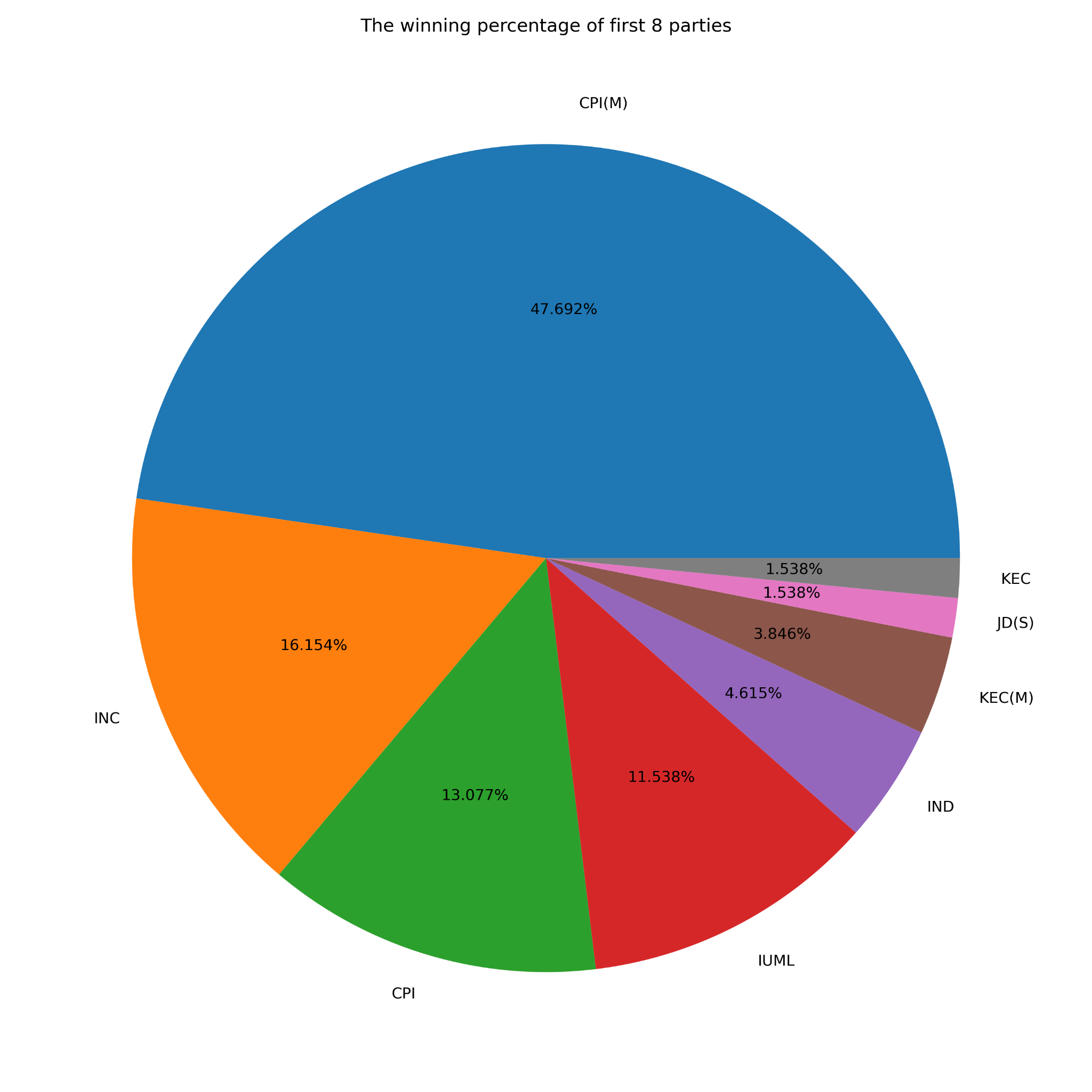
The below figure shows the visualization of criminal cases and education of the candidates. From the figure it is clear that highly educated candidates with a doctorate have less criminal cases. The large number of criminal cases are the candidates with 10th pass. Graduate Professional has the second position in criminal background. Most of the graduate professionals are advocates.



**7.9 The criminal cases on the basis of education of the candidates**

The pie diagram below shows the winning percentage of the first 8 parties in the election. By comparing the below pie diagram following inferences are found:

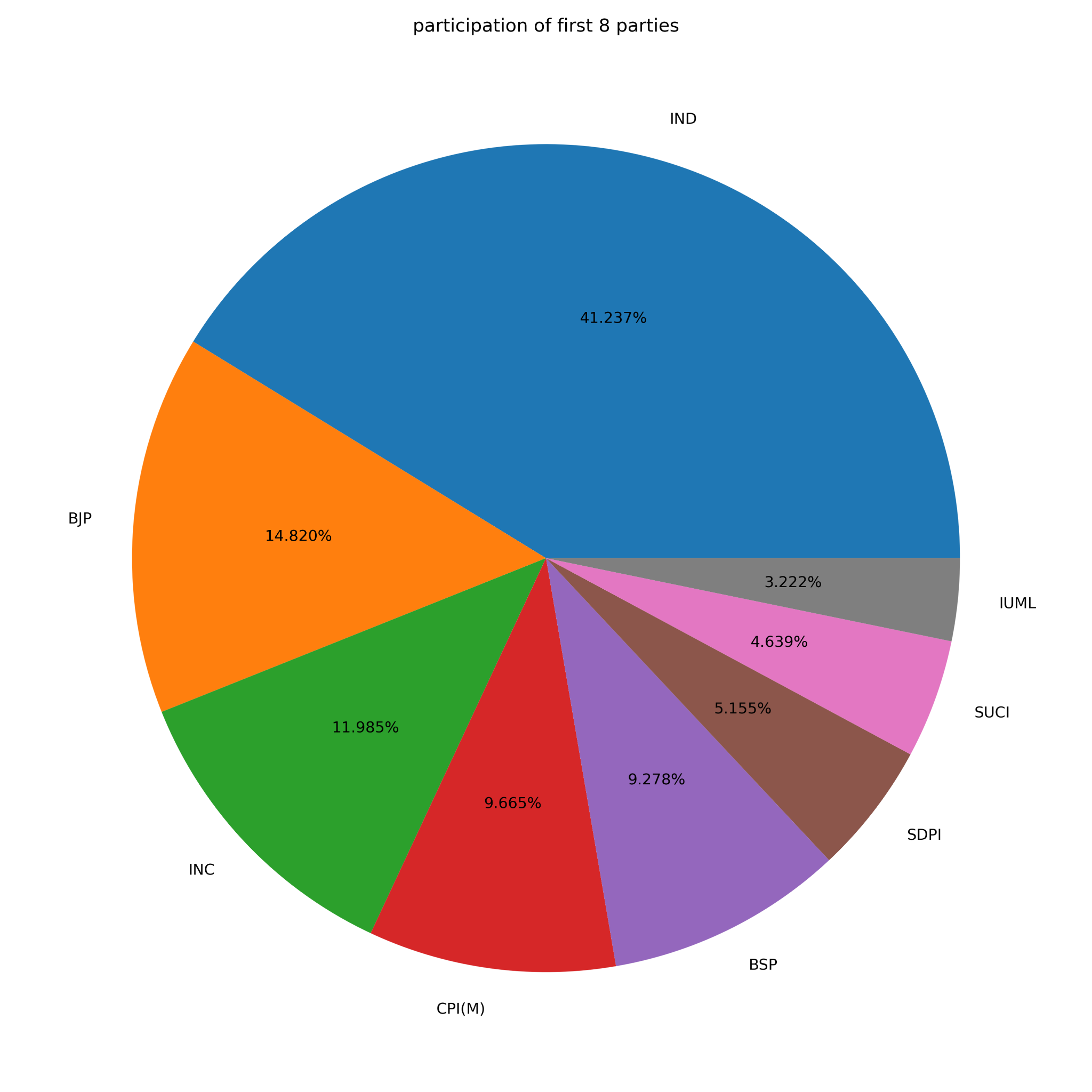
* The most of the winning candidates are in CPI(M) (47%)
* second is INC, but by comparing the winning percentage with the first party it is very low.

****

**7.10 The winning percentage of first eight parties**

The below figure shows the participation of the first 8 parties. 41% of independent candidates participated in the election. By comparing the figure 7.10 and 7.11 few interesting results are found:

* Most candidates independently participated in the election. That means they do not belong to any parties. but the fact is that only 4% of them won the election.
* BJP has the second position in participation in elections. Unfortunately they cannot win the election.
* 11% candidates participated in the INC party. 16% of them won the election. They have second position in the election
* The first position in the election was CPI(M). 9% candidates participated in the election. 47% of them won the election.

****

**7.11 The participation of first 8 parties in the election**

# 8. Data Pre-Processing

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format

**●** The preprocessing steps taken in this dataset are given as

* Data Cleaning
* Data Transformation
* Data Reduction

**8.1** **Data Cleaning**

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

#### 8.1.1 Treating Null Values

Missing values in each column of the dataset are known as null values. Null values and the percentage of null values present in the dataset of each Features are given below,

➔ **Criminal cases** - percent of data that is missing in criminal cases is: 3.4%

➔ **Education** - percent of data that is missing in education is: 3.6%

➔ **Total Assets** - percent of data that is missing in total assets is: 3.4%

➔ **Liabilities** - percent of data that is missing in liabilities is: 3.4%

The low number of values are missing in the 4 features.

➔ **Filling missing values in Criminal cases column**

● Here as the criminal cases have 33 missing values, I have just simply put the '0 in the criminal cases due to the fact that we cannot put any number in the missing column, without correct investigation.

● Since Criminal background affects the candidate’s future, we fill the missing value with 0.

➔ **Filling missing values in Education column**

* Education has 35 missing values.
* we have just simply put the Not Given in the Education.

➔ **Filling missing values in total assets and Liabilities**

* Total assets and Liabilities have 33 missing values
* Since most of the values in these two columns are 0. we fill the missing value with 0.

### 8.2 Data Transformation

Data transformation is the process of converting data from one format or structure into another format or structure. It is a fundamental aspect of most data integration and data management tasks such as data wrangling, data warehousing, data integration.This step is taken in order to transform the data in appropriate forms suitable for the mining process. That is to transfer noisy dataset into efficient dataset.

**8.2.1 Changing Variable Class**

* While checking the basic information of the dataset, the feature named criminal cases has numerical value, but the machine took it as a categorical value.
* Therefore we are changing the object datatype of the feature ‘Criminal Cases’ to the integer datatype.

**8.2.2** **Detecting and resolving data value concepts**

* Since ‘total assets’ and ‘Liabilities’ comma separated numeric values, they are considered categorical types. but in the actual sense these must be integer or float values.
* Total assets are the assets of the candidates which are represented in rupees. Similarly liabilities feature includes the total liabilities of the candidates which is also represented in rupees.
* Therefore we can directly change the datatype of these two features to int or float by eliminating the comma from the values.

**8.2.3** **Treating Outliers**

* In statistics, an outlier is a data point that differs significantly from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set.
* An outlier can cause serious problems in statistical analyses.
* There are different types of methods to check outliers and handling outliers such as using

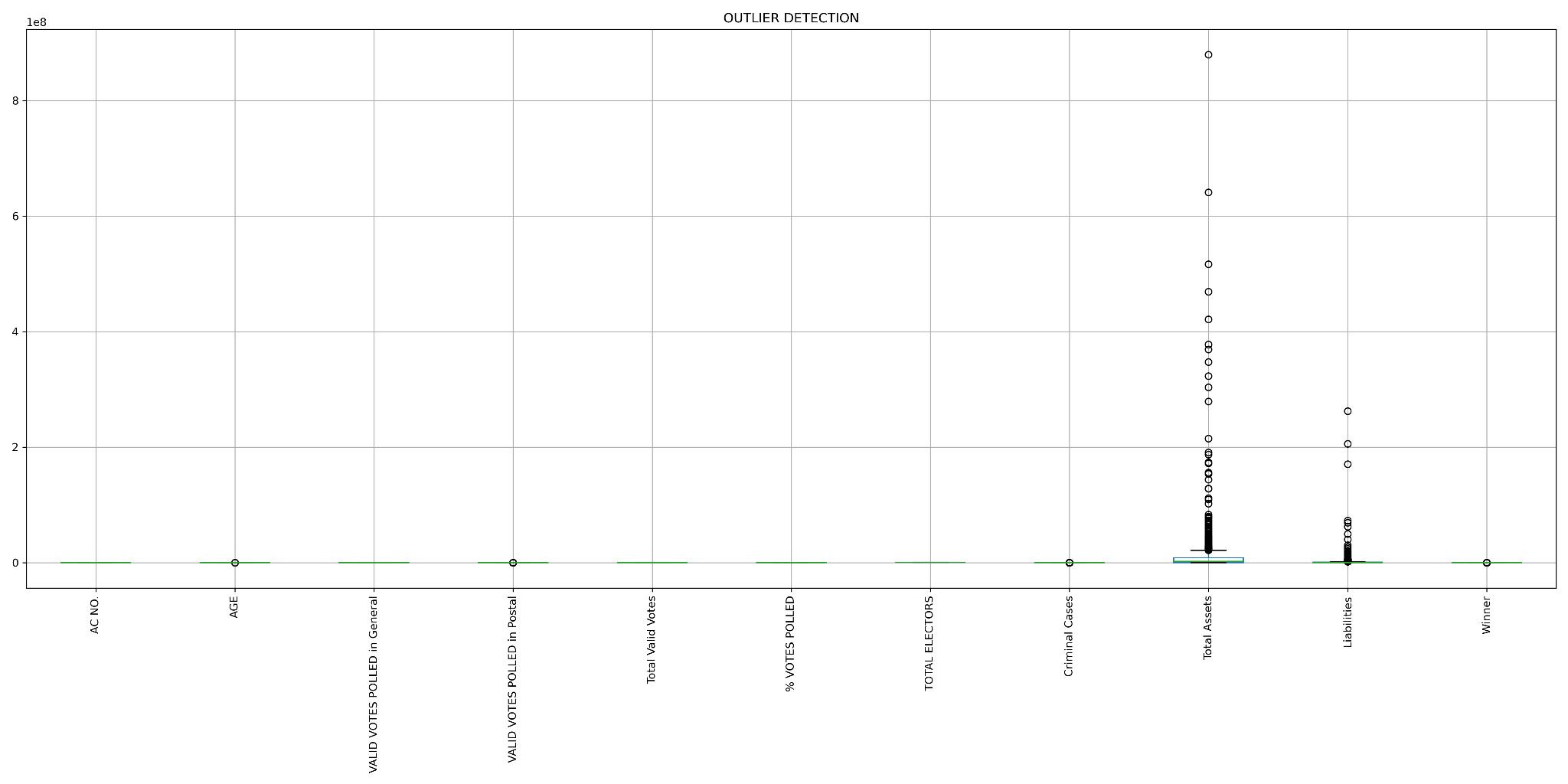
##### A. Graphing,

B. z-score,

##### C. Interquartile range,

D. Hypothesis Tests etc.

* In this, for preprocessing data we used the “Graphing” method to find and handle outliers.
* Box plot is a data visualization plotting function. It shows the min, max, median, first quartile, and third quartile.The maximum and the minimum is the max and min value of the data-set. 50 percentile is the median of the data-set. The first quartile is the median of the data between the min to 50% and the third quartile is the median of the data between 50% to max. The outliers will be the values that are out of the (1.5\*interquartile range) from the 25 or 75 percentile.



8.1 Box plot of the dataset

* The above figure shows the box plot of the dataset. It shows that Total Assets and Liabilities have outlier values.
* Since the size of the dataset is too small we are considering the outliers.

### 

### 8.3 Data Reduction

#### Since data mining is a technique that is used to handle huge amounts of data. While working with a huge volume of data, analysis became harder in such cases. In order to get rid of this, we use data reduction techniques. It aims to increase the storage efficiency and reduce data storage and analysis costs.

#### 

#### 8.3.1 Dimensionality reduction

Here, the features are overcrowded that affect the predictive model algorithm,to reduce the feature column with low variance and feature not dependent to the target value is done with “High Cardinality “ method and other similar techniques

➔ **High Cardinality method**

Almost all datasets now have categorical variables. Each categorical variable

consists of unique values. A categorical feature is said to possess high cardinality when there are too many of these unique values,hence it does not affect the modeling anyway,therefore the features above to the assigned maximum cardinality were dropped from the dataset. This helps in the reduction of the volume of the data which makes the analysis easier yet produces the same or almost the same result.

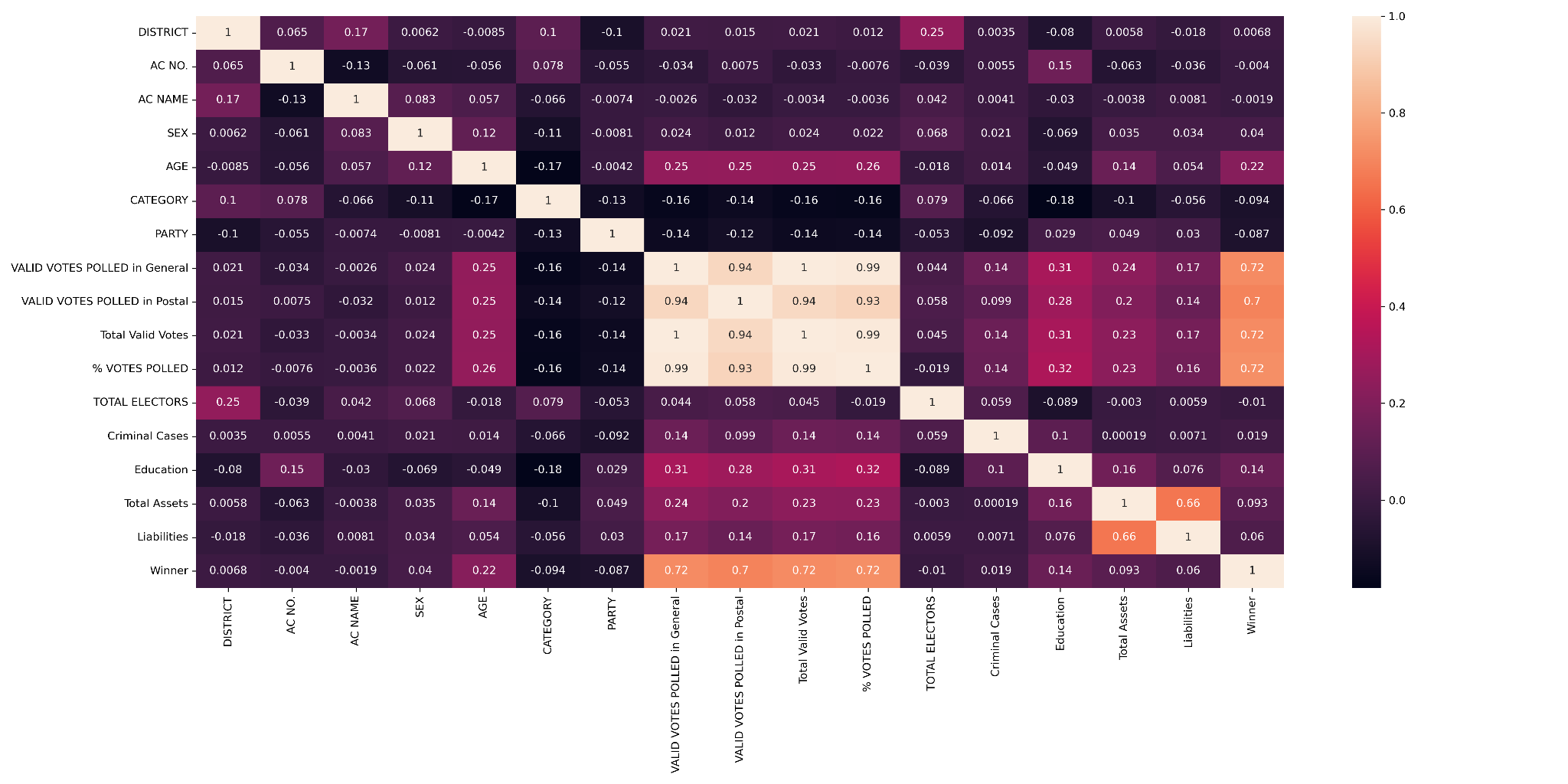
High-cardinality refers to columns with values that are very uncommon or unique. If we have more than limited categories for each of these features, it won’t be very useful to use them. It would add dimensions to our dataset and we don’t want to do that

Here all the values of column ‘Candidate Name’ are unique and most of the values in the column ‘SYMBOL’ are unique. The column with unique and very uncommon values is dropped from the dataset, thus the features 'Candidate Name'and ‘Symbol’ has removed from the dataset

➔ **Dropping features**

Since there is only one value in the ‘year’ column that is 2021, it doesn’t have any significance to the dataset.Therefore we dropped the ”Year” coloumn.

We are finding the features to drop by using correlation of the features. Correlation is a term that refers to the strength of a relationship between two variables where a strong, or high, correlation means that two or more variables have a strong relationship with each other while a weak or low correlation means that the variables are hardly related. We can drop any one of the features from the dataset if it is highly correlated.



8.2 Correlation between the features.

The correlation is visualized using heatmap, the above figure shows the correlation between the features of the dataset. From the above figure we found following matters:

* The correlation between the features ‘%votes polled’ and ‘Valid votes polled in general’ is 0.99. That is they are highly correlated .So we can drop any one of them.
* The correlation between the features ‘%votes polled’ and ‘ Total Valid votes’ is 0.99. That is they are highly correlated .So we can drop any one of them.
* The correlation between the features ‘Valid votes polled in postal’ and ‘%votes polled’ ,‘Valid votes polled in general ’ is 0.94. That is they are highly correlated .So we can drop any one of them.

Therefore we dropped the feature ‘Valid votes polled in postal’ ‘%votes polled’ ,‘Valid votes polled in general ’ and ‘Total Elector’.

# 9. Feature Engineering

Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.The features in our data will directly influence the predictive models you use and the results you can achieve.

#### 9.1 Encoding

For efficient storage of these strings, the sequence of code points is converted into a set of bytes. The process is known as encoding. This reduce the size of data by encoding mechanisms there are different types of encoding such as

➔ One Hot Encoding

➔ Label Encoding or Ordinal Encoding

➔ Dummy Encoding

Here for encoding we have used label encoding or Ordinal encoding by importing the labelencoder library from sklearn, the encoding technique is proceed. In Label encoding, each categorical label is converted into an integer value.

#### 9.2 Splitting Dataset

The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model. It can be used for classification or regression problems and can be used for any supervised learning algorithm.

● **Train Dataset**: Used to fit the machine learning model.

● **Test Dataset**: Used to evaluate the fit machine learning model.

Here we split the dataset as two subset as x and y where x and y are input and output variable respectively(output variable is considered as Winner That is winner of the election)

By importing the train\_test\_split library from sklearn.model, the splitting the train and test data is proceed.

The data is splitted by Train: 80%, Test: 20%, with the random state of 42.

● No. of training examples: 765

● No. of testing examples: 192

**9.3 Feature Scaling**

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. Techniques to perform Feature Scaling**,**Considering the two most important ones:

➔ **Min-Max Normalization:** This technique re-scales a feature or observation value with distribution value between 0 and 1.

➔ **Standardization:** It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

Here in this dataset the standardization method is used by importing the standard\_scalar library from sklearn.preprocessing, the standardization feature scaling is proceed.

· **10.** **Predictive Modeling**

Since this modeling predicts a discrete class label output therefore we use a classification predictive modeling method predictive output consisting of multi classes therefore using Multi-class classification predictive modeling.

Algorithms used for this predictive models are

**1. Logistic Regression**

2. **k-Nearest Neighbors**

**3. Decision Trees**

**4. Random Forest.**

**5.Support Vector Machine**

Necessary libraries for using predictive algorithms are the classifiers and mean\_absolute\_error from sklearn. Since it is a classification problem we should consider the accuracy of each algorithm. The maximum accuracy obtained from this algorithm is the best predictive model for this dataset. The accuracy achieved for each of these algorithms are given as

* **Accuracy for logistic regression algorithm : 76.5%**
* **Accuracy for Decision Tree algorithm : 87.5%**
* **Accuracy for Random Forest Algorithm : 87.5%**
* **Accuracy for Knn algorithm : 87.5%**
* **Accuracy for Support Vector Machine algorithm : 87.5%**

While comparing the algorithms Random Forest algorithm,Decision Tree Algorithm, Knn algorithm and Support Vector Machine Algorithm has the highest accuracy percentage at nearly 87.5%. Therefore , here we consider the best model as the Random Forest algorithm.

# 

# 11. Model Fine Tuning

Machine learning models have hyperparameters that must be set in order to customize the model to the dataset. A better approach is to objectively search different values for model hyperparameters and choose a subset that results in a model that achieves the best performance on a given dataset. This is called hyperparameter optimization or hyperparameter tuning. A range of different optimization algorithms may be used, although two of the simplest and most common methods are Random Search & Grid Search.

Here we are using voting classifier ensemble method and Grid Search optimization algorithm,

➔ **Voting Classifiers**

* A Voting Classifier is a machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on their highest probability of chosen class as the output
* The voting classifier aggregates the predicted class or predicted probability on the basis of hard voting or soft voting. So if we feed a variety of base models to the voting classifier it makes sure to resolve the error by any model.
* Here we fit the voting classifiers in all model classification algorithms we used for predicting,showing the accuracy as 87.5%. It's not much better than the classifiers, Hence we used to boost the algorithm.

➔ **Boosting/Tuning the model**

* Applying grid search optimization in Random Forest algorithm model,because Random Forest is our best predictive algorithm
* Best parameter currently in use are ['max\_features': 'sqrt', 'n\_estimators': 1200].
* By optimizing the algorithm inRandom Forest model, we get best score as Best Score: 87.5%, and Best Hyperparameters as 'max\_features': 'sqrt’' & 'n\_estimators': 1200 ,Accuracy as 87.5%
* Therefore, the best configuration achieved an accuracy of about 87.5 % which is very close to the best performance on the model and the specific values for the max\_features': 'sqrt' hyperparameters used to achieve that score.
* Hence the tuned parameter is taken as “Best Model” for predicting Winning candidate of the election.

**12. Website Hosting**

Web hosting is an online service that allows you to publish your website files onto the internet. So, anyone who has access to the internet has access to your website. In practice, it usually refers to the service you get from a web hosting provider like “pythonanywhere.com”

There are different types of hosting are offered such as

❖ Shared Hosting

❖ WordPress Hosting

❖ VPS Hosting

❖ Dedicated Hosting

❖ Cloud Hosting

In this project, after finishing the predictive modeling and tuning of the model, just confirmed whether the model is predicting or not, just checked the model with sample inputs.After creating a best model next step is to host the model in web,therefore ‘python flask’ method is adopted here to create a flask folder

➔ Python Flask

Flask is a web framework that provides libraries to build lightweight web applications in python. It is developed by Armin Ronacher who leads an international group of python enthusiasts (POCCO). It is based on the WSGI toolkit and jinja2 template engine. Flask is considered as a micro framework.

* WSGI is an acronym for web server gateway interface which is a standard for python web application development. It is considered as the specification for the universal interface between the web server and web application
* Jinja2 is a web template engine which combines a template with a certain data source to render the dynamic web pages

Here Flask where created in ‘spyder python environment”,HTML & CSS code where used to create flask webpages

* HTML, in full hypertext markup language, a formatting system for displaying material retrieved over the Internet
* CSS is a style sheet language used for describing the presentation of a document written in a markup language such as HTML

To create a flask we need to combine the model that we predicted to html & css, so that created a folder with templates and static. In templates folder consist of html code of home and the result page,and In static folder consist of stylesheet of those webpage

necessary libraries used in flask are

❖ pickle

❖ Flask,

❖ render\_template,

❖ request

After creating flask with the port of 5000, we created a webpage with IP address

IP address : “http://127.0.0.1:5000/”

lets the created code is imported in “python anywhere.com” and created a web host, to predict the primary type of the pokemon using our best model.

# 

**13. Result**

Predictive models use known results to develop (or train) a model that can be used to predict values for different or new data. Modeling provides results in the form of predictions that represent a probability of the target variable based on estimated significance from a set of input variables. Hence Predictive analytics is the use of historical data, statistical algorithms, predictive modeling, and big data machine learning techniques to help organizations predict future outcomes more accurately, plan for unknown events, and discover opportunities in future activities.

In this project by using the Kerala Legislative Election dataset we have predicted whether the given candidate will win or lose in future elections using the Classification model. By applying various classification algorithms we got the most percent accuracy model as the Random Forest algorithm.

After getting the best algorithm we have tuned the model using hyperparameter tuning in the gridsearchcv method,therefore we best model tuned for prediction.

Hence the chosen best model is imported in python flask with necessary HTML and CSS code to create a web page,Thus we have created a web hosting using Flask in “pythonanywhere.com”

The hosted website link of each participants are given below :

http://athiravipindas.pythonanywhere.com/

**14. Conclusion**

Predictive data analytics is a technology that can anticipate future trends. It is an evolution of earlier data analytics models and works by predicting what will happen in the future by analyzing historical data, discovering patterns and using that information to draw up predictions about the overall direction of the industry.In predictive modeling, data is collected, a statistical model is formulated, predictions are made, and the model is validated (or revised) as additional data becomes available.Various types of predictive models can be used to analyze data and generate predictive outputs. Examples of predictive models include Naive Bayes classifiers, k-nearest neighbour classifiers, support vector machines, and logistic regression techniques, depending upon the classification /regression method

Predictive models are the models that define the relationship between the various attributes or features of that unit.Typically, a predictive model is trained with training data that includes input data and output data that mirror the form of input data that will be entered into the predictive model and the desired predictive output, respectively. The amount of training data that can be required to train a predictive model can be large, e.g., in the order of gigabytes or terabytes. The number of different types of predictive models available is extensive, and different models behave differently depending on the type of input data. Additionally, a particular type of predictive model can be made to behave differently, for example, by adjusting the hyper-parameters or via feature induction or selection

Predictive models make assumptions based on what has happened in the past and what is happening now. If incoming, new data shows changes in what is happening now, the impact on the likely future outcome must be recalculated, too.predictive analytics reduce time, effort and costs in forecasting business outcomes. Variables such as environmental factors, competitive intelligence, regulation changes and market conditions can be factored into the mathematical calculation to render more complete views at relatively low costs.

Predictive modeling, also known as predictive analytics, and machine learning are still young and developing technologies, meaning there is much more to come. As techniques, methods, tools and technologies improve, so will the benefits to businesses and societies.

**15. References**

* https://eci.gov.in/files/file/13827-kerala-general-legislative-election-2021/
* <https://myneta.info/Kerala2021/>