



EJ Publications

International Journal Research in Applied Engineering, Science and Technology (IJRAEST) Impact Factor: 4.537(SJIF)

An International Peer-Reviewed Journal

Vol-2, Issue-1, 2020

www.ijraest.com Indexed in: Google Scholar, Cite Factor

ISSN: 2582-029X

RESEARCH ARTICLE

EXPLORATORY DATA ANALYSIS AND MACHINE LEARNING TECHNIQUES

D. Thushara (19031D0505)¹, M. Tech,

School of Information Technology JNTUH, Hyderabad

Dr. K. Santhi Sree², PhD,

Jawaharlal Nehru Technological University, Hyderabad

Abstract

Exploratory Data Analysis (EDA) plays an important role before performing the predictive analysis using suitable machine learning techniques. In this analysis, we find the type of data in the dataset; visualize the data and graphical representation of data takes place by which it is easy for us to select the desired and suitable model for obtaining the result for the dataset. After the analysis, we implement suitable machine learning technique to obtain the desired result. In this paper, we are considering a recent dataset, which consists of the data of recent Indian Elections 2019. This dataset consists of over 600 million voters voting for 8500+ candidates in 543 constituencies. In this dataset we have the details of each candidate, age, education details, party to which they belong to, about the number of votes received by each candidate and many more. The aim of this paper is to first summarize the data, perform the exploratory data analysis to know about the data that is required for getting the desired result and then select the suitable model and apply the suitable machine learning techniques for obtaining the final result. The final result tells us about the suitable technique which can be used to get the result in a perfect manner.

Keywords: Data Science, Machine Learning, Classification, Exploratory Data Analysis, Machine Learning Techniques, Decision Tree Classifier, Random Forest Classifier

1. Introduction

Data Science

Data Science is in zoom in present world, where a lot of companies are in need of data scientists for handling the structured and unstructured data. It is a process which is used for organizing, processing and analyzing huge data with scientific methods and different algorithms. It has many uses in the real world in different sectors for different purposes like making predictions, converting massive raw data & unstructured data to meaningful data and for automation. Applications of data science involve Image Recognition, Speech Recognition, Price Comparison Websites, Fraud and Risk Detection and many more.

RESEARCH ARTICLE**Machine Learning**

Machine learning is a part of the data science which is used for predicting the future. The formal definition of machine learning was given by Tom Mitchell as follows [2]:

“A computer program is said to learn from experience(E) with respect to some task(T) and some performance measure(P), if its performance on T, as measured by P, improves with experience E then the program is called machine learning program [2]”.

The main purpose of machine learning is to learn from the data, where the model should be able to do correct predictions with the new data, based on the rules learned from previous exposure to similar data.

Machine Learning Methods

Machine Learning methods are grouped into four categories. They are supervised, unsupervised, semi supervised and reinforcement learning techniques.

a) Supervised Learning

Supervised Learning is a machine learning technique where we train a model with a set of inputs and outputs. It deals with labelled data. Supervised learning is classified into two techniques they are Classification and Regression. It is mainly used in Bioinformatics, Speech Recognition, Spam Detection and Object Recognition.

b) Unsupervised Learning

Unsupervised Learning is a machine learning technique where we don't have any information of the desired output. It deals with unlabeled data. Unsupervised learning is classified into two techniques they are Clustering and Association. It is mainly used in Fraud detection in banking, Image Segmentation and Market Segmentation.

c) Semi supervised Learning

Semi supervised Learning is a machine learning technique which deals with both labelled and unlabeled data. In this learning we have large amount of unlabeled data and small amount of labelled data. Semi supervised learning is mainly used in Speech Analysis, Internet Content Classification and Protein Sequence Classification.

d) Reinforcement Learning

Reinforcement Learning is a machine learning technique which deals with the discovery of action that gives us the desired result over longer duration. Here we interact with the environment. In reinforcement learning the learning parameter is a single value. It is used mainly in



RESEARCH ARTICLE

Robotics for industrial automation and business planning.

Machine Learning Algorithms:

The machine learning algorithms that we use for solving the data problems are:

a) Linear Regression:

Linear Regression is a supervised machine learning algorithm. It is used for solving regression problems. Its main approach is to model a relationship between a scalar response (dependent variable) and one or more explanatory variables (independent variables) [3]. It is mainly used in business for making estimations or forecasting.

b) Logistic Regression:

Logistic Regression is a supervised machine learning algorithm. It is used for solving of classification problems by using the concept of probability. In this prediction of result takes place where the output will be in binary format. It is mainly used in cancer detection, diabetes prediction, spam detection etc.

c) Decision Tree:

Decision Tree is a supervised machine learning algorithm. It is used for solving both classification and regression problems. Decision Tree Classification is used for categorical variables which is a default machine learning algorithm and Decision Tree Regression is used for continuous variables.

d) Support Vector Machines (SVM):

Support Vector Machines (SVM) is a supervised machine learning algorithm. It is also used for solving both classification and regression problems but mostly used for classification only. It works on the principal of margin calculation [4].

e) Naïve Bayes:

Naïve Bayes is a supervised machine learning algorithm used for solving of classification problems. In this algorithm we apply the Bayes' theorem where one feature is independent of another feature in the same class. It is in person classification and many more.

f) k Nearest Neighbor:

k- Nearest neighbor is a supervised machine learning algorithm for solving of classification and regression problems. It is also called as lazy learning. It does not have any parameters, where it consists of available

RESEARCH ARTICLE

input and new input. By using this we can easily classify the new input data into categories based on the available input data.

g) Random Forest:

Random Forest is also a supervised machine learning algorithm used for solving both classification and regression problems. It works on the concept of ensemble learning where a multiple classifiers are combined together to solve a complex problem and improve the performance. It is used in the fields of banking, medicine, marketing, etc.

h) Dimensionality Reduction:

Dimensionality reduction is used for clustering algorithms. It is used for high dimensional data for removing of unwanted data from it. It can be both linear and non-linear. It is performed before model selection and evaluation.

i) K-means clustering:

K-means clustering is a unsupervised machine learning algorithm. It is also called as flat clustering algorithm. This is used for making of clusters based on the geometric distance between the points.

2. Literature Survey

Machine Learning Techniques:

The techniques used in machine learning are:

a) Classification

- i. Logistic Regression
- ii. SVM
- iii. Naïve Bayes
- iv. Decision Tree
- v. Random Forest
- vi. k-NN

b) Regression

- i. Linear Regression
- ii. k-NN
- iii. Decision Tree
- iv. Random Forest

c) Clustering

- i. K-Means Clustering
- ii. Hierarchical Clustering

RESEARCH ARTICLE

iii. Probabilistic Clustering

d) Dimensionality Reduction

- i. Principal Component Analysis (PCA)
- ii. SVD

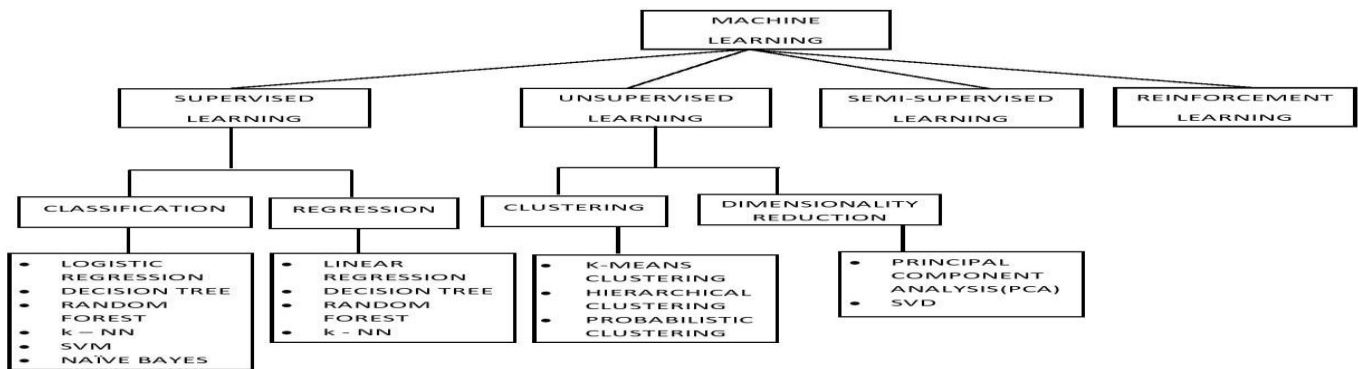


Figure 1: Classification of machine learning techniques

Applications:

Machine Learning is used in various fields for various purposes. Each application is useful in one or another way by which the works gets compact and get the desired result. It also helps in knowing the future result before the actual output is obtained. The most common applications of machine learning are:

a)Image Recognition

One of most common and heard application of machine learning is Image Recognition, which is used for identify the persons, places, etc. Face book is one of them which is using machine learning for image recognition. In this whenever we upload a photo, tagging suggestion with name is automatically generated which is one of the applications.

b)Speech Recognition

Speech Recognition is one of the natural communications, as our speech is much faster than our type. Alexa, Google Home, Siri, Amazon Echo are examples of devices which are using speech recognition. By using them we can easily find the information that is required. It is present in various platforms like Smart Speakers, Smart phones and Mobile Apps.

c) Disease identification or Diagnosis



RESEARCH ARTICLE

Machine learning is also showing its way in health sector also. It is used for diagnosis of diseases. By using machine learning in medical, we can build 3D models for predicting the exact position of the disease. It is mainly used in neurology for finding whether the person is having brain tumor or not and also for other brain – related diseases.

e) Weather Forecast

Machine learning is used for predicting, what weather condition is present now, what will be in the future so that precautions steps can be taken to reduce the loss or harm that can be caused due to that.

f) Product Recommendations

E-commerce and entertainment is the most frequently used social services. Amazon, Netflix, etc. are the e-commerce and entertainment companies using machine learning for various purposes. Whenever we search for any product or any movie etc. then we get advertisements related to that topic frequently because of machine learning. This is because, it understands the interest of the user and gives suggestions related to that as it will be easy for user to access.

Recent Trends of Machine Learning:

a) Improving attendance and workplace access using Facial Recognition:

In the present work place, the attendance of the employees is taken through biometric process, which is also one of the machine learning application, but in the future the facial recognition is going to replace the biometric process for cyber security as some of them are using other person's biometric scan. By the face recognition, the companies can keep track of the employees and solve issues related to password and access cards by saving time and costs.

b) The convergence of I o T and Machine learning:

Machine Learning and IoT are going to make a match a coming future. IoT devices collect and handle huge amount of data. These algorithms convert this data into required results which can be used by the IoT devices.

c) Rapid Adoption will Drive Growth:

There is huge growth of Machine Learning in the present environment, which leads to the replacement of workers due to the increase in demand of jobs in data science. In coming future everything is going to work with automation, where the use of mankind is going to decrease and is going to be ruled by machines only. For doing this, a lot of data science opportunities are on their way.

RESEARCH ARTICLE

3. Proposed Work:

Introduction

Exploratory Data Analysis: This is the step that is performed before evaluation of the dataset using the respective machine learning techniques. It is used for the understanding of each column present in the dataset by the visualization and graphical representation of data. We can explore each and every aspect of the dataset to get the in detail information about each attribute. Then the data is split into train data and test data, which is evaluated and the required predictions are made. Figure 1 shows the flowchart of the proposed framework.

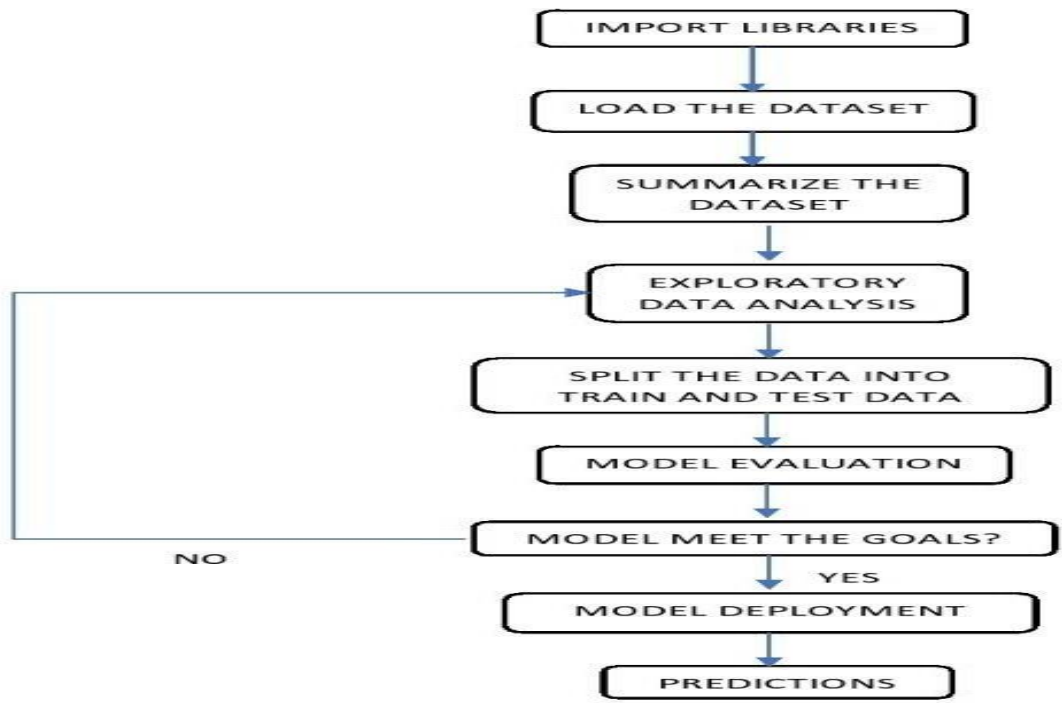


Figure 2: Flowchart of Proposed Framework

Step 1: Dataset

The dataset used here is the elections dataset which contains the details of the Indian candidates for General Election. It consists of the recent 2019 election where over 600 million voters voted for more than 8500 candidates in 543 constituencies. The columns that are present in the dataset are the details of state, constituency, name of candidates, winner, party name, party symbol, gender, candidates

International Journal Research in Applied Engineering, Science and Technology (IJRAEST) Impact Factor: 4.537(SJIF)

RESEARCH ARTICLE

having criminal cases, age of the candidates, category, education, assets, liabilities, general votes, postal votes, total votes, electors in constituency, votes polled in constituency and the total electors present. It consists of 2263 records. Firstly, we begin the process by loading this dataset. Figure 2 represents the dataset used.

State	Constituency	Criminal Cases	Gender	Party	Winner	Education	Total Votes
Telangana	ADILABAD	52	MALE	TRS	1	12 th Pass	377374
Telangana	ADILABAD	0	MALE	INC	0	Post Graduate	318814
Telangana	ADILABAD	3	MALE	NOTA	0		13030
Uttar Pradesh	AGRA	5	MALE	BJP	1	Doctorate	646875
Uttar Pradesh	AGRA	0	FEMALE	BSP	0	Post Graduate	45149
Maharashtra	AHMADNAGAR	0	MALE	BJP	1	Doctorate	704660
Maharashtra	AHMADNAGAR	1	MALE	NCP	0	Graduate	423186
Maharashtra	AHMADNAGAR	0	MALE	VBA	0	Graduate	31807
Gujarat	AMRELI	0	MALE	BJP	1	12 th Pass	529035
Gujarat	AMRELI	0	MALE	INC	0	Graduate	327604
Gujarat	AMRELI			NOTA	0		
Rajasthan	BANSWAR	0	MALE	BJP	1	Graduate	711709
Rajasthan	BANSWAR	0	MALE	INC	0	Post Graduate	406245
Kerala	CHALAKUDY	4	MALE	INC	1	Graduate	473444
Kerala	CHALAKUDY	0	MALE	CPI(M)	0	8 th Pass	341170
Kerala	CHALAKUDY	6	MALE	BJP	0	Graduate	154159
West Bengal	COOCH BEHAR	11	MALE	BJP	1	8 th Pass	731594
West Bengal	COOCH BEHAR	0	MALE	AITC	0	12 th Pass	677363

RESEARCH ARTICLE

West Bengal	COOCH BEHAR	0	FEMALE	INC	0	Post Graduate	28215
Karnataka	DAKSHINA	4	MALE	BJP	1	10 th Pass	774285
Karnataka	DAKSHINA	1	MALE	INC	0	Graduate	499664

Table 1: Sample data from election dataset

Step 2: Summarization

After loading the dataset, we summarize the data to know the details of the dataset and then we perform the cleaning of unwanted data in the dataset.

3.3.1 Number of rows and columns.

(2263, 19)

3.3.2. Statistics of the numeric data in the dataset.

	WINNER	AGE	GENERAL\nVOTES	POSTAL\nVOTES	TOTAL\nVOTES	OVER TOTAL ELECTORS \nIN CONSTITUENCY	OVER TOTAL VOTES POLLED \nIN CONSTITUENCY	TOTAL ELECTORS
count	2263.000000	2018.000000	2.263000e+03	2263.000000	2.263000e+03	2263.000000	2263.000000	2.263000e+03
mean	0.238179	52.273538	2.615991e+05	990.710561	2.625898e+05	15.811412	23.190525	1.658016e+06
std	0.426064	11.869373	2.549906e+05	1602.839174	2.559822e+05	14.962861	21.564758	3.145187e+05
min	0.000000	25.000000	1.339000e+03	0.000000	1.342000e+03	0.097941	1.000039	5.518900e+04
25%	0.000000	43.250000	2.103450e+04	57.000000	2.116250e+04	1.296518	1.899502	1.530014e+06
50%	0.000000	52.000000	1.539340e+05	316.000000	1.544890e+05	10.510553	16.221721	1.679030e+06
75%	0.000000	61.000000	4.858040e+05	1385.000000	4.872315e+05	29.468185	42.590233	1.816857e+06
max	1.000000	86.000000	1.066824e+06	19367.000000	1.068569e+06	51.951012	74.411856	3.150313e+06

Figure 3: Statistics of numeric data

RESEARCH ARTICLE

3.3.3 Complete Overview of Dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2263 entries, 0 to 2262
Data columns (total 19 columns):
STATE                2263 non-null object
CONSTITUENCY        2263 non-null object
NAME                2263 non-null object
WINNER             2263 non-null int64
PARTY              2263 non-null object
SYMBOL            2018 non-null object
GENDER            2018 non-null object
CRIMINAL          2018 non-null object
CASES             2018 non-null object
AGE               2018 non-null float64
CATEGORY          2018 non-null object
EDUCATION         2018 non-null object
ASSETS            2018 non-null object
LIABILITIES       2018 non-null object
GENERAL
VOTES             2263 non-null int64
POSTAL
VOTES            2263 non-null int64
TOTAL
VOTES            2263 non-null int64
OVER TOTAL ELECTORS
IN CONSTITUENCY  2263 non-null float64
OVER TOTAL VOTES POLLED
IN CONSTITUENCY  2263 non-null float64
TOTAL ELECTORS   2263 non-null int64
dtypes: float64(3), int64(5), object(11)
memory usage: 336.0+ KB
```

Figure 4: Dataset information

3.3.4. Checking and Dropping of Null values.

STATE	0	STATE	0
CONSTITUENCY	0	CONSTITUENCY	0
NAME	0	NAME	0
WINNER	0	WINNER	0
PARTY	0	PARTY	0
SYMBOL	245	SYMBOL	0
GENDER	245	GENDER	0
CRCASES	245	CRCASES	0
AGE	245	AGE	0
CATEGORY	245	CATEGORY	0
EDUCATION	245	EDUCATION	0
ASSETS	245	ASSETS	0
LIABILITIES	245	LIABILITIES	0
GENVOTES	0	GENVOTES	0
POSTVOTES	0	POSTVOTES	0
TOTVOTES	0	TOTVOTES	0
ELECTORS_CONSTITUENCY	0	ELECTORS_CONSTITUENCY	0
VOTES_CONSTITUENCY	0	VOTES_CONSTITUENCY	0
TOTELECTORS	0	TOTELECTORS	0
dtype: int64		dtype: int64	

Before Dropping of Values **After Dropping of Values**

Figure 5: Before and after dropping null values

RESEARCH ARTICLE

Step 3: Exploratory Data Analysis

Here we analyze the data, by which we can get know the required attributes by which we get the required result.

Education:

Number of candidates with different education qualification.

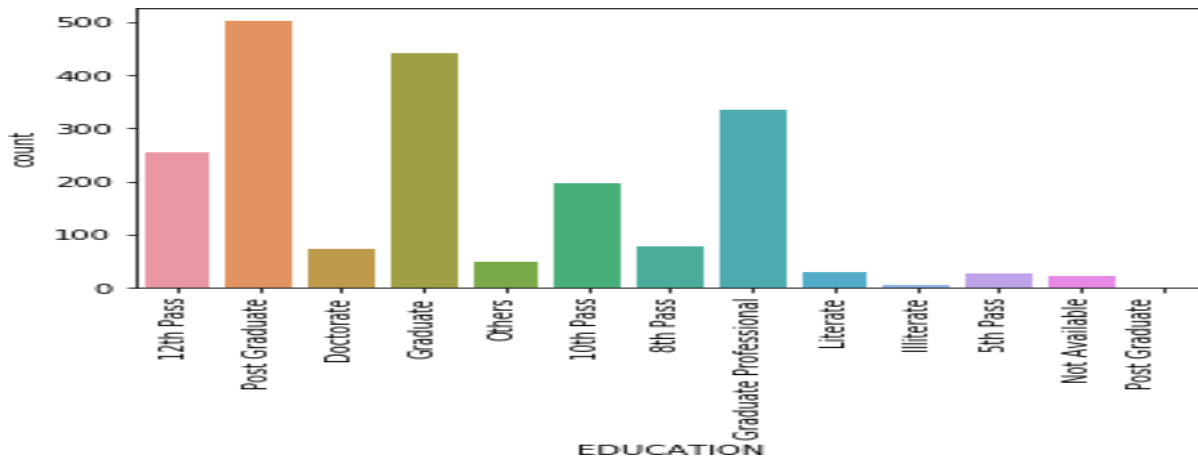


Figure 6: Education qualification of each candidate

Category:

Number of candidates of different categories.

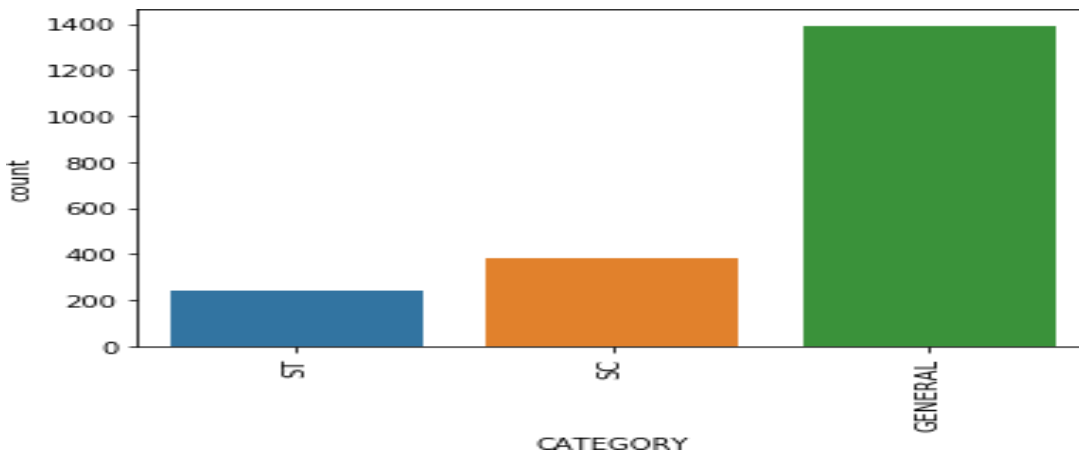


Figure 7: Category count

Gender vs Winner: Number of winners and losers in both male and female.

RESEARCH ARTICLE

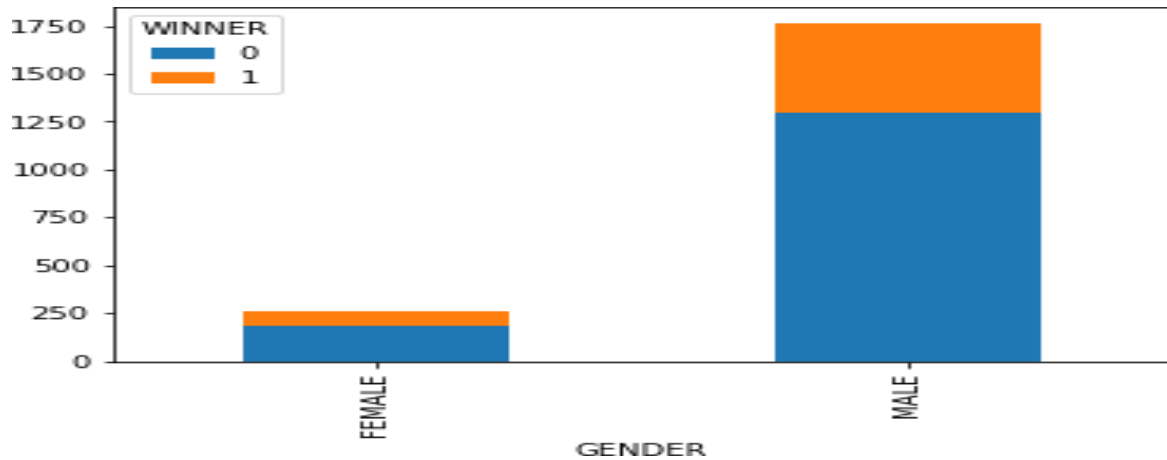


Figure 8: Gender vs Winner

Gender: Number of male and female candidates who contested in elections.

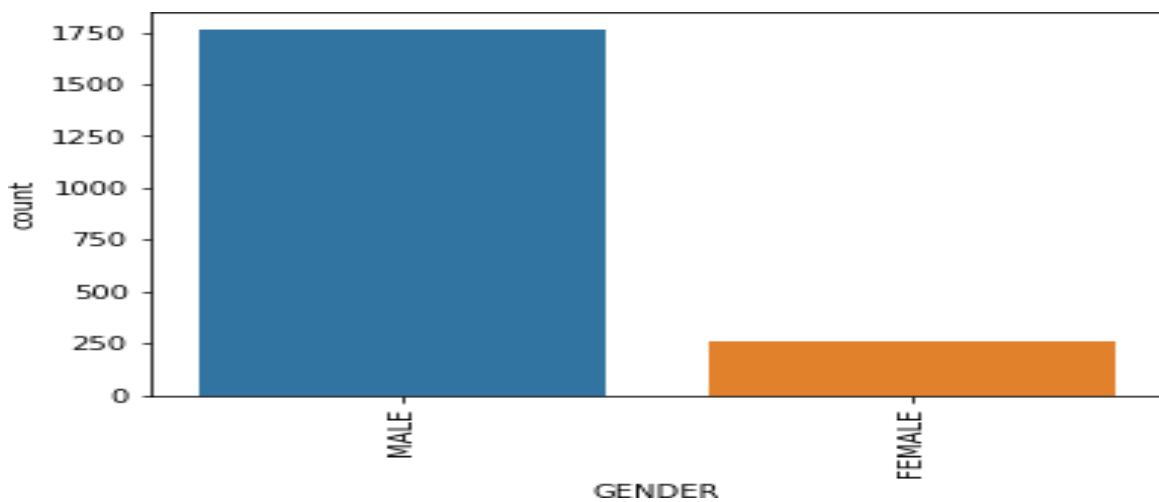


Figure 8: Gender vs Winner

Gender: Number of male and female candidates who contested in elections.

RESEARCH ARTICLE

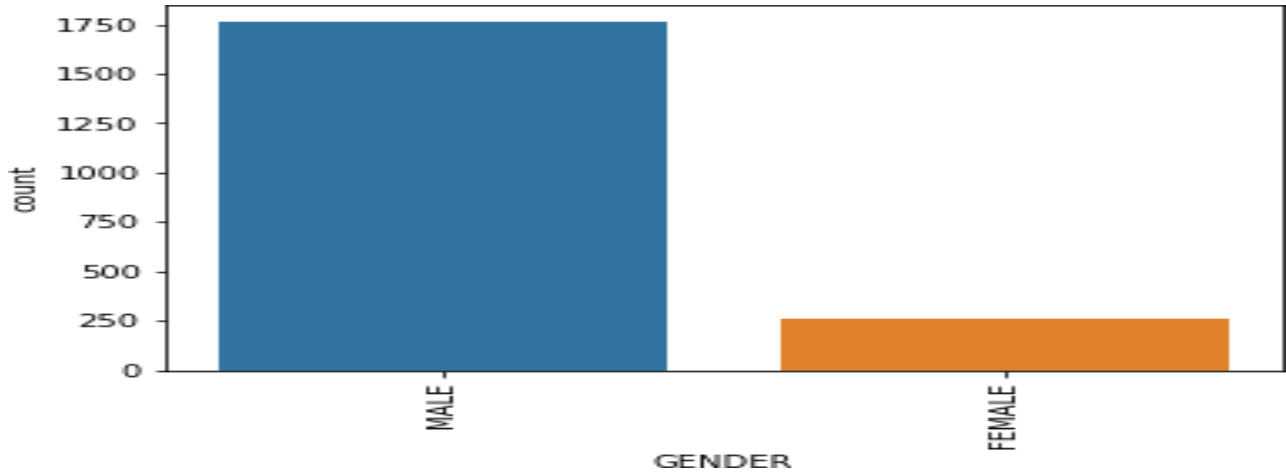


Figure 9: Gender count

State vs Winner:

To predict the number of winners and losers in each state.

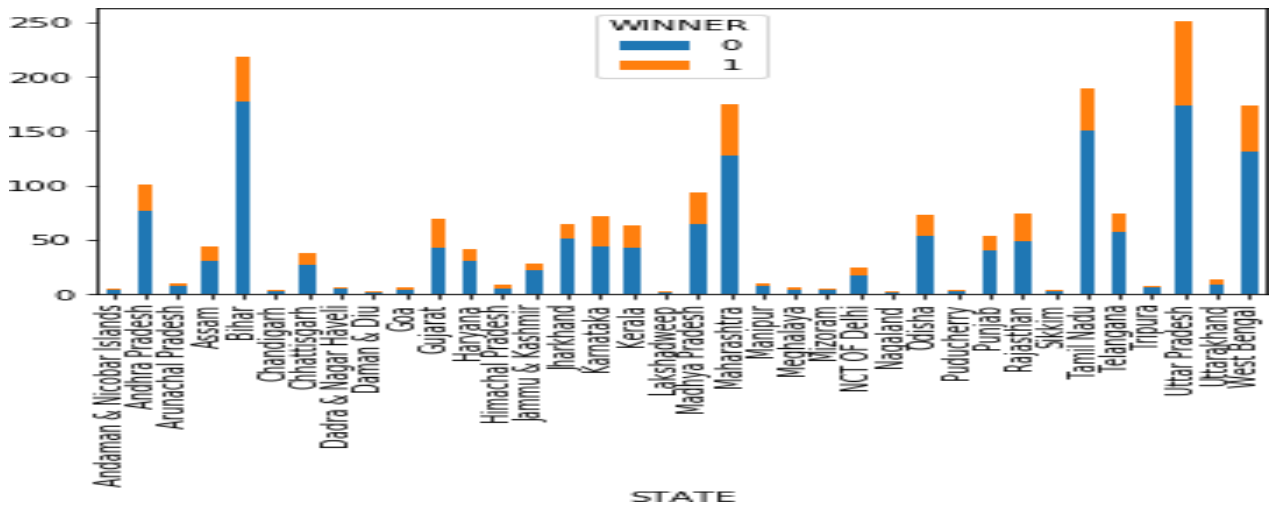


Figure 10: State vs Winne

State vs Total Votes: Total votes received by each state.

RESEARCH ARTICLE

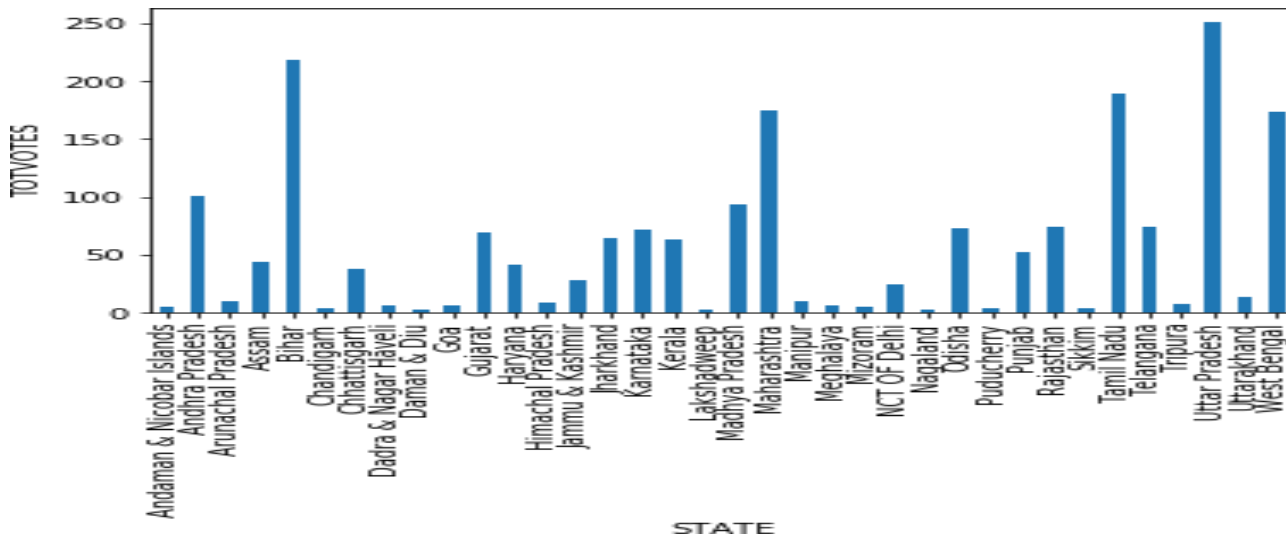


Figure 11: State vs Total Votes State vs Constituency:
To know the number of constituencies, present in each state

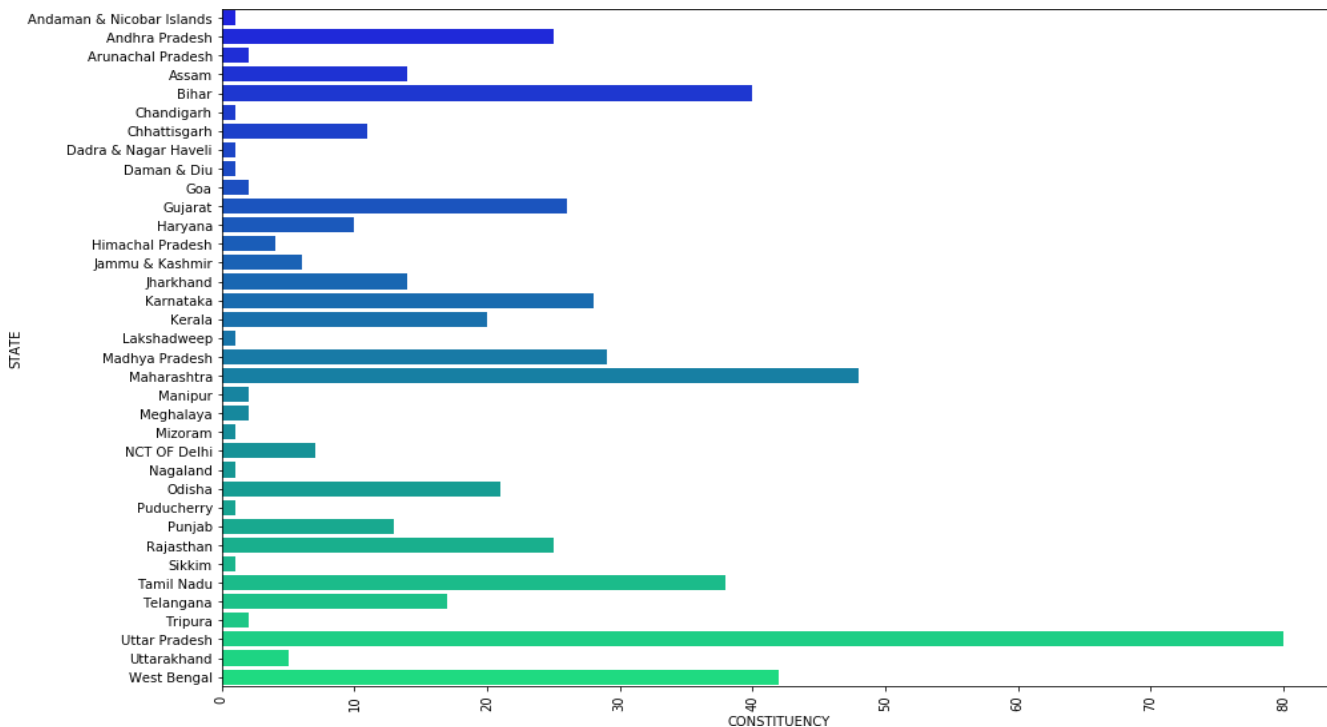


Figure 12: Constituencies per stat

RESEARCH ARTICLE

Step 4: Evaluation and Prediction

After the data analysis, we get to know the attributes that give us the result. Then using those attributes, we perform the evaluation of the dataset. First, before model evaluation, we split the data into training and testing data. The trained data is used for evaluating the dataset using respective techniques. The test data is used to test whether the technique used is giving the same result as that of train data. If the model obtained is suitable then the prediction of result takes place. For this data set we are using Logistic Regression, Naïve Bayes, k Nearest Neighbors, Decision Tree Classifier, Random Forest Classifier and Support Vector Machine. The accuracy score obtained by these techniques is useful for finding the suitable algorithm for predicting the result.

Accuracy of Train Data:

Logistic Regression:	0.9001416430594901
Naive Bayeas:	0.8711048158640227
K Nearest Neighbors:	0.9050991501416431
Decision Tree:	0.9525495750708215
Random Forest:	0.9525495750708215
Support Vector Machine:	0.9050991501416431

Accuracy of Test Data:

Logistic Regression:	0.8844884488448845
Naive Bayeas:	0.8432343234323433
K Nearest Neighbors:	0.8877887788778878
Decision Tree:	0.9438943894389439
Random Forest:	0.9471947194719472
Support Vector Machine:	0.8894389438943895

Figure 13: Accuracy of different ML techniques

4. Conclusions:

In our paper, we performed the Exploratory Data Analysis i.e. EDA for visualizing the data in graphical representation. After, the evaluation of the dataset with different algorithms techniques, we obtained to the result that the Random Forest Classification with accuracy of 95.2% for train data and 94.7% for test data and Decision Tree Classification with accuracy of 95.2% for train data & 94.3%



EJ Publications

International Journal Research in Applied Engineering, Science and Technology (IJRAEST) Impact Factor: 4.537(SJIF)

An International Peer-Reviewed Journal

Vol-2, Issue-1, 2020

www.ijraest.com Indexed in: Google Scholar, Cite Factor

ISSN: 2582-029X

RESEARCH ARTICLE

for test data, are the most suitable for predicting our result. Among them, the random forest is best technique to get accurate and perfect result. In our project we got our required result with an accuracy of 94.7%.

References:

- [1] <https://www.kaggle.com/prakrutchauhan/indian-candidates-for-general-election-2019>
- [2] Sumit Das, Aritra Dey, Akash Pal, Nabamita Roy; “Applications of Artificial Intelligence in Machine Learning: Review and Prospect”; Volume 115(9); April 2015;0975-8887.
- [3] https://en.wikipedia.org/wiki/Linear_regression
- [4] Ayon Dey, “Machine Learning Algorithms: A Review”; Vol. 7(3); 2016; 1174-1179.