Journal Homepage: -www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI:10.21474/IJAR01/19477
DOI URL: http://dx.doi.org/10.21474/IJAR01/19477

RESEARCH ARTICLE

DECODING THE BALLOT: PREDICTING INDIAN GENERAL ELECTIONS WITH MACHINE LEARNING

Gaurav R. Mandhyan[1] and Shreea Bose[2]

1. Indus International School Pune, Pune, Maharashtra, 411057, India.
2. Department of Computer Science and Information Systems, BITS Pilani, Hyderabad.

Manuscript Info

Manuscript History
Received: 15 July 2024
Final Accepted: 17 August 2024
Published: September 2024

Key words:-
Machine Learning, Election Forecasting, Indian General Elections, SHRUG, Random Forest

Abstract

This paper explores the complexities of predicting election outcomes in India, focusing on the winning party and the probability of incumbent reelection. Leveraging historical voting data and socio-economic indicators from the Socioeconomic High-resolution Rural-Urban Geographic (SHRUG) dataset and the Lok Dhaba database, the study employs advanced machine learning models to forecast electoral results. The main goal of this paper is to find these models' ability to forecast the victorious party and determine the likelihood of reelection is the main goal. Several models, including Random Forest, Gradient Boosting, and Decision Tree, were assessed to meet these goals. With an accuracy of 99.89%, the Random Forest model outperformed the rest of them. This is because of its ensemble learning strategy, which lowers overfitting and increases predictive power. Additionally successful were the Decision Tree and Gradient Boosting models, which yielded accuracies of 98.75% and 99.78%, respectively. The study faced challenges such as computational complexity and potential bias introduced by the dataset, particularly due to the historical dominance of the Indian National Congress (INC) party. Despite these challenges, the models provided valuable insights into voter behaviour and electoral trends. The implications of this study are significant for political analysts and campaign strategists. Accurate predictions can guide the development of targeted campaign strategies and enhance understanding of electoral dynamics. Future research should address dataset biases and explore more efficient algorithms to improve the robustness and applicability of these predictions in real-world scenarios.

Introduction:-

The Indian General Elections

The Indian general elections, renowned for their scale and complexity, are among the most significant democratic exercises globally. This historic event, which has over 900 million eligible voters, is held every five years to decide the makeup of the Lok Sabha, the lower house of Parliament. The election procedure is well thought out, with several steps and strict supervision to guarantee its accuracy. An overview of the many election phases, the

Corresponding Author:-Gaurav R. Mandhyan
Address:-Indus International School Pune, Pune, Maharashtra, 411057, India.

governing bodies participating, integrity-maintaining procedures, and the general significance of these elections in preserving democratic values in India are all intended to be covered in this section.

Election Process

In India, general election preparations begin many months in advance of the actual polling dates, and the Election Commission of India (ECI) is a key player in coordinating the whole process. Making sure that every eligible voter is properly registered is one of the ECI's main duties, along with updating the electoral roll with precision. This includes deleting the names of people who have passed away or are otherwise ineligible and adding new voters who have just turned 18 to the voter rolls.

The ECI carries out thorough verification drives to accomplish this. During these drives, authorized officials verify voters' information door-to-door and gather the required paperwork. Campaigns for public awareness have also started to urge people to check their information and, if necessary, register to vote. To ensure that no eligible voter is left out, the ECI also works with a variety of governmental and non-governmental groups to reach out to isolated and vulnerable areas.

The formal commencement of the election process is marked by the ECI's announcement of the election dates. This announcement triggers the implementation of the Model Code of Conduct — a set of guidelines meticulously designed and enforced by the ECI to ensure fair play among political parties and candidates. Following this, the ECI issues a formal notification detailing the schedule for the various phases of polling. This notification establishes the timeline for all subsequent activities related to the elections, including the nomination process, campaigning, and polling days.

In the nomination phase, candidates from various political parties and independents file their nominations. This phase is followed by a rigorous scrutiny process where the ECI examines the validity of each nomination. Candidates must meet specific eligibility criteria, including citizenship, age, and the absence of disqualifications as specified in the Representation of the People Act, 1951. Additionally, the nomination phase also involves the submission of candidates' affidavits detailing their assets, liabilities, criminal records (if any), and educational qualifications. These affidavits are made public, enabling voters to make informed decisions about the candidates. The scrutiny process, conducted by Returning Officers, involves verifying the information provided in the nominations and affidavits. Any discrepancies or false information can lead to rejecting a candidate's nomination.

Following the nomination phase, campaigning is a critical phase where candidates and political parties communicate their policies and promises to the electorate. Campaigning methods include rallies, door-to-door visits, media advertisements, and increasingly, social media outreach. The ECI sets expenditure limits for candidates to prevent excessive spending and ensure a level playing field. Additionally, the ECI vigilantly monitors campaign activities to detect any violations of the Model Code of Conduct, particularly hate speech, false information, or any content that could incite violence or disharmony. This includes monitoring speeches, advertisements, and social media posts. The ECI also sets up a grievance redressal mechanism to address complaints from candidates, parties, and voters regarding any violations during the campaigning phase, further ensuring a safe and fair electoral process.

Polling & Counting

Polling is conducted in multiple phases, a strategy designed to manage the vast voter base and ensure adequate security. Polling stations are set up across the country, each equipped with Electronic Voting Machines (EVMs) and Voter Verifiable Paper Audit Trails (VVPATs). These machines, which enhance transparency and efficiency, are a testament to the fairness of the process. Voters secretly cast their ballots, ensuring the confidentiality of their choices. On the polling day, a well-coordinated system ensures that voters can cast their votes smoothly. Polling personnel, including presiding officers and polling officers, are deployed to manage the process. Security personnel are stationed at polling stations to maintain order and prevent any incidents of violence or intimidation. Special provisions are made for differently-abled voters, senior citizens, and women to ensure that they can vote comfortably and securely.

After the polling concludes, EVMs are securely transported to designated counting centres. The counting process is meticulously conducted under the strict supervision of the ECI. Results are declared constituency-wise, and the party or coalition with the majority of seats is invited to form the government. VVPAT slips are counted for a randomly selected sample of polling stations to cross-verify the results from EVMs.

Constituencies
India's electoral framework is organized into parliamentary and state constituencies, enabling a structured and democratic representation system. The country is divided into 543 parliamentary constituencies for the Lok Sabha, the lower house of Parliament, each represented by a single Member of Parliament (MP). These constituencies are distributed among the states and Union Territories based on population, with some constituencies reserved for Scheduled Castes (SC) and Scheduled Tribes (ST).

Party System
India operates a multi-party system in its general elections, characterized by a plethora of national and regional parties vying for political power. The two most prominent national parties are the Bharatiya Janata Party (BJP) and the Indian National Congress (INC). These parties have historically dominated the political landscape, with the BJP advocating for conservative and nationalist policies, while the INC typically promotes a more centrist and secular agenda. In addition to the national parties, numerous regional parties play a crucial role in the Indian political arena. These parties often cater to specific state or regional interests and can wield significant influence in forming coalition governments. Examples include the All India Trinamool Congress (AITC) in West Bengal, the Dravida MunnetraKazhagam (DMK) in Tamil Nadu, and the Samajwadi Party (SP) in Uttar Pradesh. These regional entities can affect the balance of power, particularly in hung parliaments where no single party gains an outright majority.

Literature Review:-

Trying to predict elections has been a long-standing field wherein countless studies have approached the problem with different methods. One pivotal work by Andreas Graefe compares vote expectation surveys with expert judgment, traditional polls, prediction markets, and quantitative models. Analyzing 217 surveys from 1932 to 2012, Graefe found that vote expectation surveys had an 89% correct prediction rate, reducing errors by 51% compared to traditional polls and 6% to prediction markets, especially in the last 100 days before elections [1]. Another study introduced the Minimal Influence Gap (MIG) metric, which is effective in homophilic networks, with a Pearson Correlation Coefficient (PCC) of 0.8278 [2].

Social media data has become a crucial tool in predicting election outcomes. A study on Indian elections utilized Twitter data and sentiment analysis, achieving an accuracy of 86.3% in predicting the BJP's victory using various machine-learning models [3]. Another study on the 2024 Indian elections employed algorithms like Naïve Bayes, SVM, and BERT, with Random Forest Regression integrating social media analytics with traditional voter demographics [5].

A systematic review of 83 studies from 2008 to 2019 highlighted Twitter volume and sentiment analysis challenges, often finding regression methods trained with traditional polls more reliable [8]. Another study analyzed Twitter sentiment data for the Indian general elections, collecting tweets from January to March 2019. Using R for sentiment analysis, it concluded that 'Candidate-1' was more favoured than 'Candidate-2'. These predictions aligned closely with the election outcomes in May 2019 [4].

A study using frequentist and Bayesian approaches with state-level pre-election polls demonstrated that simulations could predict outcomes effectively, with a 70.3% probability of Bush winning the 2004 election. This study underlines the critical importance of quality and timing in polling data [6]. Additionally, research using social media data, including RSS feeds and Twitter, found that social media predictions closely aligned with exit polls, relying heavily on the volume and sentiment of social media data [7].

Datasets
In this study, we will utilize two datasets: SHRUG Trivedi election-level and Lok Dhaba.

Shrug
The Trivedi Election dataset, part of the SHRUG collection, is a comprehensive resource that provides detailed information on Indian elections at the constituency level. It encompasses data from both parliamentary and state assembly elections, capturing various aspects such as candidate details, party affiliations, vote shares, and election outcomes. The dataset also includes information on demographic and socioeconomic variables, used in analyzing voting patterns and election results. This comprehensive nature is designed to facilitate in-depth research into the dynamics of electoral politics in India, allowing for the study of factors influencing electoral success and the role of development in political incumbency. [9] [10] [11] [12]

Lok Dhaba

Lok Dhaba, hosted by Ashoka University, is an extensive database offering detailed electoral data for General Elections in India. This platform provides a user-friendly interface, designed to make browsing and analyzing election data across various levels, including parliamentary and state assembly elections, a seamless experience. Users can filter data by election type, state, and year, allowing for targeted exploration of voting patterns, candidate details, party performance, and constituency-level results. The database includes comprehensive information such as vote shares, margins of victory, and demographic insights, which are crucial for political analysis and research. Lok Dhaba aims to democratize access to election data, supporting scholars, policymakers, and citizens in understanding the complexities of India's electoral landscape.

The Trivedi Election dataset, part of the SHRUG framework, is designed for integration with a wide array of socioeconomic indicators, allowing researchers to explore the interplay between development and electoral outcomes. However, Lok Dhaba does not inherently integrate with broader socioeconomic datasets, focusing instead on delivering detailed electoral data in an accessible format. [13] [14]

Proposed Model

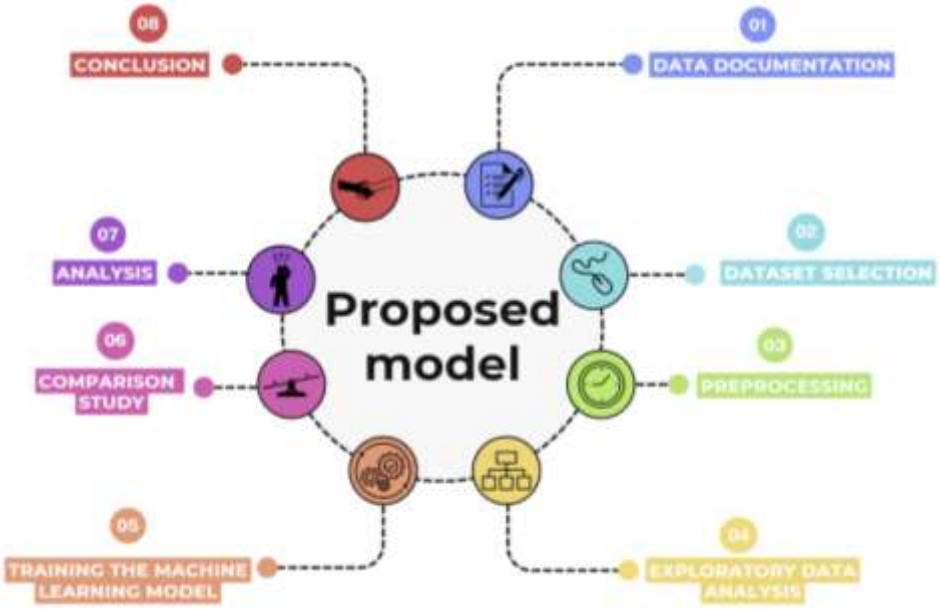

Figure 1:- A proposed workflow for the model.

Data collection

This study used secondary data collection techniques to collect comprehensive and reliable data sets. We used two main data sets: SHRUG and Lok Dhaba.

The Lok Dhaba dataset was very useful for analyzing electoral outcomes because it provided all the necessary variables for calculating incumbents' probability of re-election. This provided quantitative support for understanding electoral dynamics and voter behaviour patterns. The electoral data provided here was very detailed and formed the basis upon which the incumbent re-election probabilities were statistically analyzed and modelled.

In sharp contrast, the SHRUG data set was crucial for predicting the winning party. This prediction was made by incorporating the Trivedi Elections Data and Forest Cover Data under the SHRUG framework. The Trivedi Elections data held historical information related to past elections. Forest Cover Data incorporated historical environmental context that could have affected the trends in voting patterns and, subsequently, the success of a political party. We combined these data to make a predictive model that combined political and environmental variables to predict election outcomes.

Data Preprocessing
The first step in processing the SHRUG datasets was to merge the VCF dataset with the Trivedi dataset using the SHRUG keys given.

Once the two datasets had been joined, the same preprocessing procedure was applied to both the Lok Dhaba and the SHRUG datasets. An important part of this preprocessing was converting categorical values into numerical labels. This was done using LabelEncoder, which converted all the non-numeric columns to numeric values.

LabelEncoder is a preprocessing tool that changes categorical labels into numeric ones, i.e., if some column had categories like ['JKN', 'JKPDP', 'BJP', 'INC'], LabelEncoder would convert them into unique numeric values such as [0, 1, 2, 3], respectively. It will be transformed so that the algorithms can interpret categorical data in a numerical form.

We very selectively chose only those features relevant to our research objectives to reduce dimensionality. This is done by checking the importance of every feature we have selected and filtering out the irrelevant and redundant ones, thus ensuring that our models are efficient and effective. To add to this point, the missing data problem was catered for by dropping out rows with null values. Missing values can be a massive burden on analyses and model training.

Data Sampling
In our analysis, the datasets we were working with, namely the SHRUG and Lok Dhaba datasets, were exceptionally large, posing significant computational challenges. Given the limitations of our computational resources, processing the entire dataset in a single pass was infeasible. To address this, we employed random sampling to create a representative subset of the data. Random sampling ensures that each data point has an equal probability of being selected, which helps preserve the statistical properties and diversity of the original dataset. By using this method, the danger of overfitting is reduced and the analysis's conclusions are kept reliable and applicable to a wider audience.

Standard Scaler
As part of our preprocessing pipeline, we used the `StandardScaler`, which standardizes our data. More explicitly, the `StandardScaler` de-means and scales the data to unit **variance:** $\sigma^2 = 1$

This operation is mathematically defined as $z = \frac{(x-\mu)}{\sigma}$, where z is the standardized value, x is the original value, and μ and σare the parameters of the mean and standard deviation. This is desirable because most machine-learning algorithms, especially those based on gradient descent, tend to perform very well when the features are centred around zero with unit variance. This normalization process ensures that a bias in the learning process does not go towards those features whose values are of higher magnitude.

Methodology:-
In the following portion, we will describe and evaluate the effectiveness of various models for forecasting the winning party. We used many machine learning methods, such as Random Forest, Logistic Regression, Naive Bayes, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Gradient Boosting, Decision Tree, AdaBoost, and Neural Networks. Each model's performance is evaluated using accuracy metrics, and the results are visualized as charts indicating accuracies and predicted winning parties. Detailed assessments of the best and worst-performing models are offered, with a focus on Random Forest, Logistic Regression, and Naive Bayes.

Random Forest
Random Forest is an ensemble learning method that creates numerous decision trees during training and returns the mode of the classes for classification. This improves the model's accuracy and ability to control overfitting by averaging predictions from a large number of trees.

Random Forest fits the decision tree based on a collection of instances of the training set. Training employs bootstrapping, with each tree using roughly two-thirds of the entire dataset. The rest will be used for OOB (out-of-bag) estimation to ensure that random forest fitting is based on various parts of the data.

Mathematically, the prediction of a Random Forest for a classification task is the majority vote of predictions from individual decision trees. Let there be T trees in the forest and $h_t(X)$ be the prediction of the t-th tree for input X. The final $\hat{Y}$ prediction is:

$$\hat{Y} = \text{mode}\{h_t(X) \mid t = 1,2,\ldots,T\}$$

Each of the decision trees in the forest splits the data based on features to reduce impurity, using measures such as Gini impurity or entropy. For example, Gini impurity for a node m can be defined as:

$$\mathbf{Gini(m) = 1 - \sum_{i=1}^{C} p_i^2}$$

Where p_i is the proportion of instances of class i in node m, and C is the total number of classes. The algorithm identifies splits that minimize the impurity of the child nodes.

In our work, the Random Forest was employed to predict the winning political party according to several socio-economic indicators. The Natural Averaging within the Ensemble Method reduces the Variance further and enables better prediction, particularly in our datasets which have high dimensions and complexities. Random Forest made classification possible by finding out the mode of predictions for all trees making the model less prone to overfitting as compared with a single tree. Thus the Random Forest model trumped all other models by achieving an accuracy of 99.8%.

Logistic Regression

Logistic Regression is a statistical method for binary classification that models the likelihood of belonging to one class or the other. Logistic Regression is essentially a Sigmoid function that converts a linear collection of input features into a probability score ranging from 0 to 1.

The sigmoid function is defined as:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

where $\sigma(z)$ is the probability of the positive class, and z is the linear combination of input features, defined as:

$$z = w \cdot X + b$$

Here, w is the weight vector, X represents the feature vector, and b is the bias term. For the logistic regression model, it estimates these parameters w and b through the likelihood maximization of the observed data.

We are predicting whether the election party will win or not using the logistic regression model, based on forest cover. A $\sigma(z)$ is translated into probability output; it indicates the likelihood with which a certain party is the winner. If $\sigma(z)$ is close to 1, it indicates a very high probability for that party to win; if it is close to 0, it indicates a very low probability. However, our Logistic Regression model got an accuracy of 44%, which is lower compared to the others. This is because Logistic Regression assumes a linear relationship between the input features and the log odds of the outcome. Moreover, it does not take into account interactions between features that could be crucial in understanding the dynamics of electoral outcomes.

Additionally, patterns and relationships in our dataset are highly likely to be complex and cannot be captured well by Logistic Regression, a simple model. Better handling of this kind of complexity can be through the use of more sophisticated models like Random Forest or Gradient Boosting. This claim is also backed by our findings wherein Random Forest achieved an accuracy of 99.8% and Gradient Boosting received an accuracy of 99.7%. In addition, with significant class imbalance in data (i.e., where some parties win much more frequently than the others), the model may not be able to learn about the minority class from data using Logistic Regression.

Naive Bayes

Naive Bayes is a probabilistic classifier based on Bayes' Theorem, which holds that the presence of one feature in a class is unaffected by the presence of any other feature. This assumption is referred to as conditional independence.

Bayes' Theorem is defined as:

$$P(y|X) = \frac{P(X|y) \cdot P(y)}{P(X)}$$

where $P(y|X)$ is the posterior probability of class y given the feature vector X, $P(X|y)$ is the likelihood of the feature vector X given class y, $P(y)$ is the prior probability of class y, and $P(X)$ is the probability of the feature vector X.

The model calculates the posterior probability for each class (party) and assigns the class with the highest probability to the input data. However, our Naive Bayes model achieved an accuracy of only 49.7%. This relatively modest performance can be attributed to several factors. Firstly, the Naive Bayes assumption of conditional independence is often violated in real-world data, where features can be correlated. In our dataset, various indicators like winner votes, turnout percentage, and VCF mean are likely interdependent, making the conditional independence assumption unrealistic. Secondly, Naive Bayes can struggle with complex patterns and non-linear relationships in the data. Our dataset likely contains such complexities, which the Naive Bayes model, due to its simplicity, may not effectively capture. Additionally, Naive Bayes is particularly sensitive to the distribution of data. If the features do not follow a normal distribution, the performance of the Gaussian Naive Bayes variant can be suboptimal. Thus, Naive Bayes provided a baseline performance, helping to identify that more sophisticated models might yield better results.

Results and Discussion:-

Why we chose forest cover data

In our analysis, we chose forest cover data to predict the winning party in various electoral districts. This decision was based on the hypothesis that environmental factors, particularly forest cover, have significant socio-economic and political implications. Forest cover can influence the livelihoods of local populations, impact regional economies, and shape public opinion on environmental and developmental policies, which in turn can affect electoral outcomes.

The forest cover data used in our study spans from 2001 to 2020 and is sourced from the Vegetation Continuous Fields (VCF) dataset, a MODIS product that measures tree cover at a 250mm resolution. The VCF dataset is generated using a machine learning algorithm that processes broad-spectrum satellite images, trained with human-categorized data to distinguish between crops, plantations, and primary forest cover. This high-resolution, long-term data provides a comprehensive view of changes in forest cover over nearly two decades.

Figure 4, "State vs Sum of Forest Cover Over Years," illustrates the trends in forest cover across different states from 2001 to 2020. The y-axis represents the sum of forest cover, while the x-axis denotes the years. Each line represents a different state, showcasing the variations in forest cover over time. For instance, some states exhibit significant fluctuations in forest cover, while others maintain relatively stable levels. This variability in forest cover changes can be attributed to factors such as deforestation, afforestation, urbanization, and agricultural expansion. From the graph, it is evident that states like State 1 and State 3 have relatively high forest cover, which might correlate with different socio-economic and political dynamics compared to states with lower forest cover. By analyzing these trends, we can infer potential influences on voter behaviour and electoral outcomes. For example, states with increasing forest cover might reflect successful environmental policies, influencing voters to support the incumbent government or party responsible for these policies.

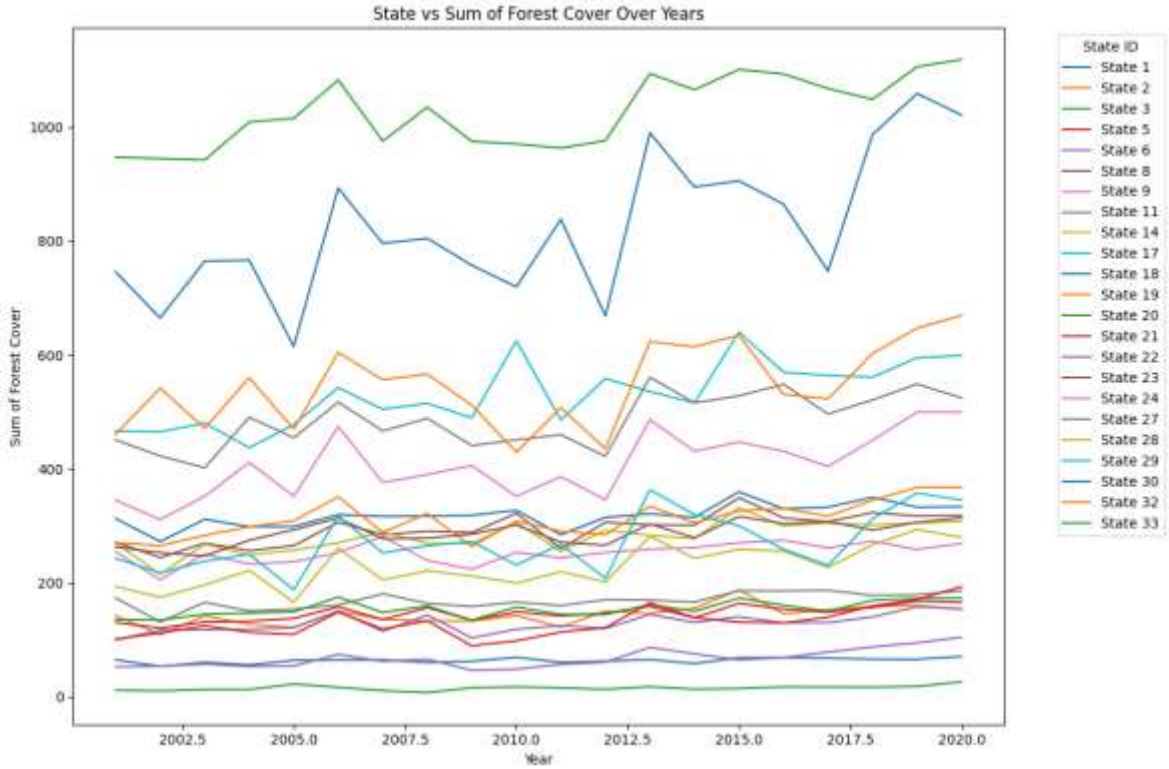

Figure 2:- A line graph showcasing the various years vs forest cover of each state.

Exploratory Data Analysis

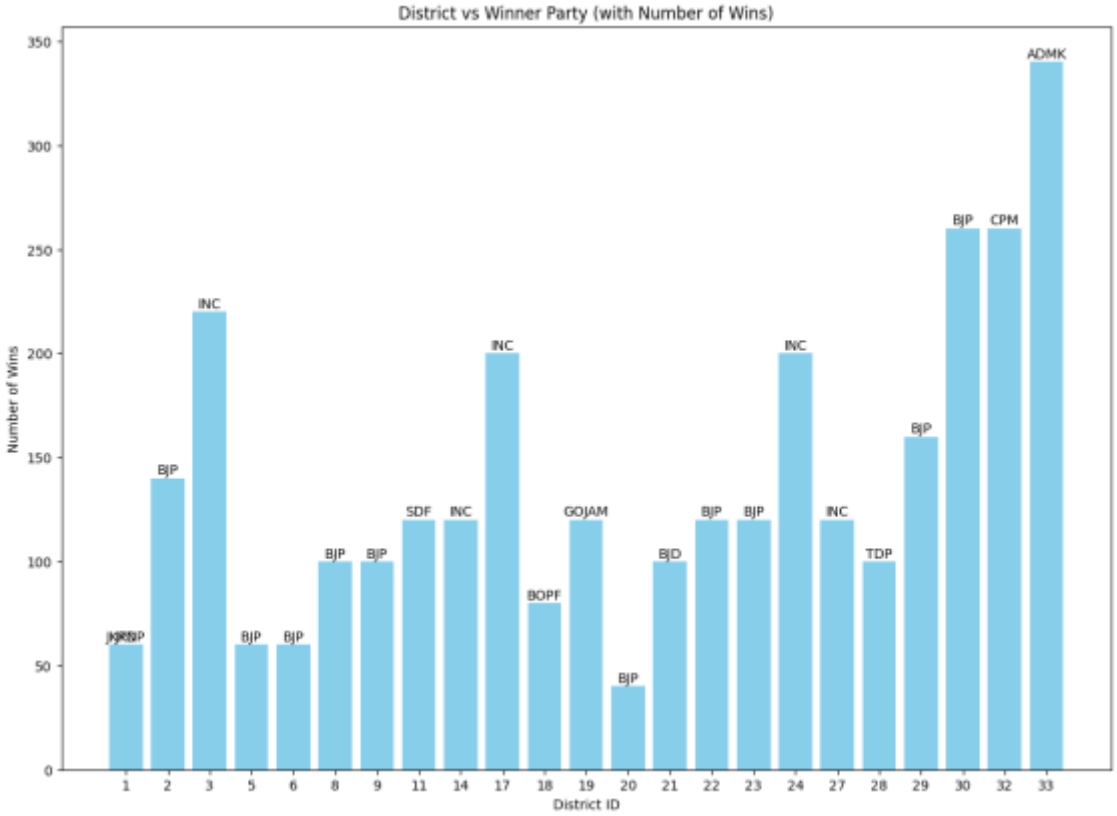

Figure 3:- A Bar Graph between District Vs Winner Party.

The graph "District vs Winner Party (with Number of Wins)" illustrates the distribution of election victories among various political parties across different districts.The x-axis indicates the district IDs, and the y-axis shows the number of wins. Each bar represents the most times a specific party has won in a given district. From the graph, we can observe that certain districts are dominated by specific parties. For example, District 33 shows a significant number of wins by ADMK, reaching over 350 victories. Similarly, INC has a strong presence in Districts 3, 17, and 24, with each district showing more than 200 wins. BJP appears to be a dominant party in multiple districts, including Districts 2, 5, 6, 8, 23, 27, and 30.

The variations in the number of wins across districts highlight the geographical strongholds of different political parties. This distribution provides insights into regional political dynamics and can be used to understand the influence and popularity of parties in various parts of the region. The regional popularity of certain parties stems from a concept known as ‘vote bank politics’.

Vote bank politics involves targeting specific voter groups based on ethnicity, religion, caste, or socioeconomic status to secure electoral support. This strategy includes targeted campaigns, patronage, and social welfare programs. While it can increase political participation among marginalized communities, it often leads to social division, policy distortion, reduced accountability, and undermines democratic principles. Vote bank politics can positively impact forecasting by establishing clear voting patterns in certain areas and providing historical data for targeted groups. This is the reasoning behind our choosing to use historical data to forecast Indian elections.

This graph plots years and assembly numbers on the x-axis against the vote share percentage on the y-axis. It shows the dominance of the Indian National Congress (INC) in the earlier years, followed by a decline over time. Conversely, the Bharatiya Janata Party (BJP) has significantly increased its vote share in recent elections. This trend indicates changing political dynamics and the rise of new forces in politics, illustrating both changes in voter choice and partisan power.

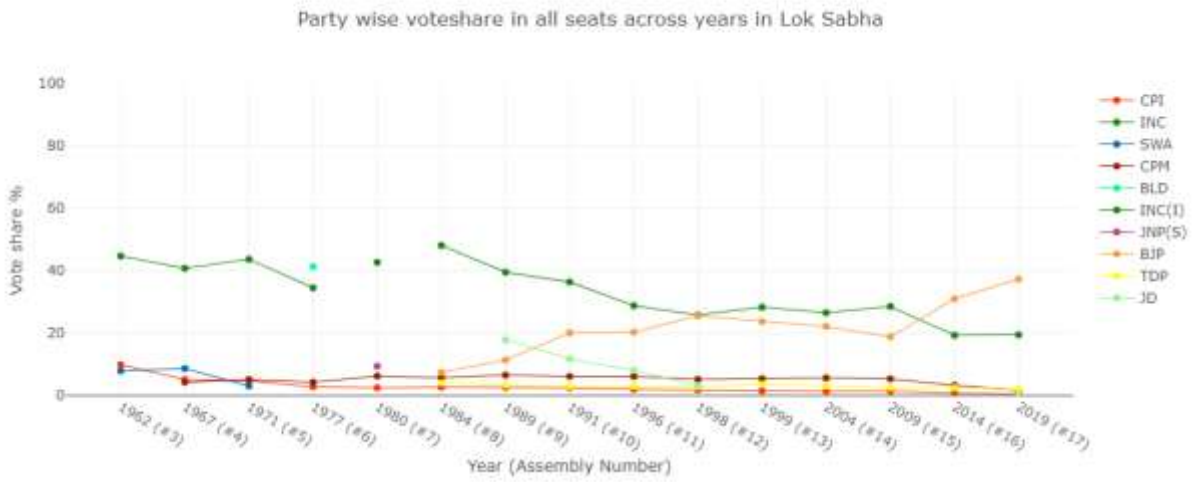

Figure 4:- Line Graph representing year-wise voter share for each party from Lok Dhaba Dataset.

This graph compares years and assembly numbers on the x-axis to the vote share percentage on the y-axis. It demonstrates the Indian National Congress (INC)'s supremacy in the early years, followed by a gradual fall. In contrast, the Bharatiya Janata Party (BJP) has dramatically grown its vote share in recent elections. This trend indicates changing political dynamics and the rise of new forces in politics, illustrating both changes in voter choice and partisan power.

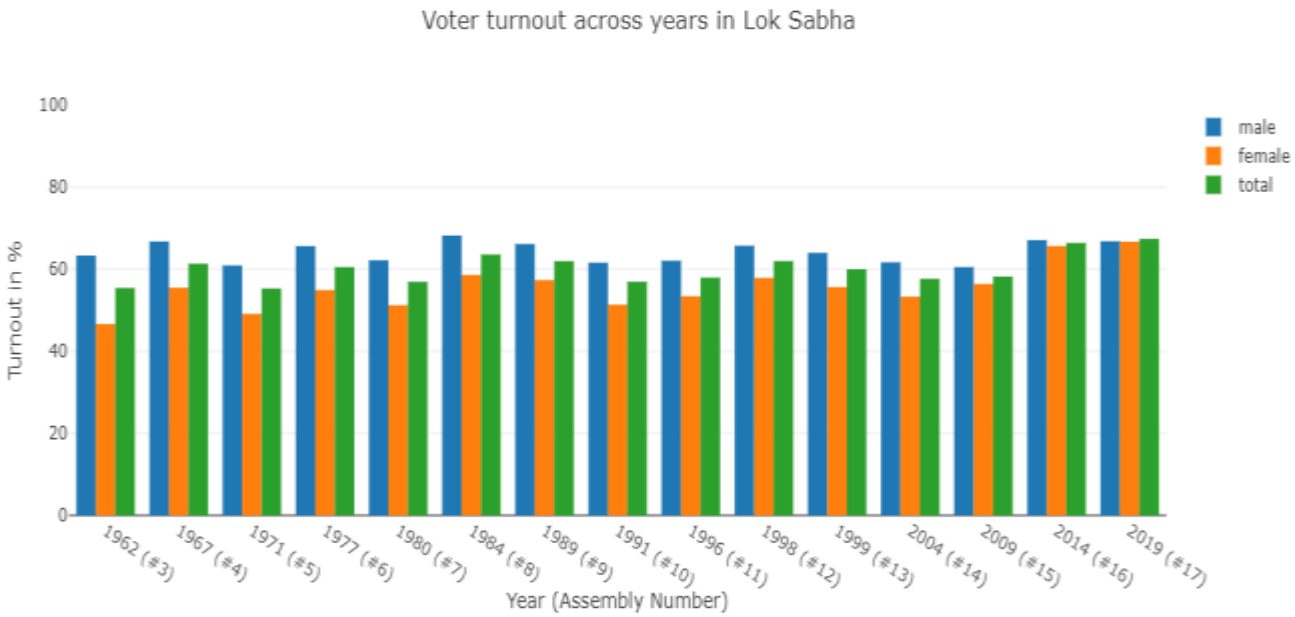

Figure 5:- Bar Graph representing Gender voter turnout from Lok Dhaba Dataset

This graph displays voter turnout percentages categorized by gender (male, female, and total) over different years. The horizontal axis reflects years and assembly numbers, while the y-axis depicts voter turnout percentages. Voter turnout has risen overall over time, with a substantial increase among female voters. For the first time in recent elections, female voter turnout has nearly reached parity with male voter turnout. This trend highlights the growing importance of gender-inclusive campaigning and active female engagement in politics.

Voter turnout percentage measures the proportion of eligible voters who cast their ballots in an election. It is calculated using the following formula:

$$\text{Voter Turnout Percentage} = \left(\frac{\text{no. of votes cast}}{\text{no. of eligible voters}}\right) \times 100$$

The number of votes cast refers to the total number of ballots submitted by voters. The number of eligible voters is the total number of individuals legally eligible to vote, defined by criteria such as age, citizenship, and residency. For example, if 10,000 votes were cast in an election with 50,000 eligible voters, the voter turnout percentage would be:

$$\frac{10000}{50000} \times 100 = 20\%$$

These graphs provide clues for predicting elections. The declining vote share of historically dominant parties like the INC, and the rising influence of the BJP, suggest a changing political landscape. Forecasting models should account for this shift to predict where elections may be headed accurately. Moreover, the increasing voter turnout, particularly among women, indicates the need for gender-inclusive campaign strategies.

The chart "Rerunning candidates across years in Lok Sabha" shows the percentage of candidates who have run for election more than once in different assembly years of the Lok Sabha. The x-axis represents the years and assembly numbers, while the y-axis represents the percentage of rerunning candidates. The chart indicates that the percentage of rerunning candidates fluctuates across different election years. Notably, there is a significant increase in the percentage of rerunning candidates in the elections of 1977, 1998, and 1999, compared to other years.

Re-election refers to the scenario where previously elected candidates or contestants who have run in previous elections stand for election again. Forecasting re-elections is crucial for several reasons. Firstly, rerunning

candidates often have established name recognition, political experience, and voter base, which can significantly influence election outcomes. Secondly, analyzing re-election trends can provide insights into the stability and continuity of political representation. High rates of re-running candidates might indicate strong voter loyalty and satisfaction with incumbents, whereas lower rates might suggest voter desire for change or dissatisfaction. Lastly, forecasting re-elections aid political parties in strategizing their campaigns and resource allocation. Knowing which incumbents are running again can help parties decide where to focus their efforts, either to support their candidates or to challenge strong opponents.

Re-election Probability of Political Incumbent

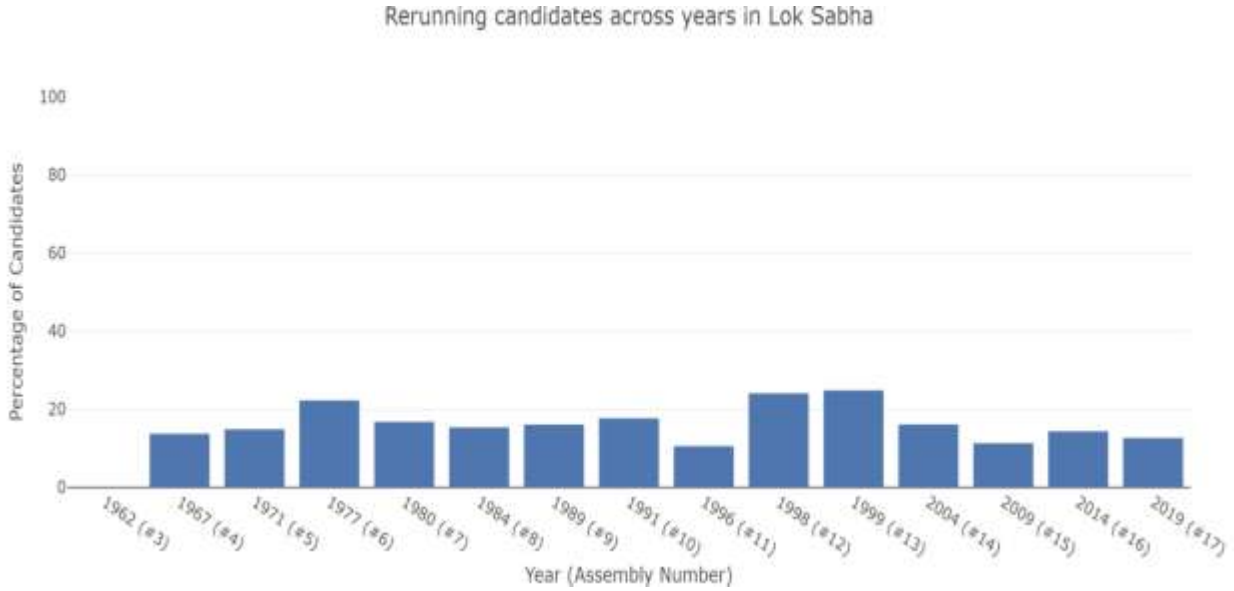

Figure 6:- Bar graph of Re-running candidates across years from the Lok dhaba dataset.

To identify rerunning candidates, the dataset was first grouped by the winner_name column to count the number of times each candidate was elected, using the size method to obtain the count of occurrences for each candidate. The resulting data frame, candidate_counts, was then filtered to identify candidates who were elected more than once. This filtering step involved selecting counts greater than one, helping to pinpoint candidates with multiple re-elections. This process elucidates the frequency of re-elections among candidates and highlights individuals who have maintained political favour over multiple election cycles.

The probability of re-election for each candidate was calculated by determining the total number of elections, represented by the length of the DataFrame df1. The formula used for calculating the re-election probability is

$$\mathrm{P(Reelection)} = \frac{\text{no.of reelections}}{\text{total no.of elections}}.$$

Candidates with high probabilities of re-election typically enjoy the incumbency advantage, characterized by increased name recognition, an established track record, and better access to campaign resources. These elements collectively bolster their electoral success over successive terms. Furthermore, political parties can leverage this information for strategic planning. By focusing support on candidates with high re-election probabilities, parties can optimize their resource allocation. Additionally, identifying incumbents who are at risk of losing can inform targeted campaign efforts to challenge them effectively. High re-election probabilities also signal strong voter loyalty and satisfaction with the incumbent's performance, which is pivotal for forecasting future election outcomes and crafting policies to sustain or boost voter approval.

The accuracy comparison reveals that Random Forest achieved the highest accuracy at 99.89%, followed by Gradient Boosting and Decision Tree at 99.78% and 98.75%, respectively. K-Nearest Neighbors also performed well at 93.31%, while Neural Network reached 81.88%. Logistic Regression and AdaBoost had lower accuracies at

44.60% and 30.74%, respectively. Overall, ensemble methods like Random Forest and Gradient Boosting demonstrated superior performance for predicting the winning party.

	winner_name	Probability	winner_party
0	anil kumar	0.000620	INC
1	ashok kumar	0.000596	TDP
2	mohan lal	0.000596	INC
3	om prakash	0.000548	INC
4	ram singh	0.000548	SAD
5	hari singh	0.000501	NaN
6	gulab singh	0.000453	TDP
7	balbir singh	0.000382	IND
8	ashok kumar singh	0.000382	TDP
9	babu lal	0.000382	TDP
10	amar singh	0.000382	AIMIM

Fig 7:- List of winners along with their re-election probability.

Model Predictions and Accuracies

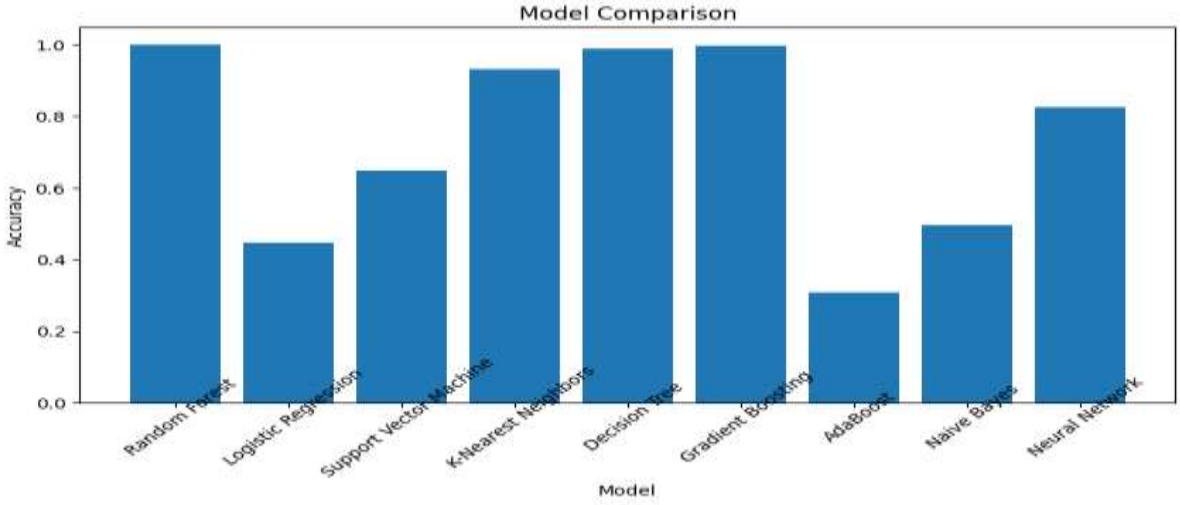

Figure 8:- Bar Plot comparing the various Machine Learning Models.

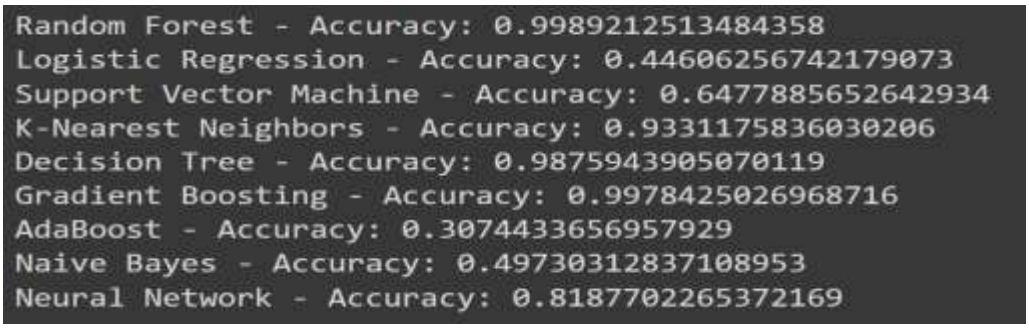

Figure 9:- Accuracy comparison chart of all Machine Learning Models.

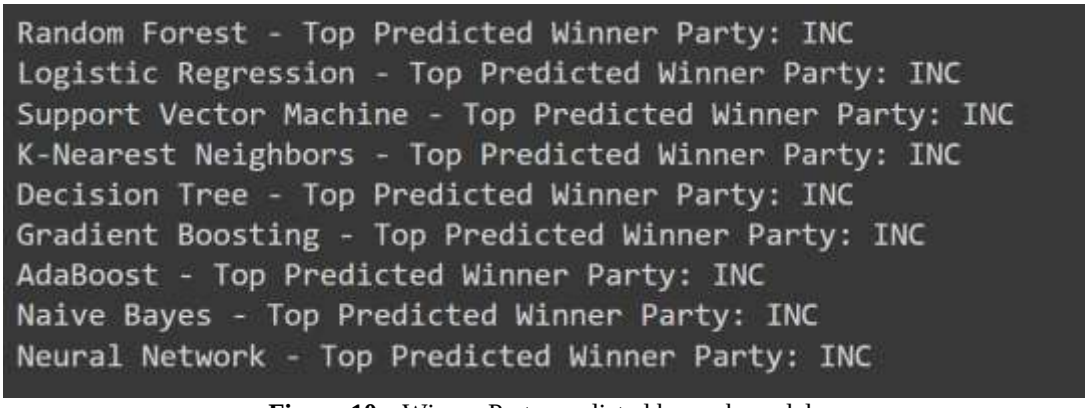

Figure 10:- Winner Party predicted by each model.

All models consistently predicted the INC (Indian National Congress) as the winning party. However, this was not consistent with the results of the 2024 Indian General. This error could be due to the INC's historical dominance and a significant number of seats held over the years. This long-term presence and consistent performance have likely resulted in a dataset heavily skewed in favour of INC, making it the most frequent label in the training data. This prevalence influences the models to favour INC predictions. The models may suffer from bias due to the imbalance in the dataset, where the historical success of INC overshadows other parties. This could lead to overfitting, where models predict the dominant class regardless of current political dynamics. Such as in the case of the high vote share of BJP in recent years. To mitigate this, future work should consider data balancing techniques and include recent political trends to enhance model accuracy and relevance. Another method could be to use EDA to determine the most recent shift in political dominance and limit the dataset to that specific timeframe.

Conclusion:-

Predicting the outcome of elections is challenging. Voting behaviour is inherently volatile and can be swayed by socioeconomic conditions, media influence, and sudden political developments. This unpredictability makes accurate modelling difficult. Additionally, the phenomenon of vote bank politics, where political parties target specific voter groups based on ethnicity, religion, caste, or socioeconomic status, adds another layer of complexity due to the highly localized and unpredictable voting patterns it creates. The dataset used in this study, though comprehensive, presented significant challenges. Its large size required substantial computational resources and meticulous handling to ensure efficient processing. Inconsistencies in the dataset, such as missing values and incomplete records, posed additional hurdles. Among the models evaluated, the Random Forest model emerged as the most accurate, achieving an impressive accuracy of 99.89%. This high performance can be attributed to the ensemble nature of Random Forest, which combines multiple decision trees to enhance predictive accuracy and reduce overfitting. The Gradient Boosting and Decision Tree models also performed exceptionally well, with accuracies of 99.78% and 98.75%, respectively. This study's implications are significant for political analysts and campaign strategists. Accurate election predictions can provide valuable insights into voter behaviour and preferences, enabling political parties to tailor their strategies more effectively. For analysts, these predictions help identify key trends and factors influencing election outcomes, contributing to a deeper understanding of the electoral landscape. However, it's important to note that the study also has several limitations that could impact the accuracy

of the predictions. One primary limitation is the potential bias introduced by the dataset. Given the historical dominance of the INC party, the models may be biased towards predicting INC as the winning party. This underscores the importance of using balanced datasets and incorporating recent political trends to ensure the models remain relevant and accurate.

Additionally, while the Random Forest model showed high accuracy, it is essential to consider that real-world election outcomes can be influenced by factors not captured in the dataset, such as campaign events, scandals, or last-minute shifts in voter sentiment. Another limitation is the computational complexity of the models, which could limit their accessibility to all researchers or analysts. However, this also presents an opportunity for future research to explore more efficient algorithms or techniques that can achieve similar accuracy with lower computational costs.

References:-

[1] Graefe, A. Accuracy of Vote Expectation Surveys in Forecasting Elections. Public Opinion Quarterly 2014, 78 (S1), 204–232. https://doi.org/10.1093/poq/nfu008.
[2] Liu, X.; Ren, F.; Su, G.; Zhang, M.; Gu, W.; Kato, S. Predicting Voting Outcomes for Multi-Alternative Elections in Social Networks. IEEE Access 2024, 1–1. https://doi.org/10.1109/access.2024.3425160.
[3] Khurana Batra, P.; Saxena, A.; Shruti; Goel, C. Election Result Prediction Using Twitter Sentiments Analysis. 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC) 2020. https://doi.org/10.1109/pdgc50313.2020.9315789.
[4] Sharma, A.; Ghose, U. Sentimental Analysis of Twitter Data with Respect to General Elections in India. Procedia Computer Science 2020, 173, 325–334. https://doi.org/10.1016/j.procs.2020.06.038.
[5] Soham Bhole; Fernandes, D.; Bajaj, P.; Batra, N.; Tewari, A. PollCast India 2024: Harnessing Data for Accurate Election Predictions. Social Science Research Network 2024. https://doi.org/10.2139/ssrn.4814788.
[6] Christensen, W. F.; Florence, L. W. Predicting Presidential and Other Multistage Election Outcomes Using State-Level Pre-Election Polls. The American Statistician 2008, 62 (1), 1–10. https://doi.org/10.1198/000313008x267820.
[7] Jain, V.; Kumar, S. Intelligent Systems and Applications. Intelligent Systems and Applications 2017, 12, 20–28. https://doi.org/10.5815/ijisa.2017.12.03.
[8] Brito, K. D. S.; Filho, R. L. C. S.; Adeodato, P. J. L. A Systematic Review of Predicting Elections Based on Social Media Data: Research Challenges and Future Directions. IEEE Transactions on Computational Social Systems 2021, 8 (4), 819–843. https://doi.org/10.1109/tcss.2021.3063660.
[9] Dimiceli, C.; Carroll, M.; Sohlberg, R.; Kim, D.; Kelly, M.; Townshend, J. MOD44B MODIS/Terra Vegetation Continuous Fields Yearly L3 Global 250 m SIN Grid V006 [Data Set]; NASA EOSDIS Land Process: 2015.
[10] Asher, S.; Lunt, T.; Matsuura, R.; Novosad, P. Development Research at High Geographic Resolution: An Analysis of Night-Lights, Firms, and Poverty in India Using the SHRUG Open Data Platform. World Bank Econ. Rev. 2021, 35(4).
[11] Jensenius, F. R.; Verniers, G. Studying Indian Politics with Large-Scale Data: Indian Election Data 1961–Today. Stud. Indian Polit. 2017.
[12] Prakash, N.; Rockmore, M.; Uppal, Y. Do Criminally Accused Politicians Affect Economic Outcomes? Evidence from India. J. Dev. Econ. 2019.
[13] Agarwal, A.; Agrawal, N.; Bhogale, S.; Hangal, S.; Jensenius, F. R.; Kumar, M.; Narayan, C.; Nissa, B. U.; Trivedi, P.; Verniers, G. TCPD Indian Elections Data v2.0; Trivedi Centre for Political Data, Ashoka University: 2021.
[14] Agarwal, A.; Agrawal, N.; Bhogale, S.; Hangal, S.; Jensenius, F. R.; Kumar, M.; Narayan, C.; Nissa, B. U.; Trivedi, P.; Verniers, G. TCPD Indian Election Data Codebook v2.0; Trivedi Centre for Political Data, Ashoka University: 2021.