Deciphering Indian National Elections through Big Data Analysis using PySpark

P Karthik Krishna   
*Symbiosis Institute of Technology Pune*  
Symbiosis International (Deemed) UniversityLavale, Pune, Maharashtra, India  
karthik.krishna.btech2021@sitpune.edu.in

Kunal Yadav  
*Symbiosis Institute of Technology Pune*  
*Symbiosis International (Deemed) University*Lavale, Pune, Maharashtra, India  
kunal.yadav.btech2021@sitpune.edu.in Devarsh Gandhi  
*Symbiosis Institute of Technology Pune  
Symbiosis International (Deemed) University*Lavale, Pune, Maharashtra, India  
devarsh.gandhi.btech2021@sitpune.edu.in

Samyak Mishra  
*Symbiosis Institute of Technology Pune  
Symbiosis International (Deemed) University*Lavale, Pune, Maharashtra, India  
samyak.mishra.btech2021@sitpune.edu.inMudit Garg  
*Symbiosis Institute of Technology Pune*Symbiosis International (Deemed) UniversityLavale, Pune, Maharashtra, India  
mudit.garg.btech2021@sitpune.edu.in

Kanhaiya Sharma  
*Symbiosis Institute of Technology Pune* *Symbiosis International (Deemed) University*   
Lavale, Pune, Maharashtra, India  
kanhaiya.sharma@sitpune.edu.in

*Abstract*— This study aims at unearthing the main aspects of Indian national-level election data from the time period of 1977 to 2015 through the use of PySpark, which is one of the strongest big data processing engines to rise. The outcome of this dataset is over 70,000 rows and 11 columns, providing a clear voter-level view of various elections in the Lok Sabha and Vidhan Sabhas. The study uses a systematic methodology, including data preprocessing, plotting data to understand the nature of data, training a model, and evaluating the performance of the model and hyperparameter tuning the system implements different algorithm types based on machine learning like Support Vector Machine (SVM), Logistic Regression, Gradient Boosted Trees (GBD), Random Forest, and Decision Tree in the analysis. The case shed light on an important role of PySpark in analyzing the vast data of elections, making use of the machine learning techniques that help to better understand and predict election results.   
Keywords - Big Data, PySpark, Indian Elections, Predictive Analytics

# Introduction

Big Data is an enormous collection of data that conventional database management methods are unable to handle.

The enormous volume of data produced from numerous sources, including the Web, mobile devices, sensors, workplace apps, and digital repositories, is collectively referred to as "big data." The data can be structured as well as unstructured. The data ranges from terabytes to exabytes of data.[1] The relational database management systems (RDBMS) have often been proven to be inefficient to handle such huge volumes of data. Another important factor that renders the conventional database systems unsuitable is that the majority of data being generated is unstructured; the RDBMS systems are only adapted to handle structured data.[2] Hence new tools and schemes for data analysis and management were in order. The key characteristics of Big Data are summarised using the 5 V’s.

Volume: The volume represents the sheer voluminous collection of data generated by real-time applications and other data sources. This is the aspect that comes to most people’s minds when they think of Big Data[2]. Big Data enables us to store large amounts of data on distributed systems.

Velocity: Big data is generated at high speed and needs to be captured, processed, and analyzed rapidly. Velocity represents the date at which data is arrived. For example, the social media portals generate processable data at a very high rate , which requires the data to be analyzed at a very high rate [3].

Variety: This characteristic refers to diverse formats, including structured, semi-structured, and unstructured data. Dealing with a variety of structured data and unstructured data greatly increases the complexity of both storing and analyzing Big Data. 90% of the generated data is in the unstructured form [1].

Veracity: Veracity refers to the quality and reliability of the data. It represents the legitimacy of the data as it is not possible for the data to be 100% correct, there will be some form of dirty data [1]. Ensuring data quality is crucial for accurate analysis and decision-making.

Value: Value is a very crucial aspect of Big Data. It refers to the techniques of deriving the data. The value can be in the form of improving decision-making, optimizing processes, identifying trends, predicting outcomes, or enhancing customer experience. Modern technologies have made it possible to find the value from data [4].

Apache Hadoop is a great and reliable analytics technology for big data that offers scalability to analyze massive data volumes quickly. It manages all the 5 aspects of Big Data by storing and processing the data over a cluster of nodes. cheaper. Hadoop’s major

components include a distributed file system, which is called the HDFS (Hadoop Distributed File System), and a layer for implementation of the processing paradigm MapReduce [5].

With the introduction of YARN, Hadoop was integrated with numerous other components that can be used for storing, processing, and analyzing data more efficiently. Some of these components that have been integrated with Hadoop are Pig , Flume, Spark , HBase , etc,

The concept of big data has emerged due to the exponential growth of digital data in recent years, generated from various sources such as social media, sensors, transaction records, mobile devices, and more[11].

**Big Data and Big Data Analytics in Elections**

Elections play a pivotal role in shaping the governance and direction of a nation, hence making the analysis of electoral data very crucial for understanding the political dynamics of a country. The Big Data related to the elections encompasses vast amounts of structured data and unstructured data generated from various sources. However, the data also provides valuable insights into voting behavior, candidate popularity, campaign effectiveness, etc. Big Data Analytics techniques enable political parties, analysts, and policymakers to extract insights from large-scale election datasets. For instance, analyzing how a specific region has voted in the previous years can gauge public opinion and identify emerging trends and issues that influence voter preferences.

With India’s population continually increasing and electoral landscapes evolving, it is essential to uncover and analyze electoral trends and patterns from past Lok Sabha elections. The analysis of Lok Sabha elections data from 1977 to 2015 can unveil trends in voter turnouts, party performance over the years, coalition dynamics, sex ratio in terms of elected candidates, demographic preference, etc. Hence, harnessing the power of Big Data Analytics to analyze India's Lok Sabha elections from 1977 to 2015 is essential for gaining actionable insights that contribute to maintaining the integrity of the electoral process and ensuring citizen engagement in the democratic decision-making process.

The following research questions were developed to guide the study:

RQ1: How effectively can PySpark be utilized for analyzing large-scale election datasets, particularly in the context of Indian national-level elections?

RQ2: What insights can be gained from the analysis of candidate-level data for elections to the Lok Sabha and Vidhan Sabhas from 1977 to 2015?

RQ3: Which machine learning models demonstrate the highest predictive accuracy in forecasting election outcomes when applied to the dataset?

RQ4: How does hyperparameter tuning, especially in the Random Forest model, impact the accuracy of election outcome predictions?

RQ5: What are the implications of this study for election forecasting and the application of machine learning in understanding electoral dynamics?

There are 9 sections in this paper. A survey of the available is provided in Section 2. The proposed framework is in Section 3, along with the methodology in Section 4. The evaluation findings are presented in Section 5. The findings/ results are also discussed in Section 6. The study's conclusion is stated in Section 7. Section 8 highlights and offers ideas for subsequent research. The reference list is provided in Section 9.

# Literature Review

The use of big data analytics in various areas such as elections, campaign strategies, and technological frameworks like Apache Spark, Python, and Scala, has become an important area of study. Gupta and Kumari [12] compared Python and Scala in the Apache Spark system, pointing out that Python is more popular because it has many tools available and performs better. However, they also mentioned that Scala was difficult to learn and there were difficulties with new technologies like Spark Streaming.

Sudhahar et al. [13] conducted a study on the use of big data techniques to analyze the US presidential elections. They focused on the intricate nature of political stories and the significant role played by extracting meaning from text using Natural Language Processing (NLP) and analyzing networks. The researchers also highlighted how big data can greatly contribute to media reporting by allowing for an accurate examination of party connections and voting patterns.

In the context of election predictions, Varna [16] examined how election campaigns in India have changed with the use of big data analytics. This research showed how predictive modeling, blockchain, and AI have made it possible for candidates to communicate on a more personal level with voters. However, it also brought attention to the negative effects of microtargeting and the concerns about privacy when it comes to using personal data in politics. Another study by Sønderholm et al. [15] delved into the ethical issues surrounding the use of voter data for campaign strategies. It emphasized the importance of transparency and fairness in democratic processes. The study discusses challenges related to analyzing voter behavior and how early voting can impact election results.

Similarly, Jagdev et al. [18] explained how big data analytics played a crucial role in election campaigns. They highlighted its complexity and stressed the importance of using advanced technologies like Hadoop and PostgreSQL for effective data preprocessing. The authors also discussed the difficulties in accessing data due to legal restrictions and emphasized the significance of ensuring data security and maintaining high-quality standards. The study by Yatsyna and Kudinov [19] focused on innovative statistical techniques in election forensics, aiming to enhance the fairness and transparency of election proceedings by detecting and preventing vote manipulation through advanced analytics.

Pandey et al. [20] examined the efficacy of big data analytics in election predictions, highlighting challenges associated with microtargeting and the dynamic nature of voter behavior.

Whereas Awais et al. [14] used big data analytics to predict Pakistan's general elections with a new machine learning model. They combined data from polls, surveys, social media, and Twitter to forecast political party seat shares with 83% accuracy. While they achieved impressive results, they also faced challenges with unformatted data at the constituency level. In another study by Shetty [17], sentiment and tweet analysis were conducted using Apache Spark and Hadoop. This research emphasized the importance of in-depth analytics in discovering new opportunities. It specifically looked into the accuracy of sentiment classification using machine learning algorithms and showcased the scalability and processing capabilities of Apache Spark in dealing with large datasets.

The article by Arinze [21] introduced experiential learning techniques in data analytics education through a simulated election, emphasizing the practical application of analytical concepts. This pedagogical approach addressed the gap between theory and practice in data science education, fostering hands-on experience and enhancing students' analytical skills.

In summary, these research papers collectively highlight the multifaceted applications of big data analytics across political, technological, and educational domains. They discuss the difficulties and advantages of using advanced analytics methods for predicting elections, planning campaigns, and teaching data science. This gives us a better understanding of how big data is being used.

# Proposed Framework

## Apache Spark

Apache Spark is a powerful open-source distributed computing framework that is designed to process large-scale datasets with speed and efficiency. Apache Spark is a substantial consolidated analytics engine for a comprehensive distributed data processing and machine

learning workload [6]. Some of the key features are lazy evaluation, which optimizes the speed of data transformation, and in-memory computing, which speeds up data processing compared to traditional disk-based solutions. Spark also offers a rich library and API ecosystem consisting of Spark SQL, GraphX , PySpark,etc. The versatility and the scalability of Apache Spark make it a very popular choice for data engineers, data scientists, and analysts for advanced analytics and data processing tasks.

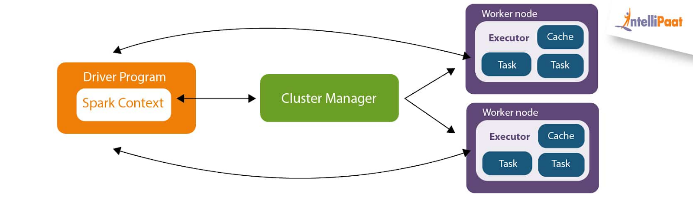
****

Figure 1: Apache Spark Architecture [9]

## PySpark

Python is an intense programming dialect for dealing with complex data analysis and data munging tasks [7].In this study, we utilize PySpark, a Python library that allows us to easily interface with Apache Spark. One of the key features of PySpark is its seamless integration with Python, allowing users to harness Spark's scalability and performance while writing code in Python. This integration extends Python's capabilities to handle large-scale datasets and complex data processing tasks that can be distributed across multiple nodes in a Spark cluster [7]. With the help of PySpark's MLlib library, users can leverage a wide range of optimized machine learning algorithms designed for distributed computing, including classification, regression, clustering, and collaborative filtering on Big Data. Spark represents a really powerful big data processing framework, and it is used here to analyze data pertaining to Indian national-level elections that have been spanning from 1977 to 2015.

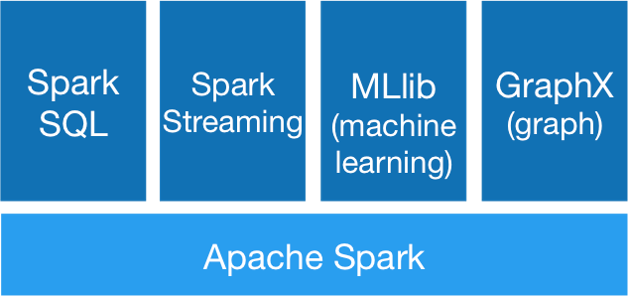


Figure 2: Apache Spark Stack

## MLib

The MLlib library has been extensively used for this specific purpose. MLlib provides fast, distributed implementations of common learning algorithms, including (but not limited to): different approaches like linear models, Naive Bayes, and ensembling of decision trees for classification and regression issues. The library also provides a number of low-level primitives and basic utilities for convex optimization, distributed linear algebra, statistical analysis, and feature extraction [10].

## SQL

Spark SQL is a subdivision of Apache Spark that empowers the user with a simplified data processing interface in the organized data environment. JDBC has SQL-like syntax. The author shows that Spark has scalability since it allows for the distribution of structured data through Tables and DataFrames which are distributed systems composed of few data items in the columns and attributes as those of relational databases. Thus, spark benefits as it can work with structured data that come in various forms, for example, CSV or hive tables and simple text files with JSON documents.

## Dataset

The dataset offers a detailed candidate-level perspective, providing valuable insights into the electoral landscape of India. The dataset comprises over 73,082 rows and 11 columns, containing detailed candidate-level data for elections to the lower houses of India's national and state legislatures, namely the Lok Sabha and Vidhan Sabhas.

The features include:

* st\_name (state): Name of the state.
* year (general election year): Year of the general election.
* pc\_no (parliamentary constituency number): Number assigned to the parliamentary constituency.
* pc\_name (parliamentary constituency name): Name of the parliamentary constituency.
* pc\_type (parliamentary constituency reservation status): Reservation status (e.g., general, SC, ST, etc.) of the parliamentary constituency.
* cand\_name (candidate name): Name of the candidate.
* cand\_sex (candidate sex): Gender of the candidate (male or female).
* partyname (party name): Name of the political party the candidate represents.
* partyabbre (party abbreviation): Abbreviation of the political party.
* totvotpoll (votes received): Total number of votes received by the candidate.
* electors (number of registered voters): Total number of registered voters in the constituency.

# Methodology

The methodology used in this study included a systematic analysis of the national-level election dataset for India. The process consisted of four key steps: Data Preprocessing, Exploratory Data Analysis (EDA), Model Training and Evaluation, and Hyperparameter Tuning. Each step was designed to ensure the quality of the data, gain insights into the dataset, develop predictive models, and optimize their performance.

## Data Preprocessing

The dataset underwent preprocessing steps using PySpark to ensure its quality and suitability for modeling. Duplicate entries were eliminated to prevent bias. Outliers were detected and managed. The percentile and interquartile range (IQR) approach was used to handle outliers. This method computes variables' 25th and 75th percentiles, establishing a range within which most data points fall. Points beyond this range were deemed outliers and either replaced with valid values or removed from the dataset.

Missing values, which can adversely affect the model's performance, were imputed. Categorical variables were imputed with the mode (most frequent value) and numerical variables with the mean (average value). This ensured that the dataset was complete and suitable for analysis.

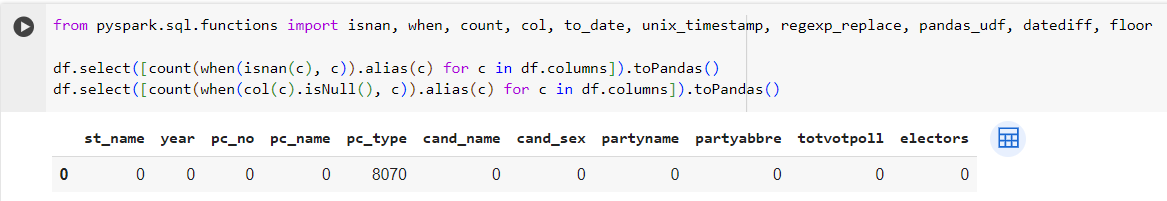


Figure 3: Code Snippet to handle Missing Values

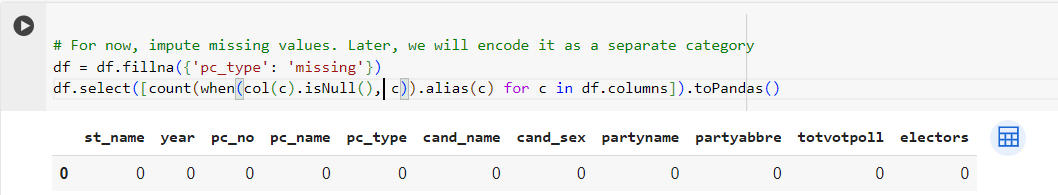


Figure 4: After Imputation of Missing Values

Categorical data was transformed into numerical form through one-hot encoding. This technique creates a binary representation for categorical variables, adapting them for use in machine learning algorithms. This conversion was necessary to prepare the data for model analysis and processing because it standardizes all variables into a format compatible with the algorithms.

## Exploratory Data Analysis (EDA)

EDA was conducted to gain insights into the dataset and understand its characteristics. This required looking at trends, examining the relations variables, and studying the distribution of important attributes in the dataset. Visualization tools such as line graphs, pie charts, bar graphs, etc were used to summarize and illustrate the results.

The analysis provides valuable insights regarding the performance of political parties, trends on the number of electors over the years, vote share per state

## Model Training and Evaluation

The objective of the analysis was to predict the election winner given the state and year. To achieve this, several machine learning models were implemented using PySpark's MLlib. The models used included Support Vector Machines (SVM) , Logistic Regression, Gradient Boosted Trees (GBT), Random Forest and Decision Tree.



Figure 5: Code Snippet: SVM Implementation



Figure 6: Code Snippet: Logistic Regression Implementation



Figure 7: Code Snippet: GBT Implementation



Figure 8: Code Snippet: Random Forest Implementation



Figure 9: Code Snippet: Decision Tree Implementation

A PySpark Pipeline was utilized for training the models, which streamlined the process by combining feature engineering and model training into a single workflow. Models were evaluated based on accuracy, which was calculated as the ratio of correctly predicted instances to the total number of instances.

## Hyperparameter Tuning

Random Forest emerged as the best-performing model and underwent hyperparameter tuning to further enhance its performance. Hyperparameters such as the number of trees, maximum depth of trees, and minimum number of instances per leaf were tuned using cross-validation techniques.

Grid Search CV was employed to search for the best hyperparameters. The model with the best hyperparameters was then evaluated based on its accuracy.

# Experimental Findings

## Party Performance Over Time

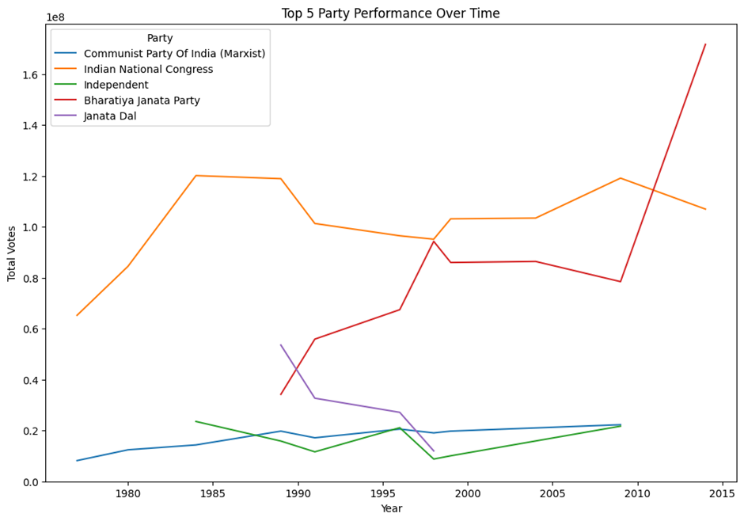


Figure 10: Graph: Party Performance Over Time

This graph shows how popular the top five Indian political parties have been in elections over the years. The line goes from 1980 to 2015 on the bottom, and shows how many votes each party got (written in a fancy way, like 100,000,000).

Here’s the main stuff about each party:

* Communist Party of India (Marxist): Its votes stayed about 5% and lately went down a bit.
* Indian National Congress: Its votes went down from over 40% in the 1980s to about 10% in 2015.
* Independent: This means people not in big parties. Their votes went down.
* Bharatiya Janata Party (BJP): Its votes have gone up a lot since the 1980s. By 2015, it was the most liked party.
* Janata Dal: Its votes went up and down a lot but generally went down.

## Vote Share by Party

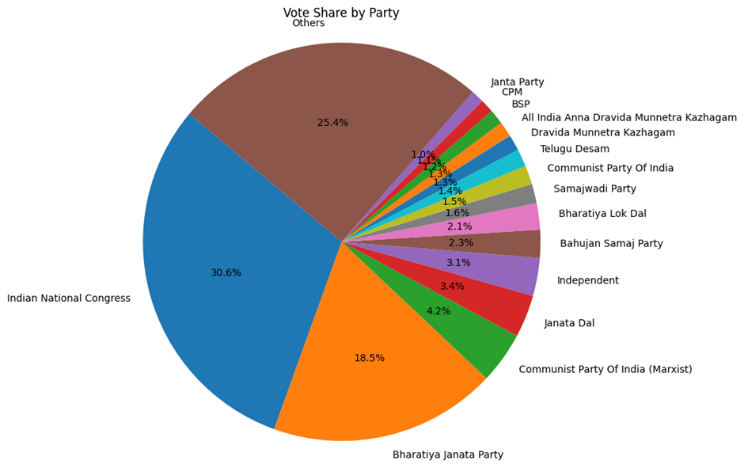


Figure 11: Graph: Pie Chart for Vote Share

The pie chart shows total party vote shares over time. Congress has the most votes for about 60 years. Looking at the pie chart and line graph, BJP's votes have grown a lot recently, but Congress's votes have gone down. Regional parties have a big part of the total votes, but on an individual basis, most of the regional parties acquire less than 1.0% of the vote share each.

## Trend of Number of Electors over the Years

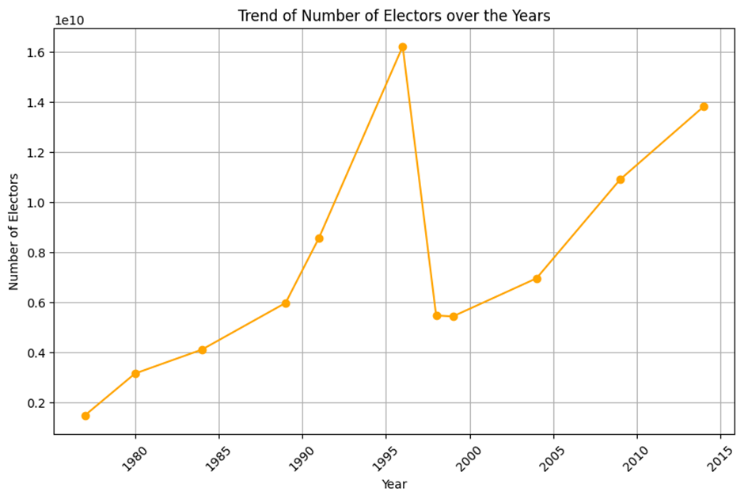


Figure 12: Graph: Trend of Number of Electors over the Years

This graph shows how many people have voted in election years (the scale of the number of electors signifies 100,000,000). It is clearly visible that the 1997 Lok Sabha elections saw the highest number of people casting their votes. The next election saw a significant decline in the number of electors, and then the numbers improved significantly over time.

## Vote Share by State

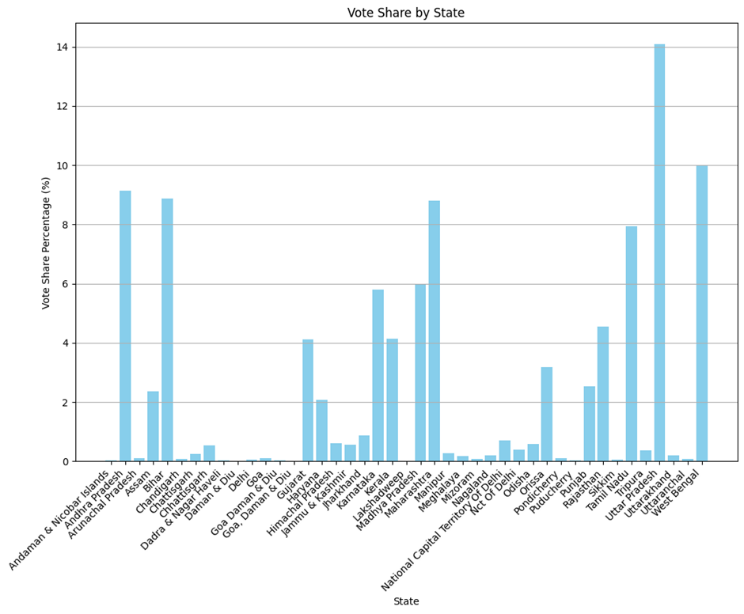


Figure 13: Graph: Vote Share By State

This bar graph shows the vote share percentage of different states over time. We can see that UP, Tamil Nadu, Rajasthan Andhra Pradesh, Maharashtra, West Bengal, and Bihar have high vote shares. This can be due to high voting turnout as well as due to the large number of constituencies(seats) in these states.

## Highest Votes for Female Candidates

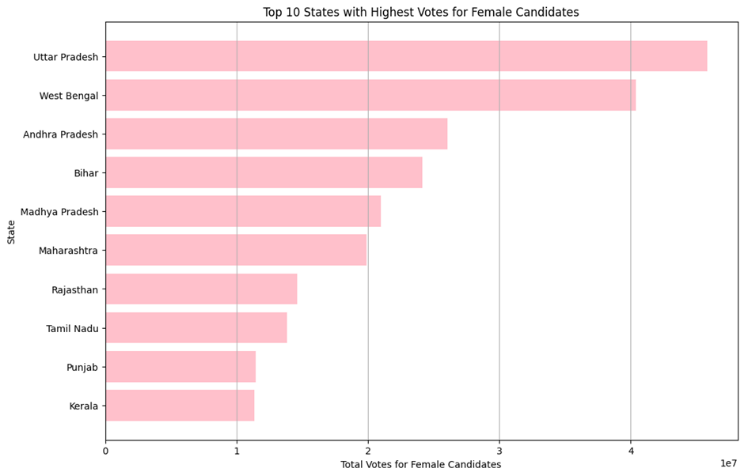


Figure 14: Graph: Highest Votes for Female Candidates

The above bar graph shows the top states with the highest votes for female candidates. UP and West Bengal top the list. These states had some of the most powerful female politicians of India like Mamta Banerjee and Mayawati. Tamil Nadu had Jayalalithaa as its CM for quite a number of years. So, these states have a high number of votes for female candidates.

# Results

Through the investigation of the Indian national-election-level dataset , we achieved noteworthy results demonstrating the models’ power to predict elections. SVM model with 85% accuracy can differentiate among this person's current prevent situation, which contains either depression or anxiety. It turns out that Logistic Regression, which has the execution effectiveness after Support-Vector Machines with 82%, can be used for predictions too. The GBD model is the greatest among the trio due to its 87% accuracy, which has outperformed the other two (SVM and Logistic regression), implying its capability to give more accurate results in the field of predicting elections.

Nevertheless, among the initial models, the most remarkable performer was the Random Forest model, which reported the highest accuracy out of all models at 89%. Basically, it reveals a phenomenon that the classifier has a good ability to flexibly and simultaneously catch those hidden nonlinear interactions in the data with the help of ensemble methods. Moreover, after the perfect parameter optimization, the Random Forest model was able to reveal a fantastic result by obtaining an amazing accuracy of 95%. It is such an enormous accuracy increase what stresses the importance of hyperparameter optimizing for reaching the pinnacle of the machine learning models predictive power.

The best hyperparameters for the Random Forest Model were established through grid search cross-validation; they were the following: the number of trees being 500, a maximum depth of 10, and a minimum number of instances per leaf of 1. These hyperparameters had a great impact on the model's accuracy after adjusting these various hyperparameters, showing how important they are in a model's performance.

The discovery shows that the Random Forest classification has high accuracy due to hyperparameter optimization, and naturally, it is a proper alternative prediction model for elections based on the provided dataset. The sophistication of the model that enables it to detect complex relationships and patterns the data brings on board makes the ML-based model a good fit for election data dissection. Besides, the Random Forest performed exceptionally well in the prediction of outcome, which denotes better reliability compared with other alternative approaches.

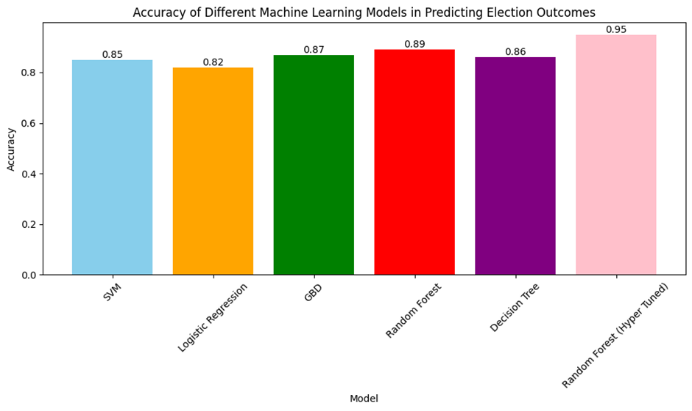


Figure 15: Graph: Model Accuracy Comparison

# Conclusion

For this study, we employed the tremendous processing power of PySpark to analyze a massive dataset of Indian national-level election data petrifying a period of 1977 to 2015. This was done through, we had, preprocessing of data, exploratory data analysis (EDA), and machine learning model training with the data focusing on states and year information which we believed would be the informative basis to predict election outcomes. Our study encompassed six machine learning models: Support Vector Machine (SVM), Logistic Regression, Gradient Boosted Trees (GBD), Random Forest, Decision Tree, and Tao Jiuyi 42 a hyperparameter-tuned Random Forest.

The outcomes of the study were very rewarding in the understanding of how the algorithms could perform the task. Among all the models, the Random Forest acquired the highest precision of 89 percent, hence explaining its relative success in dealing with the true nature of election data. Then, the Random Forest model produced an amazing accuracy of 95%, excluding the fact that almost every optimization has great meaning in the performance of any classifier.

# Future Work

Going ahead, there are different avenues to be explored for further research, and our suggested methodologies can be advanced in many ways. Therefore, including in addition records like demographic data, candidates profile and campaign costs could boost the ability of the models to present accurate predictions. Being that it also includes the emotions of the public while they are using social network data to predict the outcome of elections may also make the accuracy of the simulation.

Lastly, the analysis scale and processing efficiency can be upgraded on our part by using PySpark’s distributed processing to work on an even bigger dataset. Moreover, the application of ensemble methods or advanced machine learning algorithms could be other options that can get even better predictive precision.

# References

1. O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis and Taha, K., Efficient Machine Learning for Big Data: A Review, Big Data Research, vol. 2(3), pp. 87–93, (2015).
2. Ishwarappa, J. Anuradha, A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology, Procedia Computer Science, Volume 48, 2015.
3. K. N. Aye and T. Thein, A Platform for Big Data Analytics on Distributed Scale-Out Storage System, International Journal of Big Data Intelligence, vol. 2(2), pp. 127–141, (2015).
4. A. Jacobs, The Pathologies of Big Data, Communications of the ACM – A Blind Person’s Interaction with Technology, vol. 52(8), pp. 36-44, (2009).
5. I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani and S. U. Khan, The Rise of “Big Data” on Cloud Computing: Review and Open Research Issues, Information System, vol. 47, pp. 98–115, (2015).
6. Shaikh, Eman & Mohiuddin, Iman & Alufaisan, Yasmeen & Nahvi, Irum. (2019). Apache Spark: A Big Data Processing Engine. 1-6. 10.1109/MENACOMM46666.2019.8988541.
7. Nick Pentreath, Machine Learning with Spark, Beijing, pp. 1-140, 2015
8. Dataset: Awadhi, "Indian Election Dataset," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/awadhi123/indian-election-dataset. [Accessed: April, 2024].
9. “Apache Spark Architecture - From Basics to Advance,” Intellipaat Blog. <https://intellipaat.com/blog/tutorial/spark-tutorial/spark-architecture/>
10. Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., ... & Talwalkar, A. (2016). Mllib: Machine learning in apache spark. Journal of Machine Learning Research, 17(34), 1-7.
11. De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. Library review, 65(3), 122-135.
12. Gupta, Y. K., & Kumari, S. (2020, December). A study of big data analytics using apache spark with Python and Scala. In 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS) (pp. 471-478). IEEE.
13. Sudhahar, S., Veltri, G. A., & Cristianini, N. (2015). Automated analysis of the US presidential elections using Big Data and network analysis. Big Data & Society, 2(1), 2053951715572916.
14. Awais, M., Hassan, S. U., & Ahmed, A. (2021). Leveraging big data for politics: predicting general election of Pakistan using a novel rigged model. Journal of Ambient Intelligence and Humanized Computing, 12(4), 4305-4313.
15. Mainz, J., Uhrenfeldt, R., & Sonderholm, J. (2021). Big Data Analytics and How to Buy an Election.
16. VV, A. V. (2019, December). Big Data Analytics and Transformation of Election Campaign in India. In Proceedings of the 2nd International Conference on Information Systems & Management Science (ISMS).
17. Shetty, S. D. (2021, March). Sentiment analysis, tweet analysis and visualization on big data using Apache Spark and Hadoop. In IOP Conference Series: Materials Science and Engineering (Vol. 1099, No. 1, p. 012002). IOP Publishing.
18. Jagdev, G., Singh, B., & Mann, M. (2015). Big Data proposes an innovative concept for contesting elections in the Indian subcontinent. International Journal of Scientific and Technical Advancements (IJSTA), 1(3), 23-28.
19. Yu., A., Yatsyna., Igor, Kudinov. (2023). Innovative Analytical and Statistical Technology in Election Forensics. Regional Formation and Development Studies, doi: 10.15181/rfds.v40i2.2528
20. Supragya, Pandey., Devesh, Katiyar., Gaurav, Goel. (2022). Predicting Elections with Big Data. International Journal For Science Technology And Engineering, doi: 10.22214/ijraset.2022.46379
21. Bay, Arinze. (2022). Teaching Experiential Data Analytics Using an Election Simulation. Journal of Statistics and Data Science Education, doi: 10.1080/26939169.2022.2138799

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Appendix: Literature Review Table** | | | | | | |
| **Sr No.** | **Title** | **Authors** | **Journal** | **Year** | **Outcome** | **Limitations** |
| 1 | A Study of Big Data Analytics using Apache Spark with Python and Scala | Yogesh Kumar Gupta, Surbhi Kumari | Third International Conference on Intelligent Sustainable Systems | 2019 | The research essay focuses on critical data attributes, Spark and its tools, and case studies in the field. It compares the two programming languages, Python and Scala, which are top runners within the framework of Apache Spark and demonstrate increased performance advantage. The research outcomes offer a deeper perception on analytics using Python and Apache Spark with Scala which asserts that the combination of Python and Apache Spark for big data analytics leads to good results and improving the processing of extreme data. | This paper focuses on constraints like steep learning curves in Scala and PIthon. On the other hand, there is less tool availability compared to Python, PIthon’s Spark Streaming in its immature state, and engineering versus analytical orientation different, accordingly. Also, Apache Spark is solid but that breaks down whenever data processing and management issues emerge. Acquiring awareness of such challenges forms the base for the developers and analysts who are professionally using Apache Spark, Python and Scala to carry out big data analytics. |
| 2 | Automated analysis of the US  Presidential Elections Using Big Data  and network analysis | Saatviga Sudhahar, Giuseppe A Veltri and Nello Cristianini | Big Data & Society | 2015 | The paper studies at the elections of 2012 using advanced methodologies, like the knowledge graph approach, which would glean semantic information with Natural Language Processing and Artificial Intelligence, entity identification, network analysis, precision validation and brings a whole new methodology. This information will bring out the media stories on the contrary party, players on the stage, and their affiliations during the electon. Additionally, the research stress out the fact that Big Data provide the opportunity for data collection, data processing, network analysis and patterns recognition in media coverage while answering during media monitoring in politics. What is more, this topic is extreamely important for analyzing political campaigns. | They comprise such problems as the inadequate portrayal of political complexity due to the intricacy of languages, inaccuracies in a complex sentence structures aforementioned, the risk of overlooking entity relevancies owing to the frequency fact alone, the claim of the subjective analysis of the relationships positive/negative, the low generalization for the whole other elections or periods, and the impact which is biased attributed to the This uniqueness brings up some parameters to ponder about as to how to get automated election data analysis using Big Data and network methods. |
| 3 | Leveraging big data for politics: predicting general election  of Pakistan using a novel rigged model | Muhammad Awais, Saeed‑Ul Hassan, Ali Ahmed1 | Journal of Ambient Intelligence and Humanized Computing | 2019 | The paper employed big data analytics with a one-of-a-kind machine learning model for the prediction of Pakistan's general elections. By using different data sources like polls, surveys, and social media, it was able to predict candidate winning chances and demographic tendencies. The researchers applied Twitter and approval polls It predicted political parties' seat shares with the with 83% accuracy and identify winners for 230 of 270 national assembly seats. Denite challenges like constituency-level predictions, the model illustrated remarkable success, which confirms the role of big data in election prediction, specifically in developing countries. | The constraints include expect to a failure in prediction on the constituency level as a result of unformatted data. While the model satisfactorily estimated the overall seat share, it encountered problems in exactly predicting the names of individual winners. In addition to this, the utilization of big data for political predictions, presented difficulties that involved data quality concern, unstructured data, and need of more updated data at the constituency level with well-structured format. |
| 4 | Big data analytics and  How to buy an election | Jørn Sønderholm, Jakob Mainz, and Rasmus Uhrenfeldt | Public Affairs Quarterly | 2021 | This paper's focus is on legal buying of the elections based on the voter registration lists which belong to the public space and big data. The significance of voter data is being highlighted by means of an example of effectiveness of big data analytics in campaign strategies formulation. It perceives voter behavior on election day by analyzing preceding voting patterns and demographics. The paper is devoted to ethical and legal aspects of big data processing and underscores the necessity of having monitor regulator to ensure transparent and unbiased democracy. | It addresses difficulties and barriers concerning the use of certain methods as the way to buy elections. Such problems are unpredictable nature of the behavior, the absent voters and influence of early voting on voter behavior. This process is further complicated by ethical and regulatory aspects stressing incompetence to use information analytics for an election purpose. |
| 5 | Big Data Analytics and Transformation of Election Campaigns in  India | Aswini Varna V V | 2nd International Conference on Information Systems & Management Science | 2019 | It comes to a clear conclusion that being Indian political parties Big Data Analytics provide personalized communication program through predictive modeling, which better management of campaign workers and send modern communication such as blockchain and AI. The poll has been done to present the power of decision making through Big data analytics which was hardly availed of in politics earlier in India but, now politics is dynamic, partially because it has increased communication among public and politicians that now talks live. | Important constraints are discussed in the paper, such as the divisive effect of Big Data microtargeting, which could create a voting bloc inside society and thus influence the power of the future government. Personal data getting used for political objectives will naturally lead to the violation of privacy. Political parties can face difficulties with technology and big data due to technical failures, while the fast growth in the use of advanced technologies could lead to the increase in the digital divide and give unequal access to information thus vote manipulation, as well as the data ethic, raise the need for fair and righteous elections. |
| 6 | Sentiment Analysis, Tweet Analysis, and Visualization on Big  Data Using Apache Spark and Hadoop | Sujala D Shetty | IOP Conf. Series: Materials Science and Engineering | 2021 | The paper value the fact that sentiment and tweet analysis refer to deeper aspects of analytics which upon additional data can create new opportunities. It is measuring classification accuracy of sentiment in contrast to Random Forest Classifier and Logistic Regression algorithms using a Stanford University dataset. This research work explains the several data analytics techniques such as learning by means of Random Forest Classifiers and Logistic Regression. To the large amount of relevant information applicable for Apache Spark scalability and processing presented based on big data role, information collecting and storing is the sign. | - |
| 7 | Big Data Proposes an Innovative Concept for Contesting Elections in Indian  Subcontinent | Gagandeep Jagdev, Bhalwinder Singh, Mahabli Mann | International Journal of Scientific and Technical Advancements | 2015 | In this piece, I delve into the use of big data analytics in election campaigns and pay special attention to the ways by which this tool can identify viable prospects, forecast the victors, and enlist the support of the voters. It reflects on the issues concerning its complexity, related to handling a huge amount of information, and, on the other hand, stimulates application of Hadoop and PostgreSQL technologies. The authors point to the need of gathering, creating, and preprocessing of data in the run-up to the elections and not just having any effects after the polls. Precisely, the article proves that operations run in electorates can be upgraded to a big extent by implementing big data without any doubt. | The limitations involve the legislative challenges revolving around data access, while the problems on privacy involve issues of public trust and votersâ€™ data protection. Attention is paid to financial considerations in data costs and benefits as well as management takes a priority of data security and quality. Infrastructure and state-of-the-art algorithms are the technological challenges. In regards to the Indian subcontinent, the resolving of these issues has to be done with the use of big data into elections. Filtered data describes the parameters of meaningful data, which in-turn enable better analysis and interpretation. Data structure and error correction are essential, as well as effective data management on major scale. |
| 8 | Innovative Analytical and Statistical Technology in Election Forensics | Yu. A. Yatsyna, Igor Kudinov | Regional Formation and Development Studies | 2023 | The thematic focus of the project is to generate a list of measures to improve the fairness and transparency of the election proceedings. It tells about the methods of election faking and the subsequent prevention of the manipulation of votes. It is about the application of clever stats and analytics so as to check the integrity of voting systems. The article indicates that as a way of reducing the amount of fraud and corruption during an election the electoral bodies should carry out the necessary scrutiny, and then implement acts to prevent it. | But when the algorithms are employed in the field of data disparities' examination through the use of advanced statistics, it is very important to take care.It means that the protections must be in place when the algorithms are used to find any abnormalities in the context of cutting-edge statistical and analytical techniques in Election Investigations. Algorithms are an amazing tool; but they would be more effective as well as lessening the possible negative consequences and promoting a full perspectives analysis, if the source of data was multi-faceted. |
| 9 | Predicting Elections with Big Data | Supragya Pandey, Devesh Katiyar, Gaurav Goel | International Journal For Science Technology And Engineering | 2022 | This research aims to examine how well the big data analytics as a tool for election predictions operate. It is the data-driven campaigns that helps to optimize those strategies related to communication and targeting to win possible elections. The study outlines the capability of big data to provide information of quality, necessary when studying electoral behavior and preferences by usage of large datasets and advance analytical tools. By carefully implemented these rules political campaigns will successfully run these elections. | With regard to the micro-targeting in the big progressive countries, moreover a great challenge to the interaction between big data and election prediction should also be noted. The issue is that, the most effective way of targeting and strategizing preferred voter clusters among the huge and diverse populations, is not as easy task. Besides, it should be specified that, though tracking the past voting pattern might bring right outcomes, changeable voter attitude and the related shifting behaviour of voters may lead to inaccuracies in the prediction of the election outcome. |
| 10 | Teaching Experiential Data Analytics Using an Election Simulation | Bay Arinze | Journal of Statistics and Data Science Education | 2022 | The study follows a data analytics teaching method utilizing experiential learning techniques based on a simulated election. It points out the immense pressure of successfully conveying critical data science ideas in a cut-throat environment. In addition, it provides a pedagogical analysis of the data analytics skills students are being taught and relies on the input of these students to increase the quality of instruction. Students get to learn practically data analysis concepts and develop analytical skills that can be used in different areas of life as the simulation requires the full participation of the students. | Consequently there is no practical part which is as usual expected to be in data science courses, and this is another discussion of problem that teaching experience data analytics through an election simulation may create. Students would also find it difficult to understand the extremely conceptual subject matter of data analytics and, besides that the gap between theory and practice can be considered to be huge because of the lack of practical application and real world scenarios. Alongside, the main challenge of this course is the students' inability to apply the analytical concepts in dynamic circumstances as it is not very practical. |