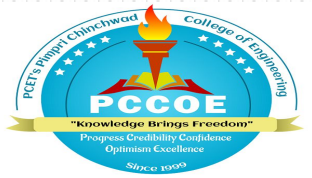
A
Project Review-I Presentation
on

Election Result Prediction

in
Partial Fulfillment of Final Year Computer Engineering Course- Project Stage-I
by

Pranav Bhople (121B1D024), Dnyaneshwari Katkar (121B1D049), Sahil Dhanwani (121B1D052),
Vaibhav Indure (121B1D053)
(A. Y. 2024-25)
Under the guidance of
Prof. Anand Birajdar



Department of Computer Engineering(Regional Language)
PCET's Pimpri Chinchwad College of Engineering

Contents

Problem Statement

Predicting election results in India based on candidate data involves analyzing various factors such as past election performance, demographic influence, political trends, and public sentiment.

Motivation

Elections play a crucial role in shaping the democratic structure of India. Traditional election predictions rely heavily on opinion polls and expert analysis, which may sometimes be biased or inaccurate. With the rise of big data and machine learning, there is an opportunity to create a more precise and objective prediction model. By leveraging historical election data, candidate profiles, and real-time public sentiment, this study aims to improve election forecasting accuracy. The insights from such predictions can help political parties in strategizing campaigns, aid researchers in studying voter trends, and enhance transparency in the electoral process.

Objectives

Objectives:

1. **Predict Election Outcomes** – Utilize candidate data and past trends to forecast election results.
2. **Analyze Key Factors** – Identify the influence of various factors such as criminal cases, assets, and education on election results.
3. **Enhance Voter Awareness** – Provide insights into candidate backgrounds for informed decision-making.
4. **Leverage Data-Driven Insights** – Use statistical models and machine learning to improve prediction accuracy.
5. **Assist Political Strategies** – Help parties understand voting patterns and optimize their campaigns.

Scope & Feasibility

Scope:

1. **Candidate Data Analysis –** Evaluating personal, financial, and political attributes of candidates.
2. **Voting Pattern Recognition –** Understanding trends based on past election data and demographics.
3. **Geographical Coverage –** State-wise and constituency-level predictions.
4. **Machine Learning Integration –** Using algorithms to refine predictions and detect anomalies.
5. **Impact Assessment –** Measuring how various candidate factors affect voter preferences.

Feasibility:

1. **Technical Feasibility –**
   - Availability of historical election data.
   - Use of AI/ML techniques for data processing and prediction.
2. **Operational Feasibility –**
   - Requires collaboration with political analysts and data experts.
3. **Economic Feasibility –**
   - Low-cost implementation using open-source tools (Python, R, etc.).
   - Can be monetized via consulting for political parties or media houses.

Project Requirements

Software Requirements

- **Languages & Tools:** Python, Jupyter Notebook, VS Code
- **Libraries:** NumPy, Pandas, Scikit-learn, Plotly, Seaborn, Matplotlib
- **OS:** Windows/Linux/macOS
- **Database:** CSV/Excel files

Hardware Requirements

- **Processor:** Intel i5 / AMD Ryzen 5 or higher
- **RAM:** 8GB (16GB recommended)
- **Storage:** 100GB+ SSD preferred
- **GPU (Optional):** NVIDIA CUDA-enabled for faster ML training

Project Requirements

Functional Requirements

- **Data Handling:** Load, clean, and preprocess election data
- **Model Training:** Train and optimize ML models (K-Nearest Neighbors, Logistic Regression, etc.)
- **Prediction & Visualization:** Forecast results and generate insights with charts
- **Evaluation:** Assess accuracy, precision, and performance metrics

Non-Functional Requirements

- **Scalability:** Handle large datasets efficiently
- **Performance:** Optimized for quick predictions
- **Reliability:** Robust model validation for accuracy
- **Usability:** Intuitive visualizations for easy interpretation
- **Security:** Secure data handling

Literature Survey

- First 5 most relevant papers, explain in detail in the format given in next slide)

Reference 1	DeLuca, Kevin. "A Model of the 2024 Presidential Election with Candidate Quality." (2024).
Objectives	Provide a more accurate forecast of the 2024 U.S. presidential election, focusing on both the national two-party Democratic vote share and the Electoral College results.
Proposed Solution	DeLuca builds upon Fair's (2009) economic fundamentals model by adding a candidate quality differential. This differential is derived from newspaper endorsements, as detailed in DeLuca's 2024 study on measuring candidate quality using local newspaper endorsements.
Results	The model predicts that Kamala Harris would secure 52.5% of the two-party popular vote, which is approximately 1.5 to 2 percentage points higher than other forecasts and polling averages at the time.
Advantages	The integration of candidate quality aims to improve the precision of election forecasts, potentially capturing dynamics that purely economic models might miss.
Limitations	The model's measure of candidate quality is based on newspaper endorsements, which may not fully capture public perceptions or the multifaceted nature of candidate appeal.

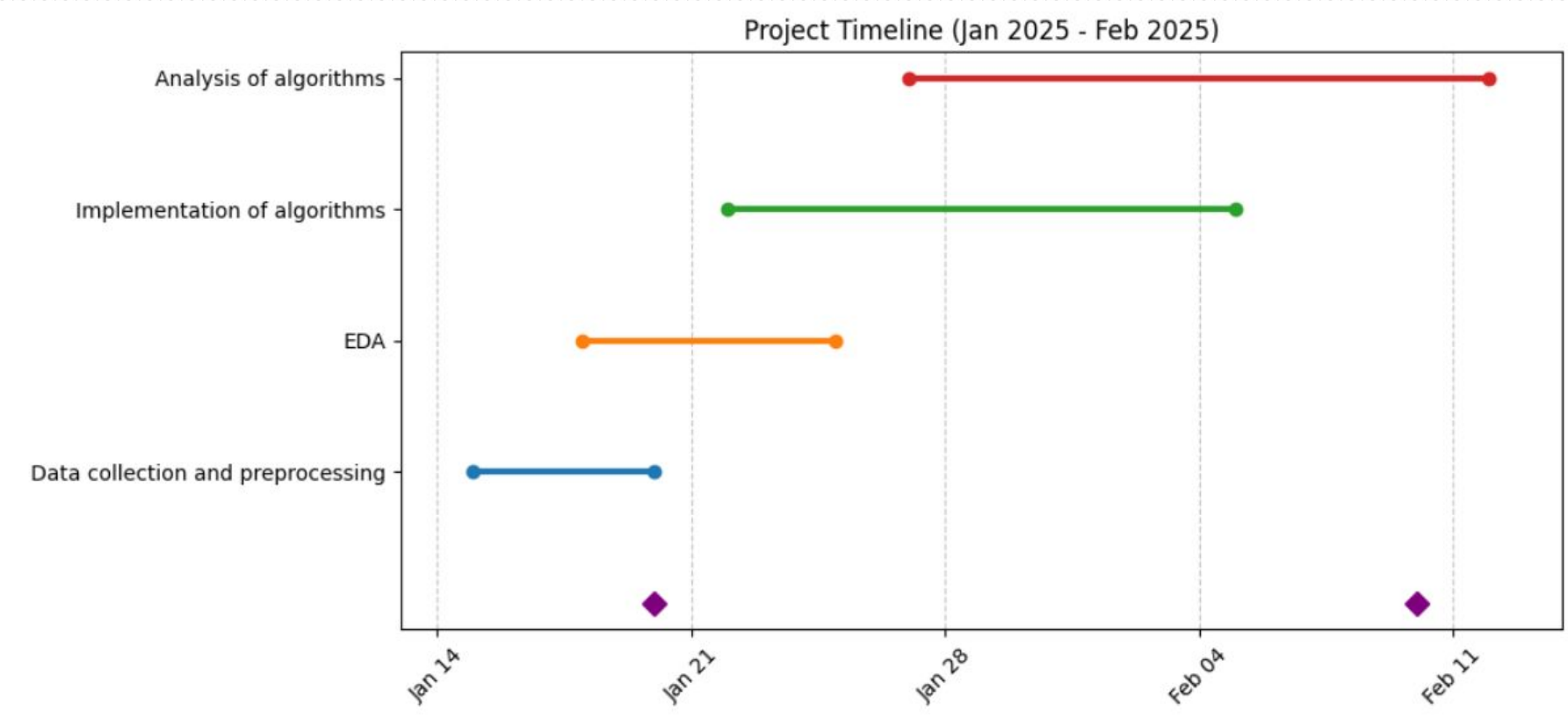
Reference 2	Basyar, Goran Fadhil. SUPPORTING PREPARATION FOR INDONESIA'S 2024 ELECTION. Diss. TILBURG UNIVERSITY, 2024.
Objectives	Analyze the existing electoral system in Indonesia to identify potential challenges and areas for improvement.
Proposed Solution	Conduct a thorough evaluation of the current electoral processes, including voter registration, campaign regulations, and voting mechanisms.
Results	By engaging various stakeholders, the research fosters a greater understanding of the importance of collaborative efforts in election preparation.
Advantages	Recommendations are specifically tailored to Indonesia's socio-political landscape, enhancing their applicability and effectiveness.
Limitations	While the study offers valuable recommendations, actual implementation may face obstacles such as political resistance or resource constraints.

Reference 3	DHADA, PUTU SUKMA DHARMALAKSANA. PREDICTING PRESIDENTIAL ELECTIONS IN INDONESIA USING MACHINE LEARNING BASED ON CANDIDATE SOCIAL MEDIA PERFORMANCE. Diss. Nusa Putra University, 2023.
Objectives	Develop a machine learning model to predict election outcomes based on social media metrics.
Proposed Solution	Employ supervised machine learning algorithms to train predictive models using the engineered features.Assess the model's performance using historical election data to ensure its predictive accuracy.
Results	The developed machine learning model achieves a notable accuracy in predicting election outcomes based on social media metrics.
Advantages	Utilizing social media data allows for timely insights into public opinion and candidate popularity.
Limitations	Social media users may not represent the entire voting population, potentially introducing bias into the analysis.

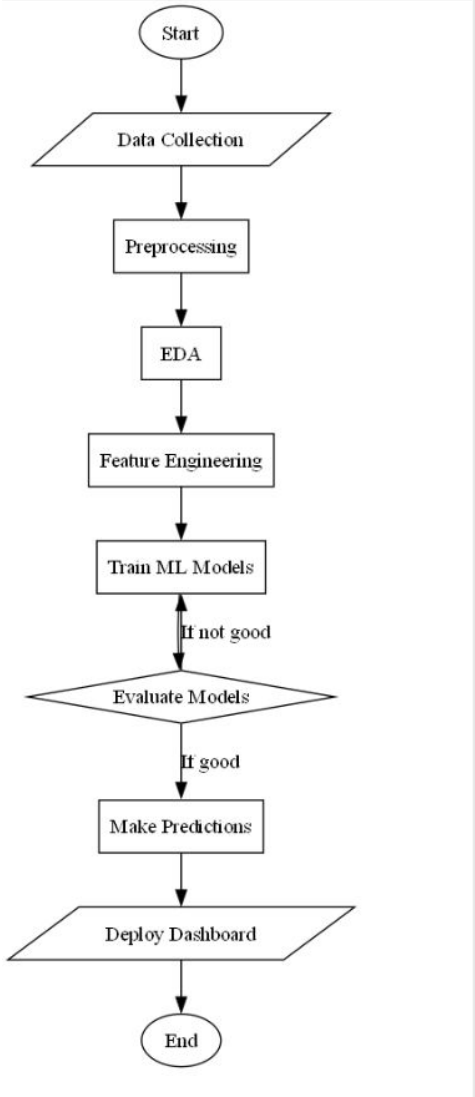
Reference 4	Mantika, Alisya Mutia, Agung Triayudi, and Rima Tamara Aldisa. "Sentiment Analysis on Twitter Using Naïve Bayes and Logistic Regression for the 2024 Presidential Election." SaNa: Journal of Blockchain, NFTs and Metaverse Technology 2.1 (2024): 44-55.
Objectives	Compare the effectiveness of Naïve Bayes and Logistic Regression algorithms in classifying sentiments.
Proposed Solution	Clean the data by removing noise such as URLs, special characters, and stop words, and perform tokenization. Convert textual data into numerical features using techniques like Term Frequency-Inverse Document Frequency (TF-IDF). Apply Naïve Bayes and Logistic Regression algorithms to classify the sentiments expressed in the tweets.
Results	Both Naïve Bayes and Logistic Regression models effectively classified sentiments in the collected tweets.
Advantages	The study offers a comparison between two machine learning algorithms, highlighting their respective strengths in sentiment analysis.
Limitations	The complexity of language, including sarcasm and slang, can pose challenges for accurate sentiment classification.

Reference 5	Hamdani, Agus Umar, et al. "Comparison of K-NN, SVM, and Random Forest Algorithm for Detecting Hoax on Indonesian Election 2024." Jurnal Nasional Pendidikan Teknik Informatika: JANAPATI 13.1 (2024).
Objectives	Compare the performance of K-NN, SVM, and Random Forest algorithms in tweets detection.
Proposed Solution	Implement K-NN, SVM, and Random Forest algorithms to classify the tweets based on the extracted features. Assess the performance of each model using metrics such as accuracy, precision, recall, and F1-score to determine their effectiveness in detecting hoaxes.
Results	Random Forest achieved higher accuracy and better performance metrics compared to the other two algorithms.
Advantages	By comparing multiple algorithms, the study identifies the most effective method for hoax detection in the given context.
Limitations	While Random Forest performed best in this study, its effectiveness may vary with different datasets or in different contexts.

Project planning: Timeline Diagram



Flowchart

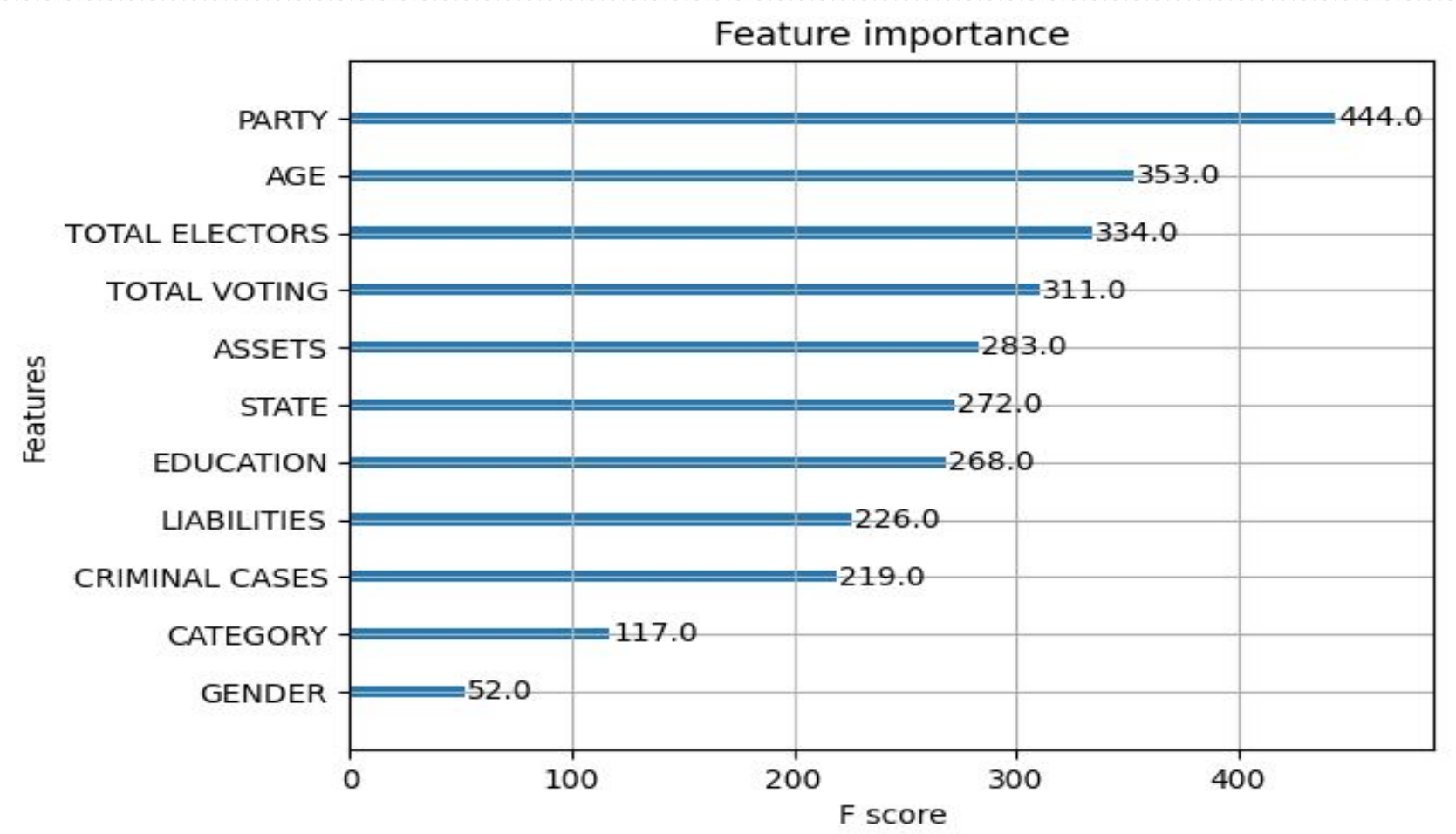

Start
Data Collection
Preprocessing
EDA
Feature Engineering
Train ML Models
If not good
Evaluate Models
If good
Make Predictions
Deploy Dashboard
End

Modern Tools & Techniques selected for project development

Programming Languages – Python (for data processing and ML modeling).
Libraries & Frameworks –

- **Data Manipulation** – NumPy, Pandas
- **Data Visualization** – Matplotlib, Seaborn, Plotly
- **Machine Learning Models** – Scikit-learn, KNN, Logistic Regression
- **Data Resampling** – SMOTE
- **Data Scaling & Encoding** – MinMaxScaler, OrdinalEncoder, LabelEncoder
- **Model Evaluation** – Sklearn metrics (accuracy, precision, recall, etc.)

Development Tools – Jupyter Notebook, VS Code
Operating System – Windows
Database – CSV/Excel files


Feature importance
PARTY
AGE
TOTAL ELECTORS
TOTAL VOTING
ASSETS
STATE
EDUCATION
LIABILITIES
CRIMINAL CASES
CATEGORY
GENDER
444.0
353.0
334.0
311.0
283.0
272.0
268.0
226.0
219.0
117.0
52.0
Features
0
100
200
300
400
F score

	Party	Winning Seats	Winning Percentage
0	BJP	86	50.000000
4	INC	30	17.441860
9	SP	13	7.558140
2	AITC	12	6.976744
5	TDP	6	3.488372
1	DMK	5	2.906977
8	SHSUBT	5	2.906977
10	Other	5	2.906977
11	YSRCP	3	1.744186
3	NCPSP	1	0.581395

Model	*Train Acc.*	*Train Prec.*	*Train Rec.*	*Train F1*	*Test Acc.*	*Test Prec.*	*Test Rec.*	*Test F1*	*CV Time (s)*
Cat Boosting	96.61	96.25	97.01	96.44	94.09	60.20	61.49	60.84	165.27
XG Boosting	96.73	96.47	97.01	96.54	93.97	60.46	55.61	57.93	4.19
LG Boosting	96.39	96.50	96.28	96.06	93.77	58.20	58.82	58.51	9.11
Random Forest	96.41	95.09	98.04	96.50	93.22	53.91	62.56	57.92	22.90
Decision Trees	94.70	94.34	95.29	94.61	92.26	48.41	57.21	52.45	1.81
Gradient Boosting	94.94	93.99	95.97	94.88	92.10	47.95	68.98	56.57	38.56
ADA Boosting	92.77	92.85	92.64	92.69	91.30	44.32	64.70	52.60	8.78
KNN	93.92	91.46	96.90	94.10	89.19	38.13	72.19	49.90	1.14
Logistic Regression	76.01	77.20	73.83	75.47	77.67	20.81	71.12	32.20	1.50
SVM	70.89	75.75	61.48	67.86	77.35	19.12	63.10	29.35	103.15

Table 8.1: Performance Comparison of Machine Learning Models

Conclusion

In this project, we analyzed election data to predict outcomes using machine learning. Key factors like criminal cases, assets, and education influenced results. Ensemble methods showed the highest accuracy, with CatBoost (94.09%) outperforming XGBoost (93.97%) and LightGBM (93.77%). Random Forest (93.22%) was competitive but resource-intensive. Overall, gradient boosting proved highly effective for election prediction, offering accuracy and robustness.

References

For example (Follow given format):

1. Gnoying Feng et.al “Experimental research on vertical axis wind turbine” IEEE school of energy and power engineering, vol. 978,no.1, 2009.

2. G.M.Hasan Shaharirar “Design and construction of vertical axis wind turbine” IEEE International form on strategic technology, vol. 978, no. 1, pp 326-329, Oct 2014.

3. S.Sathiyamoorthy et.al, “Hybrid energy harvesting using Piezoelectric materials, automatic rotational solar panel, vertical axis wind turbine”, International conference on modelling, optimisation and computing, IEEE, Vol. 38, no. 10, pp 843-852, July 2005.

4. Atif Shahzad et. Al. “ Performance of a Vertical axis wind turbine under accelerating and decelerating flows” Second international through-life engineering service conference, Elsevier, Vol. 11, no. 1, pp 311-316, 2013.

5. Sahishnu R shah et. Al., “ Design, modelling and economic performance of a vertical axis wind turbine,” Energy Reports, Elsevier, Vol 11, pp. 619-623, 2018.

Thank you