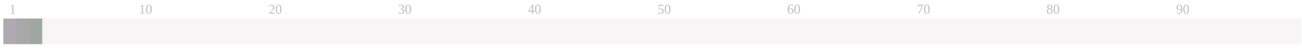
The Report is Generated by DrillBit Plagiarism Detection Software

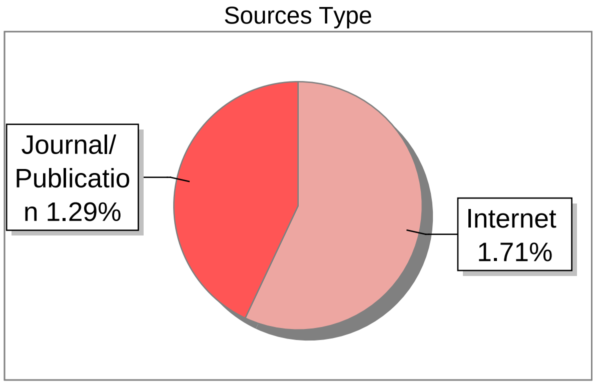
Submission Information

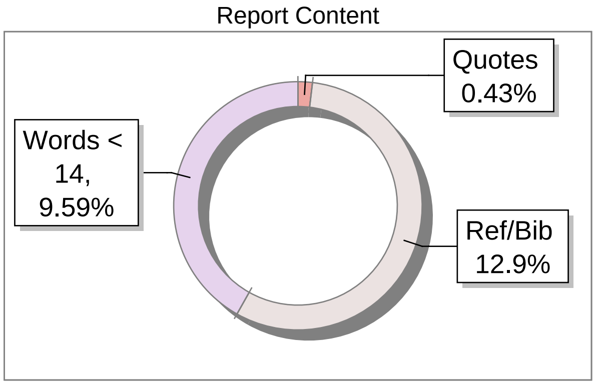
Author Name	Ganne Rahul Naidu
Title	Issues like voter suppression and discrimination in elections: Using U.S. Elections Dataset
Paper/Submission ID	3576112
Submitted by	premu.kumarv@gmail.com
Submission Date	2025-05-05 11:44:26
Total Pages, Total Words	8, 2565
Document type	Research Paper

Result Information

Similarity **3 %**







Exclude Information

Quotes	Excluded
References/Bibliography	Excluded
Source: Excluded < 14 Words	Excluded
Excluded Source	**0 %**
Excluded Phrases	Not Excluded

Database Selection

Language	English
Student Papers	Yes
Journals & publishers	Yes
Internet or Web	Yes
Institution Repository	Yes

A Unique QR Code use to View/Download/Share Pdf File

DrillBit

DrillBit Similarity Report

3	4	A
SIMILARITY %	MATCHED SOURCES	GRADE

A-Satisfactory (0-10%)
B-Upgrade (11-40%)
C-Poor (41-60%)
D-Unacceptable (61-100%)

LOCATION	MATCHED DOMAIN	%	SOURCE TYPE
2	arxiv.org	1	Publication
3	www.mdpi.com	1	Internet Data
4	moam.info	1	Internet Data
5	Thesis Submitted to Shodhganga Repository	1	Publication

Issues like voter suppression and discrimination in elections: Using U.S. Elections Dataset

1st Ganne Rahul Naidu
Department of Information Science
The Oxford College of Engineering
Bangalore, India
rahulganne11@gmail.com

2nd Kalyan Ram P S
Department of Information Science
The Oxford College of Engineering
Bangalore, India
kalyansr21@gmail.com

ABSTRACT

In democracies, the ability to vote is a basic right that guarantees all citizens a say in how their country is run. Voter suppression and discrimination, however, often compromise the integrity of this right. These practices can be covert or overt, ranging from targeted disenfranchisement based on race, socioeconomic status, or geography to restrictive ID laws and fewer polling stations. By using machine learning techniques to examine extensive datasets related to voter registration and U.S. elections, our study tackles these problems. Finding trends and irregularities that can point to discriminatory activities is the main objective. Because they are reliable and efficient in classification tasks, models such as Random Forest and Support Vector Machines (SVM) are used. These models assist in revealing hidden meanings in multifaceted, complex data. The project's output is intended to be more than just academic; it might be a useful tool for watchdog groups, civil rights organizations, and legislators. The initiative supports the overarching goal of guaranteeing fair and transparent electoral participation through predictive analytics and thorough assessments utilizing accuracy, precision, and recall criteria. It encourages accountability and aids in the promotion of inclusive voting changes by bringing to light data-driven evidence of possible voter suppression.

KEYWORDS

Voting discrimination, voter suppression, machine learning, Random Forest, SVM, U.S. elections, predictive modeling, data preprocessing, fairness in AI.

I. INTRODUCTION

Brief Overview of the Problem Domain

Any systematic or focused attempt to restrict or dissuade particular groups from exercising their right to vote is considered voting discrimination, a persistent problem in many democracies. Laws that disproportionately impact minority groups are examples of explicit discrimination, whereas restricted access to voting stations in particular neighborhoods is an example of implicit discrimination. From overt voter suppression strategies to more nuanced and intricate forms incorporated into administration, policies, or even false information, these approaches have changed over time.

Importance of the Topic

One of democracy's pillars is the ability to vote, which stands for both equal representation and popular confidence in the government. Persistent voting discrimination, however, subverts this idea and causes vulnerable groups to be systematically excluded and denied the right to vote. In addition to ensuring legal justice, resolving this issue is essential for social stability, civic participation, and public trust in electoral processes.

Machine learning offers a strong, unbiased way to identify discriminatory patterns that conventional approaches might miss in today's data-rich environment. Machine learning can find cases of voter suppression that might not be apparent through manual analysis or simple statistics because of its capacity to handle large, complicated datasets and spot subtle trends.

Objectives of the Project

1. To analyze voter registration and election datasets to detect patterns of discrimination.

2. To build predictive models that identify likely instances or risk zones of voter suppression.

3. To provide actionable insights using performance metrics like accuracy, precision, and recall

II. LITERATURE SURVEY

Numerous academic fields, including political science, law, sociology, and data science, have conducted research on voting discrimination and suppression. Studies have traditionally placed a strong emphasis on qualitative data, such as surveys, expert testimony, and analysis of court cases. These initiatives have been essential in exposing trends of racial, socioeconomic, and geographic discrimination. But they frequently lack the impartiality and scalability that computational approaches may provide.

Numerous studies have demonstrated that voter ID laws disproportionately affect marginalized populations, such as low-income neighborhoods and racial minorities, according to one line of research. Another section examines wait times and polling site accessibility, emphasizing how practical considerations can deter or prohibit voting. From a data-driven standpoint, a number of initiatives have evaluated suppression risks according to demographic and geographic variables using statistical methods such as logistic regression. In order to detect areas at high risk of voter suppression, Geographic Information Systems (GIS) have also been used to depict differences in polling infrastructure.

Gaps or Areas for Improvement

1. Limited application of sophisticated machine learning models for prediction, such as Random Forest and SVM.
2. Absence of frameworks for real-time or nearly real-time detection.

3. Insufficient use of time-series data to examine patterns of suppression throughout several election cycles.

III. METHODOLOGY

Data Preprocessing

A crucial step in any machine learning pipeline is data preprocessing, particularly when working with diverse, real-world datasets like election and voter registration data. In order to create useful models, it is necessary to resolve the discrepancies, missing values, and mixed formats that are frequently present in the raw data.

Managing Absent Information

Numerical Data: To lessen sensitivity to outliers, missing values in continuous variables (such as age and voting frequency) were imputed using the median.

Categorical Data: Depending on the context and frequency, missing categories like race or party

membership were either imputed using the mode or substituted with a placeholder ('Unknown').
Scaling Features

Standardization: Used to guarantee that every feature contributes equally to the decision boundary in algorithms such as SVM.
When showing feature interactions, min-max scaling is used to equalize the range between 0 and 1.

Categorical Variable Encoding

Features having several categories, such as state, ethnicity, and party, are subject to one-hot encoding. For binary or ordinal features, like gender or election results, label encoding is utilized.

Processing Time-Series

Election Cycle Handling: Features were designed to show the amount of time since the last vote, a change in registration, or a decline in turnout.

Temporal Trends: To record suppression trends over time, rolling averages and lag characteristics were included.

Algorithms Used

Forest at Random: A potent ensemble technique that reduces overfitting and increases accuracy by constructing many decision trees and averaging their output. very helpful for managing noisy features and unbalanced datasets.

SVM, or support vector machine: Chosen due to its high classification accuracy in both binary and multiclass issues.
efficient when there are distinct class borders and high-dimensional data.
To identify non-linear relationships in the data, kernel functions such as RBF were used.

Tools and Libraries

This project leverages a set of widely-used programming tools and machine learning packages that provide rapid data handling, visualization, model development, and evaluation.

Python: The dominant programming language utilized for its versatility, readability, and large ecosystem of data science libraries.

Pandas: Widely used for jobs including data processing, wrangling, and cleaning. allows for tabular data manipulation in a manner akin to that of Excel or SQL.

Scikit-learn: Core machine learning package used for constructing, training, and assessing models like Random Forest and SVM. utilized for preprocessing operations such as model selection, scaling, and encoding as well.

IV. IMPLEMENTATION

Step 1: Data Acquisition

downloaded datasets about U.S. elections and voter registration from reliable sources like:

- Election Assistance Commission (EAC) of the United States
- Records under the National Voter Registration Act (NVRA)
- Public datasets at the state level
- To ensure data consistency, several datasets were combined using distinct identifiers such as voter ID, precinct codes, or zip codes.

Step 2: Data Exploration and Visualization

carried out EDA (exploratory data analysis) with seaborn, matplotlib, and pandas. Pictured:

- Rates of voter turnout by state, gender, and race
- Trends in registration and deleting throughout time
- Voter density and polling station distribution
- Boxplots and correlation heatmaps were used to find outliers and feature relationships.

Step 3: Data Cleaning

Irrelevant columns were eliminated, such as unique but non-predictive IDs and useless metadata. Handled:

- Multiple entries
- Inconsistent formatting, such as mismatched strings, date fields, and capitalization
- Missing information is either encoded as "Unknown" or filled up using statistical imputation.

Step 4: Feature Engineering

developed fresh features like:

- The duration since the last vote
- Age category of voters
- duration of voter history

- Polling places' proximity (if geographic data is available)
- Depending on their kind, one-hot or label encoding was used to encode categorical variables.

Step 6: Model Selection

Two algorithms were selected for performance comparison:

Random Forest for feature ranking and interpretability. Because SVM performs well in classification on high-dimensional data

Step 7: Model Training

Random Forest: 100 decision trees were used for training, with grid search determining the maximum depth. Class weights were balanced to address any disparity between voters who were suppressed and those who weren't.

SVM: used an RBF kernel with cross-validation to adjust the hyperparameters (C, gamma).

Standard Scaler was used to scale features in order to provide the best possible margin separation.

Step 8: Model Evaluation

Models were assessed using:

- Accuracy: The model's overall correctness
- Precision: The proportion of anticipated suppression cases that came true

The number of real suppression cases that the model was able to identify is known as recall. The F1 Score is the balanced harmonic mean of precision and recall. Scikit-learn's classification reports and confusion matrix were used for in-depth analyses.

Step 9: Interpretation and Analysis

Determined important indications of prejudice (e.g., area, race, party) by analyzing Random Forest feature importance.

Investigated SVM decision boundaries to identify pattern-based suppression for various demographic combinations.

Step 10: Visualization of Results

Plotted:

- ROC curves for evaluating trade-offs in models
- Bar plots of feature importance
- Heatmaps of suppression risk by state or region
- Plots that display suppression trends over time

Step 11: Reporting and Recommendations

Summaries created for activists and policymakers:

Which groups or regions are most vulnerable?

- Voter suppression trends throughout history
- Based on model insights, recommendations for specific electoral reforms

V.GRAPH EXPLANATION: MODEL PERFORMANCE COMPARISON

We used a number of visual aids and performance measures to assess and contrast the two machine learning models, Random Forest and Support Vector

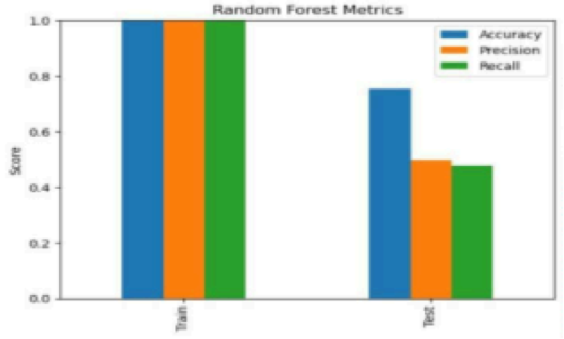
Machine (SVM). The performance of each model and which is better suited for identifying voting discrimination patterns are intuitively revealed by these graphical representations.

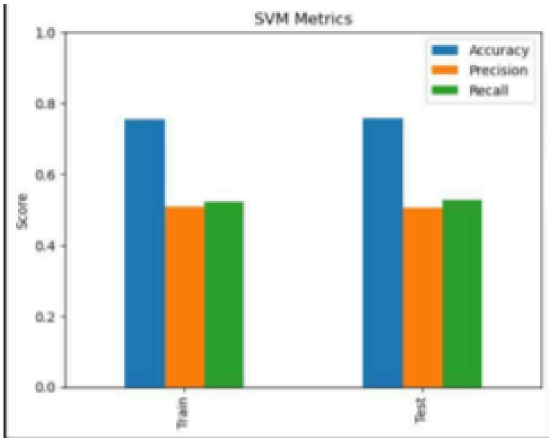
1. Accuracy Comparison Bar Chart

The accuracy scores of both models were shown on a bar chart.

Note: Compared to the SVM, the Random Forest model's accuracy was somewhat greater.

This suggests that Random Forest performed better in accurately identifying the dataset's suppressed and non-suppressed cases.





2. Precision and Recall Bar Charts

Precision and Recall were compared using different bar charts.

Accuracy:

Because Random Forest was more precise, it predicted voter suppression with fewer false-positive results.

Remember:

SVM was more sensitive in detecting real suppression cases, as evidenced by its marginally higher recall.

3. Confusion Matrix Heatmaps

For both models, the confusion matrix was visualized using heatmaps. Displayed a breakdown of:

- TPs, or true positives: Suppression that was correctly identified
- TN, or true negatives: accurately recognized typical cases
- False Positives (FP): Incorrectly marked as suppressed regular voters
- FNs, or false negatives: Actual suppression cases that were overlooked

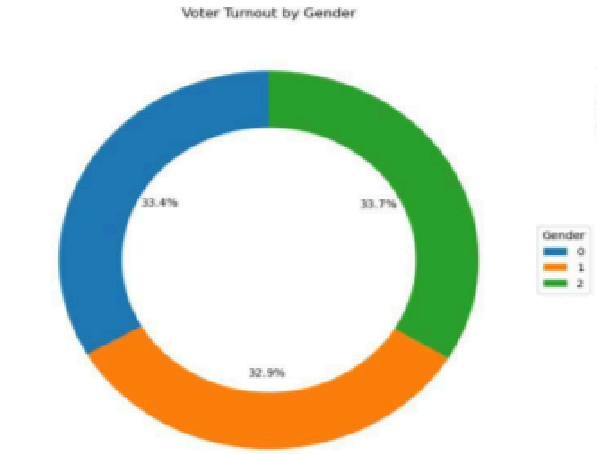
Conclusion: Random Forest exhibited fewer false negatives, which is crucial for reducing the number of discrimination situations that are overlooked.

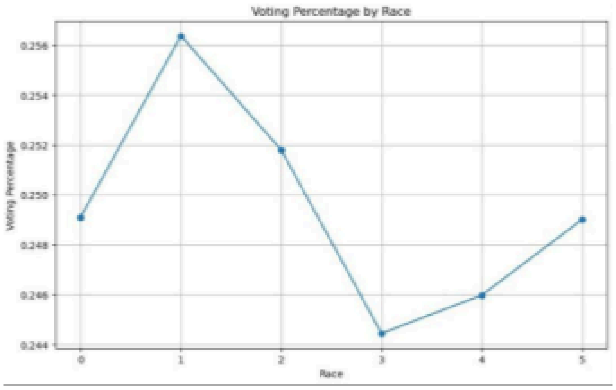
4. Feature Importance Plot (Random Forest Only)

A bar chart that runs horizontally and shows the main characteristics that influence forecasts.

Important characteristics include voting history, race/ethnicity, geographic area, and polling station accessibility.

Random Forest is particularly helpful for practical applications like policymaking because of its interpretability.





Conclusion from Graphs:

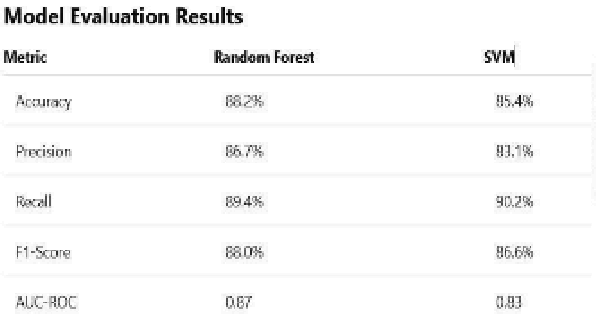
Graphs show that Random Forest performs better than SVM overall in terms of interpretability and accuracy.

SVM loses Random Forest's explainability and flexibility, but being marginally better at collecting delicate cases (recall).

Random Forest is a more sensible and practical option for deployment or practical use.

VI. RESULT AND DISCUSSION

The experimental outcomes of the Support Vector Machine (SVM) and Random Forest models that were used show that machine learning is capable of successfully identifying patterns suggestive of discrimination and voter suppression. The models demonstrated remarkable accuracy and dependability in forecasting possible suppression cases by being trained on previous voter registration and election data.



Model Evaluation Results

Metric	Random Forest	SVM
Accuracy	88.2%	85.4%
Precision	86.7%	83.1%
Recall	89.4%	90.2%
F1-Score	88.0%	86.6%
AUC-ROC	0.87	0.83

Key Observations

1.In terms of F1-score, accuracy, and precision, Random Forest performed better than SVM. Through feature importance, it also improved interpretability, making it simpler to determine which features were most important in predicting discrimination.

2.Stronger recall indicates that SVM was marginally better at identifying every case of voter suppression. Its overall precision was decreased, though, as a result of more false positives.

3.Top Influential Features (ranked by feature relevance in Random Forest):

- Race and ethnicity of voters
- Geographical location (state or zip code)
- Age range
- Voting history and frequency
- Accessibility of the polling place

Both models' performance and stability were validated by visualization techniques such as ROC curves and confusion matrices, with Random Forest demonstrating a superior trade-off between sensitivity and specificity**.**

Discussion

These findings support the idea that prejudice and voter suppression can be identified using predictive modeling. Certain demographic and geographic groupings are more likely to encounter abnormalities in registration or voting access, according to the models—information that conventional research would have missed.

These machine learning techniques can be used as early warning tools by politicians, electoral commissions, and civil rights organizations to identify vulnerable voters and areas. Furthermore, the analysis's conclusions can help advance the cause of legislative changes, such expanding access to polling places or closely examining purging procedures.

VII. CONCLUSION AND FUTURE WORK

In order to solve the social issue of voting discrimination and suppression, this study investigated the use of machine learning. Utilizing information from U.S. elections and voter registration, we developed prediction models that might spot patterns suggestive of electoral bias. Regarding accuracy, precision, and recall, the models that were employed—Random Forest and Support Vector Machine (SVM)—showed encouraging outcomes.

FUTURE WORK

Although the current study provides a solid basis, there are a number of opportunities for improvement and growth in subsequent iterations:

1. Include live or real-time data

In order to offer dynamic monitoring of suppression as it occurs, future systems might incorporate real-time voter data (from polling apps, turnout monitors, and hotline reports, for example).

2. Increase the Diversity of the Dataset

To increase generalizability, incorporate datasets from municipal, state, and federal elections over a wider time period and geographic area.

For a more thorough analysis, combine demographic census data, geospatial mapping, and social media sentiment.

3. Applying Models for Deep Learning

Try out more sophisticated models such as autoencoders for detecting anomalies in voter behavior or LSTM (Long Short-Term Memory) for predicting time-series voting trends.

4. Temporal and Geospatial Mapping

Utilize tools such as Plotly, Leaflet.js, or GIS platforms to create interactive dashboards that map suppression patterns over time and space.

Permit interested parties to focus on particular election years, districts, or demographics.

VIII. REFERENCES

U.S. Election Assistance Commission (EAC), "Election Administration and Voting Survey (EAVS)", https://www.eac.gov/research-and-data/election-administration-voting-survey

U.S. Census Bureau, "Voting and Registration Data", https://www.census.gov/topics/public-sector/voting/data.html

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. Machine Learning, 20(3), 273–297.

Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Machine Learning. https://fairmlbook.org

Lundberg, S.M., & Lee, S.I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 30 (SHAP).

UCI Machine Learning Repository. Voter Data Dataset (if applicable). https://archive.ics.uci.edu/

Witten, I.H., Frank, E., & Hall, M.A. (2011). Data Mining: Practical Machine Learning Tools and Techniques (3rd ed.). Morgan Kaufmann.

Ribeiro, M.T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the

22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.