**Political Analysis 1962-2019**

**SAIF KHAN**

**Business Analyst Trainee,**

**Uncodemy, Noida**

**Abstract:**

This research focuses on analyzing political insights from election data. With increasing complexities in elections and diverse political landscapes, understanding trends and patterns is crucial. This project uses various data analysis techniques to examine political party participation trends, gender representation, success rates in reserved seats, performance of independent candidates, and other key metrics. The primary goal is to provide a comprehensive overview of political dynamics over the years and identify significant shifts and patterns in the political arena.

**1. Problem Statement**

The project aims to predict and analyze political trends, focusing on different aspects such as party participation, female representation, success rates in reserved seats, performance of independent candidates, and more. Understanding these trends is valuable for political analysts, policymakers, and the general public to gain insights into the political landscape and its evolution over time.

**2. Introduction**

The rapid economic development and changing consumption patterns have led to significant shifts in political dynamics. People’s preferences and political affiliations have evolved, making it crucial to analyze historical data to understand these trends. This project addresses the imbalanced nature of political data and employs various techniques to provide a balanced and comprehensive analysis. The focus is on understanding how political participation, representation, and success rates have changed over time, providing insights into the evolving political landscape.

**3. Data Description**

The dataset includes various features related to political participation and performance. Key variables include:

* **Assembly, State\_Name, No\_Assemblies\_Contested, Assemblies\_Contested, Candidates\_Contested, Candidates\_Represented, Females\_Contested, Females\_Represented, SC\_Seats\_Contested, SC\_Seats\_Represented, ST\_Seats\_Contested, ST\_Seats\_Represented, BiPoll\_Contested:** No missing values.
* **Party\_Name, Party\_ID:** A few missing values.
* **Party\_Type:** Significant number of missing values.
* **Frequent\_Abbreviation, Last\_Abbreviation, Abbreviations:** A few missing values.
* **Start\_Year, Last\_Year:** Very few missing values.

**4. Steps Involved:**

**Exploratory Data Analysis**

After loading the dataset, we performed exploratory data analysis (EDA) to understand the relationships between the target variable (DEFAULT PAYMENT MONTH) and other independent variables. This helped identify key trends and patterns in the data.

**Null Values Treatment**

We addressed missing values by imputing them with appropriate techniques. Categorical columns were filled with "Unknown" and numeric columns were filled with median values to ensure data integrity and accuracy.

**Feature Selection**

We used algorithms like Extra Tree classifier to identify important features. This helped us focus on the most relevant variables for our analysis.

**Standardization of Features**

Data standardization was performed to ensure consistency and improve the performance of various algorithms. This step was crucial for achieving accurate results during model fitting.

**Balancing the Dataset**

We used oversampling techniques to balance the dataset, addressing the issue of class imbalance and improving the reliability of our analysis.

**Fitting Different Models**

We employed various classification algorithms, including:

1. Logistic Regression

2. K-Nearest Neighbors (KNN)

3. Random Forest Classifier

4. Naive Bayes

5. Decision Tree

6. Hyperparameter tuning on Random Forest algorithm

**Tuning the Hyperparameters for Better Accuracy**

Hyperparameter tuning was performed to enhance the performance of our models. Techniques like Grid Search CV, Randomized Search CV, and Bayesian Optimization were used to find the optimal hyperparameters.

**SHAP Values for Features**

SHAP (SHapley Additive exPlanations) value plots were used to determine the importance of features in the Random Forest model. This helped us understand which features had the most significant impact on the model's performance.

**5. Algorithms:**

**Logistic Regression:**

Logistic Regression is a classification algorithm that uses the sigmoid function to predict probabilities. The optimization algorithm used is Maximum Log Likelihood, and it helps in determining the likelihood of default.

**K-Nearest Neighbors (KNN)**

KNN is a simple, easy-to-implement algorithm that classifies data points based on their proximity to each other. The algorithm assumes that similar data points are close to each other.

**Random Forest Classifier:**

Random Forest is an ensemble method that creates multiple decision trees and aggregates their predictions to improve accuracy. It is a robust algorithm that reduces overfitting.

**Naive Bayes:**

Naive Bayes is a probabilistic classifier based on Bayes' theorem. It makes strong independence assumptions and is effective for classification tasks.

**Decision Tree:**

Decision Tree is a powerful tool for classification and prediction. It uses a tree-like structure to represent decisions and their possible consequences.

**6. Model Performance:**

**Confusion Matrix**

The confusion matrix summarizes the classification model's performance by showing the actual vs. predicted labels.

**Precision/Recall**

Precision is the ratio of correct positive predictions to the total positive predictions, while recall is the ratio of correct positive predictions to the total actual positive examples.

**Accuracy**

Accuracy is the ratio of correctly classified examples to the total examples. It is a measure of the overall effectiveness of the model.

**Area under ROC Curve (AUC)**

AUC represents the performance of the model by combining the true positive rate and false positive rate. It provides a comprehensive view of the model's classification performance.

**7. Hyperparameter Tuning:**

Hyperparameter tuning involves adjusting the model's parameters to optimize performance. Techniques used include:

**Grid Search CV**

Grid Search evaluates a predefined set of hyperparameters to find the best combination for the model.

**Randomized Search CV**

Randomized Search selects hyperparameters randomly within a specified range, increasing the chances of finding optimal parameters.

**Bayesian Optimization**

Bayesian Optimization uses a probabilistic model to find the best hyperparameters. It evaluates the model's performance at each iteration and adjusts the search accordingly.

**8. Conclusion:**

Using a Random Forest classifier, we achieved an accuracy of ~78.2% in predicting political trends. The most significant predictors included historical participation data, party affiliation, and demographic variables. Our analysis provided valuable insights into political dynamics, helping to understand the evolving political landscape.

This comprehensive analysis offers a detailed overview of political trends, highlighting key factors that influence political participation and representation. It serves as a valuable resource for political analysts, policymakers, and the general public to understand a