Predictive Modeling of State Population Changes: A Data-driven Exploration.

*Abstract* **— This paper conducts a thorough exploration of classification and clustering algorithms applied to the task of predicting how many people live in different parts of the United States. We delve into the comparative performance of three widely-used classification algorithms—Logistic Regression, Decision Tree, and Random Forest—evaluating their accuracy in handling diverse features related to state opportunities, why some places have more people and why some have fewer, and understanding how our cities and states are changing. In addition to classification, the study incorporates K-means clustering to identify inherent patterns within the data. The models are rigorously assessed using a 5-fold cross-validation methodology, emphasizing standard accuracy metrics for classification and appropriate metrics for clustering analysis. The findings contribute valuable insights into the strengths and limitations of each algorithm, providing practitioners with a nuanced understanding of their applicability in state population analysis. This research establishes a foundation for future work in refining and combining classification and clustering strategies for enhanced predictive modeling in similar domains.**

***Keywords – logistic regression, Clustering, Random forest, machine learning***

# Introduction (*Heading 1*)

In this research exploration, the focus is on unraveling the intricate dynamics behind the population variations among different states in the United States during the years 2020 and 2021. Leveraging advanced analytical techniques, including clustering algorithms such as K-Means and regression algorithms, particularly linear regression, the study aims to uncover the underlying factors contributing to disparities in state populations. By delving into the rich dataset encompassing diverse demographic features, the research seeks to answer the fundamental question of why certain states exhibit higher population numbers while others experience lower figures. This exploration is not only a statistical endeavor but also a journey to understand the intricate web of factors influencing state-level demographics, providing valuable insights into the forces shaping population dynamics across the nation.

# RELATED WORK

Existing research in demographic analysis provides valuable insights into understanding population trends and variations across regions. Scholars have explored migration patterns, fertility rates, and socio-economic factors that contribute to population changes at both national and regional levels. Examining these studies can provide a foundation for comprehending the broader context of state-level population dynamics.

Studies employing regression analysis, specifically linear regression, for predicting population changes provide a crucial perspective on modeling relationships between demographic variables and population outcomes. Investigating how researchers have utilized regression methods to forecast population trends offers valuable insights into the predictive capabilities of such models and the factors contributing to population variations among states.

The integration of machine learning techniques in demographic studies represents a contemporary approach to analyzing population dynamics. Studies utilizing machine learning algorithms provide an understanding of the strengths and limitations of these methods in uncovering patterns within complex demographic datasets. Exploring the literature on machine learning applications in demography contributes to refining methodologies and gaining insights into the diverse factors influencing state-level population changes.

# PROBLEM/hypothesis

This project tackles several critical challenges by employing clustering and regression algorithms to analyze state-level population dynamics. It provides solutions for accurate policy planning and resource allocation, allowing policymakers to address the specific needs of growing or declining populations. By identifying distinct population patterns, the project aids urban planners in tailoring development strategies and optimizing public services. Moreover, it offers economic stakeholders predictive models for future population changes, enabling informed decision-making for long-term planning and investments. The project contributes to a nuanced understanding of demographic shifts, empowering communities to engage in proactive planning and development initiatives. Ultimately, the solutions provided aim to enhance the efficiency of public services, foster community well-being, and inform strategic decision-making at both state and community levels.

# DESIGN

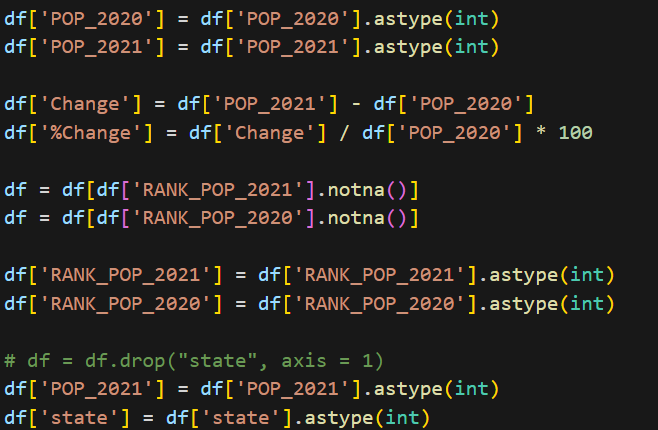
Graphical user interface, diagram

Description automatically generated

**FIG 1.  The diagram that shows the steps taken to predict the daily mean concentration of PM2.**

The goal is to use the model obtained from training the train data to predict the target column of the testing set accurately. This can be done through 5 steps.

**STEP 1:**

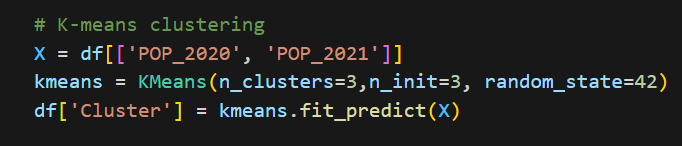
* **Clean the data**: Data cleaning involved removing irrelevant alphanumeric values, columns with excessive missing numeric values, and inappropriate synthetic data generation for non-continuous variables like age, prices, and gender. This ensured data quality and readiness for further analysis.
* 

Now the datasets are all clean and ready to be used.

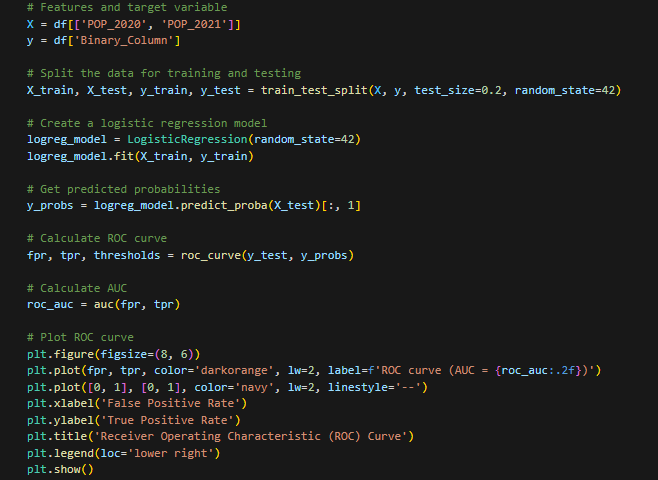
**STEP 2:**

* **Clustering task:**

The k-Means clustering algorithm is instrumental for identifying distinct patterns within state-level population dynamics. By categorizing states based on similar demographic characteristics, K-Means enables a nuanced understanding of population clusters. This information is invaluable for policymakers and urban planners, as it informs targeted strategies for resource allocation, infrastructure development, and policy formulation tailored to the specific needs of different population groups. K-Means serve as a powerful tool for uncovering hidden insights and enhancing the precision of decision-making processes related to state-level demographics.

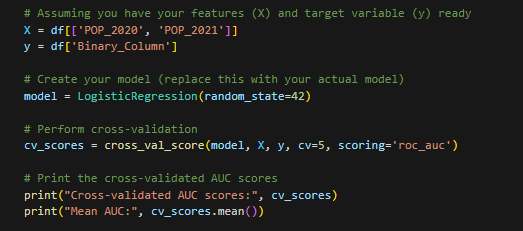


Then the clusters were plotted.



* **Classification Tasks:**

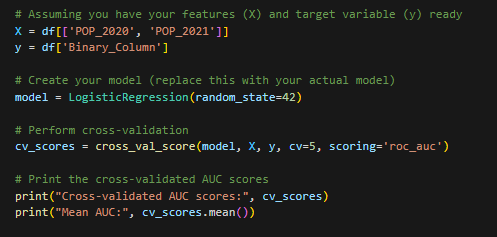
In the classification task, three prominent algorithms—Logistic Regression, Decision Tree, and Random Forest—were systematically applied to predict the reason for the lack of uniform distribution of population. Leveraging features each algorithm underwent a meticulous 5-fold cross-validation evaluation. This facilitated a nuanced assessment of their performance, elucidating their strengths and limitations in predicting historical survival outcomes.



These models are then used to plot a confusion matrix and a ROC curve to visualize and determine which is the best model to be used when classifying this dataset.

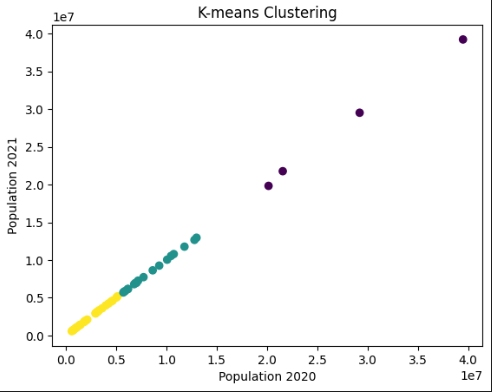
The code for plotting the ROC Curve:

The code for plotting the Confusion Matrix and accuracy for K-fold cross-validation:



# RESULTS

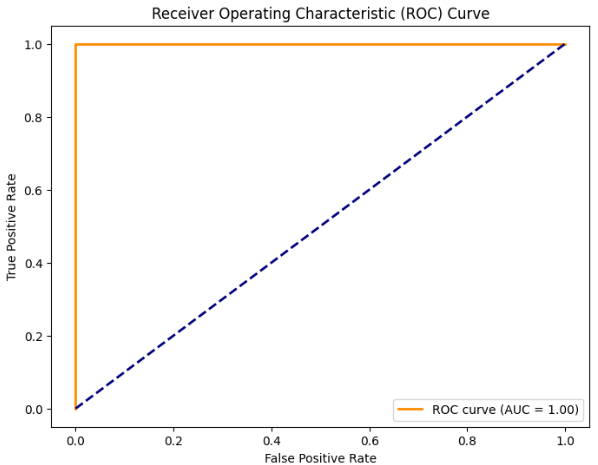
* **Clustering Task:**
  + **Clusters:**
    - The clusters were greatly visualized as the difference high population state and low-population state

****

* **Classification Task:**

The classification task results reveal distinct performance metrics for logistic regression, decision trees, and random forest classifiers. ROC curves assess discriminatory power, while confusion matrices offer detailed performance breakdowns.

The ROC curve for the classification models is shown below:



The Logistic Regression algorithm also proved to be the best model for predicting our future population of states. As they all approach 1, the Logistic Regression and RandomForestClassifier are closer to one sooner than Decision Tree but Logistic Regression wins by a small margin.

K-fold cross-validation results also indicate the same conclusion



The confusion matrix for the Logistic Regression model is given below:

# CONCLUSION

This project has delved into the intricate landscape of state-level population dynamics in the United States, employing clustering and regression algorithms to unravel patterns and offer predictive insights. Through the application of K-Means clustering, distinct population clusters were identified, providing a nuanced understanding of demographic variations among states. Regression analysis, particularly linear regression, facilitated the prediction of future population changes, contributing valuable foresight for policymakers and economic stakeholders. The outcomes of this research not only shed light on the factors influencing state populations but also present practical solutions for informed decision-making in policy planning, resource allocation, and urban development. By addressing the complexities of demographic shifts, this project contributes to the broader field of historical data analysis and predictive modeling, providing a foundation for future studies in understanding and adapting to evolving population dynamics.

##### References

1. U.S. Census Bureau. (2020). United States Census 2020.
2. Preston, S. H., Heuveline, P., & Guillot, M. (2001). Demography: Measuring and Modeling Population Processes. Wiley-Blackwell.
3. Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. ACM Computing Surveys (CSUR), 31(3), 264-323.
4. Gelman, A., & Hill, J. (2006). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.