****

**Photo credit:** [**https://www.11alive.com/article/news/crime/clayton-county-man-arrested-trafficking-case/85-329bae00-544d-424d-9efe-37379166e3b5**](https://www.11alive.com/article/news/crime/clayton-county-man-arrested-trafficking-case/85-329bae00-544d-424d-9efe-37379166e3b5)

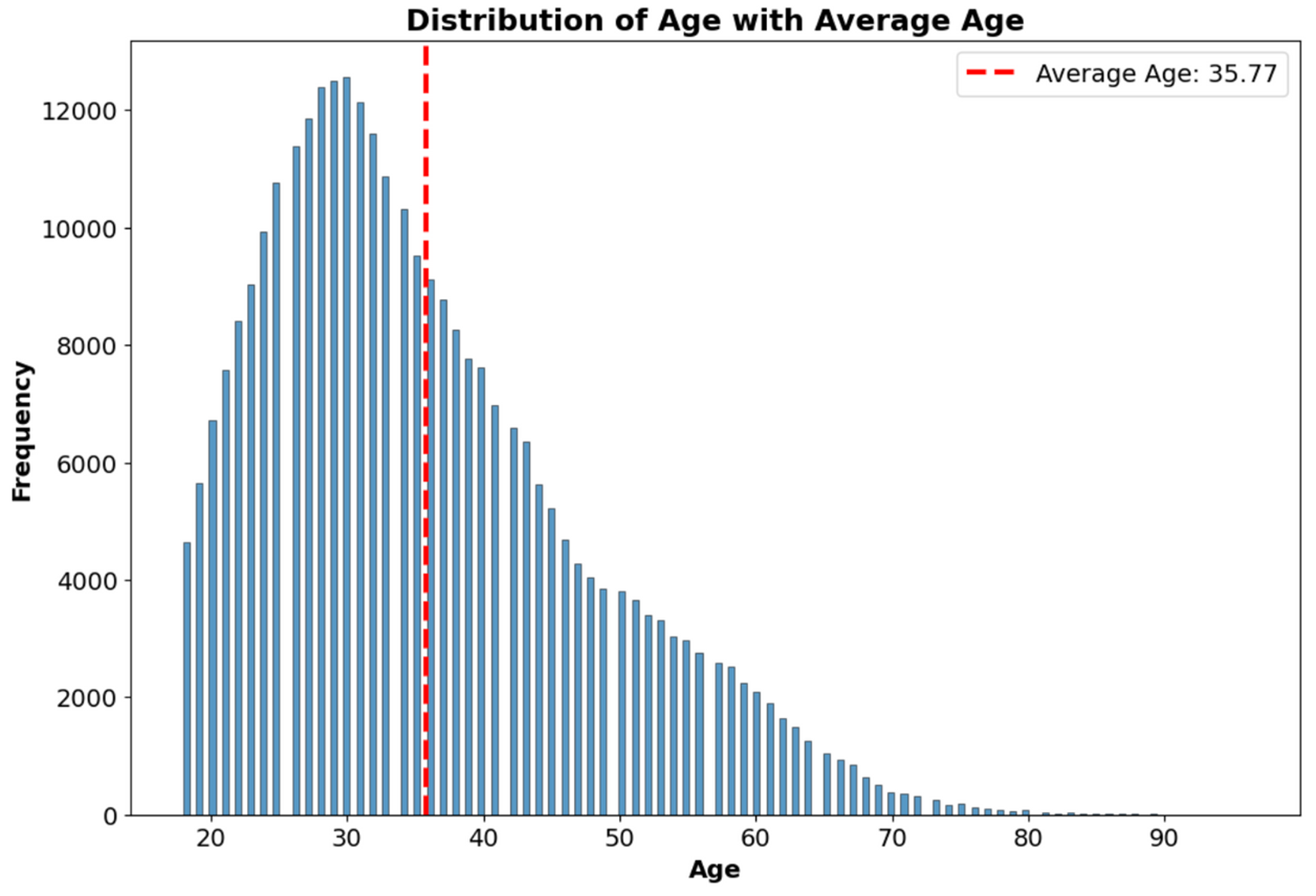
**Predicting Arrest Demographics in Los Angeles: A Classification-Based Approach**

**Introduction**The intersection of machine learning and criminal justice presents opportunities to better understand crime patterns, demographic trends, and the underlying factors influencing arrest rates. Our goal was to develop a predictive model that accurately predicts the age, gender, and ethnicity of individuals arrested based on historical data. By leveraging a Los Angeles Arrest Dataset spanning from 2020 to 2025, we sought to uncover trends and build predictive models that could provide deeper insights into criminal justice patterns.

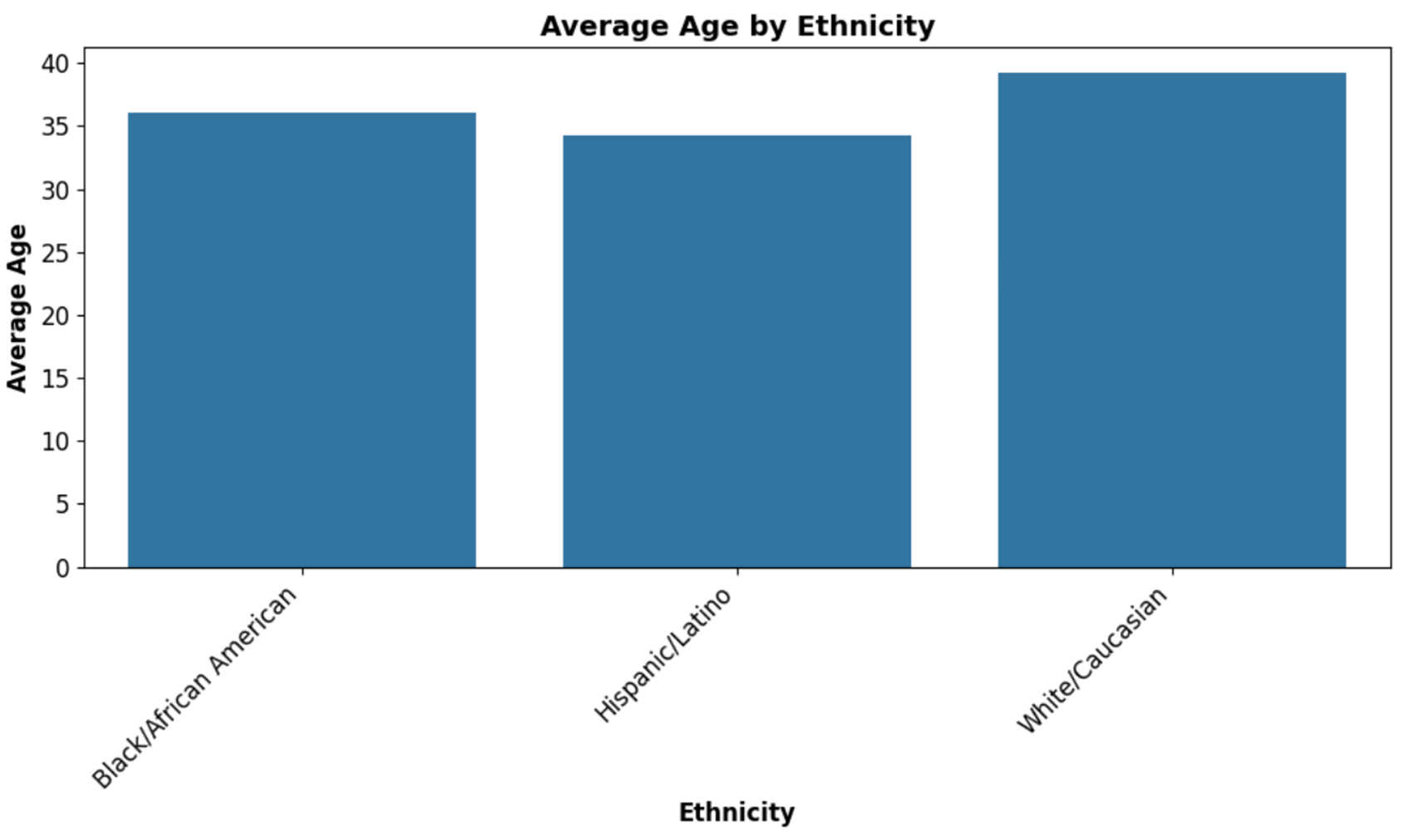
**Data Analysis**Our dataset consisted of over 297,000 arrest records, covering individuals aged 18 and older. To begin we identified key patterns within the data that allowed us to decide what we wanted to form our model around. Some observations we found were:

Seasonal Arrest Trends: Arrest rates peaked in January, May, and August, while the last quarter of the year consistently showed a decline. This could be due to shifts in law enforcement focus, seasonal crime trends, or even societal factors such as holiday-related behavioral changes.

Gender Disparity: The dataset revealed a significant gender imbalance, with 80% of arrests involving males and only 20% involving females. Understanding why this disparity exists could require additional sociological research beyond the scope of our study. The average age at arrest was 36 years for men and 34 years for women. This slight difference may be influenced by varying criminal behavior trends across age groups.



Ethnicity Trends: Different ethnic groups exhibited varying average ages at the time of arrest. Hispanic individuals had a mean age of 34, African Americans 36, and White individuals 39. These differences could be tied to socioeconomic factors, geographic distributions, or law enforcement policies.



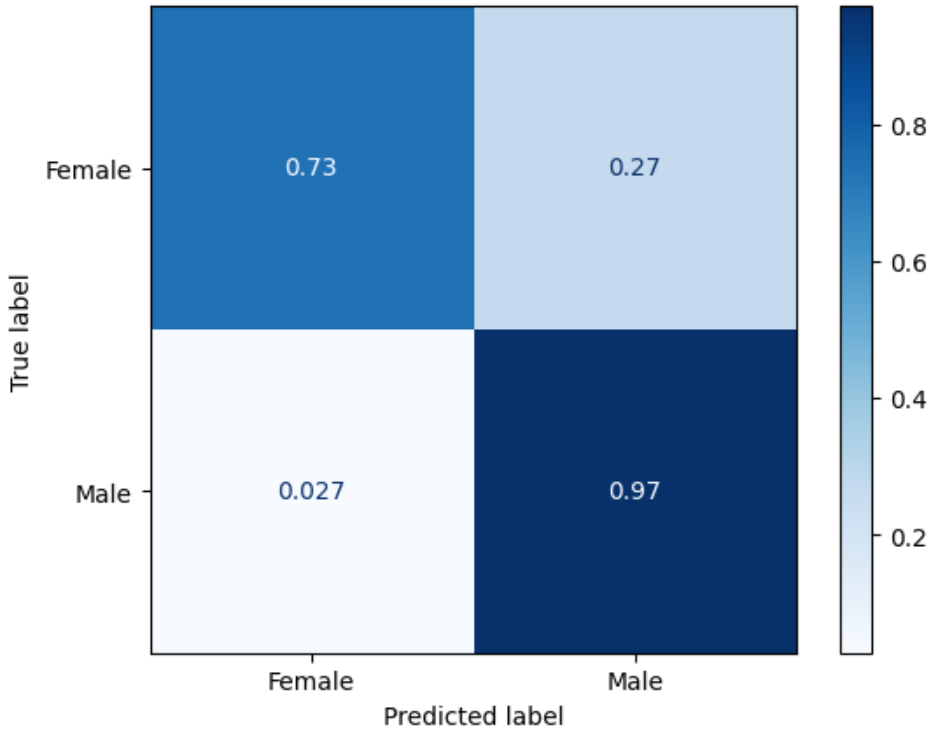
Beyond these general trends, we explored correlations between different features, such as booking time, charge descriptions, and location data. Geographic distribution of arrests proved particularly interesting, as certain neighborhoods exhibited distinct arrest patterns, indicating that location-based factors might play a critical role in law enforcement activity.

**Modeling Approach & Results**

To predict gender, age group, and ethnicity, we employed multiple machine learning models, ensuring a rigorous approach by testing different techniques. Each demographic followed the same order for Data Processing: Data Cleaning, Feature Engineering, SMOTE, GridSearchCV, Hyperparameter Tuning, & lastly training the model. We then took the best models and ensemble them to ensure better results.

**Gender Model**

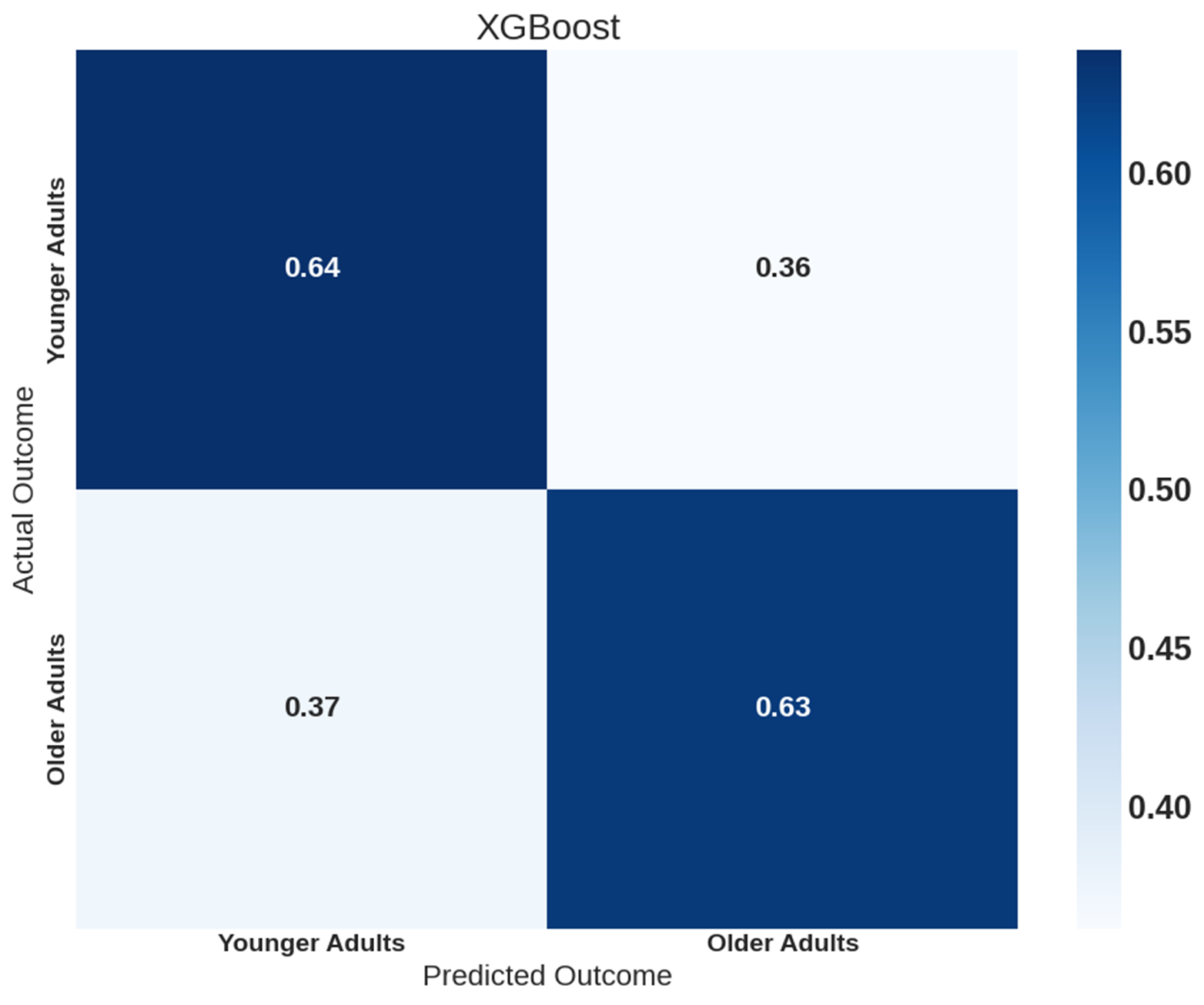
The imbalance in gender distribution (80% male, 20% female) required the use of SMOTE (Synthetic Minority Oversampling Technique) to balance the dataset. Our best-performing model was a Random Forest Classifier, fine-tuned with GridSearchCV, which achieved an accuracy of approximately 85%.



Decision Trees and XGBoost produced comparable results (~80-84%). Next we tried to ensemble the models together hoping that the combination of the models would help possibly provide us with better results. However, this model really only made it to about 83% accuracy. The biggest challenge in gender prediction was the tendency to misclassify females as males. There were certain Area ID’s and Reporting Districts that were heavily male-dominated in the training set that led the model to misclassify them. We also observed unintended patterns, such as spikes in arrests on specific days (1st, 15th, and 30th), which the model inadvertently learned. These factors played a more significant role in distinguishing gender than initially expected.

**Age Group Model**

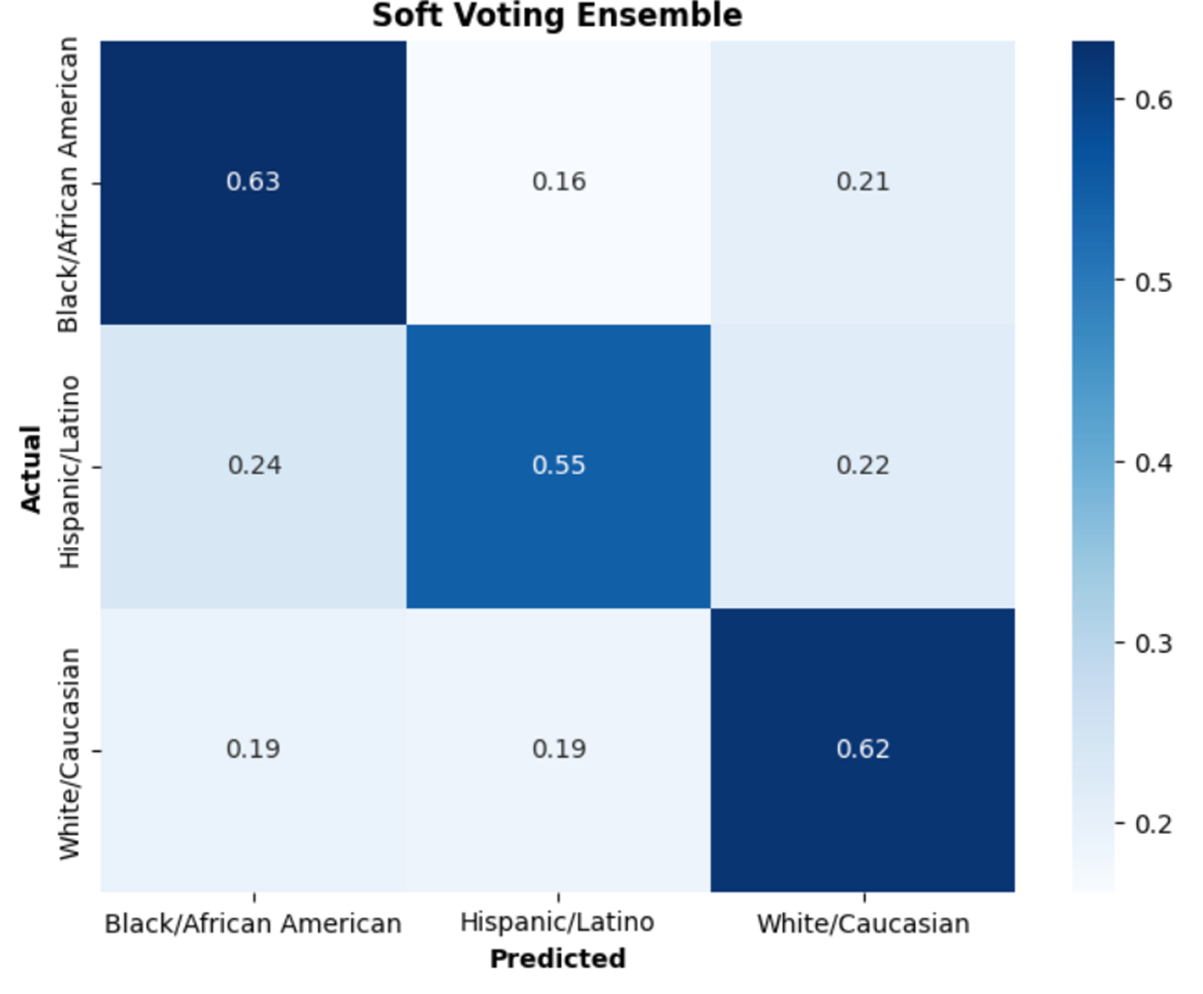
We classified individuals into two age groups: younger adults (18-33) and older adults (33+). XGBoost performed best, achieving 63% accuracy, followed by Decision Trees at 59%. We experimented with ensembling the models to improve accuracy, finding that soft voting balanced the predictions but was weaker overall. In contrast, hard voting improved predictions for younger adults at the expense of accuracy for older adults.



A major challenge was that the model struggled to differentiate between someone aged 33 and 34, leading to near-random classification when individuals were close to the threshold. Most of the features in the model were not highly correlated with the Age column, making it difficult for the model to find a strong predictive pattern within the data. This resulted in lower accuracy and more variance in the predictions. A deeper analysis of feature importance revealed that the top correlated features for ethnicity prediction included report type, ethnicity, booking time, sex code, and charge group description. However, when it came to looking at what the most important factors were for the models, the Decision Tree only used one of the top five correlated features and the XGBoost used four out of the five correlated features.

**Ethnicity Model**

This was the most challenging demographic factor to predict. First, we began by splitting the ethnicities into 3 major groups being Black/African American, White/Caucasian, and Hispanic/Latino. There were other ethnicities in the dataset, but the number of rows was minimal, leading to us dropping them from this prediction. We then tested K-Nearest Neighbors (KNN), Decision Trees, and XGBoost, with the best-performing model XGBoost reaching only 59% accuracy. Ensembling the models did not show significant improvement, remaining around 60% accuracy.



Ethnicity predictions were difficult due to weak correlations with most features, except for geographic location, which strongly influenced model predictions. The ethnicity model primarily relied on location-based trends, making it prone to bias towards the predominant ethnic group in a given area. A deeper analysis of feature importance revealed that the top correlated features for ethnicity prediction included booking location, longitude, latitude, and booking location code. The reliance on location variables suggests that without additional context the model struggles to make unbiased predictions. By categorizing neighborhoods into broader crime zones, we were able to see where the model was struggling or doing well with the predictions.

**Conclusion**While our models performed well in predicting gender, age and ethnicity proved more challenging due to weak feature correlations. The findings highlight the potential of machine learning in analyzing arrest trends but also underscore the importance of data quality and ethical considerations. Future improvements could involve incorporating richer datasets with socioeconomic and geographic variables to enhance predictive accuracy.