***TEXAS INMATES DEMOGRAPHICS AND PAROLE DECISIONS***

**1. Abstract (10 pts)**

*This research analyzed data on inmates in Texas prisons. The aim is to identify what factors if any impacts whether an inmate will be granted parole or not. It also delves into the race and gender composition of the inmates. Bar graphs, histogram and scatter plots were used to display variables characteristics. Linear model, Support Vector Machine and Decision Tree models were used in different scenarios to make predictions on inmates age and whether they were Approved or Denied at their last Parole hearing.*

**2. Introduction (20 pts)**

*Texas has a population of over 29 million. It has 3 cities with more than 1 million people: Dallas, Houston, and San Antonio. The state is not immune to crime and violent cases. These occurrences threaten our collective sense of safety and security. Individuals who are apprehended for violating the law can be sentenced to prison when deemed guilty by court processes. Research questions are:*

1. *Predicting age of inmates.*
2. *Predicting Parole decision.*

*Inmate offences include murder, rape, sexual assault, and robbery. Their sentences range from a certain number of years to Life, Death or Life Without Parole (LWOP).* Although males make up about 49.6% of the Texas population, they made up 93% of the inmates’ population. This report examines the demographics of inmates using visualization techniques and predicts certain variables using machine learning algorithms.

**3. Dataset (30 pts)**

The on-hand inmate population dataset ([**https://catalog.data.gov/dataset/high-value-dataset-december-2023**](https://catalog.data.gov/dataset/high-value-dataset-december-2023) **)** is published by the city of Austin. This downloaded data comprises on hand inmate population with relevant demographic, offense, and parole information. Data period is 6 months (June - December 2023).

**4. Exploratory Data Analysis (EDA) (60 pts)**

Dataset consists of 129,513 rows with 17 variables. The youngest inmate is 17 years and the oldest is 92.

Most of the variables in the dataset were categorical in nature, hence bar and scatter plots were used to show their characteristics as depicted in the diagrams below. The density curve of inmates age is not normally distributed, it is skewed right with a mean of 41 year. There were also more denied parole cases than approved ones. Inaddition, a large number of inmates had not had any parole hearing so their Last Parole Decision were clasified as None.

Black, Hispanic, and White race make over 90% of the inmates population. It is therefore not surprising that, they form a large portion of the Texas population. More men across all race were serving LWOP sentences than women. Inmates on LWOP were classified as Not Applicable with regards to the Last Parole Decision.

A screen shot of a graph

Description automatically generatedA screen shot of a graph

Description automatically generatedA graph of a person and person

Description automatically generatedA graph of a person in prison

Description automatically generatedA graph of a person's choice

Description automatically generatedA screen shot of a graph

Description automatically generated

**5. Models (75 pts)**

Since the variables being used for the model are categorical, as.factor() was used to convert them before fitting the model.

**Linear Model**

This model predicts inmates Age, using Gender, Race and Parole Eligibility. The model predicted the inmates’ age with an average error of 12. The correlation of 0.2 indicates there is no significant relationship between Gender, Race, Parole Eligibility and Age of inmates.

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

**Decision Tree**

This model predicts Last Parole Decision using Gender, Race, and Parole Eligibility.

Despite an 84% accuracy, the model was inefficient at predicting Last Parole Decisions that were Approved. It could not predict a single Approved Last Parole Decision. This could be a case of over-fitting as only 10% of the 129,513 dataset is Approved.

A screenshot of a computer screen

Description automatically generated

**6. Summary of learning (5 pts)**

*To achieve a very robust and reliable prediction, the nature of dataset and the desired outcome of research need to guide which model is selected. Also, when dealing with skewed data, further techniques need to be applied to enhance model performance. In addition, the dataset is biased to other gender categories like transgendered because only Male and Female gender categorizations were recorded.*