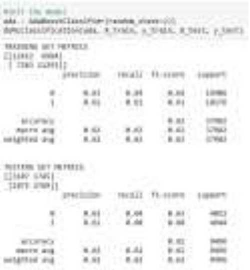
Executive Summary

I. Data Collection and Cleaning

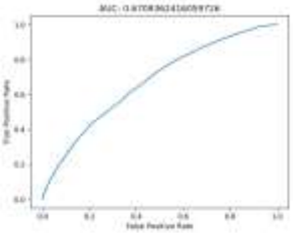
The Washington Post gathered data on criminal homicides from 50 major American cities over a decade, detailing incident locations, arrest status, and victim demographics. The dataset underwent meticulous cleaning and standardization, involving months of work to ensure accuracy. The primary dataset, obtained from Kaggle, encompassed information on over 52,000 homicides. Additionally, predicted population data from the US Census was incorporated to analyze the relationship between city size and homicide rates.

II. Machine Learning and Model Performance

We employed the LGBMClassifier algorithm to create a predictive model. The model demonstrated respectable performance metrics, with precision, recall, F1 scores, and accuracy ranging from 62% to 67%. The Area Under the Curve (AUC) score of 68% indicated the model's effectiveness in making predictions.



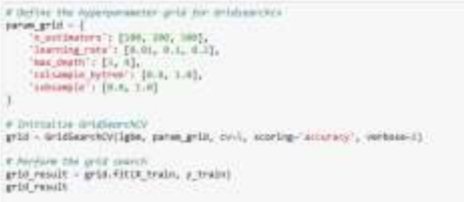
```
# Fit the model
ada = AdaBoostClassifier(random_state=[illegible])
[illegible](ada, X_train, y_train, X_test, y_test)

TRAINING SET METRICS
[illegible]
              precision    recall  f1-score   support

           0       [illegible]
           1       [illegible]

    accuracy       [illegible]
   macro avg       [illegible]
weighted avg       [illegible]

TESTING SET METRICS
[illegible]
              precision    recall  f1-score   support

           0       [illegible]
           1       [illegible]

    accuracy       [illegible]
   macro avg       [illegible]
weighted avg       [illegible]
```



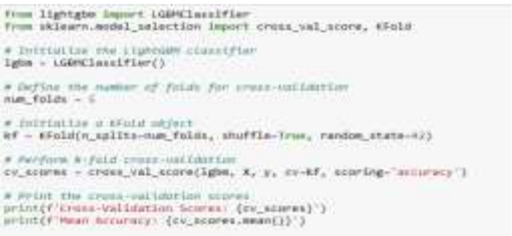
i.

III. Hyper parameter Tuning and Validation

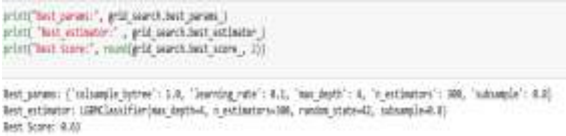
a. Despite efforts to enhance the model through hyper parameter tuning using GridSearchCV, the accuracy score remained at 63%, consistent with the original model. Cross-validation and K-fold analysis validated this finding, yielding a consistent mean accuracy of 63%.



```
# Define the hyperparameter grid for GridSearchCV
param_grid = {
    'n_estimators': [100, 200, 300],
    'learning_rate': [0.01, 0.1, 0.2],
    'max_depth': [3, 4],
    'colsample_bytree': [0.8, 1.0],
    'subsample': [0.8, 1.0]
}

# Initialize GridSearchCV
grid = GridSearchCV(lgbm, param_grid, cv=[illegible], scoring='accuracy', verbose=[illegible])

# Perform the grid search
grid_result = grid.fit(X_train, y_train)
grid_result
```

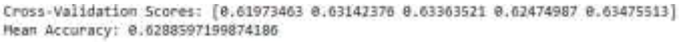


```
from lightgbm import LGBMClassifier
from sklearn.model_selection import cross_val_score, KFold

# Initialize the LightGBM classifier
lgbm = LGBMClassifier()

# Define the number of folds for cross-validation
num_folds = 5

# Initialize a KFold object
kf = KFold(n_splits=num_folds, shuffle=True, random_state=42)

# Perform K-fold cross-validation
cv_scores = cross_val_score(lgbm, X, y, cv=kf, scoring='accuracy')

# Print the cross-validation scores
print(f'Cross-Validation Scores: {cv_scores}')
print(f'Mean Accuracy: {cv_scores.mean()}')
```



```
print("Best params:", grid_search.best_params_)
print( "Best estimator:" , grid_search.best_estimator_)
print("Best Score:", round(grid_search.best_score_, 2))
```

```
Best_params: {'colsample_bytree': 1.0, 'learning_rate': 0.1, 'max_depth': 4, 'n_estimators': 300, 'subsample': 0.8}
Best_estimator: LGBMClassifier(max_depth=4, n_estimators=300, random_state=42, subsample=0.8)
Best Score: 0.63
```



```
Cross-Validation Scores: [0.61973463 0.63142376 0.63363521 0.62474987 0.63475513]
Mean Accuracy: 0.6288597199874186
```

IV. Tableau

We created two tableau dashboards.

Distribution of homicides overall rates & death per 100,000 people.

V. Website Development

a. The website aims to provide valuable insights into homicide data from 2012 to 2017. It features an 'Arrest Prediction Page' offering predictive insights and a 'Dataset Overview Page' presenting a comprehensive view of the dataset. The website utilizes Python, Flask, Tableau, and was deployed on PythonAnywhere. The website employs a dark color palette to maintain a tone of seriousness, respecting the gravity of the subject matter—homicide data representation. It prioritizes visual engagement while acknowledging the sensitive nature of the topic.

VI. Conclusions

- The dataset exhibits a right-skewed distribution of victim ages, with a mean age of 31.8. The age range spans from newborns or babies (minimum age of 0) up to a maximum age of 102. Additionally, there are notable outliers in the data.
- The dataset reveals that the most prevalent age group among victims falls within the range of 19-29 years old. Following closely, the 30-44 age group is the next most common. Conversely, the 65+ age range exhibits the lowest frequency among the victim demographics.
- Although the LGBM model holds potential for predictions, achieving an accuracy of 63% may not meet the required level of reliability. It's crucial to acknowledge the dataset's limitations: it encompasses data from only 50 US states, spanning 2007-2017, with New York contributing only two years of data. With a more comprehensive and uniformly detailed dataset, we could anticipate more precise and dependable outcomes.