**COURSEWORK SUBMISSION FORM**

|  |  |  |  |
| --- | --- | --- | --- |
| **STUDENT USE** | | **STAFF USE** | |
| Module Name | Machine Learning and Data Analytics | First Marker’s  (acts as signature) |  |
| Module Code | 6COSC017C-n | Second Marker’s  (acts as signature) |  |
| Lecturer Name | Hamid Shahbazkia | Agreed Mark |  |
| UoW Student IDs |  | **For Registrar’s office use only (hard copy submission)** | |
| WIUT Student IDs | 00010023 |
| Deadline Date | 06.06.2023 |
| Assignment Type | Individual |
| Word Count | 1287 |

**SUBMISSION INSTRUCTIONS**

**COURSEWORKS *must* be submitted in *both* HARD COPY (to the Registrar’s Office) *and* ELECTRONIC unless instructed otherwise.**

For hardcopy submission instructions refer to: <http://intranet.wiut.uz/Shared%20Documents/Forms/AllItems.aspx> - Coursework hard copy submission instructions.doc

For online submission instructions refer to: <http://intranet.wiut.uz/Shared%20Documents/Forms/AllItems.aspx> - Coursework online submission instructions.doc

|  |
| --- |
| **MARKERS FEEDBACK (Continued on the next page)** |
|  |

Table of Contents

[Introduction 1](#_Toc136339751)

[Exploratory Data Analysis Overview 2](#_Toc136339752)

[Dataset Preparation 6](#_Toc136339753)

[Implementation and Comparison of Machine Learning Algorithms 7](#_Toc136339754)

[Confusion Matrix 8](#_Toc136339755)

[Classification Metrics 8](#_Toc136339756)

[Conclusion 9](#_Toc136339757)

[References 10](#_Toc136339758)

# Introduction

This analysis is centered on fatalities in the United States involving law enforcement from 2000 to 2016. The data set reveals the number of individuals who lost their lives while resisting the police. It contains 12 principal columns featuring details such as the person's name, age, sex, ethnicity, city, state, manner of death, whether they were armed, had a mental disorder, and if they attempted to flee. The core objective of this analysis is to determine the root causes of these individuals' conflicts with the police. Given the data set includes information on the mental health status of these individuals, the study will focus on distinguishing between those who were mentally impaired and those who were not. Finally, we will offer pertinent comparisons on the number of mentally ill and mentally sound individuals who engaged in violent confrontations with the police, leading to their deaths.

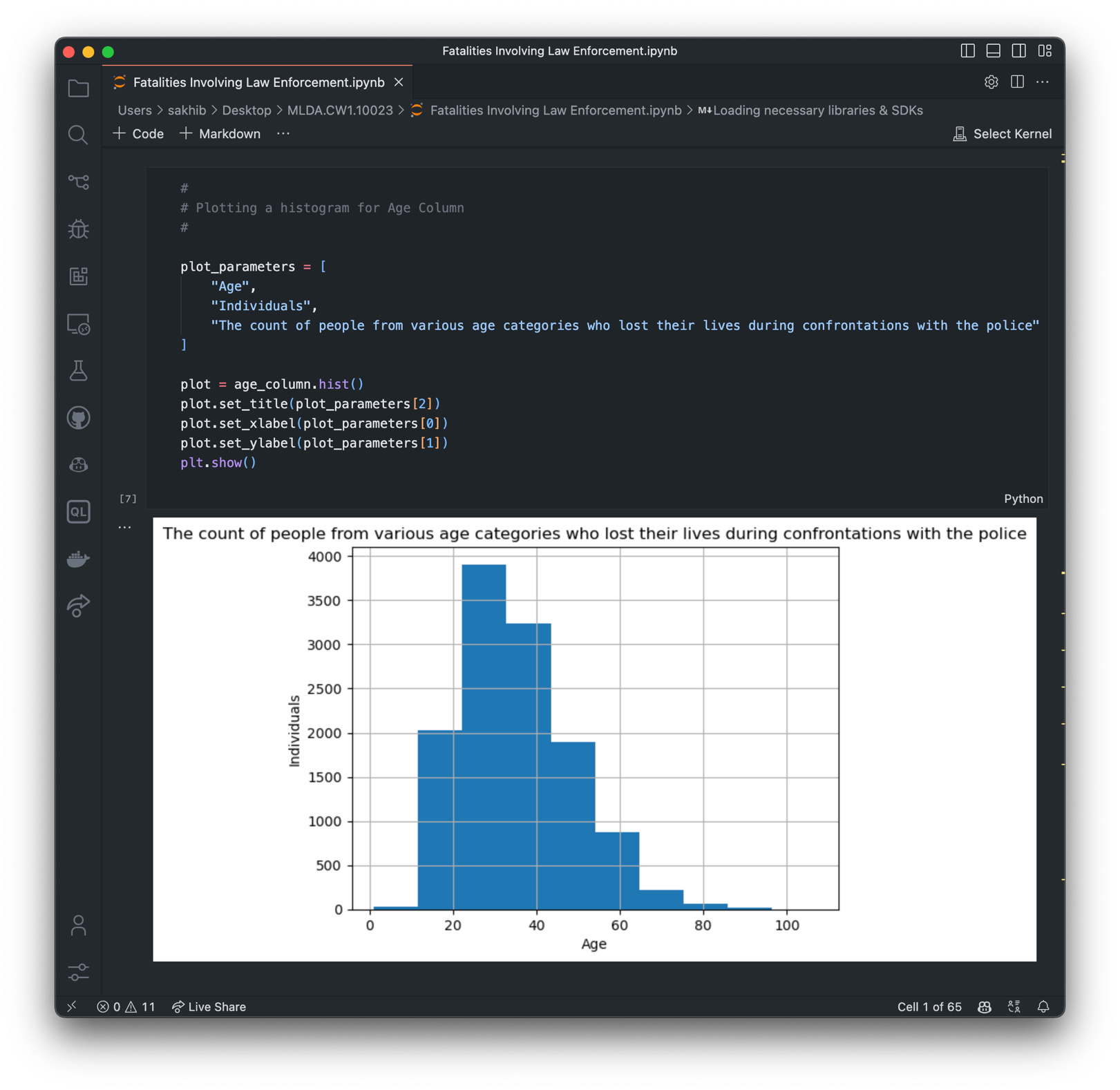
Link to the Data Set:

<https://data.world/awram/us-police-involved-fatalities/workspace/file?filename=Police+Fatalities.csv>

# Exploratory Data Analysis Overview

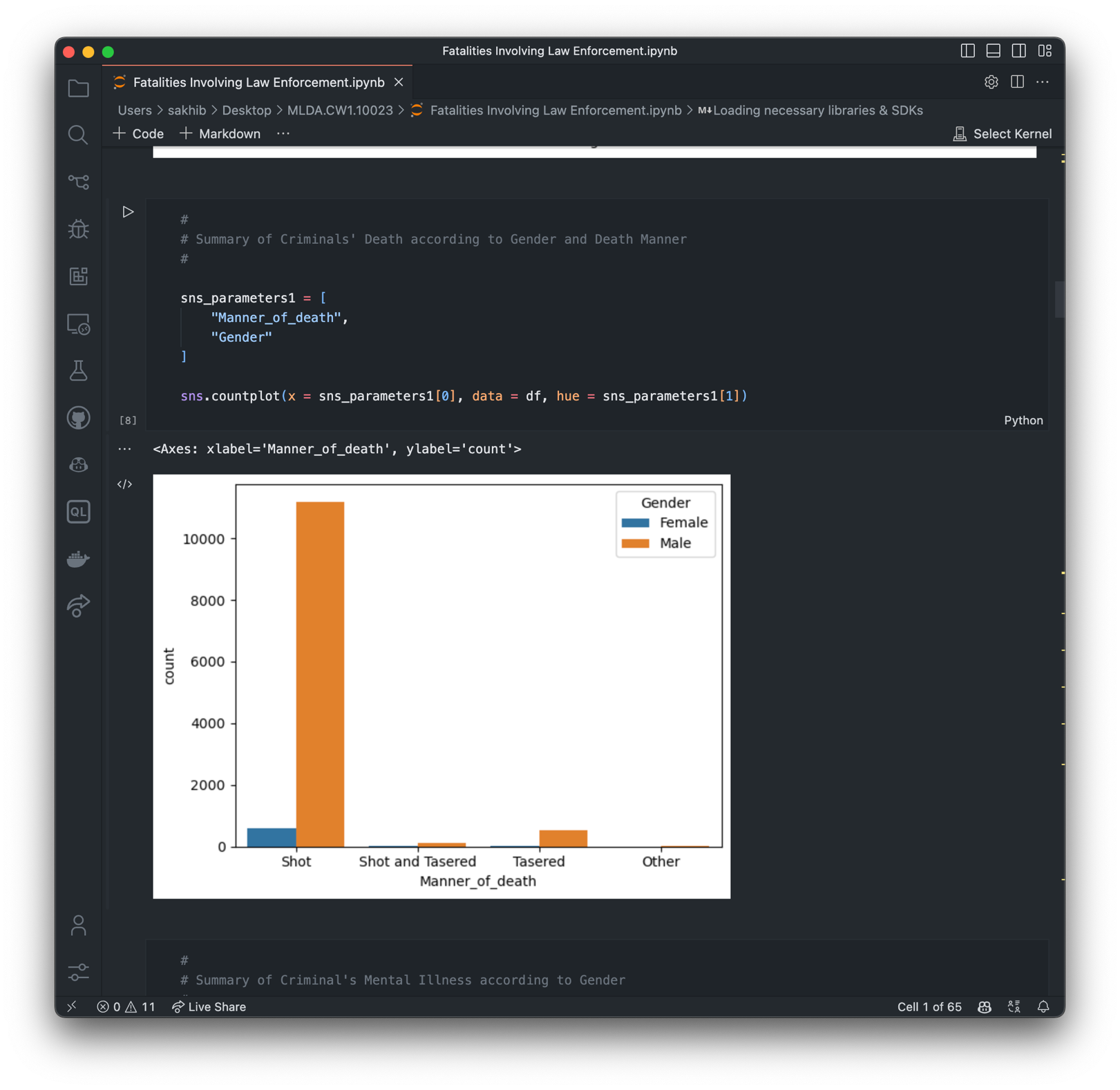
The dataset is both extensive (comprising 12491 rows and 12 columns) and somewhat unique in its initialized data types. While four data types are present in the dataset (Integer, Float, Object, Boolean), it's noteworthy that only the Age and UID columns fall under numeric data types, with the majority of the other columns predominantly comprising of Object and Boolean data types. Consequently, the dataset's central tendency is primarily oriented around the age group.

The analysis reveals that the average age of individuals who died was 35 years, with a standard deviation and variance approximately equal to 13 and 165, respectively. As indicated by the accompanying bar chart, the age group between 20 and 40 years accounted for the majority of the deaths. However, it is noteworthy that the number of fatalities among people in their 50s and beyond was significantly lower compared to their younger counterparts.

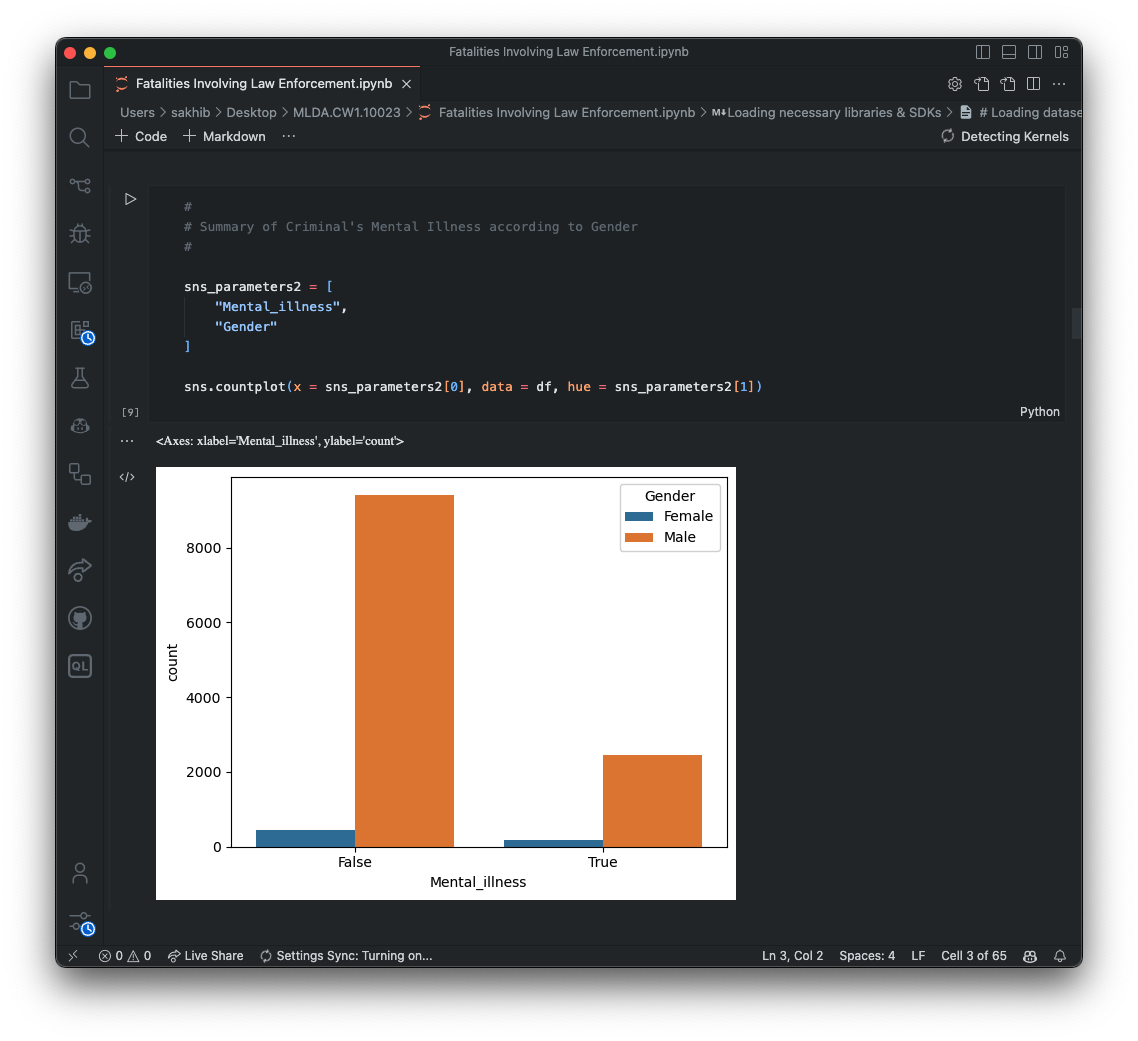


(Source: Practical Part)

In terms of the cause of death for both genders, the graph below clearly shows that a majority of individuals were killed by gunfire. Specifically, the count for men and women who were shot exceeded 10,000 and approximately 500, respectively. In contrast, the number of individuals who died from other causes, like being restrained, was considerably lower.



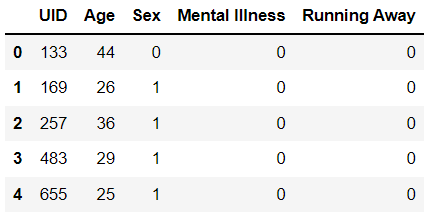
(Source: Practical Part)

The subsequent bar chart demonstrates the mental health status of both men and women. The data suggests that the majority of individuals did not have any mental health issues. To illustrate, the count of men without mental health problems was three times greater than those suffering from mental health issues. Likewise, the number of women without mental health issues was double that of women identified as mentally ill.

# Dataset Preparation

Initially, the presence of missing values was investigated. It was discovered that three columns had missing values: Age, Race, and Armed. Given that the number of null values in the Age group only accounted for 1.86%, these were replaced with the mean (average age). However, due to a substantial number of null values in the Race and Armed columns, these columns were excluded from the dataset. The next step was to assess the dataset's suitability for model training. As previously mentioned, the majority of the columns consisted of object and boolean data types. To facilitate the use of these columns in model training, it was necessary to convert the relevant columns into numeric values. Consequently, with the use of the Pandas method for conversion (get\_dummies()), the following three columns were transformed into numeric values: Gender, Mental Illness, and Fleeing. Column names were also adjusted, with Gender renamed to Sex and Fleeing changed to Running away. For some unknown reason, the Age column was of float data type, so this was converted to an integer. All other columns were removed from the dataset.

The resulting numeric dataset is displayed in the image below:

**Sex:** 0 – Females; 1 – Males

**Mental Illness:** 0 – False; 1 – True

**Running Away:** 0 – False; 1 – True

(Source: Practical Part)

After the data cleaning process, the next step involved dividing the columns into dependent and independent variables. Given the investigation's aim to determine the mental health status of the individuals, the Mental Illness column was selected as the dependent variable, with the remaining columns serving as independent variables. The train\_test\_split method was imported, facilitating the division of the dataset into training and testing sets. Lastly, the dataset was standardized using the StandardScaler() method.

# Implementation and Comparison of Machine Learning Algorithms

The dataset was exclusively subjected to supervised machine learning (ML) algorithms. Specifically, four ML algorithms were selected to analyze the data: Logistic Regression, Decision Tree, K-Nearest Neighbors, and Naïve Bayes. Interestingly, these four ML algorithms displayed both similarities and differences in their predictions when measured using two metrics: the Confusion Matrix and Classification Metrics.

For comparing ML algorithms using the Confusion Matrix, four components were considered: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). As per West (2020), these four elements represent test outcomes, where TP and FN indicate a perfect match between the actual data and prediction, while TN and FP denote a discrepancy between the actual data and prediction.

Concerning Classification Metrics, three criteria will be used to compare the ML algorithms: Precision, Recall, and Accuracy. To clarify how these metrics operate, Vujocovij (2021) provides formulas for calculating the results. For instance, Precision can be calculated by dividing the number of correct positive predictions (TP) by the total number of positive predictions (FP + TP). He indicates that the optimal score for each criterion is 1.0, with the least desirable score being 0.0.

Comparisons are illustrated in the subsequent two tables, where predictions are based on 4132 (33%) randomly selected data samples to determine whether individuals were mentally ill or not.

## Confusion Matrix

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithms** | TP | TN | FP | FN |
| Logistic Regression | 0 | 878 | 2 | 3243 |
| Decision Tree | 217 | 661 | 667 | 2578 |
| K-Nearest Neighbors | 108 | 770 | 217 | 3028 |
| Naïve Bayes | 54 | 824 | 160 | 3085 |

## Classification Metrics

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithms** | Precision | Recall | Accuracy |
| Logistic Regression | False – 0.79  True – 0.00 | False – 0.1  True – 0.0 | 0.79 |
| Decision Tree | False – 0.80  True – 0.25 | False – 0.79  True – 0.25 | 0.68 |
| K-Nearest Neighbors | False – 0.80  True – 0.33 | False – 0.93  True – 0.12 | 0.76 |
| 0.Naïve Bayes | False – 0.79  True – 0.25 | False – 0.95  True – 0.06 | 0.76 |

# Conclusion

Interpreting the above tables allows us to draw several conclusions. In the context of the confusion matrix, among 4132 randomly selected individuals, the Logistic Regression ML algorithm predicted that there were no mentally ill individuals (TP) and that 3243 individuals were not mentally ill (FN). However, there was a discrepancy with 878 individuals who were indeed mentally ill but were predicted by the ML algorithm to be healthy (TN). Nevertheless, the algorithm performed well in predicting individuals who were not mentally ill, with only 2 healthy individuals misclassified (FP).

On the other hand, the Decision Tree ML algorithm's predictions were considerably less accurate. This was demonstrated by over 1200 individuals, both mentally ill and healthy, for whom the predictions did not align with the actual results. As for the K-Nearest Neighbors and Naïve Bayes ML algorithms, these displayed somewhat similar prediction performances but were not flawless. Both algorithms inaccurately predicted over 900 mentally ill and healthy individuals.

The Classification Metrics table serves as a translation of results for all four ML algorithms. The numbers within each criterion provide the accurately predicted statistics for individuals who were mentally ill (True) and those who were not (False). Overall, the Logistic Regression algorithm appears to be the most effective, with a predicted accuracy rate of 79% from the 33% randomly selected individuals. Conversely, the least effective ML algorithm appears to be the Decision Tree, with only a 68% accuracy rate. Interestingly, both the K-Nearest Neighbors and Naïve Bayes algorithms produced identical prediction results – 76%, which is a favorable outcome compared to Logistic Regression.

# References

Vujokovic, Z. (2021). Classification Model Evaluation Metrics. *International Journal of Advanced Computer Science and Applications*. Available from <https://thesai.org/Downloads/Volume12No6/Paper_70-Classification_Model_Evaluation_Metrics.pdf> [Accessed 30 May, 2023].

West, R. (2020). Understanding the Accuracy of Diagnostic and Serology Tests: Sensitivity and Specificity. *Johns Hopkins Bloomberg School of Public Health*. Available from <https://www.centerforhealthsecurity.org/resources/COVID-19/COVID-19-fact-sheets/201207-sensitivity-specificty-factsheet.pdf> [Accessed 30 May, 2023].