**REPORT:**

**Preprocessing:**

The given dataset contains many missing values and contains string, object types of columns. So, I converted the data into numeric format. We can’t remove any column because every column has its own specification. Also, there were 2-3 columns which only contain 50 values out of 20050, So I averaged the empty data with the mean. I could have removed columns for the X variable but removing any column could have removed its role playing int the dataset. Sklearn is used for splitting the data into 3 (Train, Test, Validate).

In the X variable part, I’ve removed \_unit\_id, gender \_unit\_state column. And target column is gender which I’ve stored in y variable.

**Decision Tree Classifier.**

In this part, I’ve made decision tree for 3 different depth (4,5,6) for criteria Gini and 1 for criteria Entropy with depth of 6. Confusion matrix is of 5\*5 because the target column has 5 different types of values.

Decision Tree depth = 4, Criteria: GiniDiagram

Description automatically generated

Decision Tree depth = 5, Criteria: GiniDiagram

Description automatically generated

Decision Tree depth = 6, Criteria: GiniDiagram

Description automatically generated

Confusion Matrix - Criteria: Gini

Graphical user interface, chart, application

Description automatically generated

Confusion Matrix - Criteria: Entropy

Graphical user interface, application

Description automatically generated

Naïve Bayes:

2 models that I’ve used for naïve bayes are Gaussian Naïve bayes and Multinomial Niave Bayes for my assignment.

1. Gaussian Naïve Bayes:

Confusion Matrix:

Graphical user interface, chart

Description automatically generated

1. Multinomial Naïve bayes:

Confusion Matrix:

Graphical user interface, chart

Description automatically generated

Visualization of the Target variable:

1. Count Plot.

Chart, bar chart

Description automatically generated

1. Pie Plot.

Chart, pie chart

Description automatically generated

References:

* <https://scikitlearn.org/stable/modules/generated/sklearn.tree.plot_tree.htmlhttps://scikitlearn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html>
* <https://scikitlearn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html>
* <https://seaborn.pydata.org/generated/seaborn.heatmap.html>
* <https://www.kaggle.com/getting-started/61148>