Problem 1: we have classify all the person based on the drinking habit by observing the all the independent features (binary classification)

| | Recall | Precision | Accuracy | F1 score |
|---|---|---|---|---|
| Logistic regression | 0.72 | 0.72 | 0.7163 | 0.72 |
| Knn classifier | 0.69 | 0.74 | 0.7213 | 0.71 |
| Random forest classifier | 0.73 | 0.73 | 0.7277 | 0.73 |

Problem 2 : we have classify the smoking stats of the particular person based on the observed features that Smoking state, 1(never), 2(used to smoke but quit), 3(still smoke) (multi-class classification problem)

| | Recall | Precision | Accuracy | F1 Score |
|---|---|---|---|---|
| Knn classifier | 0.6608 | 0.6412 | 0.6705 | 0.6475 |
| Logistic regression | 0.6608 | 0.6412 | 0.69 | 0.6475 |
| XG boost | 0.7081 | 0.7026 | 0.7081 | 0.7042 |
| Ada Boost | 0.6981 | 0.6892 | 0.6981 | 0.69011 |

Problem 3: another problem we have classify the age interval [Younger, Middle Age , Older] based on the feature like drinking or not or smoking habit

| | Recall | Precision | Accuracy | F1 score |
|---|---|---|---|---|
| Knn classifier | 0.613197 | 0.616600 | 0.6131 | 0.612649 |

Observations and Learning:
Problem 1:

- From the above problem we conclude that random forest will perform best in these dateset when we compare with logistic and knn classifier
- In these datasets we cannot perform SVM because it show the less accuracy and more computational time because of the large datasets that's why we perform on the three model which show best performance on the datasets

Problem 2:

- As observed from all the model we found that ensemble learning based model has better accuracy as compared to the classical machine learning model

Problem 3 :

- I used KNN for categorizing age groups because it works well with diverse body signal data. It's good for sorting things into multiple categories. I didn't remove unusual data points because in some diseases, body values can change a lot, and we can't assume a fixed range for these values.

Contribution as per tasks:

EDA: Aditya, Dhyey, Keyur, Palak, Harshil ,keyur

Problem 1 :
Logistic regression- Harshil
Knn classifier and random forest- Dhyey

Problem 2:

XGBoost and Ada boost- Aditya
Knn classifier and logistic regression- Keyur

Problem 3:
Knn classifer- Palak