Group Activity

Aryan Pahari (23BAI10337)
Parth Chilwerwar (23BAI10601)
Kalpak Dorlikar (23BOE10020)

Smoking and Drinking Dataset with body signals

Introduction

Exploring body signals provides valuable insights into smoking and drinking habits. By analyzing these cues, healthcare practices can be informed and public health policies can be shaped. Smoking, with its 4,000 chemical compounds, poses risks like lung cancer, heart disease, and organ damage. Meanwhile, excessive drinking can lead to high blood pressure, liver issues, and memory impairment. Awareness about these dangers is crucial.

Smoking and Drinking

Smoking and drinking are common but harmful habits that affect our health and communities. Smoking is a leading cause of many serious diseases, like lung cancer and heart problems. It not only hurts the person smoking but also those around them who breathe in the smoke. Drinking too much alcohol can also cause health issues, such as liver problems and mental health issues. It can also lead to accidents and problems in families. These habits don't just affect individuals; they also cost a lot of money for healthcare and cause problems in society. To help, we need rules and support to encourage people to quit smoking and drink less. By making healthier choices and getting help when needed, we can improve our health and make our communities safer and happier places to live.

How Smoking and Drinking related with body signal ?

- Smoking and drinking can impact various body signals, as reflected in factors such as age, blood pressure, cholesterol levels, and more. For instance, individuals who smoke or drink heavily may exhibit higher blood pressure readings (SBP and DBP), increasing their risk of cardiovascular issues like heart disease and stroke. Additionally, smoking and drinking habits can influence cholesterol levels, with elevated levels of total cholesterol, LDL cholesterol, and triglycerides often observed in those who engage in these behaviors. Such lipid imbalances can contribute to the development of atherosclerosis and other vascular conditions.
- Moreover, smoking and drinking may affect markers of kidney function and liver health. Higher levels of serum creatinine and liver enzymes (such as SGOT_AST, SGOT_ALT, and gamma_GTP) could indicate potential damage to these organs, particularly in chronic smokers and heavy drinkers. Furthermore, smoking and excessive alcohol consumption may impact hemoglobin levels and urine protein excretion, potentially indicating adverse effects on blood oxygenation and kidney function.

The relationship between smoking, drinking, and body signals underscores the importance of considering lifestyle factors in assessing overall health and disease risk. By understanding these connections, healthcare providers can better tailor interventions to mitigate the adverse effects of smoking and drinking on individuals' health.

Previous Approach in Kaggel with Confusion Matrix and Classification Report

```
# Define the parameter grid for hyperparameter tuning
param_grid = {
    'n_estimators': [100, 200, 300],
    'learning_rate': [0.1, 0.5, 1.0],
#     'max_depth': [3, 4, 5],
    'subsample': [0.3, 0.6, 0.9]
}

# Initialize GridSearchCV with cross-validation
grid_search_xgb = GridSearchCV(model, param_grid, cv=3, scoring='roc_auc', verbose=1)

# Fit GridSearchCV on your training data
grid_search_xgb.fit(X_train, y_train)

# Get the best estimator from the grid search
best_xgb_classifier = grid_search_xgb.best_estimator_

# Make predictions on the train and test set
y_train_pred_prob = best_xgb_classifier.predict_proba(X_train)[:, 1]
y_test_pred_prob = best_xgb_classifier.predict_proba(X_test)[:, 1]

# Calculate ROC AUC score on the test set
train_auc = roc_auc_score(y_train, y_train_pred_prob)
test_auc = roc_auc_score(y_test, y_test_pred_prob)

print("Best Hyperparameters:", grid_search_xgb.best_params_)
print("Train ROC AUC Score:", train_auc)
print("Test ROC AUC Score:", test_auc)
```

```
Fitting 3 folds for each of 27 candidates, totalling 81 fits
Best Hyperparameters: {'learning_rate': 0.1, 'n_estimators': 200, 'subsample': 0.9}
Train ROC AUC Score: 0.831275122105966
Test ROC AUC Score: 0.8228113163827739
```

Setting up options: A variety of options are being set up to explore different configurations for a model. These configurations include different numbers of trees (100, 200, 300), various learning rates (0.1, 0.5, 1.0), and different fractions of the data to utilize (0.3, 0.6, 0.9).

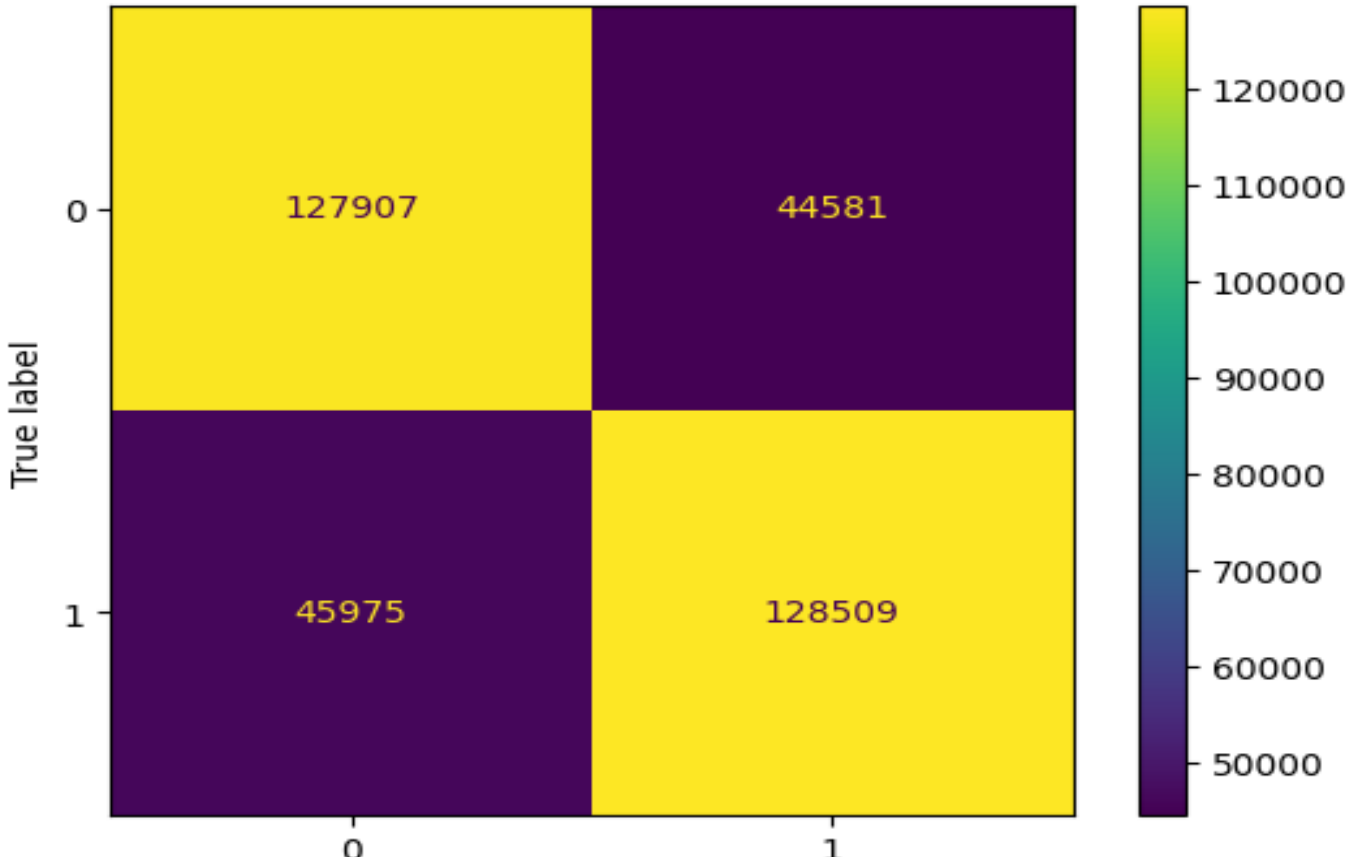
Choosing the best: GridSearchCV is employed to systematically try out all the specified options and identify the most effective combination of settings. The process involves assessing the performance of each combination through cross-validation, which divides the training data into subsets to gauge overall performance.

Training the model: Following the identification of the optimal combination of settings, the model is trained using this configuration.

Making predictions: Once trained, the model is utilized to make predictions on both the training and test datasets.

Evaluating performance: The performance of the model is evaluated using the ROC AUC score, which measures its ability to distinguish between classes in the data. This evaluation is conducted separately for both the training and test datasets to assess the model's performance on unseen data.

Confusion Matrix:



My Approach with Confusion Matrix and Classification Report and Different Machine Learning and Deep learning Model.

➢ Models used to compare the performance :-

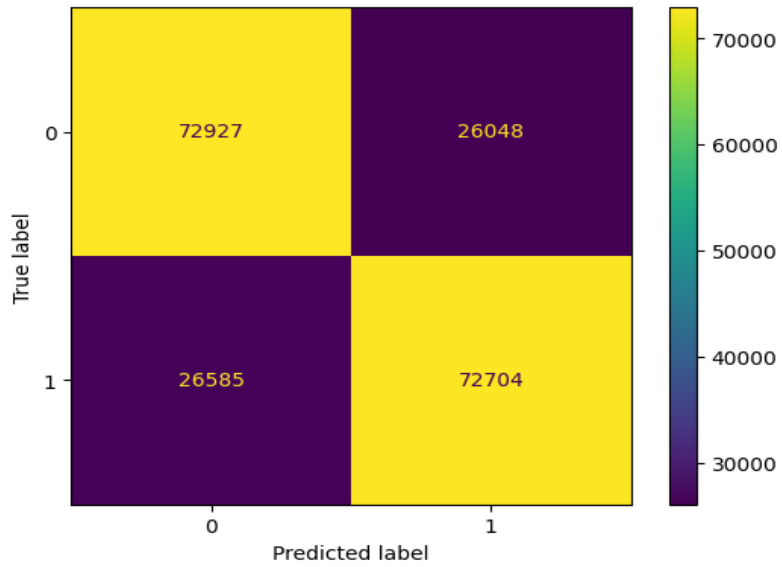
- Random Forest Classifier
- XG Boost Classifier
- Artificial Neural Network
- Voting Classifier

➢ Random Forest Classifier

- **Classification Report:-**

	precision	recall	f1-score	support
0	0.73	0.74	0.73	98975
1	0.74	0.73	0.73	99289
accuracy			0.73	198264
macro avg	0.73	0.73	0.73	198264
weighted avg	0.73	0.73	0.73	198264

- **Confusion Matrix:**



➢ Artificial Neural Network :-

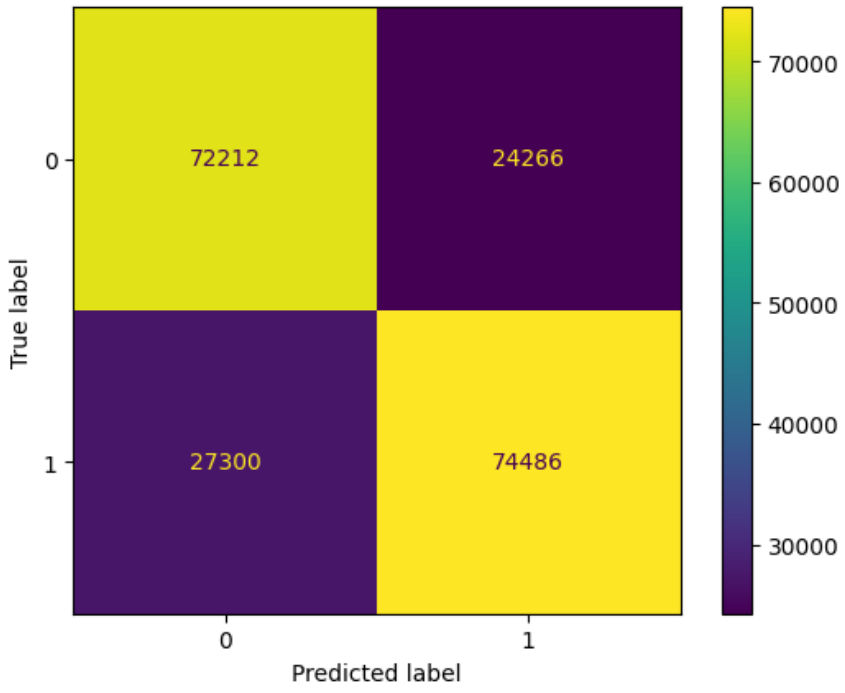
- **Parameter for ANN model :-**

  - Learning Rate = 0.00001
  - Loss Function = Binary Cross Entropy with Logist
  - Optimizer = Adam
  - Epochs = 50

- **Classification Report:**

	precision	recall	f1-score	support
0.0	0.73	0.75	0.74	96478
1.0	0.75	0.73	0.74	101786
accuracy			0.74	198264
macro avg	0.74	0.74	0.74	198264
weighted avg	0.74	0.74	0.74	198264

- **Confusion Matrix :**



➢ Voting Classifier :-

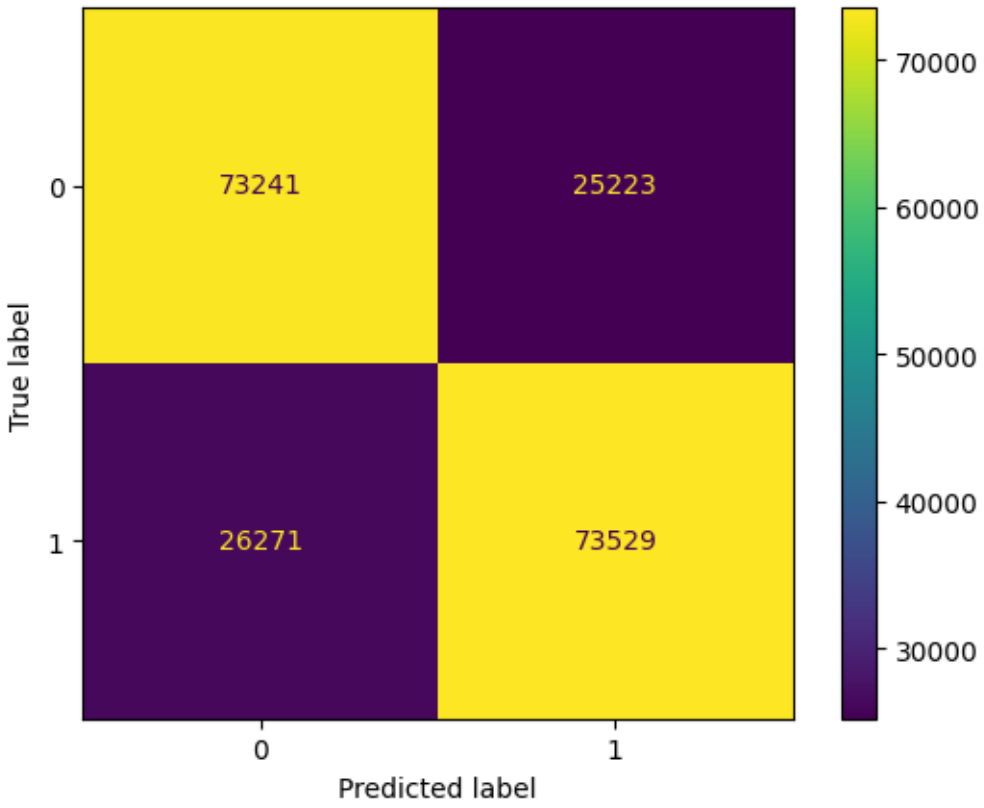
- **Models Used For ensemble learning :**
  - XG Boost with hypertune parameter (Cv = 5)

    'learning_rate': 0.1, 'max_depth': 6, 'n_estimators': 200, 'subsample': 0.6
  - XG Boost Default parameters.

- **Classification Report:**

	precision	recall	f1-score	support
0	0.74	0.74	0.74	98615
1	0.74	0.74	0.74	99649
accuracy			0.74	198264
macro avg	0.74	0.74	0.74	198264
weighted avg	0.74	0.74	0.74	198264

- **Confusion Matrix :**



➢XG Boost Classifier :-

- **Classification Report:**
  - XG Boost hypertune parameters (Cv = 5)

    'learning_rate': 0.1, 'max_depth': 6, 'n_estimators': 200, 'subsample': 0.6

...	precision	recall	f1-score	support
0	0.74	0.74	0.74	98615
1	0.74	0.74	0.74	99649
accuracy			0.74	198264
macro avg	0.74	0.74	0.74	198264
weighted avg	0.74	0.74	0.74	198264

- **Confusion Matrix :**