Model Development Phase

Date	4th June 2024
Team ID	SWTID1720175375
Project Title	Prediction and Analysis of Liver Patient Data
Maximum Marks	6 Marks

Model Selection Report:

Model-1(Logistic regression)

Description:

One of the simplest and best ML classification algorithms is logistic regression. LR is a supervised ML binary classification algorithm widely used in most applications. It operates on a categorical dependent variable, the result can be a discrete or binary categorical variable 0 or 1.

Logistic sigmoid function:

$$prob(Y=1)=\frac{e^z}{1+e^z}$$

Hyperparameters:

```
{'C': 0.0001, 'max_iter': 1000, 'solver': 'lbfgs'}
```

```
LogisticRegression(C=0.0001, max_iter=1000)
```

Accuracy, Precision and Recall:

	precision	recall	f1-score	support
0	0.44	0.46	0.45	147
1	0.78	0.77	0.78	378
accuracy			0.68	525
macro avg	0.61	0.61	0.61	525
weighted avg	0.69	0.68	0.69	525

```
Logistic Regression Training Score:
 81.03
Logistic Regression Test Score:
 68.38
```

Model-2(KNN):

Description:

Using supervised machine learning, the K-Nearest Neighbors (KNN) technique is used to solve regression and classification issues.

Hyperparameters:

```
knn_params={
    "n_neighbors":range(1,20,2),
    "weights":["uniform","distance"],
    "algorithm":["auto","ball_tree","kd_tree","brute"],
    "metric":["euclidean","minkowski","manhattan"],
    "leaf_size":range(1,30,5)
}
from sklearn.model_selection import GridSearchCV,RepeatedStratifiedKFold
grids=GridSearchCV(estimator=model,param_grid=knn_params,n_jobs=1,cv=3,scoring="accuracy",error_score=0)
res=grids.fit(X_train,y_train)
par_model=model.set_params(**res.best_params_)
```

Accuracy, Precision and Recall:

	precision	recall	f1-score	support
0	0.41	0.47	0.44	147
1	0.78	0.74	0.76	378
accuracy			0.66	525
macro avg	0.60	0.60	0.60	525
weighted avg	0.68	0.66	0.67	525

Model-3(Random Forest):

Description:

A random forest is a meta estimator that employs averaging to increase prediction accuracy and manage over-fitting after fitting several decision tree classifiers on different subsamples of the dataset. The best split strategy, or passing splitter="best" to the underlying DecisionTreeRegressor, is employed by the trees in the forest. If bootstrap=True (the default), the sub-sample size is managed using the max_samples argument; if not, each tree is constructed using the entire dataset.

Hyperparameters:

```
RandomForestClassifier(criterion='entropy', max_depth=15, max_features=0.75, min_samples_leaf=7, min_samples_split=3, n_estimators = 130)
```

Accuracy, Precision and Recall:

	precision	recall	f1-score	support
0	0.48	0.44	0.46	163
1	0.76	0.79	0.77	362
accuracy			0.68	525
macro avg	0.62	0.61	0.62	525
eighted avg	0.67	0.68	0.68	525