IIS ASSIGNMENT-LOGISTIC REGRESSION

1)Collect data
Import various libraries to analyse data,using pandas,student_data saves the data.
2)Analyse data
Checking if there are any null values because if there are present we would have to replace them with dummy values.There are no null values.
Various columns that don't contribute to high salary are dropped like ID,DOB,CollegeCityID etc.
Various graphs are plotted to check the correlation between High salary and various columns.
For graph plotting sns library was used.
3)Data Wrangling
All the columns which have string values have to be converted to implement logistic regression.
New dummy variables are created by pandas library.
Therefore,columns which had string values were converted by one hot encoding and dummy variables are concatenated to data set(data is converted to categorical data) and columns with string values are dropped.
3)Training and testing data
Data set is split into training and testing data sets.
y is the column which needs to be predicted. (dependent)
X (independent)
To split data into testing and training subset we use sklearn.
To get the correct results,I have scaled the values so they can be compared properly using standard scaler.

Then LogisticRegression is imported and we create an instance of it called logmod and we fit the data.

To evaluate how the model is performing we import the classification report,confusion matrix and accuracy score.
To get classwise accuracy ,I have normalised the confusion matrix and the diagonal values give value of classwise accuracy
For train-test split(90-10)
Accuracy_score:0.7425
Accuracy:74.25

Confusion matrix

	Predicted Negatives	Predicted Positives
True Negatives(0)	130	61
True Positives(1)	42	167

Confusion matrix(Normalised)

	Predicted Negatives	Predicted Positives
True Negatives(0)	0.68062827	0.29186603
True Positives(1)	0.21989529	0.79904306

Classwise accuracy
Class 0 :0.68062827
Class 1 :0.79904306

For train-test split(80-20)
Accuracy_score:0.70625
Accuracy:70.625

Confusion matrix

	Predicted Negatives	Predicted Positives
True Negatives(0)	251	125
True Positives(1)	110	314

Confusion matrix(Normalised)

	Predicted Negatives	Predicted Positives
True Negatives(0)	0.66755319	0.29481132
True Positives(1)	0.29255319	0.74056604

Classwise accuracy
Class 0 :0.0.66755319
Class 1 :0.74056604

For train-test split(70-30)
Accuracy_score:0.7041666666666667
Accuracy:70.41666666666667

Confusion matrix

	Predicted Negatives	Predicted Positives
True Negatives(0)	363	185
True Positives(1)	170	482

Confusion matrix(Normalised)

	Predicted Negatives	Predicted Positives
True Negatives(0)	0.66240876	0.28374233
True Positives(1)	0.31021898	0.7392638

Classwise accuracy
Class 0 :0.66240876

Class 1 :0.7392638

OBSERVATIONS

Removing gender increased the accuracy by around 1% (as seen from graph,that gender does not have a huge role to play in salary)

Removing Logical increased the accuracy by around 0.5%.

For the train-test ratios that I used (90-10) was the most accurate as it had a larger data set to train on as compared to 70-30 and 80-20,Because of the larger data set ,the results are more accurate as more data to train on is present.