Report for lab 4 using logistic regression

In the first cell I started loading required modules.

In the second cell, I only read the data from text file using `pd.read_csv('file name')`

In the third cell, I cleaned the data from NULL values, I filled every null value in each column with the average of the column.

In the fourth cell, I converted the output of binary classification problem to 1's and 0's values using `LabelEncoder.fit_transform(column)` and assign it to variable called y.

In the fifth cell, I assign to X variable all input features.

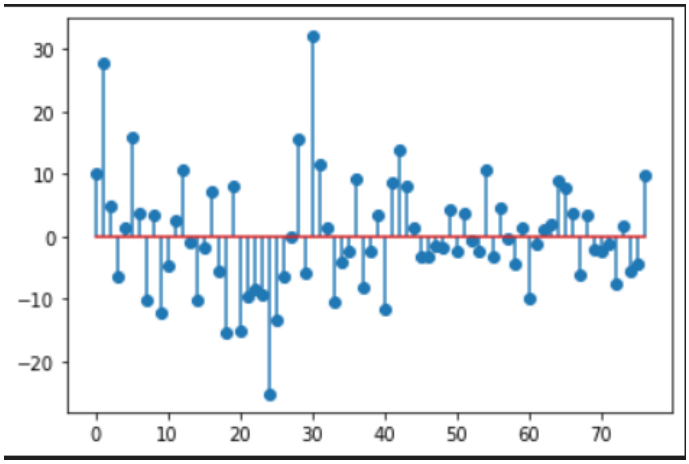
In the sixth cell, I split the data to Xtrain, Xtest, Ytrain, Ytest.

In the seventh cell, I scaled Xtrain and Ytrain using `StandardScaler().fit_transform(data)` and `StandardScaler().transform(data).`

In the eighth cell, I built the model using `sklearn.linear_model.LogisticRegression()`, after that I fitted the scaled data, using `LogisticRegression().fit(Xtrain, Ytrain)`

In the ninth cell, I tested the model using `LogisticRegression().predict(Xtest)`, and the accuracy was around 96.000

In the tenth cell, I tested the model using `LogisticRegression().predict(Xtest)`, I interpreted the coefficients and what's the important two gene in the data who affect to training and they was [`'APP_N', 'ITSN1_N'`].



In the eleventh cell, I created ten folds each folds have training sets and one validation set, to minimize the overfitting and the values compared to the trained model without using cross validation technique was a little bit better.

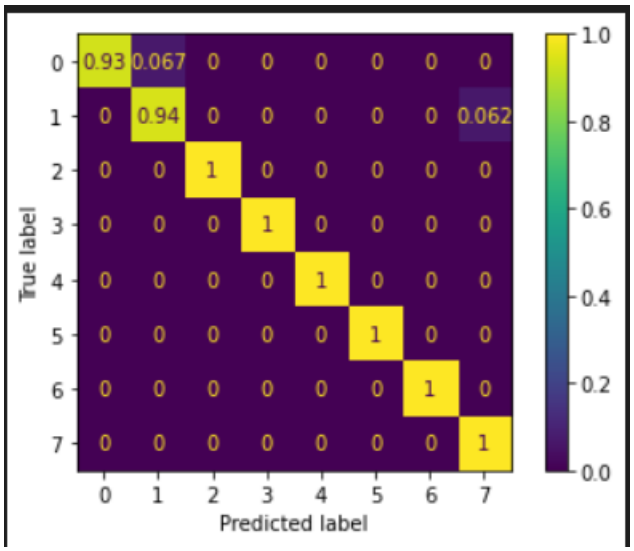
Mean precision	95.914%	**SE**	0.0101
Mean recall	96.130%	**SE**	0.0088
Mean f1 score	95.990%	**SE**	0.0065
Mean accuracy	97.5555%	**SE**	0.0063

Note: it will change in every run.

I have created all above steps with multi classification also.

But in cross validation I created normalized confusion matrix whereas the summation of each row equal to one as required.

Sample of one of folds.



Mean precision	98.914%	**SE**	0.0089
Mean recall	98.130%	**SE**	0.0105
Mean f1 score	98.990%	**SE**	0.0074
Mean accuracy	98.5555%	**SE**	0.0039

Note: it will change in every run.

In the bonus section, I created twenty values of alpha and I tried each of the on 10 folds and I took the mean accuracy of each 10 folds with each value of alpha (for binary classification and multi classification), and I got the following results.

the last cell has

all results go and see it and see what's the best value of C.

accuracy in binary classification without regularization is 97.5555 while with regularization with best value of C is lays between 96.0000 and 98.22% but without overfitting.

accuracy in multiple classification without regularization is 98.5555 while with regularization with best value of C is lays between 96.0000 and 97% but without overfitting.