CSE 4062 – Introduction to Data Science and Analytics
Spring 2021
Delivery #4 - Predictive Analytics

Project Report - Group #8

Group Members:
Caner Dağdaş | 150716001 | Electrical and Electronics Engineering | canerdagdas@hotmail.com
Ceyhun Vardar | 150317022 | Industrial Engineering | vardarceyhun13@gmail.com
Büşra Gökmen | 150116027 | Computer Engineering | busragokmen67@gmail.com
Cem Güleç | 150117828 | Computer Engineering | cem.ggulecc@gmail.com
Ömer Faruk Çakı | 150117821 | Computer Engineering | omerfarukcaki@gmail.com

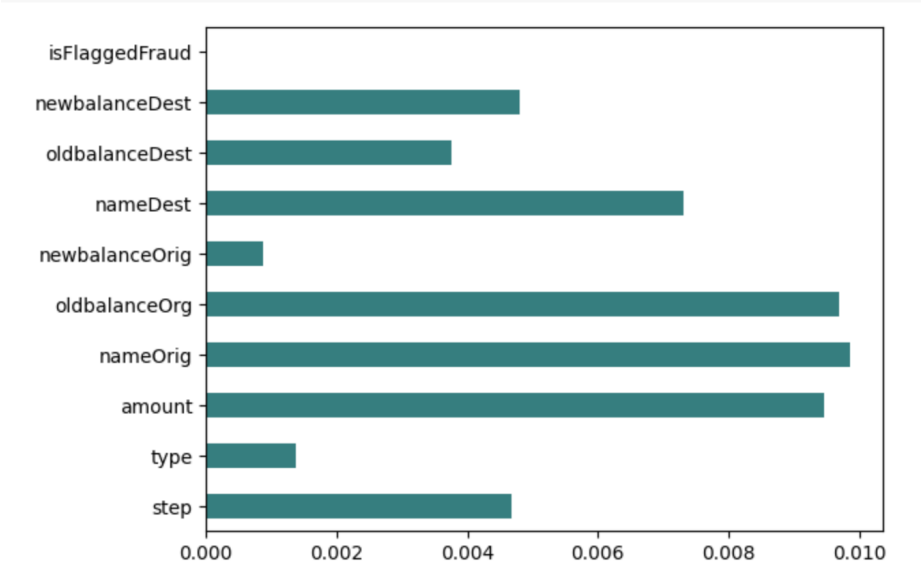
Project Title: Fraud Detection on Financial Data

Lecturer: Assoc. Prof. Murat Can Ganiz

21.05.2021

1- Feature Selection Methods

Information Gain Importance of Attributes Graphic:



Correlation Coefficient Importance of Attributes Graphic:

Chi-Square Test Importance of Attributes Graphic:

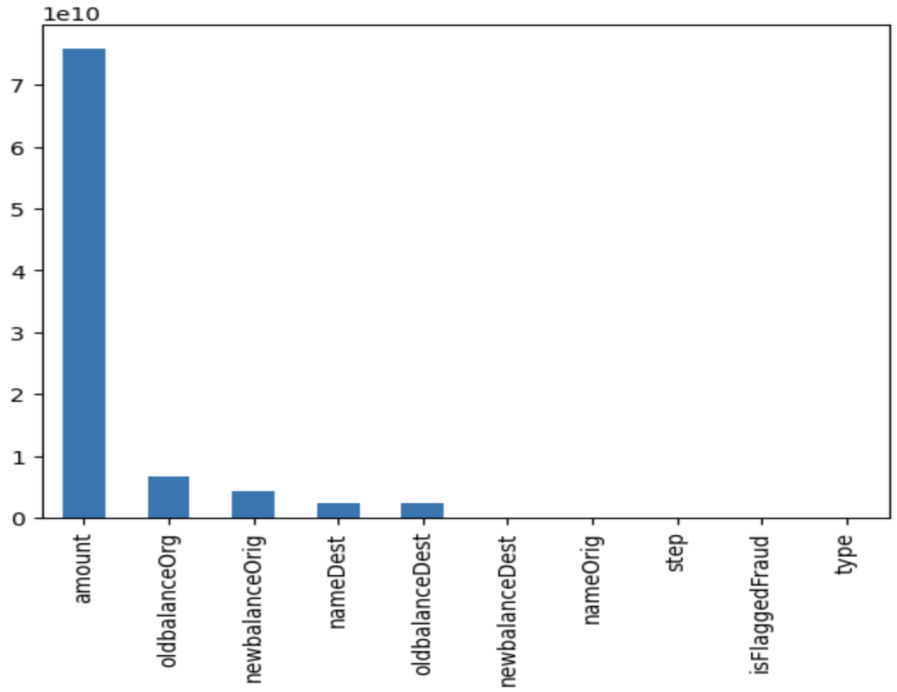


Table of Feature Extraction Values

Feature Name	Description	Information Gain Importances	Correlation Coefficient Importances	Chi-square Test Importances
step	maps a unit of time in the real world. In this case 1 step is 1 hour of time. Total steps 744 (30 days simulation).	4.68873089e-03	0.032	5.28059079e+05

type	CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER.	1.38039937e-03	0.021	2.93663069e+0
amount	amount of the transaction in local currency.	9.45706635e-03	0.077	7.58623617e+10
nameOrig	customer who started the transaction	9.87103998e-03	-0.00046	1.44913424e+06
oldbalanceOrg	initial balance before the transaction	9.68981785e-03	0.01	6.56309030e+09
newbalanceOrig	new balance after the transaction	8.84693967e-04	-0.0081	4.22377158e+09
nameDest	customer who is the recipient of the transaction	7.31022523e-03	-0.022	2.31449109e+09
oldbalanceDest	initial balance recipient before the transaction. Note that there is not information for customers that start with M (Merchants).	3.75719688e-03	-0.0059	2.31338860e+09
newbalanceDest	new balance recipient after the transaction. Note that there is not information	4.80100350e-03	0.00054	2.00946581e+07

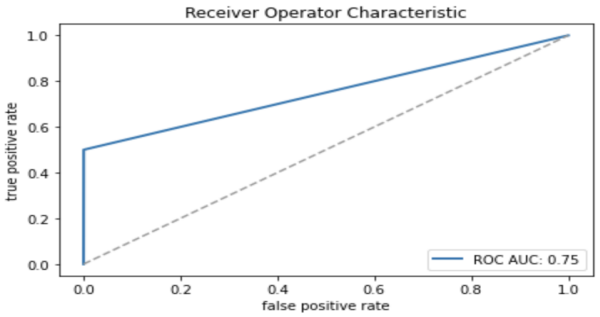
	for customers that start with M (Merchants)**.**			
isFlaggedFraud	The business model aims to control massive transfers from one account to another and flags illegal attempts. An illegal attempt in this dataset is an attempt to transfer more than 200.000 in a single transaction.	1.67313483e-05	0.044	1.23792173e+04

2- Classification Experiments

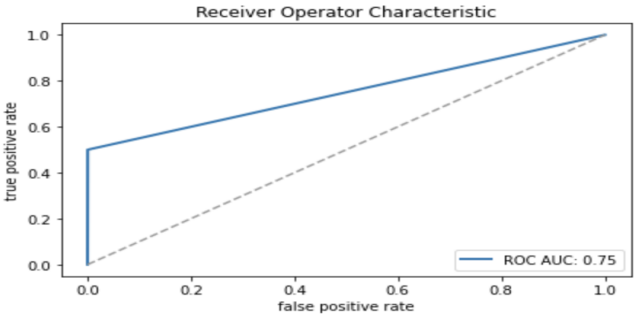
#	Experiments	F1-Score	Precision Score	Recall Score	Accuracy Score	AUC Score
		Avg	Avg	Avg	Avg	Avg
1	KNN5 IG	**0.8154848101738795**	**0.9409437379980392**	**0.7498914223669924**	**0.9983399494767231**	**0.7498914223669924**
2	KNN5 CC	**0.8408855628310885**	0.9312901166551205	0.7873552298226565	0.9984843016961386	0.7873552298226565
3	KNN5 CT	**0.8215544496522468**	0.9409617754689756	0.7561414223669924	0.998376037531577	0.7561414223669924
4	KNN30 IG	**0.7710929375909077**	0.9641328694779251	0.6999457111834962	0.998159509202454	0.6999457111834962
5	KNN30 CC	**0.7898426224385825**	0.9233715881332109	0.7248914223669924	0.9981955972573078	0.7248914223669924
6	KNN30 CT	**0.7768712017749420**	0.9666509134581999	0.7061957111834962	0.9981955972573078	0.7061957111834962
7	NB IG	**0.5794363740446421**	0.5505534666310475	0.7044239780979186	0.9874285755537989	0.7044239780979186
8	NB CC	**0.5624448718361692**	0.5388830197375291	0.6927692310168482	0.9845297266469584	0.6927692310168482
9	NB CT	**0.5624611556892887**	0.538892527825947	0.6928297703264119	0.984529365689555	0.6928297703264119
10	DTREE ENTROPY IG	**0.8912501843652410**	0.9868855956461922	0.8272586391401371	0.9989261517248349	0.8472586391401371
11	DTREE ENTROPY CC	**0.8912213289903487**	0.9868017626103001	0.8272584581250042	0.9989257907674315	0.8472584581250042
12	DTREE ENTROPY CT	**0.8912213289903487**	0.9868017626103001	0.8272584581250042	0.9989257907674315	0.8472584581250042
13	DTREE GINI IG	**0.8950951903423794**	0.9893396080954238	0.8315264393216549	0.9989608036355629	0.8315264393216549
14	DTREE GINI CC	**0.8950951903423794**	0.9893396080954238	0.8315264393216549	0.9989608036355629	0.8315264393216549
15	DTREE GINI CT	**0.8950951903423794**	0.9893396080954238	0.8315264393216549	0.9989608036355629	0.8315264393216549
16	MLP IG	**0.7879411340823692**	0.9913596052077562	0.7039245085201133	0.9982060417050184	0.7039245085201135
17	MLP CC	**0.8264447652988590**	0.987274904015663	0.7456663619457062	0.9984394608258946	0.7456663619457062
18	MLP CT	**0.8240480349705437**	0.9926672311279199	0.7416603575767171	0.9984322416778263	0.7416603575767172

3- ROC Curves

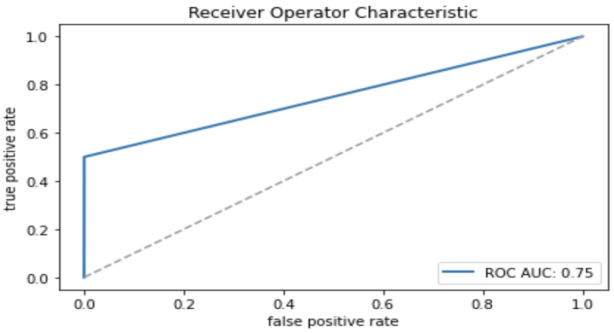
KNN5-IG



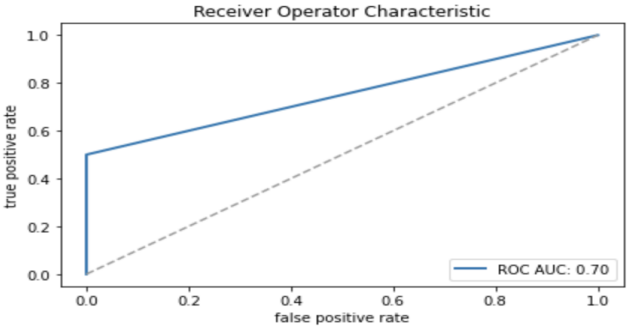
KNN5-CT



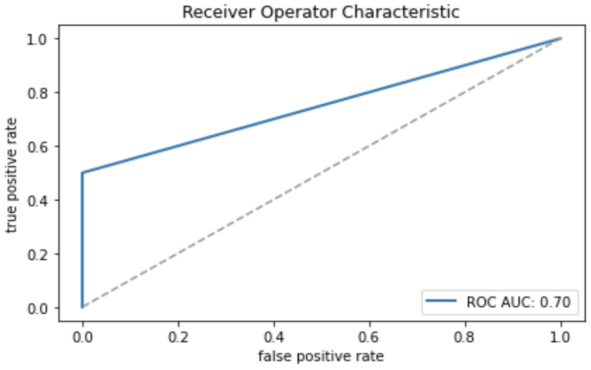
KNN5-CC



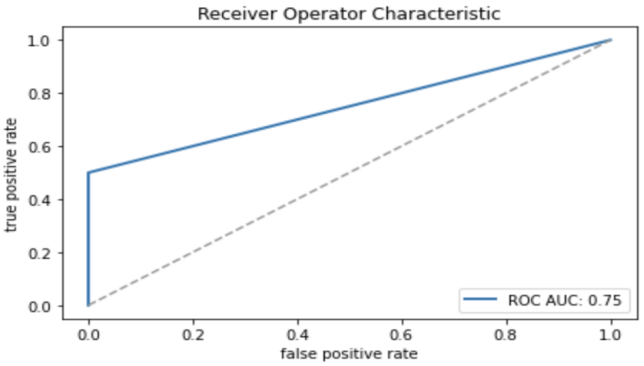
KNN30-IG



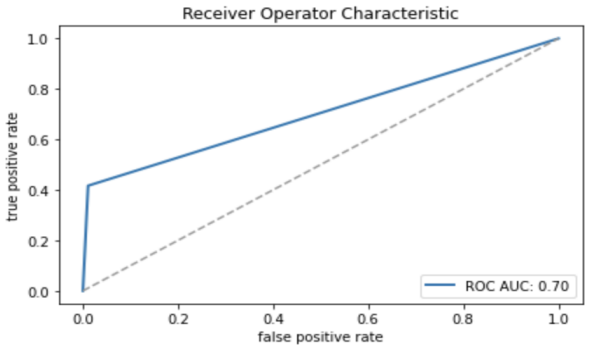
KNN30-CC



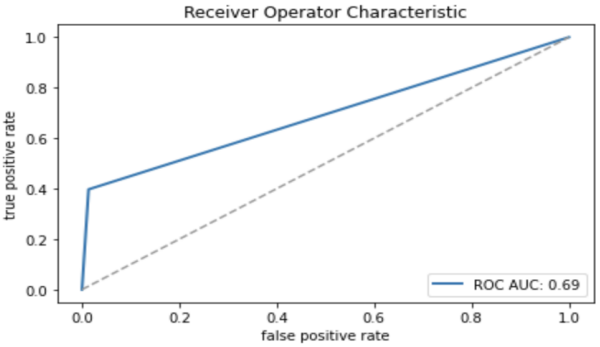
KNN30-CT



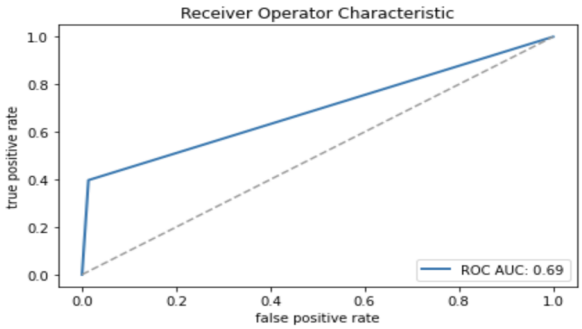
NG-IG



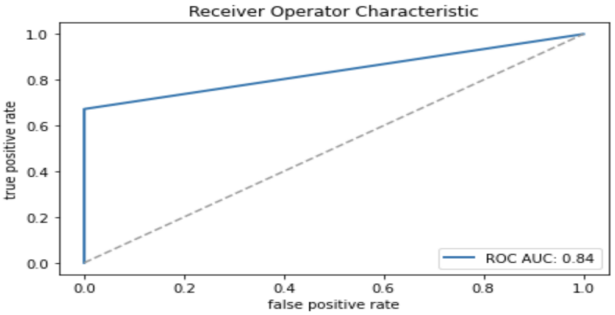
NG-CC



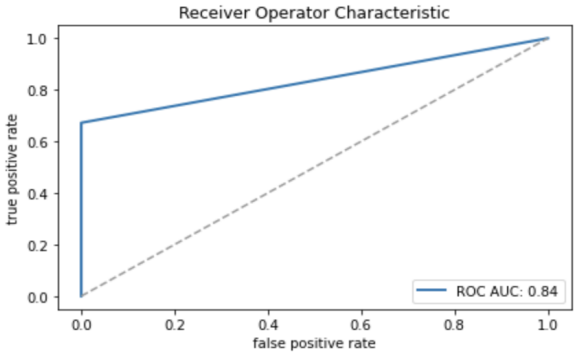
NG-CT



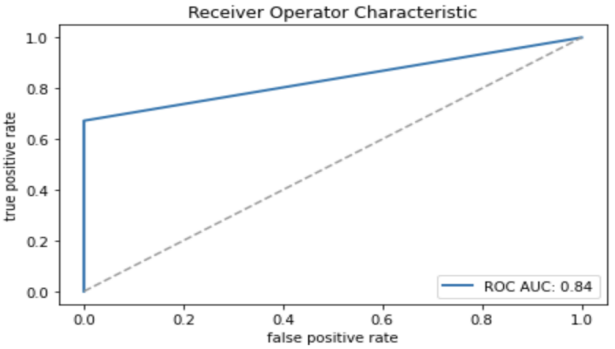
DTREE-ENTROPY-IG



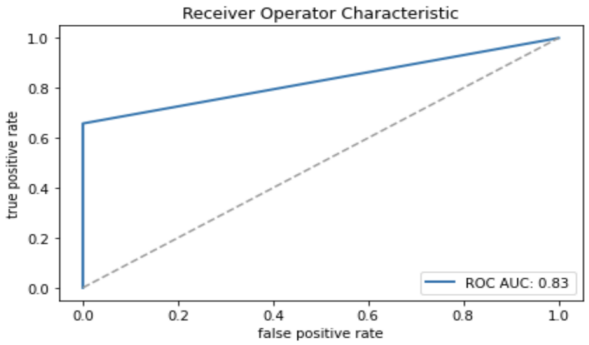
DTREE-ENTROPY-CC



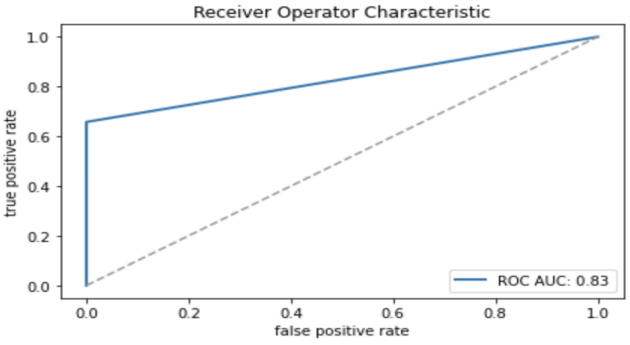
DTREE-ENTROPY-CT



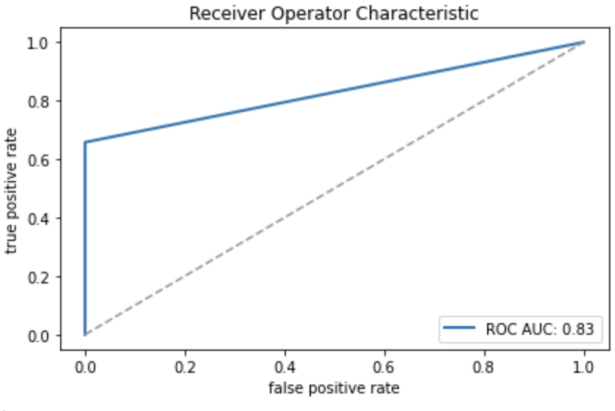
DTREE-GINI-IG



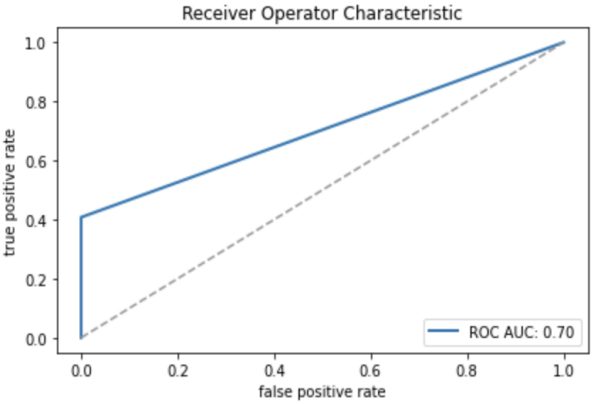
DTREE-GINI-CC



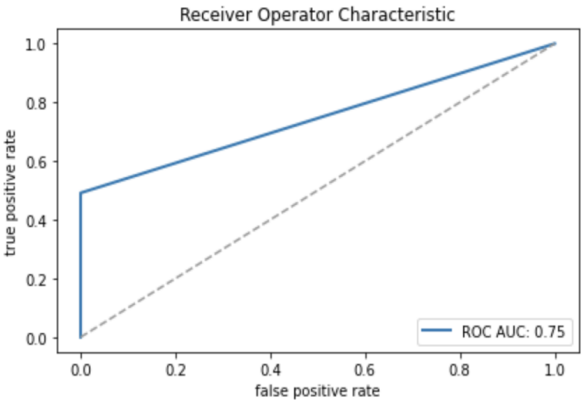
DTREE-GINI-CT



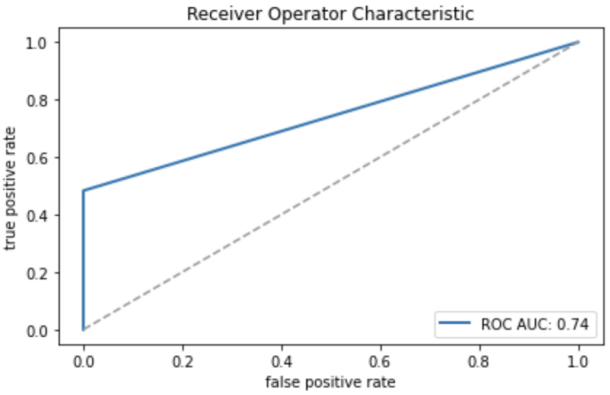
MLP-IG



MLP-CC



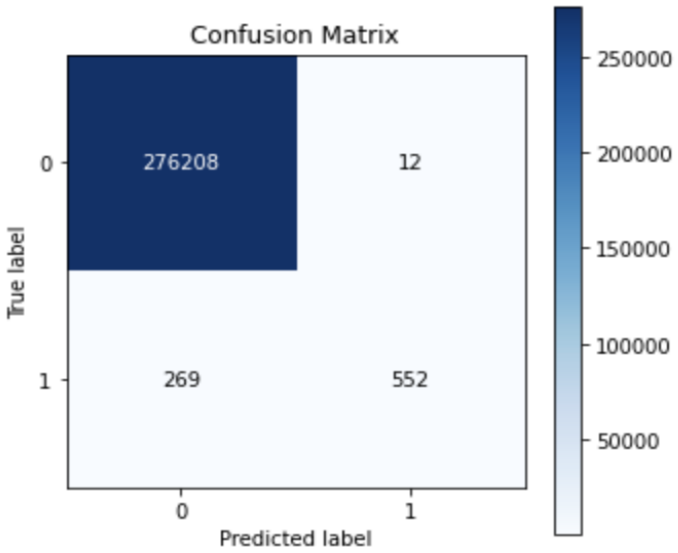
MLP-CT



4- Confusion Matrices

Confusion matrix of the best fold performance of our best performing model:

Decision Tree-Entropy-IG



5 - T-Test

- **Accuracy T_Test** between Decision Tree-Gini-IG & Decision Tree-Entropy-IG:

  T Value = [7.2228622946768155], P Value = [5.091482790930968e-11]

  p-value<=0.05 so there's a significant difference between models.

- **Precision T_Test** between Decision Tree-Gini-IG & Decision Tree-Entropy-IG:

  T Value = [6.129260759394118], P Value = [8.83001671780903e-08]

  p-value<=0.05 so there's a significant difference between models.

- **Recall T_Test** between Decision Tree-Gini-IG & Decision Tree-Entropy-IG:

  T Value = [6.018732092097544], P Value = [1.758104772875413e-07]

  p-value<=0.05 so there's a significant difference between models.

- **F1 T_Test** between Decision Tree-Gini-IG & Decision Tree-Entropy-IG:

  T Value = [7.561632004604 1725], P Value = [3.9745984281580604e-12]

p-value<=0.05 so there's a significant difference between models.

- **AUC T_Test** between Decision Tree-Gini-IG & Decision Tree-Entropy-IG:

  T Value = [6.018732092097659], P Value = [1.758104772875413e-07]

  p-value<=0.05 so there's a significant difference between models.

6 – Methods

For this iteration of the project, we first determined 3 different feature subsets:

1. Feature selection with feature_selection method of sklearn. Scoring function we chose Information Gain.
2. Feature selection with correlation matrix.
3. Feature selection with the chi2 method of sklearn. Scoring function we chose the Chi-Square Test.

And we determined 6 different classifying models:

1. KNeighborsClassifier model of Sklearn with 5 neighbours
2. KNeighborsClassifier model of Sklearn with 30 neighbours
3. DecisionTreeClassifier model of Sklearn with gini
4. DecisionTreeClassifier model of Sklearn with entropy
5. GaussianNB model of Sklearn.
6. MLP deep learning model of Sklearn

And finally we determined 5 metrics to calculate t-test on:

1. accuracy_score metric of Sklearn.
2. precision_score metric of Sklearn.
3. recall_score metric of Sklearn.
4. f1_score metric of Sklearn.
5. auc_score metric of Sklearn.

7 – Implementation

We used 10 fold cross-validation technique for all classification models. In every fold, we made feature elimination with 3 different methods. Also we used an extra deep learning model(MLP) as a classifier for our 2.770.409 data points. We trained 6 different models and we used information gain, correlation coefficient, chi-square test feature selection methods. We run our 18 different classifiers and run predictions on corresponding test-sets. And finally we evaluate models by using 5 metrics(Accuracy, P, R, F1, AUC) we determined above. In total, we run 18 different experiments in 10 folds.

8 – Conclusion

For t-test, we chose our top two experiments according to F1-Score and ROC-AUC metric :

1. Decision Tree-Entropy-IG

- F1-Score: 0.89
- ROC-AUC metric**:** 0.84

2. Decision Tree-Gini-IG

- F1-Score: 0.89
- ROC-AUC metric: 0.83

Then run scipy's t.cdf() method on their 10-Fold CV scores. We compared them by using 5 different metrics(Accuracy, P, R, F1, AUC) according to **T-Test.**

According to all score metric types there's a significant difference between models. So we can concluded that, **Decision Tree-Entropy-IG** is our best experiment.

Also, our decision tree classifier model performed much better than our deep learning model. In general, the Correlation coefficient and Chi-square test feature selection methods resulted in higher accuracy than the Information gain method.

9 – References

[1]https://medium.com/analytics-vidhya/using-the-corrected-paired-students-t-test-for-comparing-the-performance-of-machine-learning-dc6529eaa97f