CSE 4088

INTRODUCTION TO MACHINE LEARNING

PROJECT MIDTERM REPORT

1. Project Members: *Zafer Emre OCAK* *150113075*
Ozan GÜLHAN *150114013*

2. Abstract:

In machine learning algorithms, one of the biggest problems is scarcity of time and requirement of huge datasets. Feature selection provides optimization for these matters. Reducing time and amount of data required can change the fitness of the model significantly. By aiming this, feature selection algorithms are very important.
In order to see those effects, we consider implementing the algorithms which are "consistency-based" and " correlation-based" feature selection. In addition to that, we can also compare the required amount of time and data to reach the same fitness levels for the model.

To pick up a model, we have had a lot of different options but among them, we have decided on Bank Marketing Dataset. One of the most desired features of our project was finding an appropriate data. We have searched on the internet and it has two essential properties that first, high number of features and second, a lot of instances available.
In the model, the classification goal is to predict if the client will subscribe (yes/no) a term deposit.

Overall, we would like to reduce number of attributes and to decrease least amount of time to train the model for the same fitness levels.

3. Accomplishment:

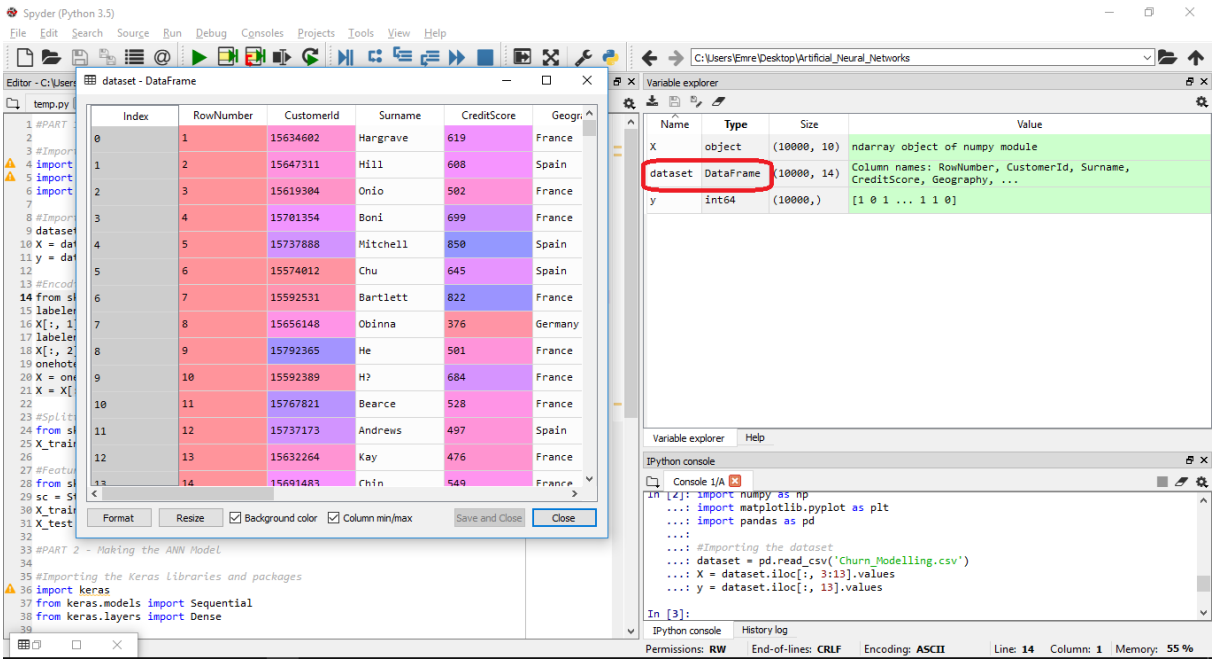
a. Installing essential libraries

In this very first step of the project, we determined we need to certify the essential libraries to be used in our model.

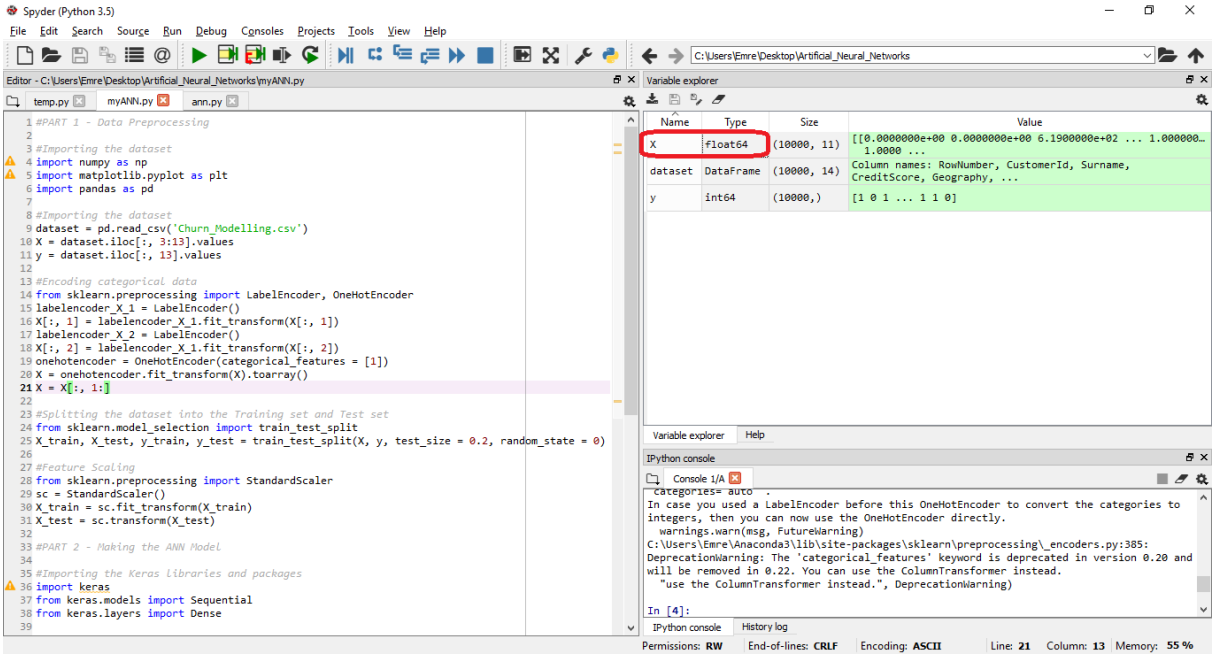
Imported libraries is going to be explained briefly. The libraries were:

- *Numpy Lib of Python*
  Basically, this library serves as a mathematical optimizer in Python. We had an idea that in machine learning algorithms, capability of arithmetic operations is a must. Thus, since we already implement Artificial Neural Networks(ANN) in our model, we knew we would need that library beforehand.
- *Pandas Lib of Python*
  Our both instance and complete dataset is in .csv format so we need to use this library to easily read the excel file.

- *LabelEncoder & OneHotEncoder Libs of Python*
  These libraries serves as fit the model dataset suitable for training. Via importing them, we can change the dataset format to array format. Prior to, as soon as we read our dataset from the .csv file, it was in Object format.



Notice that, now we have floating data format so later, the model can manipulate the data.



- *Keras Lib of Python*
  Keras is the most vital and important library in our program. Without it, we would not able to create a suitable model and train it easily.

b. Getting a relatively small dataset for training and testing

In the former subtask, the adaption of the training and testing dataset have already been mentioned. In addition to that, array of X, stands for our input data and array of Y stands for our given output to adjust the weights in our model. We have 10000 data of accounts and 8000 of them have been used for training part and the remaining 2000 data of accounts have been used for testing the model.

c. Constructing the neural network

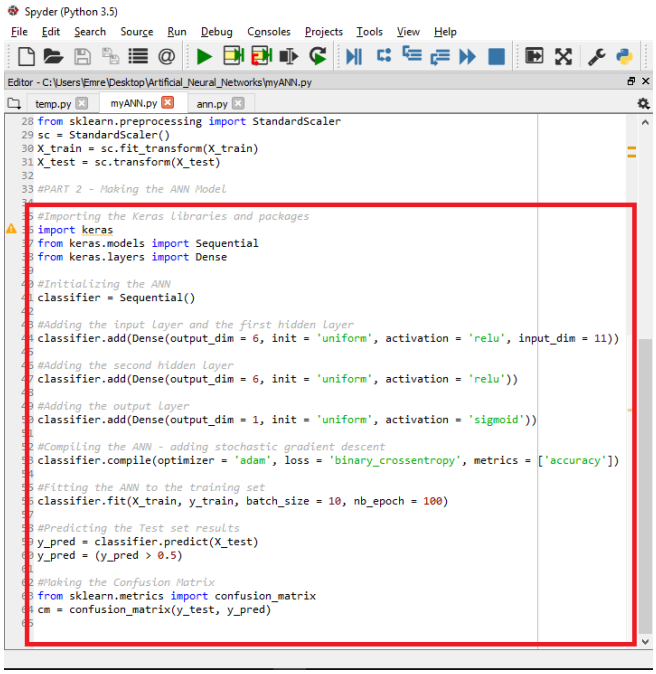
As Keras library has been already mentioned, it gave us the oppurtunity to easily construct layers and splitting up the training and testing dataset.

```
#Splitting the dataset into the Training set and Test set
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)

#Feature Scaling
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```

In first two lines, %20 of the sample dataset is dedicated to testing part and other %80 of the sample dataset has been given to training part.

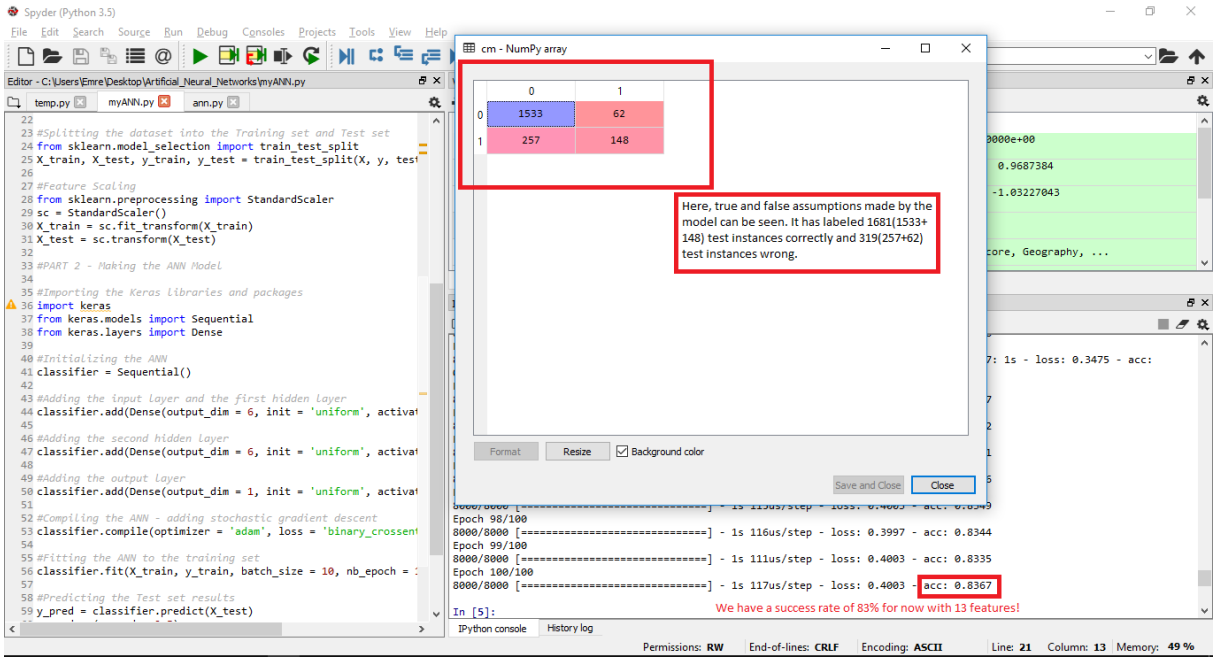
Under the label of feature scaling, X_train array has been used to altering weights and getting model to become ready for testing. X_test is obviously assigned to testing process.



This part of our program code, priorly, the corresponding import process has been done and afterwards, we determined the layers and the particular weight adjustion methods have been determined which are relu and sigmoid functions. In this model, we have 3 layers and 13 total of neurons. The weight's initial values are set according to uniform distribution.

With a batch size of 10 which means the data is going to be grouped by 10 for each iteration and will be given. Also, number of epoch is 100 so this means that our model will be trained by that dataset for 100 times.

d. Observing the initial performance of the model



We have 2 different possible outcomes of predictions, 1 or 0... The total correct and false answers can be shown with an explanation. Our current success rate is 83% with 13 different features.

4. Remaining Work:

As we said, our current success rate is 83% with **13 different features**. We are going to try training the "raw" model with fewer total number of features and observe whether we will still have an **equal or even higher success rate**. That is going to be our next goal to achieve. Selection algorithms will be implemented as well.

Over all, our schedule does not change at all nor the list of references for this project.

5. Project Schedule:

- **Finding Dataset**

Considering finding the convenient dataset, we have a decent progress. This step is not going to be a big issue for us, we believe in.

- **Train data without Feature selection**

Training the model would be a struggling but if we overcome this step then next subtasks will get easier for us in the future. It will probably take five to six days.

- **Run feature selection algorithms**

This is the hardest and the most dangerous step ahead of our plan. If we cannot deal with the algorithms, the project can be stressful. We consider allocating most of the work and time for this subtask.

- **Reducing features based on the results**

Basically, by a guidance of the selection algorithms, the corresponding features will be thrown away and re-train the model with the new dataset.

- **Re-train the model with new dataset attributes**

This task is quite alike the second one because, at this time, we will already have great experience about training our model. It will take at most for one to two days.

- **Interpret and compare results**

This is the final step and it is going to illustrate the improvement. We are planning to show how the required time and number of instances reduced after.

6. References:

[1] Manoranjan Dash, Huan Liu, Consistency-based search in feature selection, Artificial Intelligence 151 (2003) 155–176

[2] Mark A. Hall, Correlation-based Feature Selection for Machine Learning, 1999