Purchase probability using KNN.

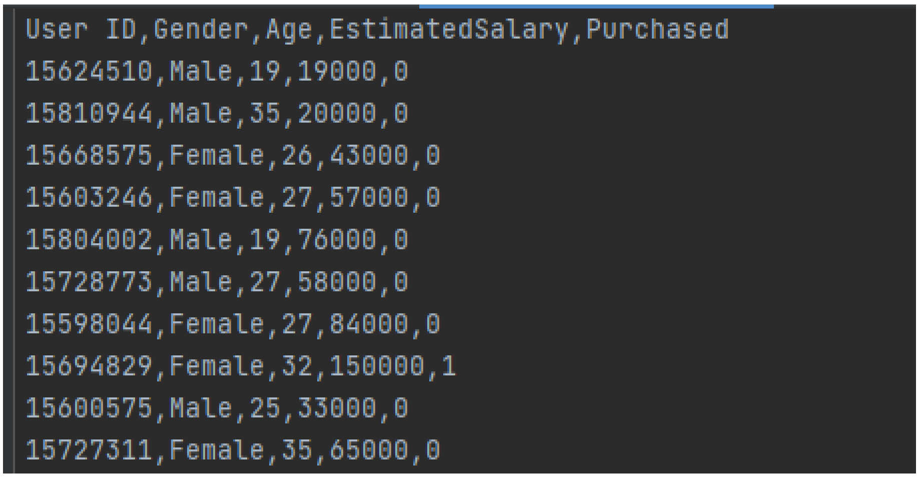
Abstract – This document shows the implementation of a KNN from scratch and the same project but using sklearn framework to determine if a user is going to purchase something based on their age and salary.

I. Introduction.

Sometimes people can decide if they are going to purchase something if their earnings are good. Also is thought that older people have better administrative skills than a younger person [4]. So, the purpose of this project is to determine if the user is going to purchase at the online store or not.

II. DATASET.

The dataset used to test this project was obtained from Github [1]. It has the following features: user, genre, age, estimated salary and if they have purchased something. It is around 400 instances, there are not a lot but enough to work with. This dataset was built for classification, so it was unnecessary to do any data preprocessing



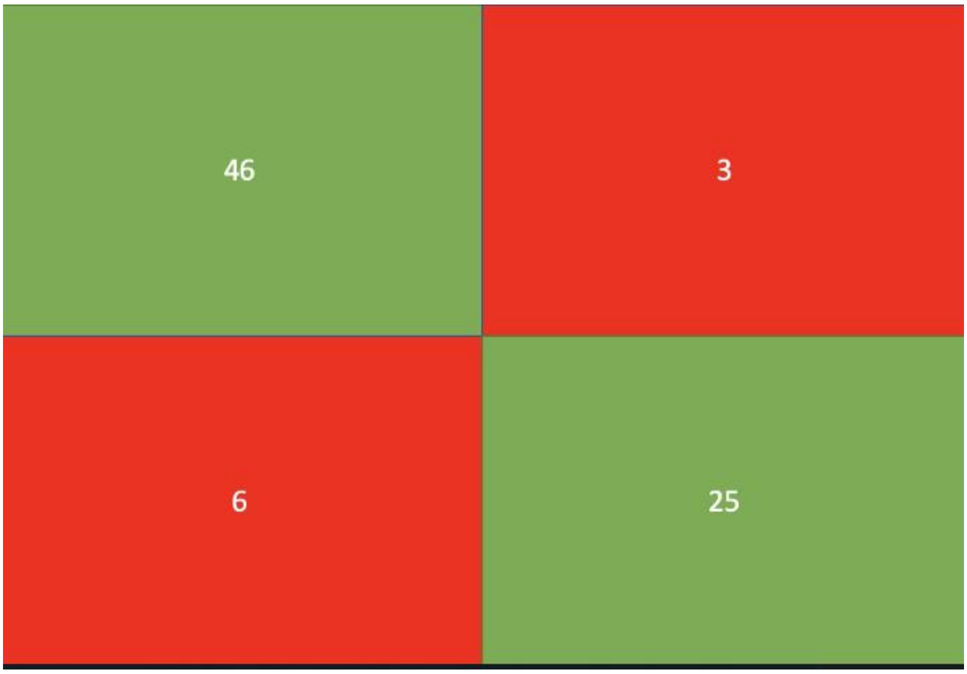
```
User ID,Gender,Age,EstimatedSalary,Purchased
15624510,Male,19,19000,0
15810944,Male,35,20000,0
15668575,Female,26,43000,0
15603246,Female,27,57000,0
15804002,Male,19,76000,0
15728773,Male,27,58000,0
15598044,Female,27,84000,0
15694829,Female,32,150000,1
15600575,Male,25,33000,0
15727311,Female,35,65000,0
```

Figure 1: Part of the data taken for the project.

III. Implementation

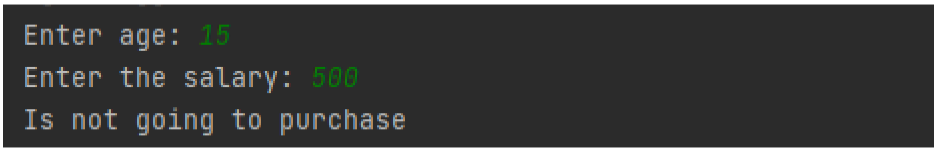
This prediction can be done with a KNN to get a classification between the selected data: age and salary with the previous values to determine if the user is going to make the purchase.

Figure2: Shows the confusion matrix



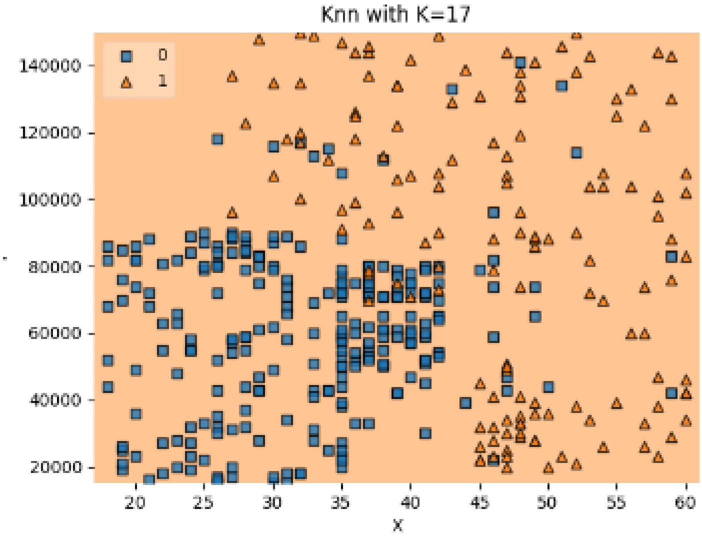
46	3
6	25

Figure3: Shows if someone of 15 years with a salary of 500 will purchase something



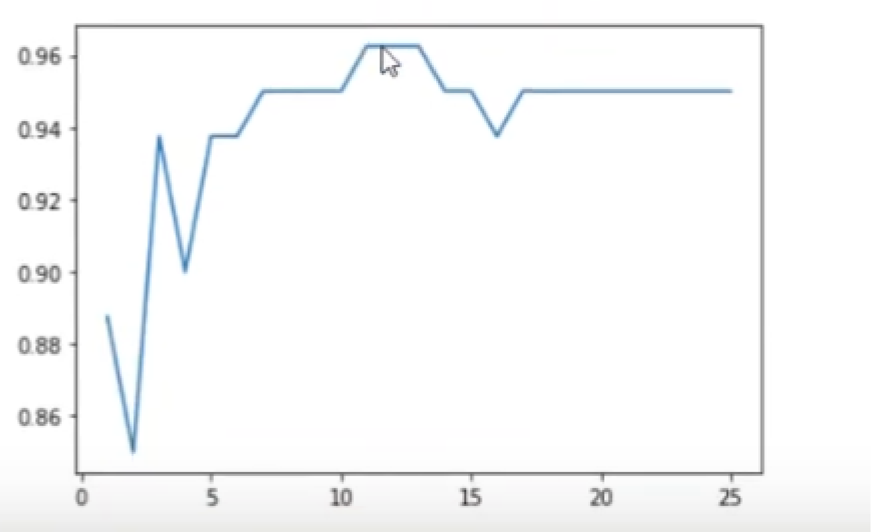
```
Enter age: 15
Enter the salary: 500
Is not going to purchase
```

Figure4: shows the distribution of the data



Our K is going to be the iteration of the dataset with more accuracy

Figure5: shows our K calculation



Comparative table KNN from scrath vs sklearn model

Scratch		SKlearn library	
Input	Output	Input	Output
27, 57000	Is not going to purchase	27, 57000	Is not going to purchase
32, 150000	Is going to purchase	32, 150000	Is going to purchase
48, 28000	Is going to purchase	48, 28000	Is going to purchase

IV KNN

The KNN algorithm is a type of lazy learning, where the computation for the generation of the predictions is deferred until classification. Although this method increases the costs of computation compared to other algorithms, KNN is still the better choice for applications where predictions are not requested frequently but where accuracy is important.[3]

The data was divided in 20% for testing and 80 for training.

KNN was applied to calculate purchase, with the help of the sklearn linear model to corroborate if it works as, it might.

V. Results

Variance was used to measure how different were the real values from the average predicted, it returned a value of 90%

Since the dataset was limited, the cross-validation score was needed to estimate the model predicted data with the test data, it reached a value of 74.9%

Additionally, the mean squared error was 0.35, which means that models are slightly close to the real world classification

VI. Conclusion

The difficulty to predict behavior of someone is extremely difficult, even people who are experts on this topic struggle while doing this. So, this was only for learning purposes. But it may be a little bit better than trying to guess.

VII. References

1. Social networks ads.csv[Online] https://github.com/shivang98/Social-Network-ads-Boost/blob/master/Social_Network_Ads.csv [Accessed 25 October2021]
2. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html HYPERLINK "https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html"[Accessed 10 November 2021]
3. https://towardsdatascience.com/knn-algorithm-what-when-why-how-41405c16c36f
4. https://digitalcommons.uri.edu/cgi/viewcontent.cgi?article=1021&context=hdf_facpubs