Beijing Normal University-Hong Kong Baptist University United International College

Machine Learning (Dr. Rui MENG)

Semester 2 of 2021-2022

1001 and 1002

User Layer Model Of Unsupervised Learning and Supervised Learning

Author:

Xinyao Han 1930026040

Yichun Huang 1930026050

Dixin Li 1930026063

LaTeX

User Layer Model Of Unsupervised Learning and Supervised Learning

Han Xinyao 1930026040 Huang Yichun 1930026050 Li Dixin 1930026063

1 May, 2022

Abstract

With the continuous development of the platform, more and more users have been found, and many operational difficulties have been found, such as how to arouse old users in a targeted manner without disturbing new users. Another example is how to accurately send the discount of promotional activity to some users who need stimulation to generate consumption. The reality of all this depends on the accurate identification of the type of user. Adopting different policies for different users is a user layer model.

Key words: *Random Forest KNN BP Logistic-Regression python K-means*

1 Motivation

Nowadays, a common issue is increasingly being considered and widely applied, which is the issue about how businesses use different schemes for different user types in order to maximize your interests. With the rapid development of technology, it is available to access the user's consumption time, consumption times, and consumption amount, which is much easier to extract information to do further analysis. A good grouping of the users layers model is essential as it can reduce business promotion costs and increase total consumption. Currently, many researchers focus on consumption data by considering several main factors such as frequency of buying and amounts.However, in reality, especially in the e-commerce store due to the totally different patterns of human buying and choosing.Therefore, by using several techniques, we do comparison and evaluation among them to make prediction be effectively exploited and to figure out relevant features for the issue.

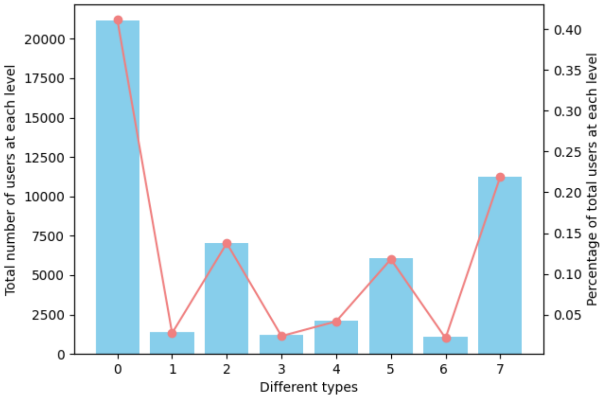


Figure 1: *Types of the users*

2 Related Works

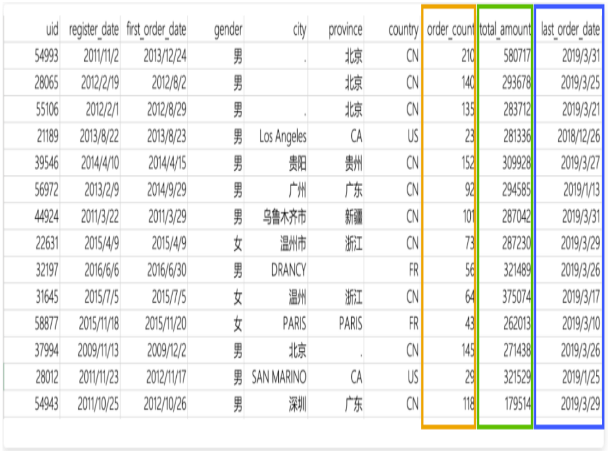
We review some previous works on crowd flow prediction.Many researchers have proposed model to tackle the issues in **RFM** model.The calculation of R (the number of days from the current date of each customer's last order) is divided into two steps, creating a calculation field for the customer's last order time, and then creating the days from the latest order to the current date, and subtracting the number of days between two days. Calculate F (the cumulative number of orders per customer. Calculate M the cumulative transaction amount per customer. Then we can get the graph for different user layers(total have 8 types of customers after using this **RFM** model.

But its model has drawbacks. This classification is very subjective, because it is divided according to binning and divided according to people's wishes, and there is no universality. Secondly, this is very expensive, because there are eight customer types, and each customer type has to do a strategy, so it consumes a lot of the cost of making the strategy.

3 Data Observation

3.1 prepossessing A

The data comes from an e-commerce platform. There are more than 50,000 pieces of data here. Among them, 51395 items remain after removing missing data.



uid	register_date	first_order_date	gender	city	province	country	order_count	total_amount	last_order_date
54993	2011/11/2	2013/12/24	男	.	北京	CN	210	580717	2019/3/31
28065	2012/2/19	2012/8/2	男		北京	CN	140	293678	2019/3/25
55106	2012/2/1	2012/8/29	男	.	北京	CN	135	283712	2019/3/21
21189	2013/8/22	2013/8/23	男	Los Angeles	CA	US	23	281336	2018/12/26
39546	2014/4/10	2014/4/15	男	贵阳	贵州	CN	152	309928	2019/3/27
56972	2013/2/9	2014/9/29	男	广州	广东	CN	92	294585	2019/1/13
44924	2011/3/22	2011/3/29	男	乌鲁木齐市	新疆	CN	101	287042	2019/3/31
22631	2015/4/9	2015/4/9	女	温州市	浙江	CN	73	287230	2019/3/29
32197	2016/6/6	2016/6/30	男	DRANCY		FR	56	321489	2019/3/26
31645	2015/7/5	2015/7/5	女	温州	浙江	CN	64	375074	2019/3/17
58877	2015/11/18	2015/11/20	女	PARIS	PARIS	FR	43	262013	2019/3/10
37994	2009/11/13	2009/12/2	男	北京	.	CN	145	271438	2019/3/26
28012	2011/11/23	2012/11/17	男	SAN MARINO	CA	US	29	321529	2019/1/25
54943	2011/10/25	2012/10/26	男	深圳	广东	CN	118	179514	2019/3/29

Figure 2: *Data set*

The brand's consumers come from all regions of China.

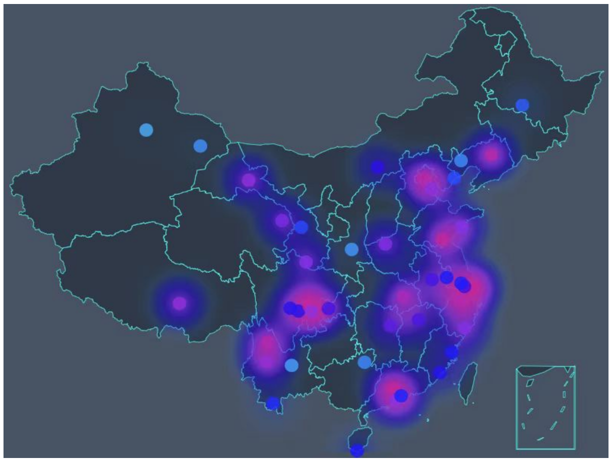


Figure 3: *Location of customers*

Then, we need to used the time gap,order count,total amount to do the **kmeans(unsupervised)**. Because the value in total amount column are much bigger than the time gap column and the order count column. So before doing the unsupervised learning, we need to do standardization for threeUndoubtedly, we need find the useful features in the data set. so we do the correlation for all the features. And using the Python drew the hot graph.

3.2 prepossessing B

We apply the standardization formula on the attribute, after transformation, the feature distribution does not change, and the new mean of the data is 0, and the standard deviation is 1

$$Z = \frac{X - \mu}{\sigma}$$

Figure 4: *formula*

So that this transformation data can help us find a best k for the kmeans algorithm. Using the eblow law, we can find that the best k is when k equals to 3.

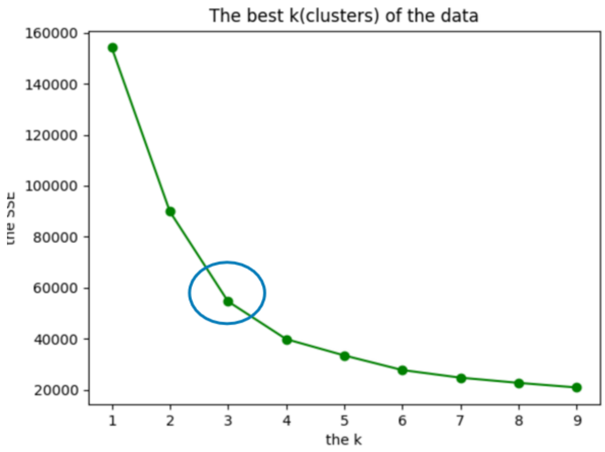


Figure 5: *the best k*

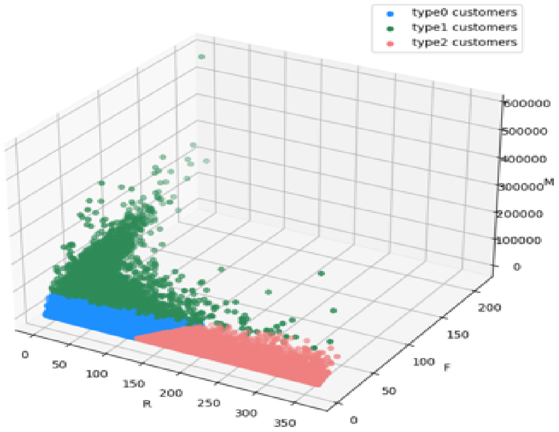


Figure 6: *3D graph*

Also using the KMEANS algorithm to clustering the different groups. There are four steps. We used python codes to write this methods without using the packages. 1. Select k objects from the data as the initial cluster centers. 2. Calculate the distance from each cluster object to the cluster center to divide. 3. Calculate each cluster center again. 4. Calculate the standard measure function. If the method reaches the maximum number of iterations, stop, otherwise, continue to operate.Finally, we draw the 3D graph with F,R,M three axis-es,in order to display the three clusters like that.

The customers of label0 are loyal customers of the brand, and we need to provide them with some VIP services so that they can continue to consume the brand. The customers of label1 are relatively active. Usually, you can give more discount information to this group, so that they can increase the consumption amount. The customers of label2 are customers who need to be awakened. This user is not active enough. It is necessary to increase brand promotion and rearouse this group of users for consumption.

4 Methodology

We use the label coming from the unsupervised learning to do the supervised learning, which means the labels is 0,1,2 generating by the Kmeans algorithms(the unsupervised learning).

4.1 K-nearest

Neighbors Regressor: Similar with KNN Classification, the idea of KNN classification is to calculate the average of the numerical target of the K nearest neighbors. In the preprocessing, we only have continuous variables. For this, we have to apply Euclidean distance to calculate them.

$$D_E = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

Figure 7: *Euclidean distance*

Therefore, we choose the optimal K value by first inspecting the data. When having K-nearest neighbors, we iterate all tuples in training data set by computing distance based on the above equations, then we figure out the top K=3 neighbors.

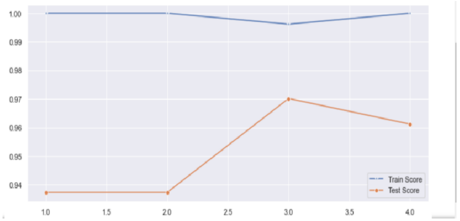


Figure 8: *The best k*

Eventually, we take the average of labels value for those K items to obtain average value as predicted value.

4.2 Logistic Regression

After iteration for different times, can get the graph of the accuracy.

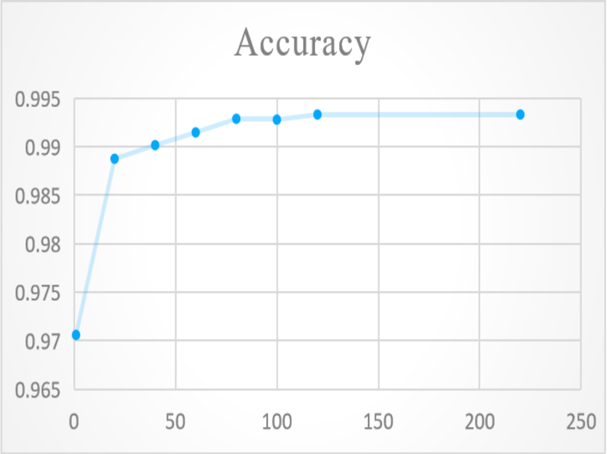


Figure 9: *iteration*

Since the logistic regression studied in class can only classify between 2 categories at a time, we used the idea of One-Vs-All for this classification problem, turning a multi-classification problem into a multiple binary classification problem. For this three-classification problem, we then obtain 3 binary classifiers.

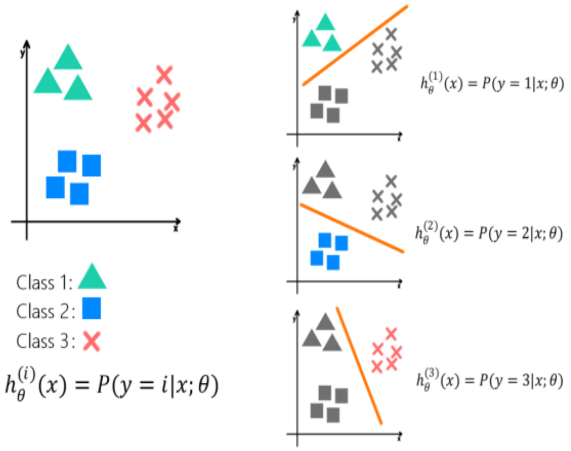

Class 1: ▲
Class 2: ■
Class 3: ✕
$h_\theta^{(i)}(x) = P(y = i|x;\theta)$
$h_\theta^{(1)}(x) = P(y = 1|x;\theta)$
$h_\theta^{(2)}(x) = P(y = 2|x;\theta)$
$h_\theta^{(3)}(x) = P(y = 3|x;\theta)$

Figure 10: *3 classification*

In this process, we train a classifier for each category i, where the corresponding category is positive and all other categories are negative, and determine in turn whether the predicted values belong to the labels of the current cycle.In each loop, the optimize.minimize function is used to learn the parameters and obtain the optimal parameters for each category. The final parameter theta after I loops is a matrix of (i, n+1) For our optimize function, in which we can adjust the number of iterations to obtain different values of theta, and then make predictions.

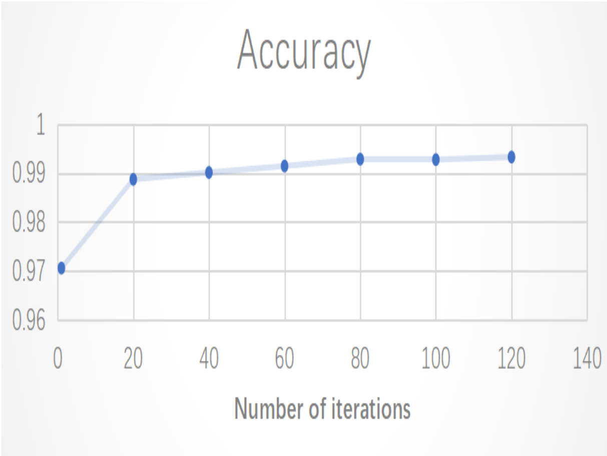


Figure 11: *iteration graph*

By increasing the number of iterations, the accuracy increases, but after the number of iterations exceeds 120, there is no need to increase it any more to avoid wasting space. After heat map of every feature pair, and we comparison of the score of different feature selections.

4.3 BP

Too many neurons may lead to overfitting and otherwise, it may lead to underfitting, we find the empirical formula:

$$\sqrt[2]{m+n+a}$$

Figure 12: *formula*

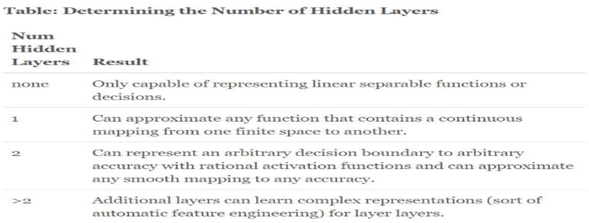


Table: Determining the Number of Hidden Layers

Num Hidden Layers	Result
none	Only capable of representing linear separable functions or decisions.
1	Can approximate any function that contains a continuous mapping from one finite space to another.
2	Can represent an arbitrary decision boundary to arbitrary accuracy with rational activation functions and can approximate any smooth mapping to any accuracy.
>2	Additional layers can learn complex representations (sort of automatic feature engineering) for layer layers.

Figure 13: *hidden layers*

where m represents the number of neurons in the input layer, can represents the number of neurons in the output layer, and a represents an integer from 1 to 10. Based on our data, we finally chose 13 as the number of neurons.The second challenge is to define the number of layers of hidden layers, again, too many layers may lead to overfitting and too few layers may lead to underfitting. We have found that for neural network models, without hidden layers, it is possible to represent linearly differentiable functions or decisions as best as possible, and if the number of layers is 1, it is possible to fit any function that "contains a continuous mapping from one finite space to another finite space". If the number of layers is 2, with an appropriate activation function, any decision boundary of arbitrary precision can be represented, and any smooth mapping of any precision can be fitted, and if the number of layers is greater than 2, then the extra hidden layers can learn complex descriptions.

We compared the accuracy of 2-layer neural network and 3-layer neural network with different number of iterations of the test set, and we can find that the 3-layer neural network improves accuracy faster at the beginning, however, as the number of iterations increases, the accuracy improvement is limited and more prone to overfitting.

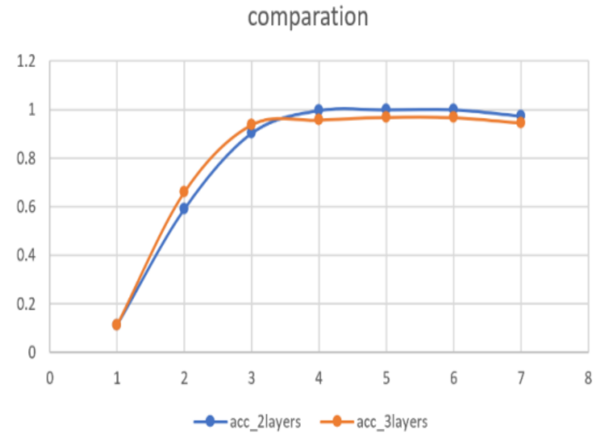


Figure 14: *compartion*

4.4 Random Forest

Random Forest classification: Random Forest is a bagging technique. with a single decision compared to the tree in the adjusted training data,decision trees may differ, resulting in different predictions. Therefore, the idea of random forest is to train at the same time constructing multiple decision trees and outputting mean predictions for a single tree. All trees in random forest also allow them to run in parallel. The general procedure of the algorithm is shown. We first split the training data-set into N subsets using guided sampling. Then, we train different decision trees for various sub-samples. We take the predictions for each decision tree and take the average of all predictions as the predicted value.

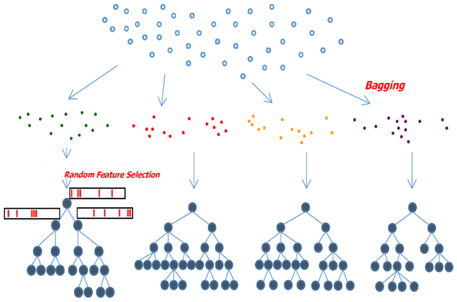


Figure 15: *Random forest Process*

Random forest has many parameters, so we used grid search to find the best parameters, which is a picture drawn by finding suitable parameters.

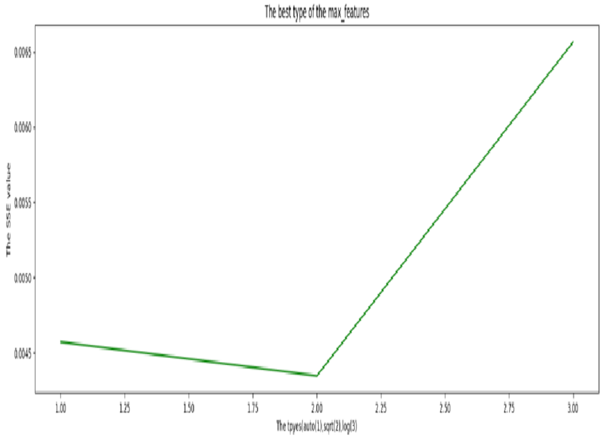


Figure 16: *find the max feature*

For many data sets, random forest algorithm can generate a highly accurate classifier and run efficiently on large data sets. It can help to improve the accuracy.

5 Result Analysis

After model selection, we need to choose the useful feature data as the input of the model. In order to find out which feature contribute to our prediction model, firstly, draw a heat map graph to have an overview of the correlation between every numeric feature pair. The significant features and the labels also have the strong relationship.

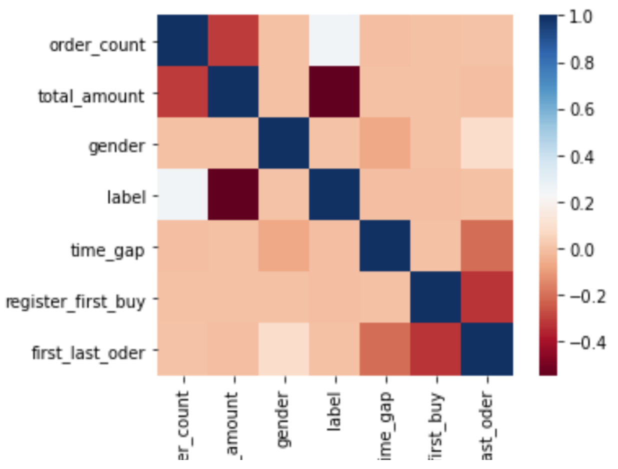
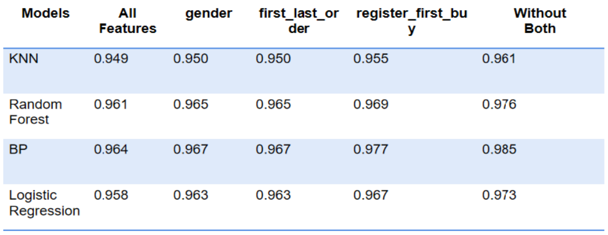


Figure 17: *Hot Map*

After heatmaps for each feature pair, we selected the three variables with the highest correlation among them and tested the the rest of 3 features which relatived lower correlation with using four models and obtained accuracy rates for each. Based on this table, we compared the scores of the different feature choices and, based on the final comparison, we decided to use the previous three variables as input to the model without adding any further variables.



Models	All Features	gender	first_last_order	register_first_buy	Without Both
KNN	0.949	0.950	0.950	0.955	0.961
Random Forest	0.961	0.965	0.965	0.969	0.976
BP	0.964	0.967	0.967	0.977	0.985
Logistic Regression	0.958	0.963	0.963	0.967	0.973

Figure 18: *Comparison of the score of different feature selections*

All for the two tables are using the label0 as positive. After training the models, we can obtain four classification models, which are KNN, Random Forest, BP, and the logistic regression. Then, by making use of the three evaluations on the test data set.Table1 is the standardization data set.

Table 2 is the evaluation value of three model results of the normalized data. In terms of the score and R square evaluation, there is no big difference between these three data types in the Random Forest model BP and Logistic Regression,all are about 0.995, 0.996 and 0.984 respectively, which means these three models are not sensitive to different types of data. In the KNN Regressor model, the score of accuracy value is0.966, when using the standardized data, and 0.947 in normalized data.

Although the score of accuracy does not change a lot in these four models, it is not hard to find out that the precision and F1-score reach the maximum in standardization data. And the high F1-score and accuracy in standardization data might cause by the scaling of the standardization, however, to obtain a model that can do prediction with higher accuracy, we decide to use the standardization data as the input data of our models. THE(P:Precision;R:Recall;F:F1-score)

Methods	P	R	F	Accuracy
KNN	0.927	0.934	0.930	0.966
Random Forest	0.988	0.992	0.990	0.995
Logistic Regression	0.933	0.939	0.936	0.984
BP	0.994	0.993	0.993	0.998

Table 1: standardization evaluations outcomes

Methods	P	R	F	Accuracy
KNN	0.922	0.932	0.927	0.947
Random Forest	0.96	0.972	0.966	0.988
Logistic Regression	0.951	0.968	0.959	0.959
BP	0.97	0.982	0.976	0.973

Table 2: normalization evaluations outcomes

6 Conclusion

In this project, we first used unsupervised learning (K-means) to classify the customer types, and then, based on the classification results, we used three types of supervised learning: logistic regression, backpropagation algorithm, and random forest algorithm to learn the data and perform the classification, and finally, we came up with a model that can classify customers based on attributes. The obtained model prediction results are more accurate and simpler than the previous FRM model prediction results.

References

[1] (2011-06-01). *research on graph-based semi-supervised learning and its applications*. China Knowledge Network, 5st edition.

[2] (2017-02-22). *research on semi-supervised learning and its application*. China Knowledge Network, 9st edition.

[3] (2022). *semi-supervised learning methods*. Journal of Computer Science [cited 2015-08 19, 1st edition.

[4] Al-Masri, A. (2019-JAN-30). *How Does Back-Propagation in Artificial Neural Networks Work?* Ediciones Tepito.

[5] Sanjay.M (2018-10-26). Machine-learning — knn using scikit-learn. `https://towardsdatascience.com/knn-using-scikit-learn-c6bed765be75`.

[6] Zhao, X. (2016-12-8). Awesome random forest. `https://github.com/zhaoxingfeng/RandomForest`.