International Journal of Advanced Science and Technology
Vol. 29, No. 9s, (2020), pp. 2796-2806

Analysis of bank customers and prediction of bank marketing strategies

G.Saranya, *Assistant Professor, Department of Computer science and Engineering, SRM institute of science and technology, Chennai, India*
saranyag@srmist.edu.in.

K.Kumaran, *Assistant Professor, Department of Information Technology, Easwari Engineering college, Chennai, India*
kumaran.me.cse@gmail.com

J.Dheeraj, *UG scholar, Department of Computer science and Engineering, SRM institute of science and technology, Chennai, India*
dheerajj500@gmail.com

S.Gurubharan, *UG scholar, Department of Computer science and Engineering, SRM institute of science and technology, Chennai, India*
gurubharan1998@gmail.com.

Mohammed K Hashim, *UG scholar, Department of Computer science and Engineering, SRM institute of science and technology, Chennai, India*
mohdkhashim07@gmail.com.

Abstract

Machine learning Visual Data Analysis is the method of using the computer graphics to explore and display data. The use of Machine learning algorithms has seen widespread use cases in the modern Data analysis field. The proposed system focuses on commercial banks' customer data, and focuses on the interaction techniques to predict the customer involvement and investment with the bank. We will be using EDA for the analysing part. Highlights of the findings suggest the strategies like Months of Marketing Activity, seasonality of the client's deposit, the number of campaign calls and target clients with high duration with the bank. Finally By combining all these strategies and simplifying the market audience the next campaign should address, it is likely that the next marketing campaign of the bank will be more effective than the current one.

Keywords: *Machine learning, Support Vector machines, Gradient Boosting, Random Forest*

I. INTRODUCTION:

The method used in this project focuses on the methodology to follow the customer figuration that can be followed by the Indian banks. The terminologies, business practices, transaction types, database fields, objectives In effective usage of technology, etc..are the same as the Indian public bank sectors.The banks in order to obtain greater profits and achieve greater cross selling of financial products. they not only need more customer resources. The banks must provide customers with their priority needs and service at all cost. If the bank is getting no return from the customer interaction with the bank it means the bank is at loss. According to the banks the clients who bring more profit for the bank are the valuable clients. The two sources of bank profits are term deposit and loan interestfee.

The client's data plays an important role in analyzing, however the attributes like age, gender, marital

status, education, housing, loan, balance, the last contact with the bank and the duration of the call is noted. The output from the client data will suggest the mean age, the mean balance and the client with more duration with the bank. The initial step analysis is done by occupation, marital status, clustering marital status with education and campaignduration.
By using stratified sampling we can divide the dataset population into groups to avoid overfitting and implement the groups in cross fitting so we get more accurate outputs. The best alternative to avoid overfitting is touse cross validation. We take the training test and split it. For instance, if we split it by 5, 3/5 of the data or 66% will be used for training and 2/5 33% will be used for testing and we will do the testing process five times. This algorithm will iterate through all the training and test sets and the main purpose of this is to get the overall pattern of the data. With the data we get, we use a confusion matrix. The main purpose of a confusion matrix four terms the true positive, false positives, true negatives and false negatives. With the help of the ROC curve for visualizing the data we get the analytic part of our finding and we can classify that the duration, month and the contact with the client plays an important role in the classifying. Finally to get the potential customer we use the gradient model to get the best accuracy in our findings. The accuracy is high in this model.

II. DATASET OVERVIEW:

1- Job: Type of job with variouscategories.
2- marital: marital status (categorical: 'divorced','married','single','unknown';)
3- education: (categorical: primary, secondary, tertiary andunknown)
4- default: has credit in default? (categorical:'no','yes','unknown')
5- housing: has a housing loan? (categorical: "no","yes")
6- loan: has a personal loan? (Categorical: "yes","no")
7- balance: Balance of the individual.
8- contact: contact communication type(categorical:'cellular','telephone')
9- month: last contact month of the year.
10 - duration: last contact duration, in seconds
11- campaign: number of contacts performed during this campaign and for this client (numeric, includes the last contact)
12- pdays: number of days that passed by after the client was last contacted from a previous campaign

13- previous: number of contacts performed before this campaign and for this client (numeric)
14- poutcome: outcome of previous campaign
15- campaign(categoricalyes/no)

III. RELATED WORKS:

The studies have been carried out earlier in Customer Profiling using Classification Approach for Bank Telemarketing. The dataset consists of customer information that will be used as an input feature during the data mining task. The input feature plays a key role in the prediction processes. The research was carried out to find the minimum set of features that is close enough to represent the original dataset but gives a good result. The data mining procedure was carried out by a Classification model called as the classifier. The goal of the research is to customer profiling through classification as well as identifying a group of customers who have a high probability to subscribe to a long termdeposit.

IV. SYSTEM ARCHITECTURE:

The main objective of our model is to analyze the data thoroughly and find interesting patterns in the data that could impact the opening of a term-deposit. We use supervised machine learning algorithms for this purpose. The problem in hand is a two-class classification problem. This is also called a binary classificationproblem.

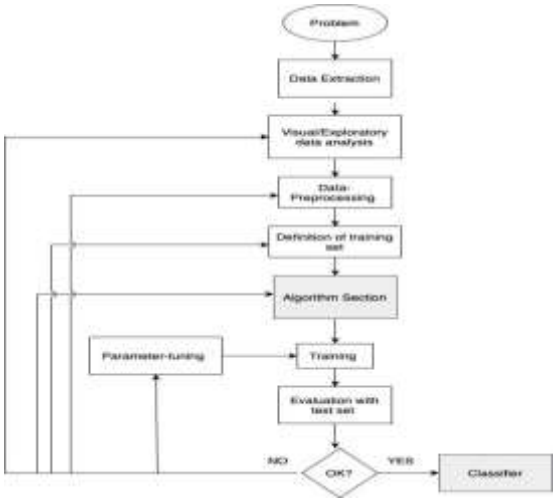

Fig 1: Supervised ML model architecture

1. Retrieval of data: Data retrieval means obtaining data from a database management system in our case it isthe data set. The retrieved data is used for further analysis.
2. Data cleaning: It is the process of detecting and correcting corrupt or inaccurate records from a Data set, it helps by increasing the accuracy.
3. Visual data analysis: By using visuals like graphs and maps, data visualization tools provide an easy way to see and understand trends, outliers, and patterns indata.
4. Data preprocessing: This technique is used to standardize the data and to convert it into a suitable form before feeding it to the machine learningmodel.
5. Model building: The process is to understand the data and paying attention to what is important in the data. We use supervised
6. machine learning model in our paper to identify the importantstats.
7. Model Evaluation: Evaluating the final model with the help of performancemetrics.

Pipelines:

The class Pipeline allows stacking multiple processes into a single scikit-learn estimator. The pipeline class has fit, predict and score method just like any other estimator (ex. Linear Regression). Pipeline helps to enforce the desired order of application steps which in turn helps in reproducibility and creating a convenient work-flow. But, there is something more to pipeline, by using grid search cross- validation, we can understand it better. We can bypass oversimplification by using pipeline. In our model, we perform four functions with the help of pipelines. We first encode the categorical features in our data with the help of one-hot encoding. We then standardize the values and build our models on the standardized and encoded data. We also perform hyperparameter tuning with the help ofgrid-search.

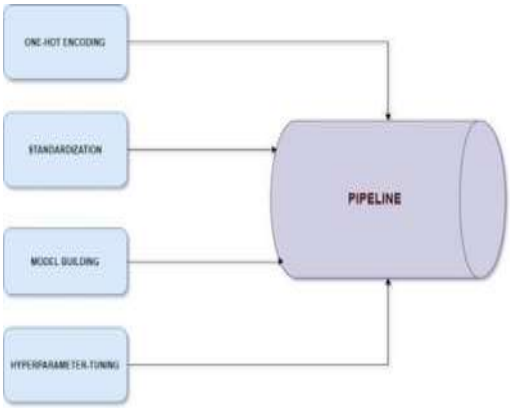

Fig 2: Pipeline architecture

V. DATA CLEANING:

Dealing with missing values:

There are various ways of dealing with NaN values. We can fill those values with the mean/median of the values in the particular column. The fillna() method from pandas helps in filling missing values with the desired value. Another way to deal with missing values is to remove them. The dropna() method from pandas helps in removing missing values from a particular column. Fortunately in our dataset, there ain't ant missing values.

We remove the unknown values in the job column and we bin the balance column into three categories. They are negative balance, low balance, and highbalance.

Negative balance: When the balance is less than or equal to zero the balance is classified as a negative balance

Low balance: when balance<=30000 &balance>0 the balance is classified as low balance. High balance: When the balance>30000 the balance is classified as highbalance. Hence a new vector called balance status is created with values of negative, low and high.

Correlation matrix:

A correlation coefficient is a metric that measures the extent to which numeric variables are associated with one another. It ranges from-1 to +1.

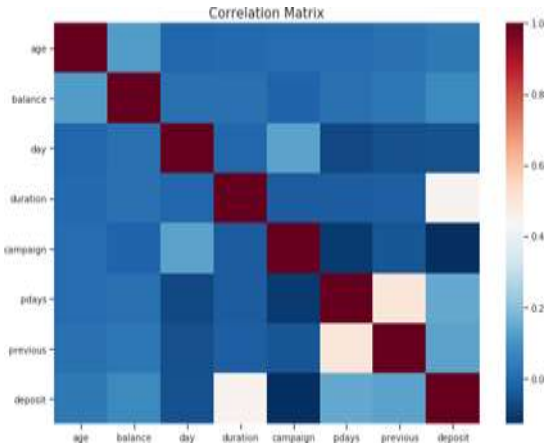

Fig 3: Correlation Matrix

A correlation matrix is a table where the variables are shown on both rows and columns and the cell values are correlations between the variables. A heatmap is used to represent the correlation between the variables in the dataset. A heatmap is a graphical representation of data in which data values are represented as colors. Here we can see that the duration column isthe most co-related to the deposit column. Thus the duration is an important variable in determining the depositstatus.

VI. DATA PRE-PROCESSING:

K-fold cross-validation:

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample as training data in order to estimate how the model is expected to perform in general when used to make predictions on unseen data or the test data. There are various types of cross-validation such as stratified, k-fold,etc.

The general procedure is as follows:

1. The dataset is shuffledrandomly.
2. It is then split into kgroups

3.Foreachgroup:
a.Take the group as a unseen or test data set b.Take the remaining groups of data as a training data set
c. Fit a model on the training data and evaluate it on the test data
d. Store the evaluation score and discard themodel
4.Summarize the model using the sample of model evaluation scores

We apply k-fold cross validation technique to our dataset as it reduces the effect of outliers in thedata.

One-hot encoding:

The phrase one-hot encoding comes from digital circuit terminology where it describes circuit settings in which only one bit is allowed to be positive. It is basically creating multiple dummy variables for a single variable. In linear and logistic regression one-hot encoding creates problems due to multicollinearity. It is in the form of binary 0/1 values. Each categorical variable is split into multiple variables depending upon the number of categories and values of 0/1 are given to each new variable. One-hot encoding can cause problems when there are multiple categories for a variable and also when there are alarge number of categorical variables are one-hot encoded.It increases the dimensions of the data. It also helps to better understand the patterns in the data. Standardization of ourdata:

Since most of our machine learning models use some form of distance to predict the out variable we need to standardize the Xi values. Standardization (or Z-score normalization) is the process of rescaling the features so that they'll have the properties of a Gaussian distribution with a mean of 0 and a standard deviation of 1. Standardization is robust to outliers as compared to normalization. They are represented by a variable Z and are called z-scores. Measurements are then stated in terms of standard deviations away from the mean. In this way, the variable doesn't influence a model simply due to its scale of its original measurement.

VII. EXPLORATORY DATA ANALYSIS:

Exploratory data analysis (EDA) is a forward step in analyzing data sets to summarize their primary characteristics, basically involves with visual methods(visual datasets). It also involves the process of performing statistical tests on the data. EDA helps in finding patterns in data. It can be performed on structured or unstructured data. It helps in the conversion of raw data into useful data. In our dataset,

we have performed multiple visual analysis techniques and have come up with some insightful information about our data.

Observation: In this bar plot the x-axis denotes the duration status of the call and the y-axis denotes the percentage of customers who have opened a term-deposit. From this plot, we understand that the longer the communication between the customer and the bank the more chances that a customer might open a term deposit.

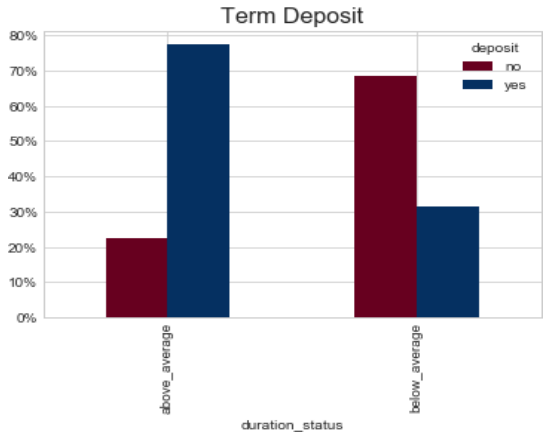

Fig 4: Barplot of duration status vs count

Observation: This is a violin plot which denotes the balance amount in the x-axis and the type of Job in the y-axis and the color- palette represents the deposit status with blue representing yes (term-deposit) and orange denoting No (no term- deposit).It is clear from the graph that the retired people have the most balance in their accounts followed by customers with blue-collar jobs. Most customers with blue- collar jobs haven't opened aterm-deposit.

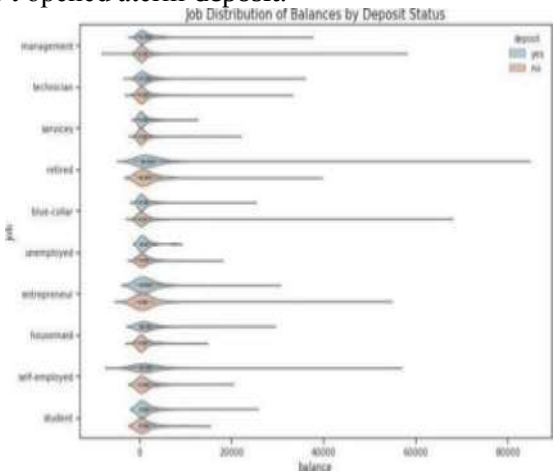

Fig 5: Violin plot of job distribution vs balance

In the below image, X-axis consists of whether the customer has a hosing or not while the y-axis displays the count of the respective categories. The unstacked bars are depicted in order to represent the deposit status with the blue bars indicating Yes (term-deposit) while the orange ones represent No (no term-deposit). It is clear from the above plot that customers who own a house don't' have a term-deposit as compared to those who don't own a house.

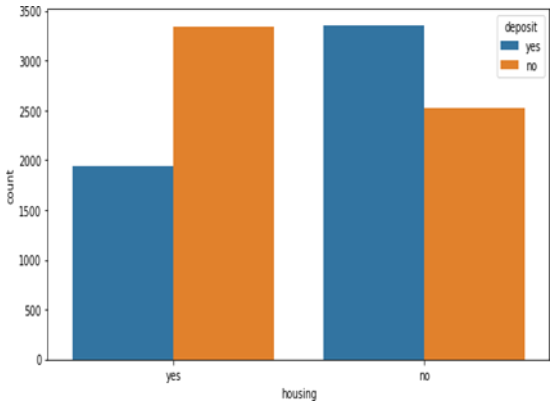

Fig 6: Barplot of housing vs count

Observation: Here the x-axis represents the balance status in terms of High, Low, Moderate and Negative respectively. It depicts the relationship between balance and the opening of a term-deposit. Customers with a high balance are most likely to open a new term-deposit as compared to the customers with low or negatives balances.

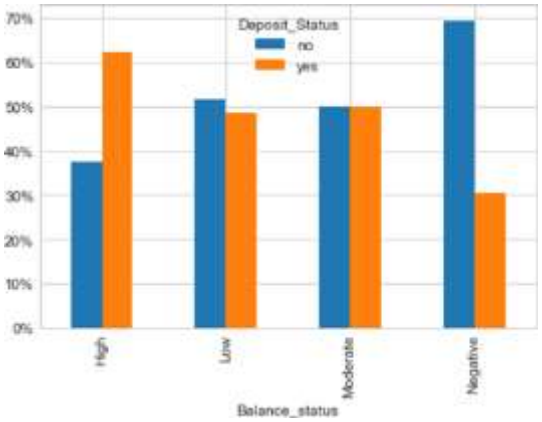

Fig 7: unstacked barplot of Balance Status vs count percentage

VIII. SYSTEM IMPLEMENTATION:

The main objective of the problem is to predict if the customer will open a term deposit or not. It is a two-class classificationproblem.

1. K-NearestNeighbors:

It uses Euclidean distances to determine the nearest neighbors to a specific data point. The number of neighbors to be chosen is determined by brute-force. The most optimal k is selected and the data is trained on that k. The most optimal k is 17.

2. Logistic Regression:

Finds the right hyperplane that classifies the data appropriately. It solves an optimization equation to minimize the vector w of the hyperplane. Uses L1 or L2 regularizer. L1 regularizer creates sparsity. Logistic regression works well on two- class classification data. Grid search is performed to find the right hyper-parameter lambda. The logistic regression is implemented inside thepipeline

.

3. Random-forest:

In random forest, we combine several decision trees with the help of bootstrap aggregation. We give some amount of data to each decision tree and combine each of its outputs. Decision trees are used as base models and the column sampling is done without replacement. The decision trees are usually deep. It uses the majority vote to determining the output class. Row sampling, as well as column sampling, is performed.

4. Gradient Boosting DecisionTrees:

Gradient Boosting trains many models in a gradual, additive and sequential manner. Gradientboostingis a machine learningtechnique for regression and classification problems, which also uses decision trees as base models to produce a prediction model in the form of an ensemble of weak prediction models. The learning rate is an important parameter and the higher the learning rate chances of overfitting increases. It builds the model in a stage-wise fashion like other boosting methods do, and is not parallelizable as each stage depends on the previous stage to perform the optimization. It increases the time complexity as it trains on every single decisiontree.

5. Linear-SVM:

It maximizes the margin-distance from the hyperplane. The dual form of SVM can be used in the case of a similarity matrix as it reduces the time complexity and space complexity. The greater the margin distance greater the generalization accuracy that is the accuracy on unseen data. The hyperparameter used here is c which is the inverse of lambda. The hinge loss is minimized in the loss minimization interpretation of SVM. Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. The linear SVM doesn't work well with non-linear data but since our data is linear it does a good job. In case of non-linear data kernel SVM transforms the data into a higher dimension with the help of kernal trick and converts in into a linearstructure.

IX. MODEL EVALUATION:

1. Accuracy:

It is a metric widely used for classification problems. It compares the predicted class label with the actual class label. It is usually represented in a probabilistic form ie it lies between 0-1. It can be imported from the scikit- learn library. The accuracy was determined for various models in our project. From the below table we can conclude that the gradient boosting algorithm produces the best accuracy.

MODEL	ACCURACY SCORE
K-nearest neighbor	0.81 (or) 81%
Logistic Regression	0.82 (or) 82 %
Linear SVM	0.84 (or) 84%
GBDT	0.83 (or) 83%
Random Forest	0.85 (or) 85 %

Table 1: Accuracy-score for various models

2. Precision:

True positives: Number of points that are positive class and are correctly classified by the model.
False positives: Number of points that are negative class but are classified as positive class by the model. True Negative: Number of points that are negative class and are correctly classified as negative class by the model.
False Negative: Number of points that are positive class but are classified as negative class by the model. Precision measures the accuracy of a predicted positive point. It is the ratio of number of true positive points to the sum of true positive and false positive points. When the number of false negatives is really imported precision value is extremely important. It describes the percentage of the predicted results that are actually relevant.

3. Recall:

Recall also known as sensitivity measures the strength of the model to predict a positive outcome. Recall is highly relevant when the false negatives points are important.

4. Confusion Matrix:

The confusion matrix is a table showing the number of correct and incorrect predictions categorized by type of response. It represents the number of correctly classified points and the number of wrongly classified points by our model.

Results:
TP: 1012
FP: 250
FN: 163
TN: 807

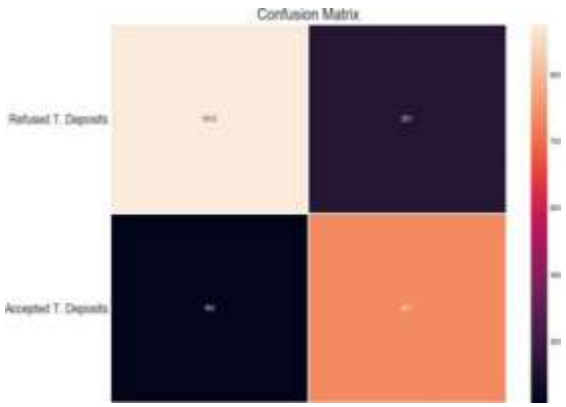

Fig 8: Confusion Matrix

X. FUTURE WORKS:

Future works involve the use of unsupervised machine learning techniques to predict customer behavior. Density-based clustering technique can be used for this. This technique does not require a particular specification of the number of clusters. DBSCAN algorithm is used to find arbitrarily size and it is also capable of finding arbitrarily shaped clusters. The potentiality of the clients is to point other customer who is in the same line of business can collaborate within the territory of the enterprise for their mutual benefits. Hence this value is independent of any outside enterprise, for any companies that have similar domain. The customer and enterprise transaction facilities will be highly enhanced for the customer's fulfillment, best efforts will be put to make customer potentiality into real value. The potential value ultimately aims the customer to opt their own attribute i.e., decision

orientation from the liberty that potential value provide each and every customers. Hence this sort of facilities are ought to provide customers just to satisfy their needs in order meet their benefits legitimately, they can use this as a platform and take genuine decisions for their potential development and cooperation between two aspects of similar line businessventures.

XI. CONCLUSION:

We have used supervised machine learning models for predicting if a customer wouldopen a term deposit or not. After evaluating the model it is clear that the gradient boosting algorithm provides the best results. During our visual data analysis stage, we have discovered some useful solutions for the next marketing campaign As per the analysis of marketing activity of any common bank, we could see a pattern in which most customers fall into the category of opting to open a term deposit, such pattern is taken as the data of analysis. Potential Customers opted to subscribe term deposits are more likely to follow some set of similar patternsspecifically Duration of customer calls, responses of customers during the phone call, age category, house loans and balances, occupation and period of customer interest. Duration of customer calls: duration of the call is the virtue that most likely correlates with whether a potential customer will open a term deposit or not.Responses of customer: by providing a questionnaire in order to increase the level of engagement of the potential customers leading to an increase in the probability of opening a term deposit. Age category: Most of the marketing officials say they prefer to target potential customers in their 20s or younger and 60s or older. Through our visual data analysis, we saw 60% chance in the youngest category and a 76% chance in the eldest category to take the term deposit offer. Occupation: Potential customers who were students and retired are most likely to take a term deposit offer Period of customer interest: As of the data analysis potential customers preferred to subscribe term deposits during the seasons of fall and winter. Implementation of these marketing strategies would help the bank in increasing its customer profiling.

XII. REFERENCES:

1. Link: http://ataspinar.com/2017/05/26/classification-with-scikit-learn/
2. Link: http://ataspinar.com/2017/05/26/classification-with-scikit-learn/
3. Link: https://www.analyticsvidhya.com/blog/2015/
4. Link: https://www.kaggle.com/janiobachmann/bank-marketing-dataset
5. Link: https://www.kaggle.com/randylaosat/predicting-employee-kernelover
6. Link: https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-018-0165-5
7. Link: https://analyticsindiamag.com/10-model-evaluation-techniques-every-machine-learning-enthusiast-must-know/
8. Link: https://towardsdatascience.com/various-ways-to-evaluate-a-machine-learning-models-performance-230449055f15?gi=e0025a587543
9. Link: https://towardsdatascience.com/normalization-vs-standardization-quantitative-analysis-a91e8a79cebf
10. Link: https://datascience.stackexchange.com/questions/57953/what-is-the-purpose-of-standardization-in-machine-learning
11. Link: https://mlcheatsheet.readthedocs.io/en/latest/logistic_regression.html
12. Link: https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab
13. Link: https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/
14. Link: https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d
15. Link: https://www.coursera.org/lecture/python-machine-learning/gradient-boosted-decision-trees-emwn3
16. Link: https://sefiks.com/2018/10/04/a-step-by-step-gradient-boosting-decision-tree-example/
17. Link: https://www.frontiersin.org/articles/10.3389/fgene.2019.00459/full
18. Link: http://www.linearsvm.com/

19. Link: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47
20. Link: https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm