ST3189- Machine Learning

LOKESH VARSHNEY
STUDENT NUMBER- 190301677

Table of Contents

Part1- Cluster Analysis

We have been provided a dataset of people's responses from the European Working Conditions Survey 2016, where 7813 individuals were asked to fill surveys about their mental health, work-life balance, and job satisfaction. Our dataset consists of mixed data type variables with 11 dimensions such as numerical variable Age and Categorical variable Gender. Our task is to explore the dataset and provide a description of the data by unsupervised learning methods such as Principal Component Analysis, Clustering.

I will be using Clustering Unsupervised Learning Algorithm for exploring the dataset. There are different types of Clustering analysis such as K-means clustering, Hierarchical clustering, etc. K-means clustering algorithm is an elegant approach for partitioning the dataset into K distinct clusters where an analyst first has to define the hyper-parameter K which is the required number of centroids and then K clusters are created by allocating each data point to the nearest centroid based on Euclidean distance. Since this distance is valid only for continuous variables and our dataset is of mixed data type variables, we won't be using Euclidean distance methodology instead we will be measuring similarity across individuals using Gower Distance. Gower distance fits well with the PAM algorithm i.e. Partitioning Around Medoids which is very similar to K-means but it is more robust to noise & outliers and produces clusters of very similar individuals which are useful during interpretation.

The optimal number of clusters

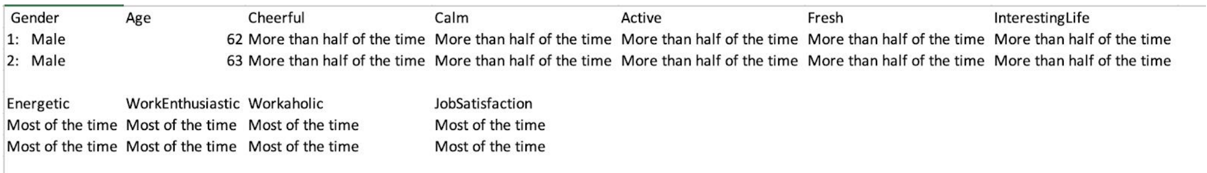
To find the optimal number of clusters, we are going to use the Silhouette coefficient. "Silhouette refers to a method of interpretation and validation of consistency within clusters of data. The technique provides a succinct graphical representation of how well each object has been classified" (Wikipedia, 2019) [1].

Finally, we have visualized our clusters in low-dimensionality by using the Rtsne function of R which stands for "t-Distributed Stochastic Neighbor Embedding. tsne is a technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets" (Filaire 2018) [2].

DataPreprocessing-

1) There were some unusual values in the dataset such as -999 which have been replaced with the mode value of that particular column. 2) For better readability, column names have been changed intuitively such as Q2a to 'Gender' and Q87a to 'Cheerful'. 3) Since the structure of discreet answers of people for every categorical question has been interpreted as an integer in R, they have been converted into factors i.e. categories. 4) For better interpretability, I have assigned given Labels to every discreet number such as for 1 it is "All of the time", for 2 it is "Most of the time" and so on. 5) Gower distance has been calculated between every individual, male & female in our case, using the daisy function of R. Through the matrix, based on Gower distance, we can find that the distance between same variables is intuitively zero. To view the most similar and dissimilar individuals we can use min and max function on this matrix of the dataset and here are the results-

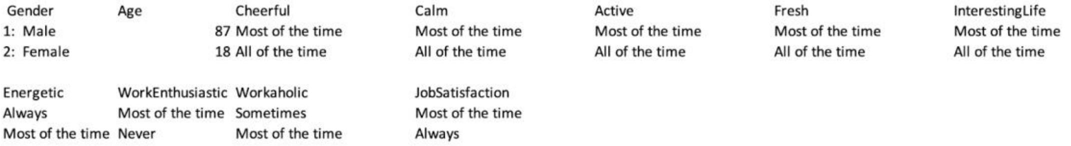
Most Similar case:



Gender	Age	Cheerful	Calm	Active	Fresh	InterestingLife
1: Male	62	More than half of the time	More than half of the time	More than half of the time	More than half of the time	More than half of the time
2: Male	63	More than half of the time	More than half of the time	More than half of the time	More than half of the time	More than half of the time

Energetic	WorkEnthusiastic	Workaholic	JobSatisfaction
Most of the time	Most of the time	Most of the time	Most of the time
Most of the time	Most of the time	Most of the time	Most of the time

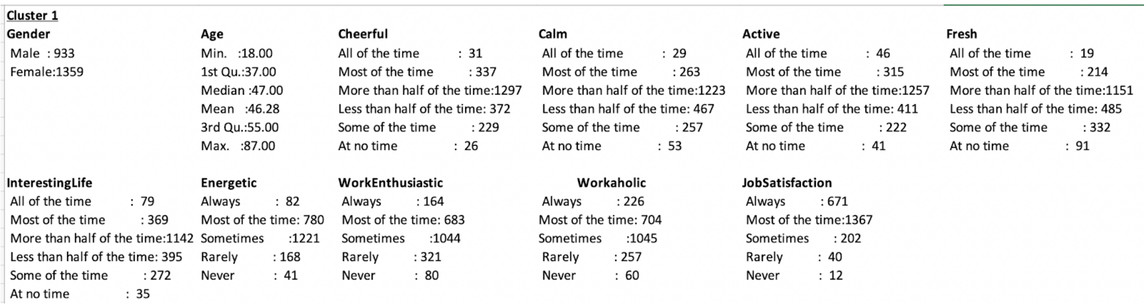
Most Dissimilar case:



Gender	Age	Cheerful	Calm	Active	Fresh	InterestingLife
1: Male	87	Most of the time	Most of the time	Most of the time	Most of the time	Most of the time
2: Female	18	All of the time	All of the time	All of the time	All of the time	All of the time

Energetic	WorkEnthusiastic	Workaholic	JobSatisfaction
Always	Most of the time	Sometimes	Most of the time
Most of the time	Never	Most of the time	Always

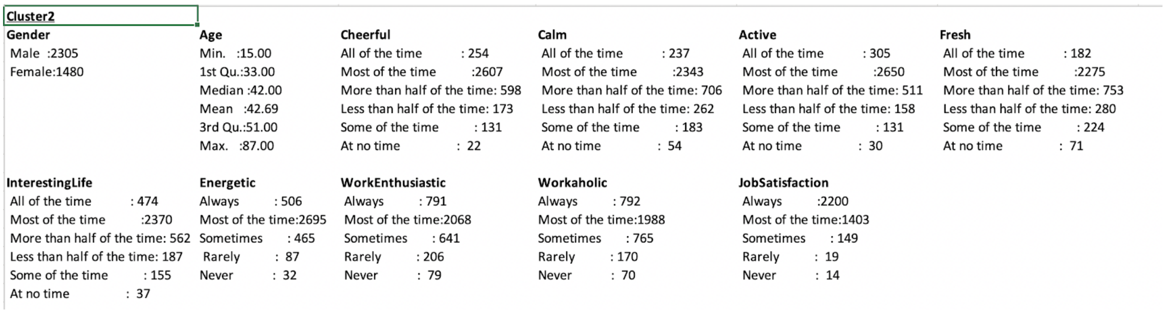
In the most similar case, the distance between two male individuals, aged 62 and 63, found to be the lowest among all and their responses to every question of the survey found similar to each other. 6) Next, using the Silhouette coefficient we found the optimal number of clusters which is 3 in our case. 7) We fit the model using the PAM function at K=3 and below is the visualization and patterns discovered in each cluster.



Cluster 1

Gender	Age	Cheerful	Calm	Active	Fresh
Male : 933	Min. :18.00	All of the time : 31	All of the time : 29	All of the time : 46	All of the time : 19
Female:1359	1st Qu.:37.00	Most of the time : 337	Most of the time : 263	Most of the time : 315	Most of the time : 214
	Median :47.00	More than half of the time:1297	More than half of the time:1223	More than half of the time:1257	More than half of the time:1151
	Mean :46.28	Less than half of the time: 372	Less than half of the time: 467	Less than half of the time: 411	Less than half of the time: 485
	3rd Qu.:55.00	Some of the time : 229	Some of the time : 257	Some of the time : 222	Some of the time : 332
	Max. :87.00	At no time : 26	At no time : 53	At no time : 41	At no time : 91

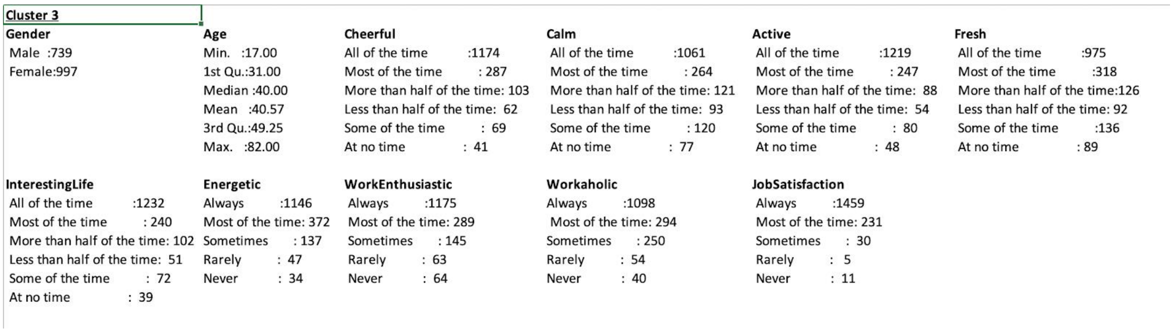
InterestingLife	Energetic	WorkEnthusiastic	Workaholic	JobSatisfaction
All of the time : 79	Always : 82	Always : 164	Always : 226	Always : 671
Most of the time : 369	Most of the time: 780	Most of the time: 683	Most of the time: 704	Most of the time:1367
More than half of the time:1142	Sometimes :1221	Sometimes :1044	Sometimes :1045	Sometimes : 202
Less than half of the time: 395	Rarely : 168	Rarely : 321	Rarely : 257	Rarely : 40
Some of the time : 272	Never : 41	Never : 80	Never : 60	Never : 12
At no time : 35				



Cluster2

Gender	Age	Cheerful	Calm	Active	Fresh
Male :2305	Min. :15.00	All of the time : 254	All of the time : 237	All of the time : 305	All of the time : 182
Female:1480	1st Qu.:33.00	Most of the time :2607	Most of the time :2343	Most of the time :2650	Most of the time :2275
	Median :42.00	More than half of the time: 598	More than half of the time: 706	More than half of the time: 511	More than half of the time: 753
	Mean :42.69	Less than half of the time: 173	Less than half of the time: 262	Less than half of the time: 158	Less than half of the time: 280
	3rd Qu.:51.00	Some of the time : 131	Some of the time : 183	Some of the time : 131	Some of the time : 224
	Max. :87.00	At no time : 22	At no time : 54	At no time : 30	At no time : 71

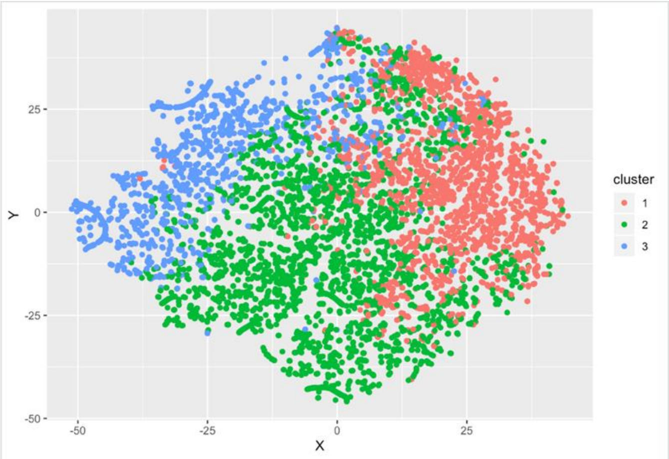
InterestingLife	Energetic	WorkEnthusiastic	Workaholic	JobSatisfaction
All of the time : 474	Always : 506	Always : 791	Always : 792	Always :2200
Most of the time :2370	Most of the time:2695	Most of the time:2068	Most of the time:1988	Most of the time:1403
More than half of the time: 562	Sometimes : 465	Sometimes : 641	Sometimes : 765	Sometimes : 149
Less than half of the time: 187	Rarely : 87	Rarely : 206	Rarely : 170	Rarely : 19
Some of the time : 155	Never : 32	Never : 79	Never : 70	Never : 14
At no time : 37				



Cluster 3

Gender	Age	Cheerful	Calm	Active	Fresh
Male :739	Min. :17.00	All of the time :1174	All of the time :1061	All of the time :1219	All of the time :975
Female:997	1st Qu.:31.00	Most of the time : 287	Most of the time : 264	Most of the time : 247	Most of the time :318
	Median :40.00	More than half of the time: 103	More than half of the time: 121	More than half of the time: 88	More than half of the time:126
	Mean :40.57	Less than half of the time: 62	Less than half of the time: 93	Less than half of the time: 54	Less than half of the time: 92
	3rd Qu.:49.25	Some of the time : 69	Some of the time : 120	Some of the time : 80	Some of the time :136
	Max. :82.00	At no time : 41	At no time : 77	At no time : 48	At no time : 89

InterestingLife	Energetic	WorkEnthusiastic	Workaholic	JobSatisfaction
All of the time :1232	Always :1146	Always :1175	Always :1098	Always :1459
Most of the time : 240	Most of the time: 372	Most of the time: 289	Most of the time: 294	Most of the time: 231
More than half of the time: 102	Sometimes : 137	Sometimes : 145	Sometimes : 250	Sometimes : 30
Less than half of the time: 51	Rarely : 47	Rarely : 63	Rarely : 54	Rarely : 5
Some of the time : 72	Never : 34	Never : 64	Never : 40	Never : 11
At no time : 39				



Cluster1: Female Cluster

The majority of individuals in cluster 1 are females with a median age of 47 years and more than half of the time they have been feeling Cheerful, Calm, Active, Fresh over the last two weeks including the things that interest their daily life. It seems individuals in this cluster are pretty satisfied with their job though they don't seem very workaholic, enthusiastic and energetic. We can call this cluster of individuals, satisfied people.

Cluster2: Male Cluster

People in this cluster are mainly males and this Male dominating cluster is happier and diligent than the Female dominating cluster. Most of the time individuals in this cluster feel Cheerful, Calm, Fresh in their routine life and they seem passionate about their work with a high level of job satisfaction. An interesting pattern to note is, this cluster has more number of individuals than any other one. We can call this cluster of individuals, happy and hard-working people.

Cluster3- Amalgam of Males and Females

This is the smallest cluster, with 1736 individuals in total out of which 739 are males and 997 are females. This cluster has a better male to female ratio than the other two clusters. With a median age of 40, the majority of the participants in this cluster are in good mental health and enjoy their work. A whopping 1459 (84%) people are satisfied with their job all the time and over 1230 people (71%) consider their life interesting all the time whereas over 1150 (68%) people feel cheerful all the time. In each survey question, there are less than 90 people with a negative response to either mental health or work satisfaction related questions. These impressive numbers tell us that people in this cluster are generally happy with both their life and work with a higher degree of satisfaction.

Part2- Regression

In this part of the project, we have been provided with real-world data of secondary education of two Portuguese schools with an aim for us to predict variable G3, the final grade of the student, by extract high-level knowledge from the two datasets using multiple algorithm techniques such as Linear Regression models and Data Mining models(Decision Tree, Random Forest, etc). Both of the datasets contain student's performance in two subjects i.e. Mathematics and Portuguese. We have been asked to predict G3 without G1 and G2, 1st and 2nd-period grades, as there is a strong correlation that exists between the final and last year grades. Although G1 & G2 are strong predictors in evaluating a student’s final marks, there are other relevant features also such as failures, absences, whether the student is getting support from the school or not and so on. We will be further exploring this rich dataset by finding valuable information for predicting G3.

A summary of datasets:

The dataset includes 395 and 649 students samples in Mathematics and Portuguese subjects respectively by using school reports and questionnaires with a total of 33 features including the target variable i.e. G3. The input variable includes student's demographic, school-related, social and family information like sex, age, school support, absences, family size, mother's job, parent's status whether they are living together or apart, alcohol consumption, etc. By running different linear and non-linear algorithms we found that all features available in the dataset are not significant in predicting student's final grades in both the subjects.

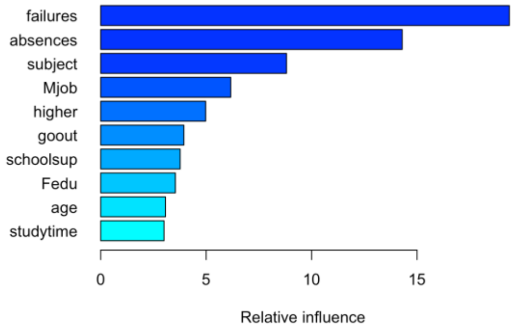
Before fitting any model, data preprocessing is very important such as checking missing value, standardization, etc. "Data preprocessing is that step in which the data gets transformed, or Encoded, to bring it to such a state that the machine can easily parse it" (Pandey 2019) [6]. I have performed below data Pre-processing steps before fitting the models. 1) Both subject files have been combined into one file and a new variable named "Subject" has been added with categories, Mathematics and Portuguese. 2) Continuous features like age, failures, and absences have been standardized as some machine learning algorithms are sensitive to the magnitude. 3) A few of the variables like Medu, Fedu, etc. have been converted as factors as they are categorical. In statistical modeling, implementing variables with the correct data type is very important. 4) Last, the dataset has been divided into training and test sets with a split of 75/25. We will be using the training dataset to train our model and test set will be used to evaluate model accuracy

After running Linear and Tree-based supervised learning algorithms, below model accuracy have been found on test data test:

Models	Train Error (rmse)	Test Error (rmse)
Linear Regression	3.135	3.4425
Hybrid Stepwise Selection	3.1991	3.4719
Lasso Regression	3.3166	3.3307
Decision Tree Regression	3.5068	3.136
Random Forest Regression	1.4519	3.0975
Gradient Boosting Machine	2.6446	3.0491

Conclusively, we can say Gradient Boosting has outperformed all other techniques and the best test dataset results were found with a root mean squared error i.e. RMSE of 3.04. RMSE is a standard way to measure the accuracy of the model. "It’s the square root of the average of squared differences between prediction and actual observation" (JJ, 2016) [5]. From the above table, it is clear that tree-based models i.e. Decision Tree, Random Forest & Gradient Boosting have better accuracy than the Linear models. "Tree-based algorithms empower predictive models with high accuracy, stability and ease of interpretation. Unlike linear models, they map non-linear relationships quite well" (Vidhya 2016) [9]. Boosting is a method of creating an ensemble. "Ensemble methods involve a group of predictive models to achieve a better accuracy and model stability"(Vidhya 2016) [9]. Ensembling starts by fitting a model, a tree in our case, to the data and then the next model is built based on the performance of the previous model. Let's say if the first model has performed poorly then the later will focus on improving the performance of the first one. In general, the combined accuracy of models is expected to perform better than the single model alone [4].

"Gradient boosting is a type of machine learning boosting" (Hoare) [4]. "It is a machine learning technique that builds an ensemble of shallow and weak successive trees with each tree learning and improving on the previous" (Github) [7]. It is based on the intuition that the best possible new model when combined with the previous model, minimizes the overall error and each next model takes a step in the direction to minimize the prediction error[4]. To improve the prediction accuracy, I have used a combination of parameters in GBM function and the lowest rmse was found when the algorithm uses 110 trees. I have plotted the top 10 variables chosen by the algorithm and past failures are by far the most important predictor in predicting the final grade of the student.



The key point to note is, subject, a new variable added in the algorithm, plays an important role in predicting final grades. Mathematics and Portuguese are two very different subjects, they can't be measured on the same scale when we talk about a student's interest in subjects. Student's final grades also depend upon how many times they have missed the classes, what is their aim in their career and the occupation of their parents.

Part3- Classification Models

In this part of the project, we are dealing with the data related to direct marketing campaigns of a Portuguese banking institution and our task is to predict whether the client will subscribe to a term deposit or not using Classification Regression and Data Mining techniques. Classification models such as Generalized Linear Model(GLM), K-Nearest Neighbours(k-NN), Support Vector Machines(SVM), Naive Bayes(NV) and Gradient Boosting Machine(GBM) have been used to classify customers who will and will not be willing to buy the term deposit.

A summary of the dataset:

There are 4521 records in the dataset which is composed of 17 features along with the binary outcome variable. Based on the information that the banks typically run these kinds of campaigns via phone calls, I am assuming the unknown category of contact categorical variable as cellular. Categorical variable 'duration' has been excluded from the dataset because when a client talks for a relatively long time, it is very likely that he/she may subscribe to the term deposit. In data pre-processing, pdays variable has been converted into a factor and bins have been created for better interpretability. For example, -1 has been categorized as "Not contacted" means the client wasn't contacted in the previous campaign and bins for the rest numeric variables have been created with an interval gap of 3 months.

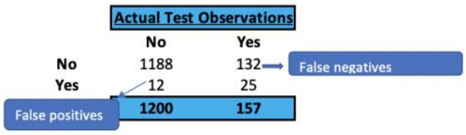
Train-Test Dataset Split:

From the data, we see that only 11.5% of the people have subscribed to the term deposit while the remaining 88.5% didn't. To handle the imbalanced classes, the training set has been created by randomly sampling 70% observations from 'yes' class and 70% observations from the 'no' class. The same strategy has been applied for the rest 30% test set split.

To model this dataset effectively, Linear and Non-Linear models were fitted and an appropriate model has been selected based on the Classification accuracy. The below table depicts each model classification accuracy on the test dataset and clearly, Gradient Boosting(GBM) has outperformed other techniques with a classification accuracy of 89.39%.

Models	**Classification Accuracy**
Generalized Linear Model	0.8887
Gradient Boosting Machine	0.8939
K-Nearest Neighbors	0.8917
Support Vector Machines	0.8931
Naïve Bayes	0.8519

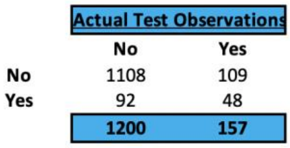
The GBM confusion matrix shows that out of 1357 observations of the test data, the model has correctly classified 1213 observations.



	Actual Test Observations	
	No	**Yes**
No	1188	132
Yes	12	25
	1200	**157**

False negatives
False positives

But since our dataset is not balanced, evaluating model performance through classification accuracy is not enough. False negatives are probably worse than False positives. What I mean by that is, predicting a customer won't subscribe to a term deposit when in reality the client subscribed for it is worse than predicting "yes" as the outcome and that customer ended up not subscribing. This trade-off would help the bank not miss out on potential subscribers. Out of 157 observations of "yes" class, 132 have been incorrectly classified by GBM. So we can see that even though the accuracy of GBM is highest, GBM is probably not the best for our use case. Considering this point in mind, I tried to model the problem differently and Naive Bayes performed the best. Here are the False positives and False negatives classification done by the Naive Bayes algorithm and here the ratio of the false negatives is low as compared to GBM.



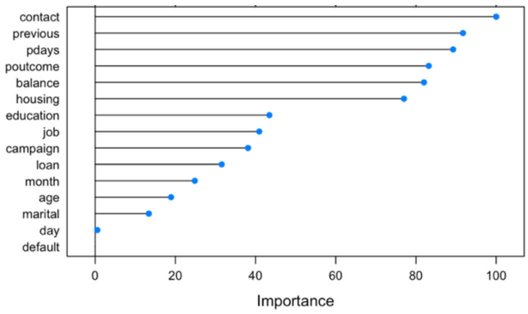
	Actual Test Observations	
	No	Yes
No	1108	109
Yes	92	48
	1200	157

I propose this model can help the bank in targeting potential customers with a better conversion rate.

Naive Bayes Algorithm

"Naive Bayes classifiers calculate the probability of a sample to be of a certain category, based on prior knowledge. They use the Naive Bayes Theorem, which assumes that the effect of a certain feature of a sample is independent of the other features" (Garcia, 2018) [3]. In the presence of continuous variables, this algorithm makes one more assumption that the observations have been sampled from Gaussian distribution [8]. Since continuous variables present in our dataset aren't following the normal distribution, this assumption of the algorithm doesn't fit here. To make it work, I have converted each continuous feature into categorical. Bins were created where each bin was provided with an approximately equal number of observations.

From the variable importance graph, we can see that “contact” is the most important variable according to the Naive Bayes algorithm. Intuitively, features based on the information of previous campaigns are helping the algorithm most in classifying classes correctly. How many times the client was contacted in the previous campaign, what was the outcome of the previous campaign, is the client maintaining a good bank balance or not, are some of the crucial features in classifying whether a client will subscribe to a term deposit or not.



Bibliography

[1] En.wikipedia.org. 2019. *Silhouette (Clustering)*. [online] Available at: <https://en.wikipedia.org/wiki/Silhouette_(clustering)> [Accessed 1 April 2020].

[2] Filaire, T., 2018. *Clustering On Mixed Type Data*. [online] Medium. Available at: <https://towardsdatascience.com/clustering-on-mixed-type-data-8bbd0a2569c3> [Accessed 1 April 2020].

[3] Garcia, S., 2018. *Easy And Quick Explanation: Naive Bayes Algorithm*. [online] Medium. Available at: <https://medium.com/@montjoile/easy-and-quick-explanation-naive-bayes-algorithm-99cb5f3f4e9c> [Accessed 1 April 2020].

[4] Hoare, J., n.d. Gradient Boosting Explained – The Coolest Kid on The Machine Learning Block. [Blog] *DISPLAYR*, Available at: <https://www.displayr.com/gradient-boosting-the-coolest-kid-on-the-machine-learning-block/> [Accessed 1 April 2020].

[5] Medium. 2016. *MAE And RMSE — Which Metric Is Better?*. [online] Available at: <https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d> [Accessed 1 April 2020].

[6] Pandey, P., 2019. *Data Preprocessing : Concepts*. [online] Medium. Available at: <https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825> [Accessed 1 April 2020].

[7] Uc-r.github.io. n.d. *Gradient Boosting Machines · UC Business Analytics R Programming Guide*. [online] Available at: <http://uc-r.github.io/gbm_regression> [Accessed 1 April 2020].

[8] Uc-r.github.io. n.d. *Naïve Bayes Classifier · UC Business Analytics R Programming Guide*. [online] Available at: <https://uc-r.github.io/naive_bayes> [Accessed 1 April 2020].

[9] Vidhya, A., 2016. *Tree Based Algorithms : A Complete Tutorial From Scratch (In R & Python)*. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2016/04/tree-based-algorithms-complete-tutorial-scratch-in-python/> [Accessed 1 April 2020].