Data Mining Report  
CP3403

**BY- ADITI AILAWADHI**

Contents

[Business scenario: 3](#_Toc71118540)

[Introduction: 3](#_Toc71118541)

[Data Description: 3](#_Toc71118542)

[Dataset Overview: 3](#_Toc71118543)

[Justification for the chosen dataset: 3](#_Toc71118544)

[Dataset Attributes: 4](#_Toc71118545)

[Pre-processing details: 4](#_Toc71118546)

[Justification for the Data Mining Approaches Chosen: 6](#_Toc71118547)

[Classification: 6](#_Toc71118548)

[Clustering: 7](#_Toc71118549)

[Association Rule Mining: 7](#_Toc71118550)

[Chosen Algorithms- 7](#_Toc71118551)

[Classification: 7](#_Toc71118552)

[ Decision Tree (J48): 7](#_Toc71118553)

[ Support vector machines (SVM): 7](#_Toc71118554)

[ Naïve Bayes: 7](#_Toc71118555)

[Clustering: 8](#_Toc71118556)

[ K-Means: 8](#_Toc71118557)

[Association rule mining: 8](#_Toc71118558)

[ Apriori algorithm: 8](#_Toc71118559)

[Discussions and comparison: 8](#_Toc71118560)

[Classification: 8](#_Toc71118561)

[Clustering: 9](#_Toc71118562)

[Association Rule Mining: 10](#_Toc71118563)

[Patterns and Findings: 11](#_Toc71118564)

[Conclusion and possible future work: 11](#_Toc71118565)

[References: 12](#_Toc71118566)

[Appendix: 12](#_Toc71118567)

[Appendix 1: Raw dataset Attributes 12](#_Toc71118568)

[Appendix 2: The implementation of ‘NominalToBinary’ algorithm 14](#_Toc71118569)

[Appendix 3: Redundant attributes 14](#_Toc71118570)

[Appendix 4: The Silhouette analysis for KMeans. 15](#_Toc71118571)

[Appendix 5: The best rules found by Association Rule Mining and visualisation of the association rules, produced using ‘associationRulesVisulaiser’. 15](#_Toc71118572)

# Business scenario:

We have been approached by a telecommunication company to find out for ways to boost the company’s customer retention and increase customer satisfaction in order to prevent customers from churning and moving to rival telecommunications corporations, which is considered as one of the obstacles that most businesses face.

# Introduction:

Based on the fact, it should be taken into consideration that it costs between 10 to 25 times more to find a new customer through marketing, outreach, campaigns, advertising, and the cost adds up, than it does to retain an existing one, hence, increasing long term company revenues.

According to the raw data provided to us, the platform has already lost over 26% of its customers, which is labelled to as churn (churn being 1869 whereas total instances being 7043). With such high attrition rates, it is assumed that the company is likely to lose majority of its customers only in a matter of several months, if no action is taken in a timely manner and we do not intervene; this is what the company is concerned about.

The purpose of this report was to use a variety of pre-processing strategies and conducting data mining testing in order to identify trends and patterns in the dataset as well as obtaining understanding of customer behaviour on a deeper level to determine which customers are the most likely to churn, which is crucial when making business decisions. We aim to help company to come up with and make some targeted decisions to diminish consumer churn rates, which will be achieved by data mining.

# Data Description:

### Dataset Overview:

The data used is collected from Kaggle Dataset – ‘Telco Customer Churn’. The dataset consists of 7043 instances, out of which 11 instances have missing values for TotalCharges attribute. The dataset comprises of 21 attributes, with most having nominal values, except for Tenure, MonthlyCharges and TotalCharges being numeric.

### Justification for the chosen dataset:

The chosen dataset is suitable for data mining purposes as it contains sufficient records of customers and it was believed that such a dataset would have several relevant novel and understandable patterns to be identified from, which may be useful to collect relevant information in order to take effective business actions in the future. The data will require various methods of pre-processing to be performed.

### Dataset Attributes:

The dataset attributes can be categorized into the following groups:

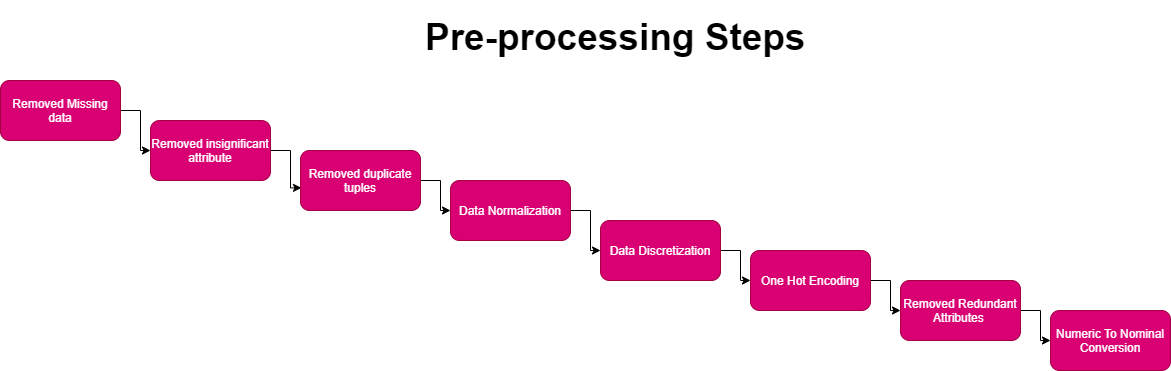
1. Demographic attributes about customers include - gender, SeniorCitizen, Partner, customerID and Dependents.
2. Telecom Services that were signed up by the customers during the tenure include – PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV and StreamingMovies.
3. The dataset contains Customer account information including - Tenure, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges and TotalCharges.
4. **Prediction Label - Churn:** Yes = Customers who left and moved on to another operator within the last month. No = the customer remained with the company.

Refer Appendix 1 for raw dataset Attributes with description.

# Pre-processing details:

It is a vital strategy in data mining in which the raw data is encoded and transformed into an effective and useful format in such a way that can be easily parsed by a model, as the data quality and the valuable patterns and knowledge that can be derived from it directly affects the ability of our model to learn.

The steps taken to pre-process the telecom company dataset are displayed in the following figure.



Once the dataset was collected, it was immediately run through WEKA and due to Weka being compatible with the .csv format file, insertion was successful.

The pre-processing algorithms provided by WEKA were applied before conducting classification, clustering, and association rule mining techniques on the dataset.

Firstly, it was noticed that the TotalCharges attribute contained 11 missing values were needed to be edited, which would have had an impact on the experiment's outcomes. The ‘RemoveWithValues' algorithm within WEKA was applied to remove the instances that had missing values in order to resolve this issue from the dataset, and the total number of instances reduced to 7032 with no missing values.

Secondly, attribute ‘customerId’ was removed that is a sequential key which does not contain any sematic meaning and along with it, duplicate instances were removed using ‘RemoveDuplicates’ algorithm with WEKA.

Thirdly, it was discovered that the values of the following numeric attributes: ‘tenure’, ‘MonthlyCharges’ and ‘TotalCharges’ were not distributed on the same scale which may lead to poor data models and clustering. Therefore, it was necessary to fix it to produce an unbiased and consistent baseline. This problem was fixed using ‘Normalise’ algorithm with WEKA to transform all the instances and rescale them in the dataset into a same range of [0,1] numeric values.

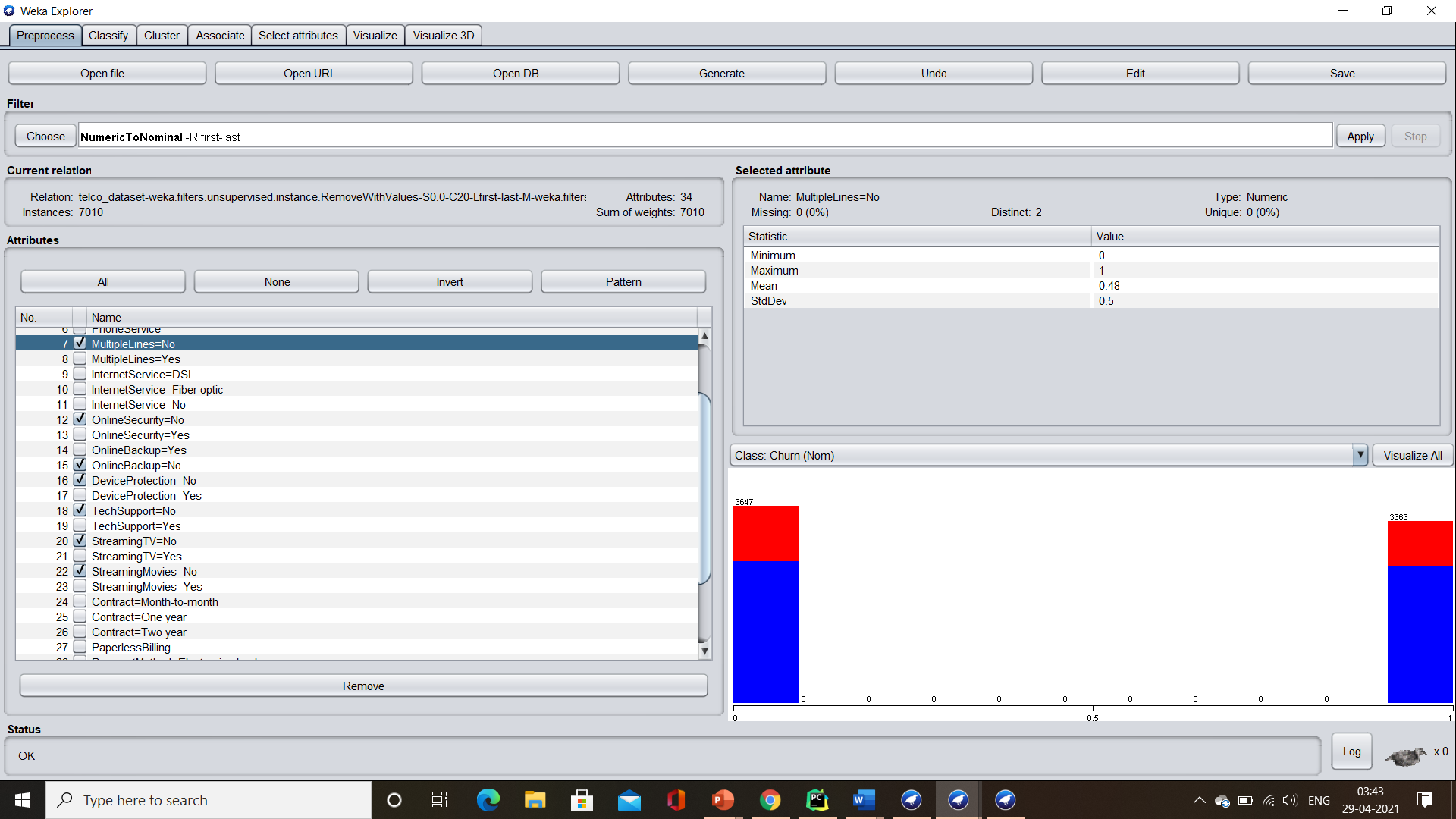
For the classification experiments to be performed, the following attributes: tenure, Monthly Charges and Total Charges were discretized which is a method in which numeric data is divided into intervals (bins) depending on the density of values surrounding it. This leads to knowledge-level representation of mining results that is easy-to-use and concise. The improvement of the signal-to-noise ratio is one justification behind discretization of continuous features. Fitting a model to bins diminishes the effect that minor data fluctuates, and noise has on the model. The fluctuations and noise in various sections of the data get "smoothed" out by each bin.

Whereas for the clustering purposes, as clustering is a distance metric and thus, it handles and works better for numerical dataset. Hence, the decision was taken to keep the following attributes: tenure, Monthly Charges and Total Charges as numeric type attributes only after they had been normalised.

The next issue discovered was that there were some polynomial types of categorical attributes containing more than two nominal values/ labels in the dataset. One hot encoding approach was employed within the dataset to resolve the issue. One hot encoding is a popular technique of generating additional attributes of every unique label in the category as dummy variables. It transforms categorical variables into several binary columns with the values of 0 and 1, where 1 indicates the presence of the row belonging to that category. Thus, categorical variables are transformed into such a format that machine learning algorithms can use to build a better model and do a better job in prediction. This was done due to the fact that many data mining algorithms does not work well with polynomial dataset and it is necessary to convert such attributes to binominal data and create models for nominal datasets that can yield better results. ‘NominalToBinary’ algorithm within WEKA was used for One hot encoding technique implementation on the dataset. It was applied on the following 10 attributes: ‘MultipleLines’, ‘InternetService’, ‘OnlineSecurity’, ‘OnlineBackup’, ‘DeviceProtection’, ‘TechSupport’, ‘StreamingTV’, ‘StreamingMovies’, ‘Contract’ and ‘PaymentMethod’. It was all done in one go due to attributeIndices. This brought the total number of attributes up to 41.

Refer Appendix 2 for the implementation of ‘NominalToBinary’ algorithm.

Following that, the new additional attributes created as a result of one hot encoding method became obsolete and redundant, such as No internet service with a count of 6, which were identical to attribute ‘InternetService=No’, and ‘MultipleLines=No phone service’ exactly opposite to ‘PhoneService’ attribute were all removed within Weka. Refer Appendix 3 for redundant datasets. Additionally, the following selected attributes were also removed which were opposite to attributes already present in the dataset reducing the count of total attributes to 27.



Finally, numeric data was converted to nominal data in order to maximise the results generated by the machine learning algorithms. ‘NumericToNominal’ algorithm was employed within the dataset to convert all numeric attributes to nominal. This was done due to the fact that many data mining algorithms, especially association rule mining and classification algorithms only works and produces models for the nominal dataset and are capable of producing better models when presented with nominal data.

# Justification for the Data Mining Approaches Chosen:

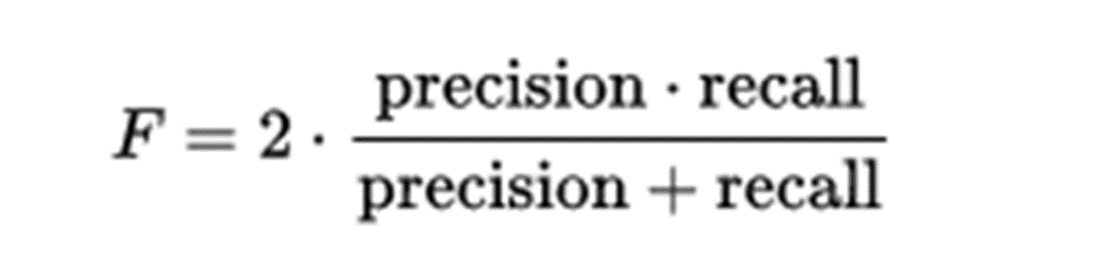
### Classification:

The concept of supervised learning comes into play when the algorithms are applied on the labelled training data, indicating the class of the observations to learn parameters of the model from the training data to make predictions for the new data. Classification was chosen as it gives the ability to view the false positive (FP), true positive (TP), the confusion matrices and several other statistical metrics that would be required in taking a business decision.

One of the goals of this report was to find the best classification classifier which is an algorithm that maps the data to a particular category and to decide what type of classification model would work best on the new unseen data to accurately classify the customers into groups (churn).

**Instead of using accuracy metric, as we have heavily imbalanced training dataset**, where the distribution is not equal across the classes and there is a class disparity, when observations of one class in the dataset are higher (majority class) than the other class (minority class), and we mostly care about the customers who will churn, the positive class (YES) which is a minority class, I considered using **F- measure** and Efficiency as the metrics that were analysed to form the basis of the critical analysis of the algorithms in evaluating the classification model. The higher the value of F- measure, the better the classification model will be for the new data.

Accuracy is used when the True Positives and True negatives are more important which just tells about the correctly classified values **which works well for evenly distributed datasets but** in our case, it can be problematic as the model will mostly predict all **observations belonging to the majority class**, being created on an **imbalanced dataset** but, in our case, the False Negatives and False Positives are crucial, that’s why F-Measure was used for evaluating the model where F-Measure is the harmonic mean of Precision and Recall.



A good F- measure indicates that we have low false negatives and false positives, implying that we are correctly detecting real threats while avoiding false alarms.

### Clustering:

For unsupervised learning with no predefined classes, Clustering was chosen to identify similarities in the dataset based on data characteristics and grouping similar data objects into multiple clusters in such a manner that the data points have high similarity within a cluster but are rather dissimilar to objects in other clusters.

In business intelligence, clustering plays an important role to facilitate the development of business strategies for enhanced and improved customer relationship management.

### Association Rule Mining:

Association rule mining is another business tool used that was examined for its analysis of patterns and trends within a dataset that occur frequently and is useful for making business decisions by discovering about what customer services offered by telecom company are being purchased together by the customer.

# Chosen Algorithms-

### Classification:

* Decision Tree (J48):It is one of the best machine learning algorithms to examine categorical and continuous data. In the field of machine learning, it is one of the most widely used supervised learning algorithms. It generates a decision tree.
* Support vector machines (SVM): They are linear classifiers based on the concept of structural risk minimization to improve the classifier's complexity with the aim of achieving excellent generalization performance.
* Naïve Bayes:It is one of the simplest and most efficient Classification algorithms for building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts based on the probability of an object.

### Clustering:

* K-Means:It is a tool for partitioning data into clusters. When k is defined, partitioning constructs several partitions, which are then evaluated using a k-clustering criterion. Each cluster is represented by the cluster's centre and objects are partitioned into k, the cluster's mean point. It also cannot handle nominal data and the number of clusters must be specified in advance.

### Association rule mining:

* Apriori algorithm:It is used for extracting item sets (groups of items) that occurring frequently in a database (frequent item sets). This data mining strategy iteratively achieves the most frequently occurring itemset by following the join and prune steps. The key goal of using it was to determine which of the telecom company's services were purchased together by consumers.

# Discussions and comparison:

### Classification:

To avoid the problem of overfitting, which is a modelling error when a function is closely fit to the dataset available and to avoid re-substitution error when a classifier is built on all the data available, and the same data is used as a test set which will make the model biased to the dataset and hence, such a model gives good performance on the training set, but the model is not a perfect fit for the new data and will produce wrong predictions for the new data set and gives poor generalisation and as a result has detrimental impact on its performance.

In order to construct an unbiased platform for the algorithms to perform, Cross validation and Splitting methods were deemed appropriate to be established on the available dataset to partition it into a training set / test set in which, a training set is implemented to build up a model using classification algorithms, while the test set is tuples used to validate the model built to test the F-measure and efficiency metrics.

The results of the classification tests are summarised in the table below. It demonstrates the differences in algorithms effectiveness with regards to the processing time.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Test** | **Algorithm** | **Validation**  **Type** | **Algorithm**  **Settings** | **Processing**  **Time (seconds)** | **F-Measure** | **PRC Area** |
| 1 | Decision Tree | Cross 10 Validation | defaults | 0.19 | 0.559 | 0.528 |
|  | SVM | Cross 10 Validation | Defaults | 4.7 | 0.518 | 0.438 |
|  | Naïve Bayes | Cross 10 Validation | defaults | 0.02 | 0.629 | 0.634 |
|  |  |  |  |  |  |  |
| 2 | Decision Tree | Cross 100 Validation | defaults | 0.08 | 0.555 | 0.538 |
|  | SVM | Cross 100 Validation | Defaults | 3.72 | 0.519 | 0.438 |
|  | Naïve Bayes | Cross 100 Validation | defaults | 0 | 0.629 | 0.634 |
|  |  |  |  |  |  |  |
| 3 | Decision Tree | Split 70% | defaults | 0.05 | 0.562 | 0.526 |
|  | SVM | Split 70% | Defaults | 0.93 | 0.503 | 0.437 |
|  | Naïve Bayes | Split 70% | defaults | 0.01 | 0.631 | 0.656 |

It can also be seen that there was a minimal shift in predicting F-Measure using cross validation with variation 10 folds and 100 folds regardless of all the classification Algorithms selected but the latter took much less processing time.

Another notable anomaly was that while both the Naïve Bayes and J48 Decision Tree algorithms produced results that were approaching 0 seconds (0.02 seconds), however the SVM algorithm produced results in an extremely contrasting elapse of time.

It is crystal cleared from the table that the SVM Algorithm took the longest to produce results for all the above tests as compared to other algorithms at an average of approximately 2.61 seconds over all the tests. Additionally, it can also be noticed that SVM Algorithm did not perform well in producing high F-Measure with an average of 0.510 over all the tests. Due to all these facts, SVM algorithm is recommended not to be used to classify telco customers dataset due to its inability of producing high F-Measure and not producing low false positives and low false negatives taking long processing times.

The Naïve bayes algorithm produces the most overall accurate results on the telecom customers dataset, yielding an estimated 0.631 F-Measure with a minimum time of 0.01 seconds on the test 3, with Naïve bayes being the fastest of the other two algorithms for all the tests conducted.

This makes Naïve bayes the best choice for quick computations for making business decisions.

### Clustering:

The number of clusters had to be defined in advance for the KMeans Algorithm. KValid package was installed from WEKA package manager which is a simple clustering evaluation package for clustering to choosing the optimal value of K. It clusters the instances using the SimpleKMeans algorithm as a backend and uses a variety of algorithms, including the Elbow and Silhouette-Index methods to evaluate the optimal number of clusters, K- value.

Except for the minimum K, which was set to 2, KValid was performed on the default values, using the Euclidean metric as the distance function and Silhouette-Index as the validation method. The Silhouette-Index plot displays a measure of how close a point is to its own cluster compared to other clusters. The location of the maximum is considered as the appropriate number of clusters which came out to be 3 in our case.

Refer Appendix 3 for the Silhouette analysis for KMeans. The following table shows the results from the KMeans test.

|  |  |
| --- | --- |
|  | **K= 3** |
| **Cluster 0** | **2107** |
| **Cluster 1** | **2227** |
| **Cluster 2** | **2676** |

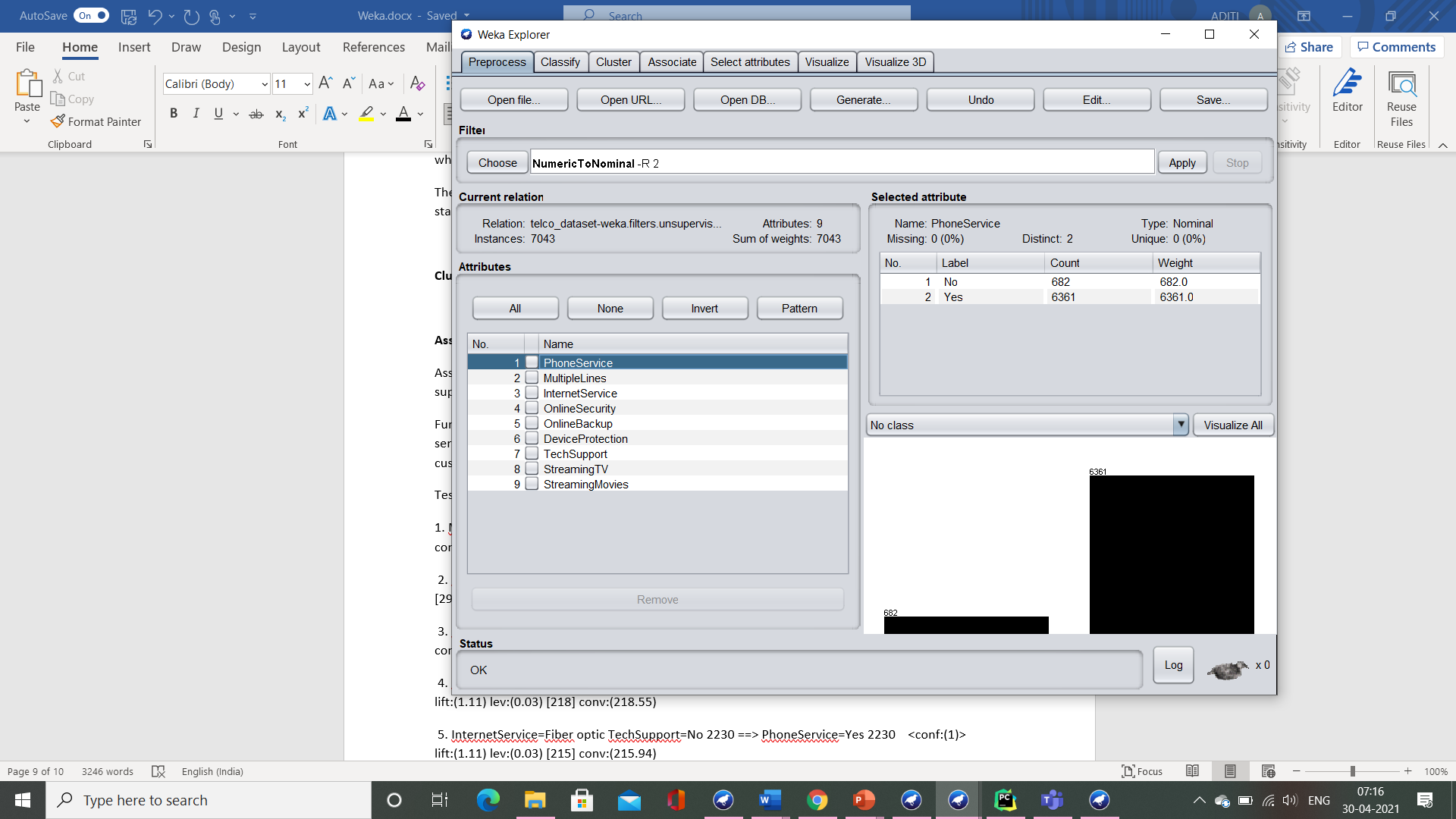
The results produced on applying KMeans to the dataset with k value set to 3 and the class attribute (Churn) selected to be an ignored attribute so that similarity is measured only upon other normal attributes, were very confusing and no pattern was found but a minimal increase in classification results were noticed when a new cluster attribute was created and added as a new column through the use of the Add Cluster filter in the Pre- process.

### Association Rule Mining:

Association rule mining parameters are support and confidence which are needed to be defined and are used to find out the probable associations and create the association rules.

Numeric attributes, ‘tenure’, ‘MonthyCharges’ and ‘TotalCharges’ were removed from the dataset as they are deemed incompatible or inappropriate with the algorithm as ARM is only compatible with nominal data. Additionally, all attributes except for the ones representing a service provided by the telecom company were removed from the dataset as the main purpose of this algorithm is to study about what services provided by the telecom company are purchased together by the customers.

Refer the figure below to check the 9 attributes on which ARM algorithm was performed.



with the maximum number of rules (numRules) set to 20, ARM was run on the default settings to observe what information could be extracted from the data set.

Refer Appendix 5 to check the best rules found and visualisation of the association rules, produced using ‘associationRulesVisulaiser’.

All of the results generated, yielded some intriguing findings regarding the services that the customers purchased in conjunction with the telecom company's Phone Service such as Fiber optic Internet Service, Multiple Lines, Streaming Movies and Streaming TV which may contribute to business decisions in unexplored and new areas, possibly opening up new opportunities and possibilities for growth and expansion. Surprisingly, Association rule mining produced all the rules with the confidence level of 100% accuracy. Some of the interesting rules and finding from the best rules found through ARM are as follows:

InternetService=Fiber optic ==> PhoneService=Yes <conf:(1)>

MultipleLines=Yes ==> PhoneService=Yes <conf:(1)>

MultipleLines=Yes InternetService=Fiber optic ==> PhoneService=Yes <conf:(1)>

InternetService=Fiber optic StreamingMovies=Yes ==> PhoneService=Yes <conf:(1)>

InternetService=Fiber optic StreamingTV=Yes ==> PhoneService=Yes <conf:(1)>

MultipleLines=Yes StreamingMovies=Yes ==> PhoneService=Yes <conf:(1)>

MultipleLines=Yes StreamingTV=Yes ==> PhoneService=Yes <conf:(1)>

# Patterns and Findings:

From the ‘Visualize All’ within Weka, the useful patterns found are as follows:

* **Monthly contract subscribers tend to churn more when compared to the customers with one year or two years contract with the telecommunication company, however, it should be noted that most of the customers opt for monthly contract over yearly subscriptions, resulting in high churn rate.**
* **The percent of churn is lower for the consumers with a longer tenure as compared to the freshly attracted customers, where tenure refers to the number of months the customer has stayed with the company.**
* **Churn rate is much higher in customers purchasing Fiber Optic Internet Services indicating that the particular service is needed to be improved.**

# Conclusion and possible future work:

Customer churn prevention is a crucial business tool. Ideally, the issue of unsatisfied customers could be nipped in the bud and keep the revenue flowing by building a model to predict customer churn with data mining algorithms.

Model predicting the amount of customer churn, which is predictive enough to inform company policy and point to potential ways to retain customers by finding in advance about a specific customer who is likely to churn, understanding the reason behind churn numbers and tackle those factors, with **reactive action plans** in time to prevent it from happening**.**

Based on the fact, it should be taken into consideration that it costs between 5 times and 25 times as much to find a new customer by Marketing, ads, campaigns, and outreach — the cost adds up than it does to retain an existing one, hence, increasing company profits in long term. This could be done through discounts, improving service (upping their internet speed, for example), or offering perks. Bundle of options could be created from the Association rule mining rules, the services not much purchased by the customers should be offered to purchase at low cost with improved service quality along with the other services purchased together by the customers in order to increase retention rate of the existing customers, hence increasing tenure rate yet allow the company to keep more customers than one-time interventions, which will make the customers more reluctant to switch by exploring the new services at cheaper rates.

If the goal is to engage and reach out to the customers to prevent them from churning, it is appropriate to engage with those that have been wrongly labelled as ‘not churned,’ (FNR) as it has no negative consequences, where FNR is the probability that a true positive will be missed by the test. It could potentially make them even more satisfied with the customer care service. This is the type of model that can add value from day one if the right actions are taken out of meaningful information it generates.

This way, company can act to keep its customers with a compelling deal and offer to prevent them from bailing out.

# References:

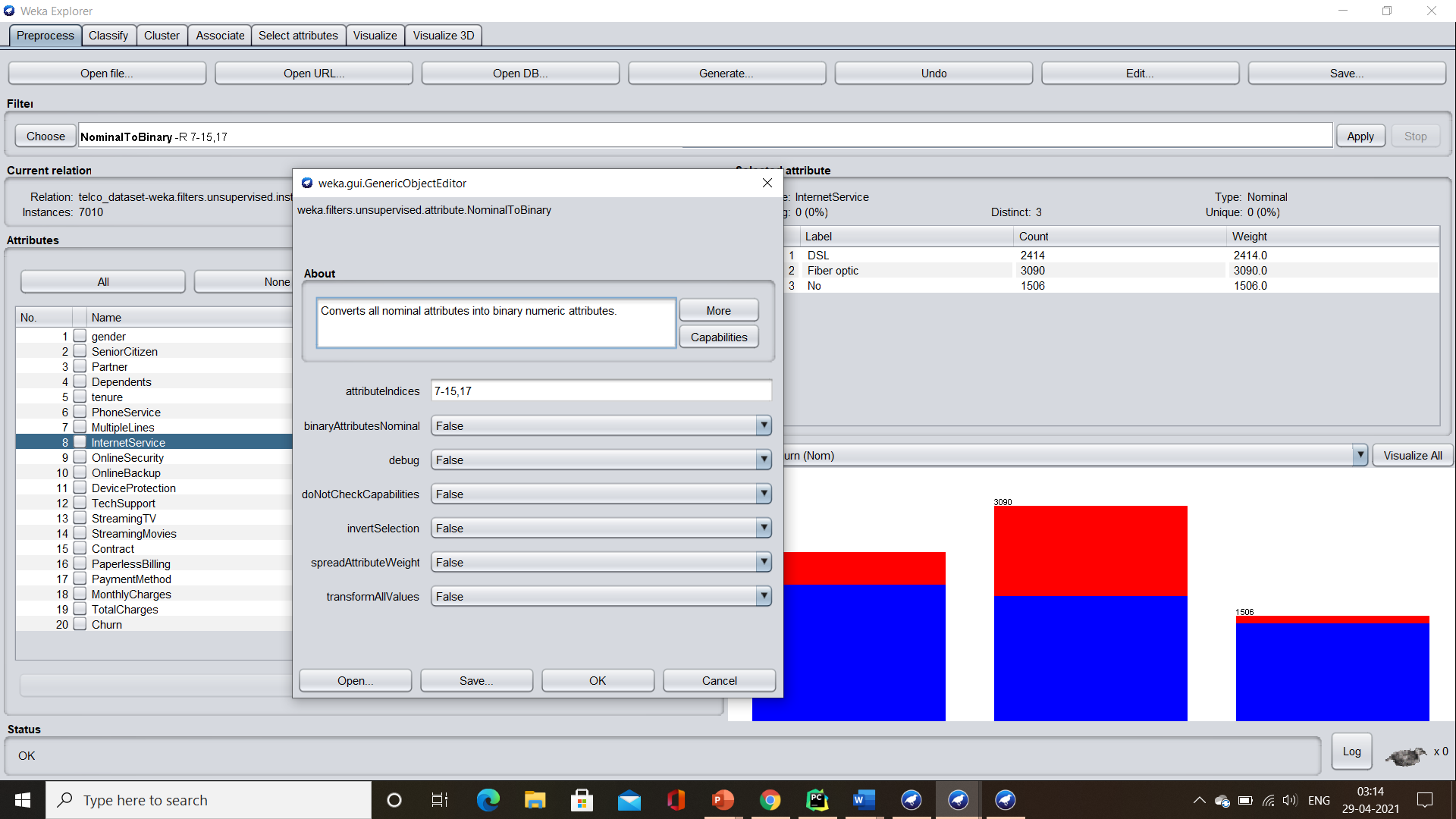
Telco Customer Churn. (n.d.). Kaggle, Retrieved April 10, 2021, from: <https://www.kaggle.com/blastchar/telco-customer-churn>

# Appendix:

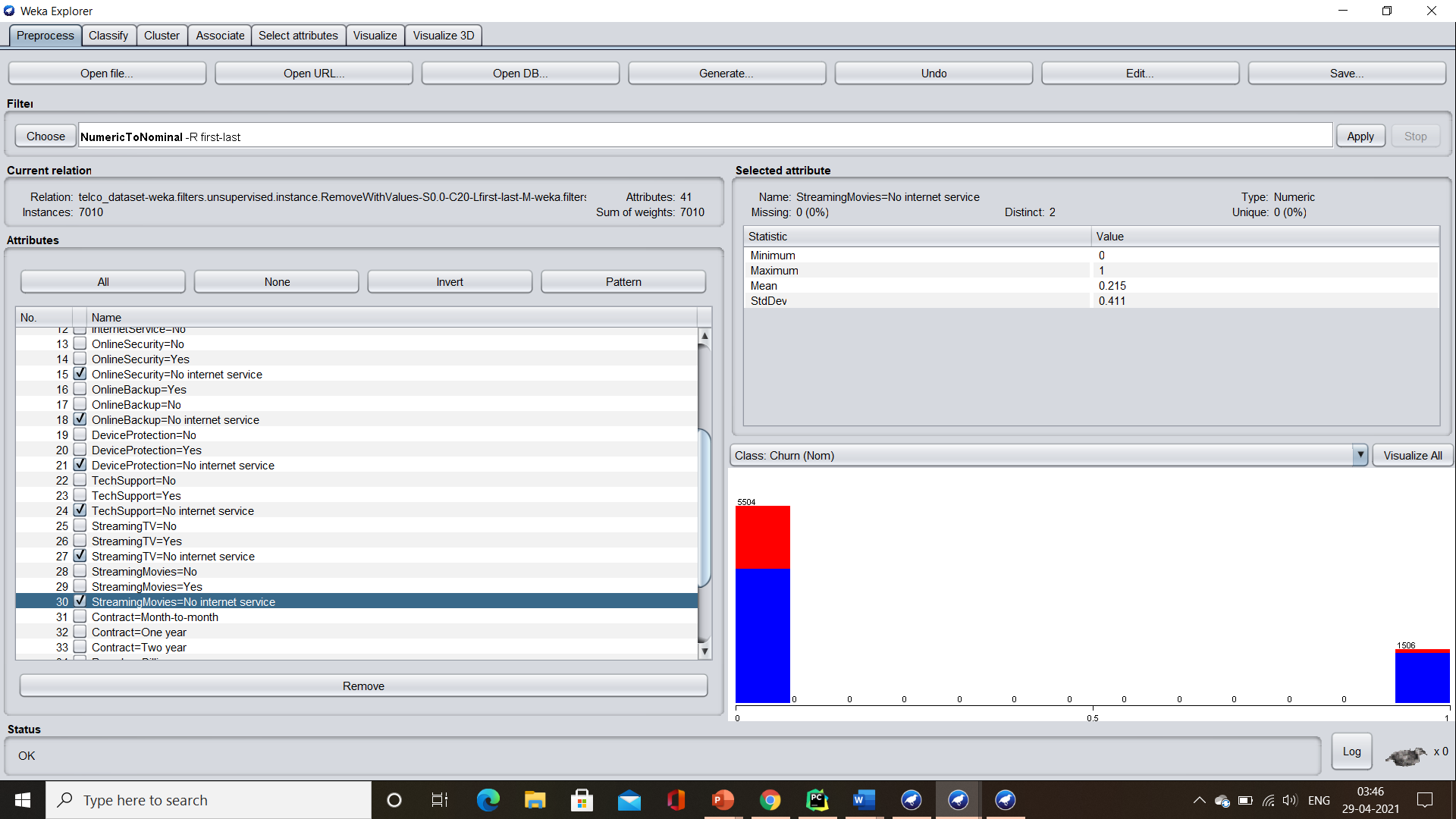
### Appendix 1: Raw dataset Attributes

|  |  |  |
| --- | --- | --- |
| **No.** | **Attribute** | **Description** |
| 1 | customerID | Customer identification number |
| 2 | gender | Customer gender (Female, Male) |
| 3 | SeniorCitizen | 1 = Yes; 0 = No (customer is a senior citizen or not) |
| 4 | Partner | Customer has a partner or not (Yes, No) |
| 5 | Dependents | Customer has dependents or not (Yes, No) |
| 6 | tenure | Number of months the customer has stayed with the company |
| 7 | PhoneService | Whether the customer has a phone service or not (Yes, No) |
| 8 | MultipleLines | Are multiple lines purchased by the customer?  Label 1: No phone service (same as No. 7 attribute)  Label 2: No  Label 3: Yes |
| 9 | InternetService | Which internet service provider is purchased by the customer-  Label 1: DSL  Label 2: Fiber optic  Label 3: No (Customer has not subscribed to internet service) |
| 10 | OnlineSecurity | Whether the customer has online security or not  Label 1: Yes  Label 2: No  Label 3: No Internet service (same as No. 9 attribute’s label 3) |
| 11 | OnlineBackup | Whether the customer has online backup or not  Label 1: Yes  Label 2: No  Label 3: No Internet service (same as No. 9 attribute’s label 3) |
| 12 | DeviceProtection | Whether the customer has device protection or not  Label 1: Yes  Label 2: No  Label 3: No Internet service (same as No. 9 attribute’s label 3) |
| 13 | TechSupport | Whether the customer has tech support or not  Label 1: Yes  Label 2: No  Label 3: No Internet service (same as No. 9 attribute’s label 3) |
| 14 | StreamingTV | Whether the customer has streaming TV service purchased or not  Label 1: Yes  Label 2: No  Label 3: No Internet service (same as No. 9 attribute’s label 3) |
| 15 | StreamingMovies | Whether the customer has streaming Movie service purchased or not  Label 1: Yes  Label 2: No  Label 3: No Internet service (same as No. 9 attribute’s label 3) |
| 16 | Contract | The contract term of the customer-  Label 1: Month-to-Month  Label 2: One year  Label 3: Two year |
| 17 | PaperlessBilling | Whether the customer has paperless billing or not (Yes, No) |
| 18 | PaymentMethod | The customer’s payment method-  Label 1: Electronic check  Label 2: Mailed check  Label 3: Bank transfer (automatic)  Label 4: Credit card (automatic) |
| 19 | MonthlyCharges | The amount charged to the customer monthly |
| 20 | TotalCharges | The total amount charged to the customer |
| 21 | Churn | Whether the customer churned or not (Yes or No) |

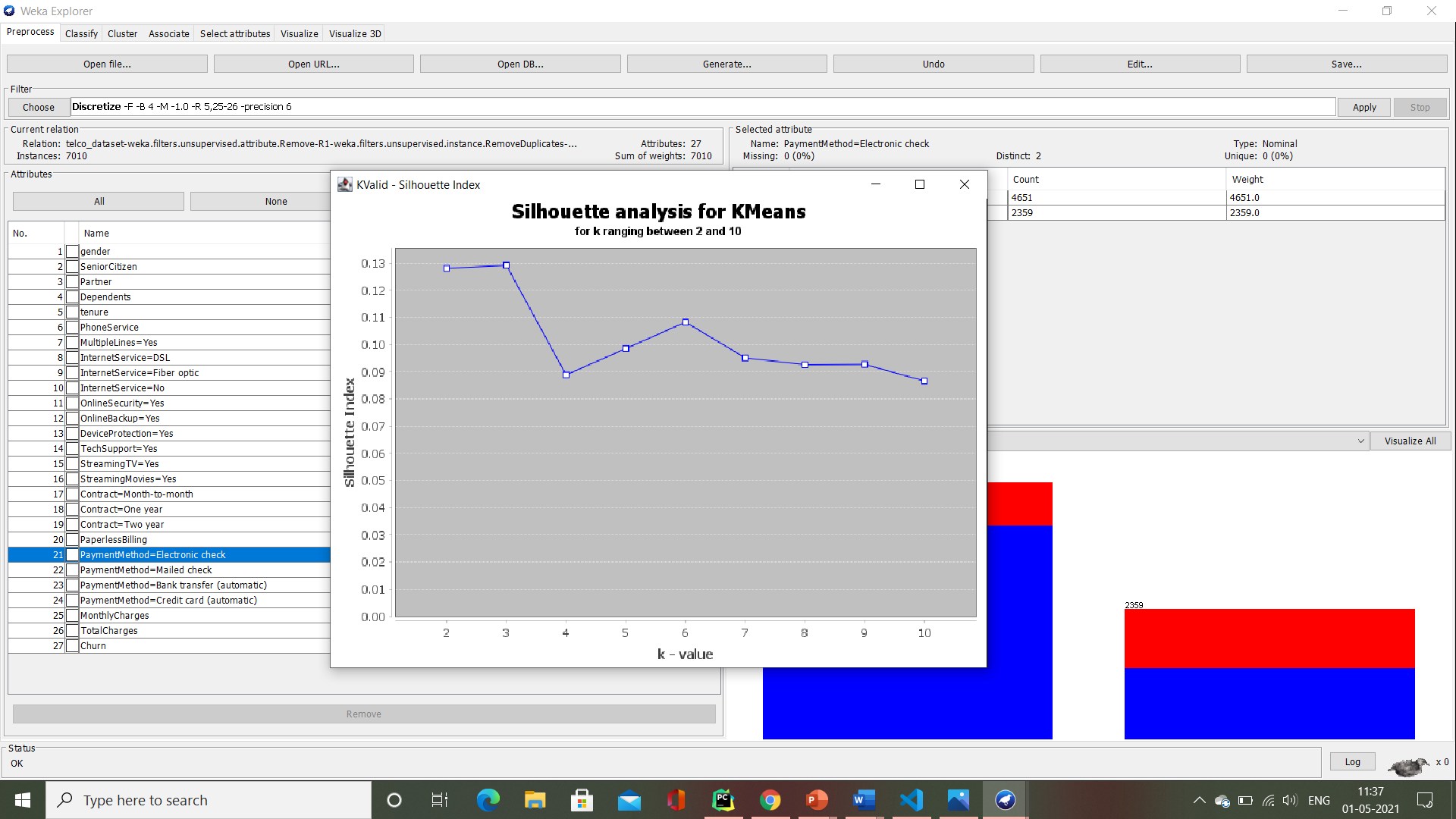
### Appendix 2: The implementation of ‘NominalToBinary’ algorithm



### Appendix 3: Redundant attributes



### Appendix 4: The Silhouette analysis for KMeans.



### Appendix 5: The best rules found by Association Rule Mining and visualisation of the association rules, produced using ‘associationRulesVisulaiser’.

