

Assessment Cover Sheet

Assessment	Project(Group)		
Assessment	Uncontrolled	Group	Not must-pass
Due Date	22-May 2024	Course Code	IT8416
Course Title	Data Mining		
Internal Moderator's	Dr. Shomona Jacob		
External Examiner's			

Instructions: Given in the Project description

1. This cover sheet must be completed (section in red below) and attached to your assessment before submission in hard copy/soft copy.
2. The time allowed for this assessment is till Wednesday, 22nd May 2024 by 11:55 p.m..
3. This assessment carries 100 marks with a weightage of 35% with 6 tasks assessing LO1,2,3
4. The materials allowed for use in this assessment are Moodle, Instructor informed website resources.
5. The **use of generative AI tools is strictly prohibited.**
6. References consulted (if any) must be properly acknowledged and cited.
7. The assessment has a total of 5 pages.

Learner ID	202002219 – 202000187 - 202001012	Date Submitted	May 2024
Learner Name	Hind Busandal – Fatema Ali – Zainab Abbass		
Programme	Bachelor of Information and Communication Technology – Database Systems		
Programme	BICT -Database		
Lecturer's	Sini Raj Pulari		

By submitting this assessment for marking, I affirm that this assessment is my own work.

	Learner	Sini Raj Pulari
Do not write beyond this line. For assessor use		
Assessor's Name	Sini Raj Pulari	
Marking Date		Marks Obtained

Comments:

Table of Contents

Task 1 - Problem Statement Formulation and definition	3
- Motivation	3
- Objectives	3
- Problem Statement / Project Definition	3
- Expected Result.....	3
Task 2 - Selection of an Appropriate Data Set (Data Collection)	4
- Source and Selection.....	4
- Selection of a suitable Data Set	4
- Essential information about the Data Set.....	4
- Dataset characteristics and Observations	6
Task 3 - Preparation and Pre-processing of Selected Data	9
- Preprocessing and preparing data	9
- Data Mining dimensionality reduction techniques.....	11
- Visualization After Data Preparation	14
- Principal component analysis (PCA)	15
- Clustering	16
Task 4 - Building Data Mining Models	17
- Models	17
Task 5 - Evaluating Data Mining Models	23
- Models Evaluation	23
Task 6 - Inferences, Recommendation and Reflection.....	26
- Inferences	26
- Recommendation.....	26
- Reflection	26
- Issues faced during the project.....	27
Extra Project Details	28
- Log files	28
- Video Demonstration.....	29
- GitHub	29

Task 1 - Problem Statement Formulation and definition

- Motivation

Social media platforms have become such a stable part that is integrated with our lifestyle as individuals and as a society. Understanding trends, influencing culture, and online businesses can help in many aspects including marketing and promotional purposes, maintaining brand and target audience, predicting future trends, and recognize dynamics of users and over all online communities.

- Objectives

Gain knowledge and understanding of concepts and techniques of Data Mining and apply what is learned throughout the course to get a full understanding of the impact of social media influencers. In addition, perform data mining techniques using the provided software R Studio or Rapid Minor and produce data results and visualization. And finally, evaluate numerical and graphical data from social media to recommend the best prediction model.

- Problem Statement / Project Definition

In this generation, social media is growing rapidly every day and creating more and more data that accumulates over the years. This great quantity of data can be utilized to study the changes and nature of social media influencing culture and how it impacts individuals, businesses, and the market. The problem at hand is selecting a data set that suits our purpose, preparing the data, building data mining models and evaluating them and analyze the results. The aim is to develop a data mining model to hopefully predict which people are influential in a social network.

- Expected Result

The successful implementation of the Data Mining models will help us:

- 1) Identify influential users.
- 2) Recognize factors that contribute to influence.
- 3) Evaluate model performance.

Task 2 - Selection of an Appropriate Data Set (Data Collection)

- Source and Selection

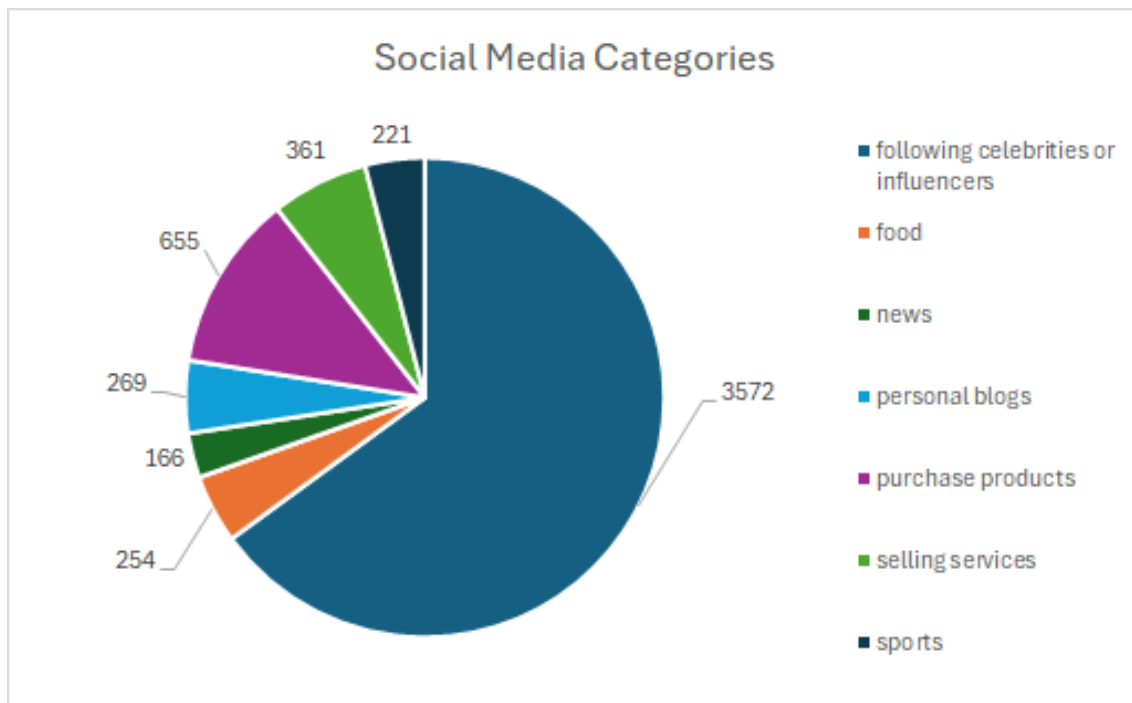
Source of Dataset

The project topic we used “Social Media Influencer Prediction” was chosen because social media is a huge part of this generation’s lifestyle, and a topic like this can help predict changes that affect different categories of people whether individuals, organizations, and society as a whole. The topic was chosen from an article provided in the course material and it came with its datasets: test, train, and sample predictions. Kaggle is one of the best sources of quality datasets that are well-studied and picked by scientists and researchers, so we expect the data to be reliable and suitable for the Data Mining techniques we aim to perform. The dataset aligns with our topic and provides the data required to examine and detect who is the most influential user between users A and B. We chose the dataset “train” as it is used for building the Data Mining Models and training them.

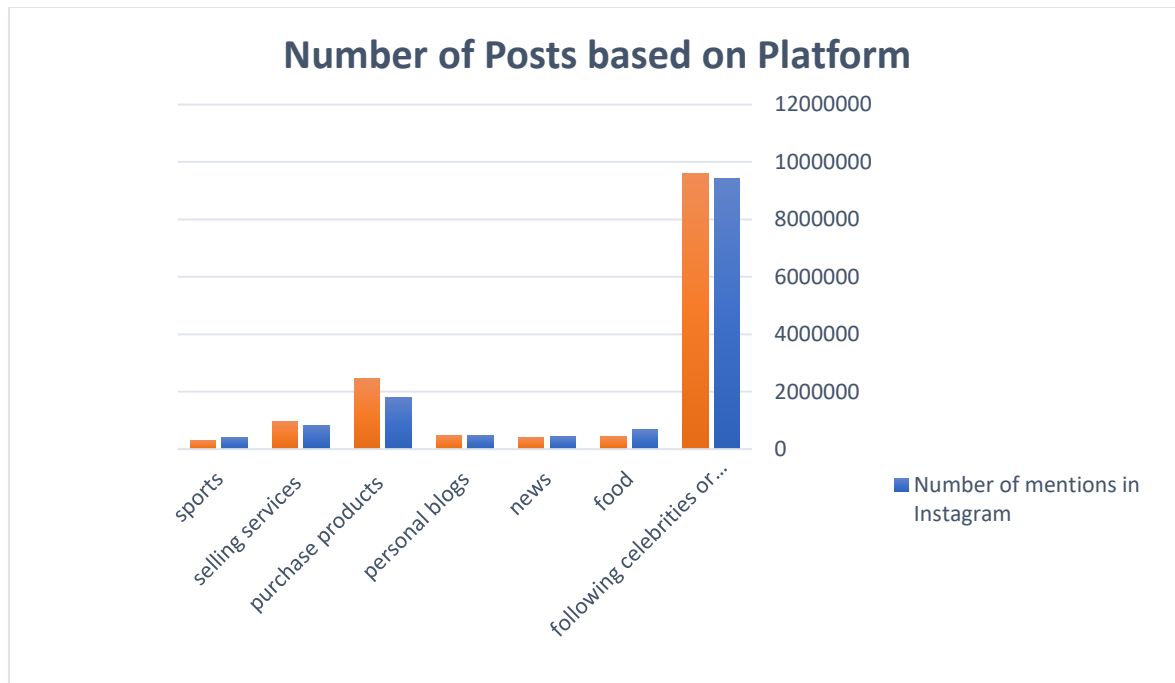
- Selection of a suitable Data Set

The data set selected contains information related to different categories of social media like following celebrities, sports, news and selling products and services. This data set includes different types of data and different formats related to number of posts, retweets, mentions in each platform like Instagram and Twitter. The sample contains 5500 records collected to be utilized for analysis purposes.

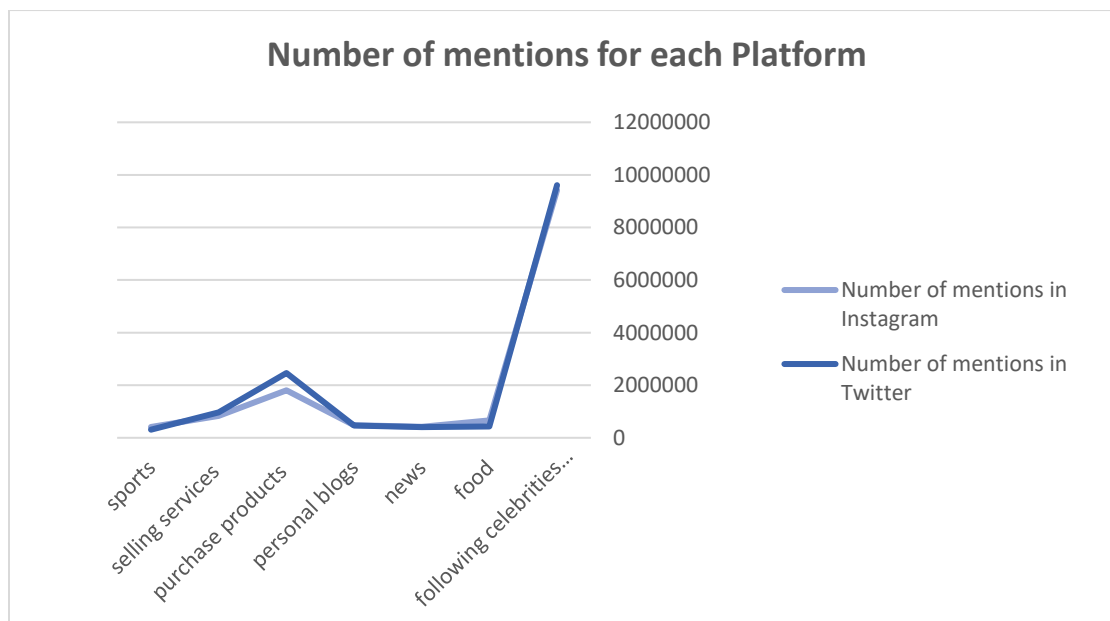
- Essential information about the Data Set



This Pie chart shows the number of posts interactions based on the category. It can be observed that most people in the study utilize social media to follow a celebrity or influencers, the second category people interact with is to purchase products and the third category is selling services. The least favorite category in social media is to watch the news.



This Chart illustrates the number of posts for each category on different platforms like Instagram and Twitter. It can be observed that celebrities and influencers have the highest number of posts on Instagram and Twitter and the least posts comes from the news. It can be noticed that people have more posts on Instagram than Twitter because people like to see photos, reels, and videos.



This line Graph demonstrates the number of mentions in Instagram and Twitter. It can be seen that the mention in Twitter has the highest compared to Instagram mentions especially in following celebrities and influencers while the least mentions related to news and sports.

- Dataset characteristics and Observations

Attributes

The dataset has 26 attributes:

- Choice
- A_follower_count
- A_following_count
- A_listed_count
- A_mentions_received
- A_retweets_received
- A_mentions_sent
- A_retweets_sent
- A_posts
- A_network_feature_1
- A_network_feature_2
- A_network_feature_3
- B_follower_count
- B_following_count
- B_listed_count
- B_mentions_received
- B_retweets_received
- B_mentions_sent
- B_retweets_sent
- B_posts
- B_network_feature_1
- B_network_feature_2
- B_network_feature_3
- Platform A
- Platform B
- Category

Name	Type	Missing	Filter (26 / 26 attributes)	Search for Attributes
Category	Nominal	2	Least news (166) Most followin [...] rs (3572)	
Platform A	Nominal	0	Least T (4) Most Twitter (5496)	
Choice	Integer	0	Min 0 Max 1	
A_follower_count	Integer	0	Min 16 Max 36543194	
A_following_count	Integer	0	Min 0 Max 1165830	
A_listed_count	Integer	0	Min 0 Max 549144	
A_mentions_received	Real	0	Min 0.101 Max 1145218.988	

Showing attributes 1 - 26 Examples: 5,500 Special Attributes: 0 Regular Attributes: 26

Example Instances


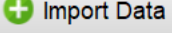
There are 5500 instances in the data set.

Row No.	Choice	A_follower_...	A_following...	A_listed_co...
5490	0	592	193	41
5491	1	2208080	240	39725
5492	0	708	631	26
5493	0	44	79	1
5494	0	268170	2087	7870
5495	1	25360	20872	461
5496	0	41765	185	1356
5497	1	112	243	5
5498	0	15385	673	747
5499	0	265258	209	551
5500	0	628	921	6


ExampleSet (5,500 examples, 0 special attributes, 23 regular attributes)




Missing Data

Using the filter option to only show attributes with missing data, the results came back with 11 missing attributes.

Filter (0 / 26 attributes):  

☒ Show categorical attributes
☒ Show numeric attributes
☒ Show object attributes
☒ Show only attributes with missing values
☒ Show special attributes
☒ Show regular attributes

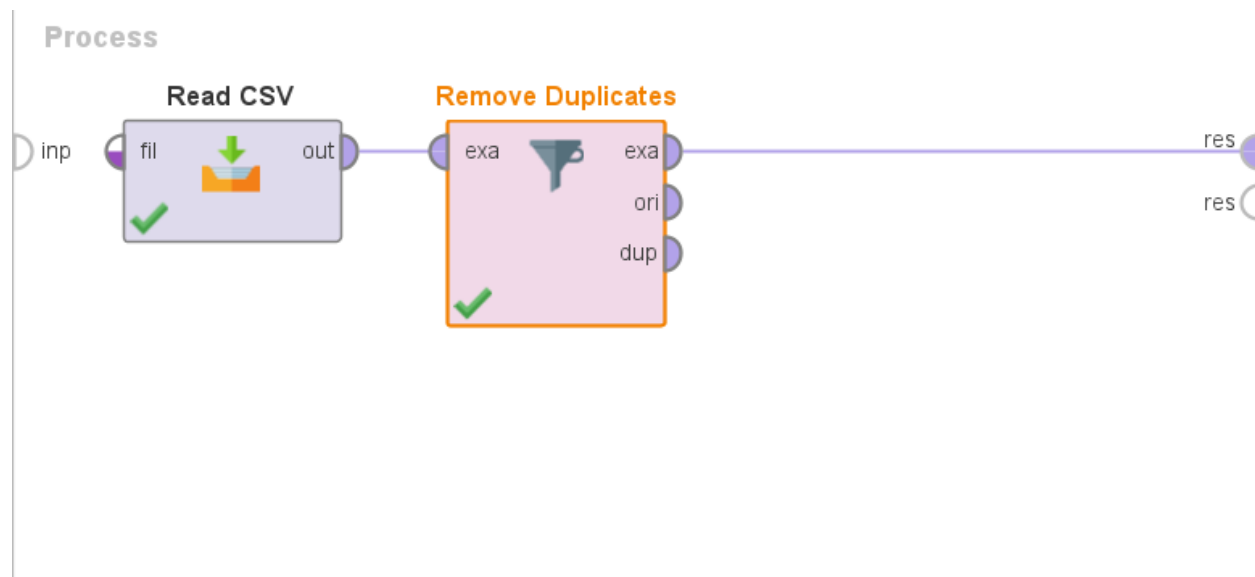
ExampleSet (Read CSV) 

Open in  Turbo Prep  Auto Model  Interactive Analysis Filter: 11 / 5,500 examples: missing_attributes ▼

Row No.	Category	Platform A	Choice	A_follower_...	A_following...	A_listed_co...	A_mentions...	A_retweets...	A_mentions...	A_retweets...
17	following cel...	Twitter	0	4760	425	96	5.012	1.408	0.101	0.101
24	following cel...	Twitter	0	127	36	2	0.101	0.101	0.101	0.101

Redundant values/ attributes

Using the 'Remove Duplicates' operator and choosing all attributes in the data set.



The data set shows 5441 examples which means that there are **59 duplicate values** that count as redundant data.

Turbo Prep

Interactive Analysis

Filter (5,441 / 5,441 examples):

Row No.	Category	Platfor...	Choice	A_follo...	A_follo...
1	food	Twitter	0	228	302
2	sports	Twitter	0	21591	1179
3	following...	T	0	7310	1215
4	purchas...	T	0	20	7
5	selling s...	T	1	45589	862
6	news	T	0	285735	276251
7	personal...	Twitter	0	285735	276251
8	following...	Twitter	1	9512	12
9	following...	Twitter	1	2273871	4524
10	following...	Twitter	0	182598	1402
11	following...	Twitter	0	3200	3256

ExampleSet (5.441 examples, 0 special attributes, 26 regular attributes)

Outliers

Any data collection whose values deviate from the usual range is called an outlier. Human mistakes might be the reason it appears. For example, if we have a range number of posts on social media.

Number of posts: {2,15,3,5,6,66,9}

Since number 66's is much higher than the other number of posts in the group, it is seen as an outlier. If an outlier seriously affects the statistical analysis, it may be eliminated from the data set. In this dataset, we found that there are 10 outlier instances, as shown below using the detect outlier operator.

<div>Outlier</div> <div>outlier</div>	<div>Binominal</div> <div>0</div>	<div>Negative</div> <div>false</div>	<div>Positive</div> <div>true</div>	<div>Values</div> <div>false (5490), true (10)</div>
---------------------------------------	-----------------------------------	--------------------------------------	-------------------------------------	--

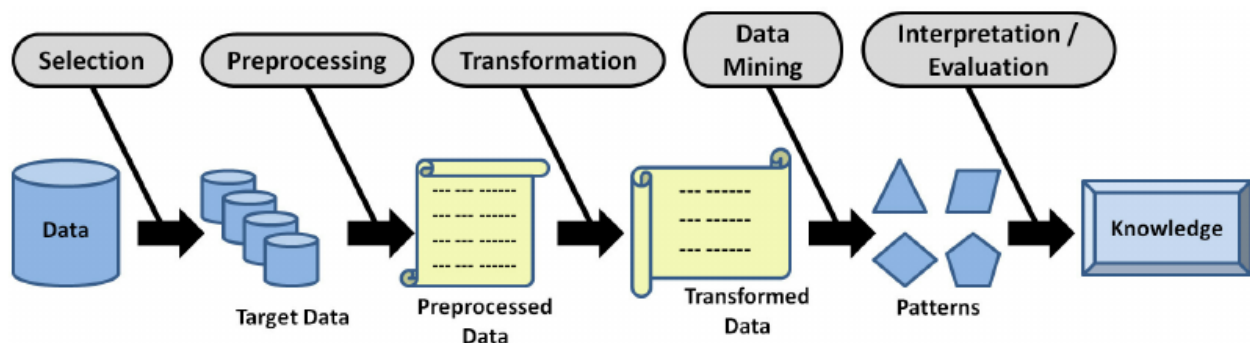
Task 3 - Preparation and Pre-processing of Selected Data

- Preprocessing and preparing data

A number of procedures known as "data pre-processing" are used to solve different problems that arise from unprocessed datasets. Missing values, superfluous or unnecessary features, outliers, and inconsistent data formats are a few examples of these problems. Pre-processing carefully addresses these issues in an effort to improve the dataset's general quality and integrity, setting the stage for studies that are more trustworthy and accurate.

The ideas of transformation and cleaning are essential to the concept of data pre-processing. Cleaning is the process of locating abnormalities in the dataset and fixing them so that inconsistent or incorrect data points are eliminated. In the meantime, transformation involves transforming the data into a consistent and logical format, which makes it easier to analyze different datasets together.

In order to reduce the possibility of incorrect conclusions and inadequate insights, our approach acknowledges the need of data pre-processing. Our goal is to maximize the value from our datasets while reducing the impact of data flaws by carefully preparing our data using pre-processing procedures.



Data Selection:

In this step, relevant statistics from social media sites are chosen, including user profiles, social connections (like followers, likes), engagement metrics (likes, shares, comments), and content qualities (like posting frequency, type of content). These datasets are essential to comprehending the social network's dynamics and structure.

Pre-processing:

Before being analyzed, the unprocessed social media data is cleaned, transformed, and integrated. This entails classifying categorical variables, eliminating duplication, and handling missing data. Stopword elimination, stemming, and tokenization may also be necessary for text data. Feature engineering is another aspect of pre-processing that is used to extract pertinent aspects including network priority, user involvement, and content characteristics.

Data Mining:

In order to find influential people within social networks, data mining techniques are used. Using machine learning algorithms like logistic regression, decision trees, random forests, or neural networks, this entails creating predictive models. The measure of impact (e.g., number of followers, engagement rate) is the target variable, and features retrieved during pre-processing are used as input variables.

Interpretation/Evaluation:

The predictive performance of the models is determined by interpreting and assessing the data mining process's output. The model's predictive power for prominent users is measured using metrics including accuracy, precision, recall, and F1-score. To verify the results and pinpoint the key elements influencing effect prediction, statistical significance tests could be run.

Knowledge Presentation:

Stakeholders are supplied with the data mining process's insights in an easily comprehended format. Visualizations that show the distribution of prominent users and their attributes, including bar charts, scatter plots, or heatmaps, may be used in this. To further enable stakeholders to examine the data and model outputs, interactive dashboards or reports may be developed.

Feedback:

The predictive models are improved and tuned by input from domain experts, stakeholders, and model performance indicators. During this iterative phase, the effect prediction model's accuracy and robustness may be improved by modifying model parameters, investigating new features, or utilizing alternative techniques.

Overall, the KDD diagram shows how to anticipate prominent users in a social network in a methodical manner. This method starts with data selection and pre-processing, moves through data mining and interpretation, presents information, and refines the predictive models based on feedback. The objective of comprehending and utilizing social media data for influencer prediction is advanced at each step of the procedure.

- Data Mining dimensionality reduction techniques



ExampleSet (Read CSV)

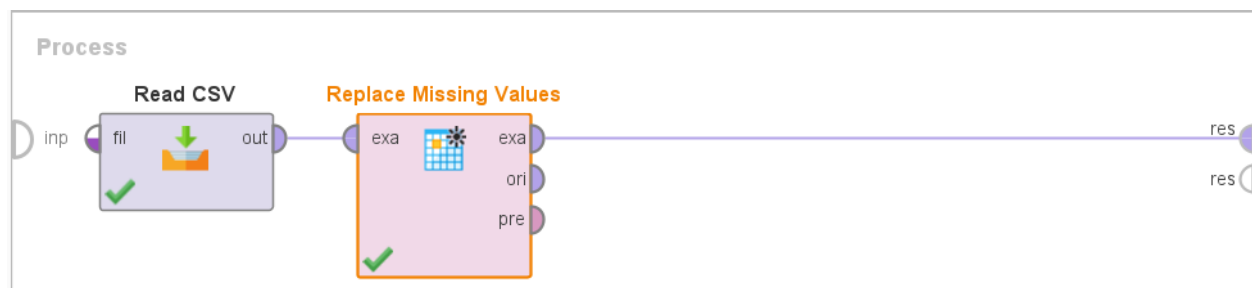
Open in [Turbo Prep](#) [Auto Model](#) [Interactive Analysis](#)

Filter (11 / 5,500 examples): [missing_attributes](#)

ExampleSet (Read CSV)

Name	Type	Missing	Statistics	Filter (26 / 26 attributes):	Search for Attributes
✓ A_retweets_sent	Real	0	0.101	16.291	1.110
✓ A_posts	Real	0	Min 0.101	Max 193.072	Average 9.091
✓ A_network_feature_1	Integer	4	Min 0	Max 920838	Average 5271.4
✓ A_network_feature_2	Real	5	Min 0	Max 1121	Average 84.86
✓ A_network_feature_3	Real	4	Min 0	Max 144651.462	Average 3749.4
✓ Platform B	Nominal	0	Least Instagram (5500)	Most Instagram (5500)	Values Instag
✓ B_follower_count	Integer	0	Min 20	Max 36543194	Average 68548

As you can see in our datasets, there are some missing values, and we need to fix them by using replace missing values operators in the image below.



Open in Turbo Prep Auto Model Interactive Analysis Filter (0 / 5,500 examples): missing_attributes

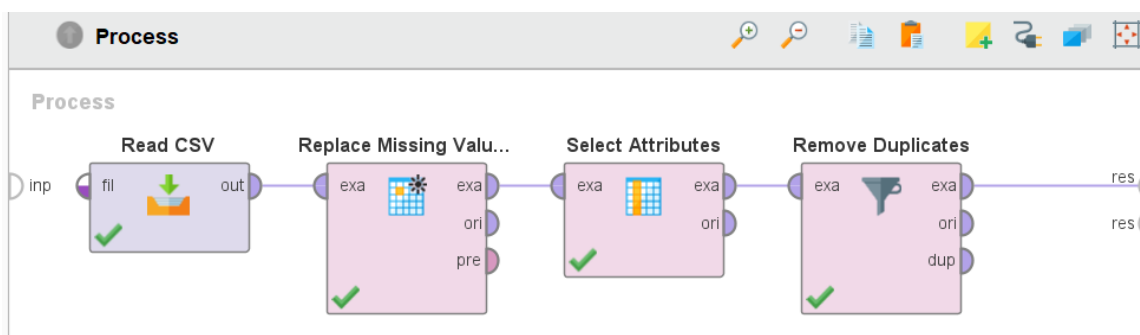
Row No.	Category	Platform...	Choice	A_follo...	A_follo...	A_liste...	A_ment...	A_retw...	A_ment...	A_retw...	A_posts	A_ne
---------	----------	-------------	--------	------------	------------	------------	-----------	-----------	-----------	-----------	---------	------

ExampleSet (Replace Missing Values)

Open in Turbo Prep Auto Model Interactive Analysis Filter (5,500 / 5,500 examples): all

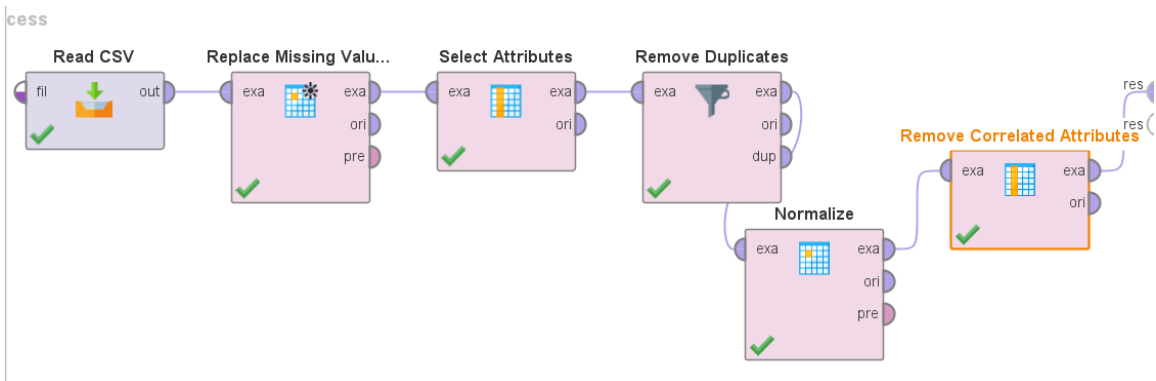
Row No.	Category	Platform A	Choice	A_follower_...	A_following...	A_listed_co...	A_mentions...	A_retweets...	A_...
1	food	Twitter	0	228	302	3	0.584	0.101	0.1
2	sports	Twitter	0	21591	1179	228	90.457	25.798	5.1
3	following cel...	Twitter	0	7310	1215	101	25.504	9.556	5.1
4	purchase pro...	Twitter	0	20	7	2	7.691	0.277	1.1
5	selling servic...	Twitter	1	45589	862	2641	148.854	36.999	27.1
6	news	Twitter	0	285735	276251	3417	19.328	7.292	0.1
7	personal blogs	Twitter	0	285735	276251	3417	19.328	7.292	0.1
8	following cel...	Twitter	1	9512	12	213	52.167	23.182	0.1
9	following cel...	Twitter	1	2273871	4524	11946	6782.405	2944.524	12.1
10	following cel...	Twitter	0	182598	1402	3831	145.845	74.003	23.1
11	following cel...	Twitter	0	3200	3256	146	0.290	0.101	0.1
12	following cel...	Twitter	0	3914	1439	165	2.404	1.162	0.1
13	following cel...	Twitter	1	23230	195	826	118.052	56.668	6.1

As you can see, in order to minimize the influence of missing values on the distribution of the data as a whole, we have substituted the column average for each missing value.

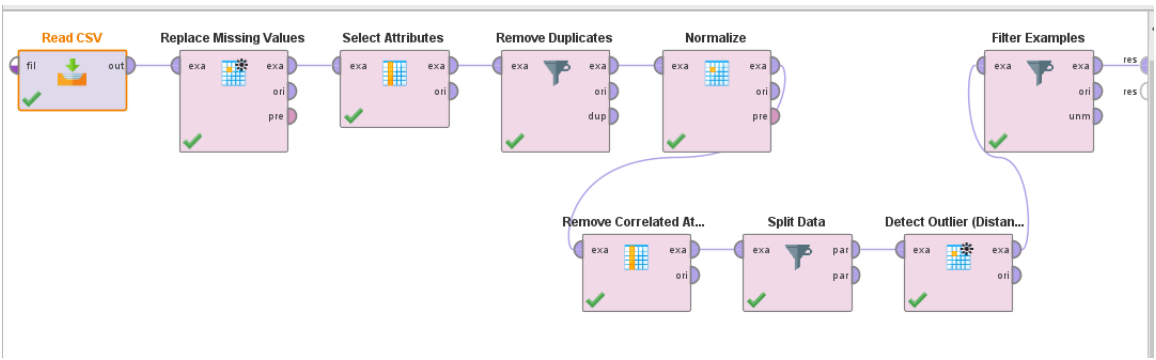


In order to decrease dimensionality and eliminate any unnecessary attributes that are unrelated to our research, we have implemented the "select attributes" operation. This allows us to choose which attribute is most significant. Furthermore, we have eliminated some qualities by manual means and by utilizing the "Remove duplicates" operator to minimize duplicate data that might adversely affect our research, as well

as the "Remove Correlated Attributes" operator, which will be demonstrated in subsequent figures. Including both operators improves the analysis's quality.




As any operation would take a long time to perform without normalizing, we can observe the usage of the Remove Correlated Attributes operator to lower the danger of overfitting and the normalization operator utilizing the Z-transformation approach to minimize the range of data.



Filter examples is the final operator in the model. Having this operator has the goal of returning a data collection clear of outliers. If the user chooses to return data if the outlier is false, it will offer an example set depending on that selection made from the filter.


Create Filters: filters


 Create Filters: filters
Defines the list of filters to apply.

outlier

equals

false





Outlier	Binomial		Negative	Positive	Values
outlier		0	false	true	false (3799), true (0)

As you can see, there are no outliers in this picture (0 true values), indicating that we utilized the detect outlier operator with Euclidian distance as the distance function. Outliers might represent errors in data entry, thus maintaining the quality of the analysis is important.

ExampleSet (Filter Examples)

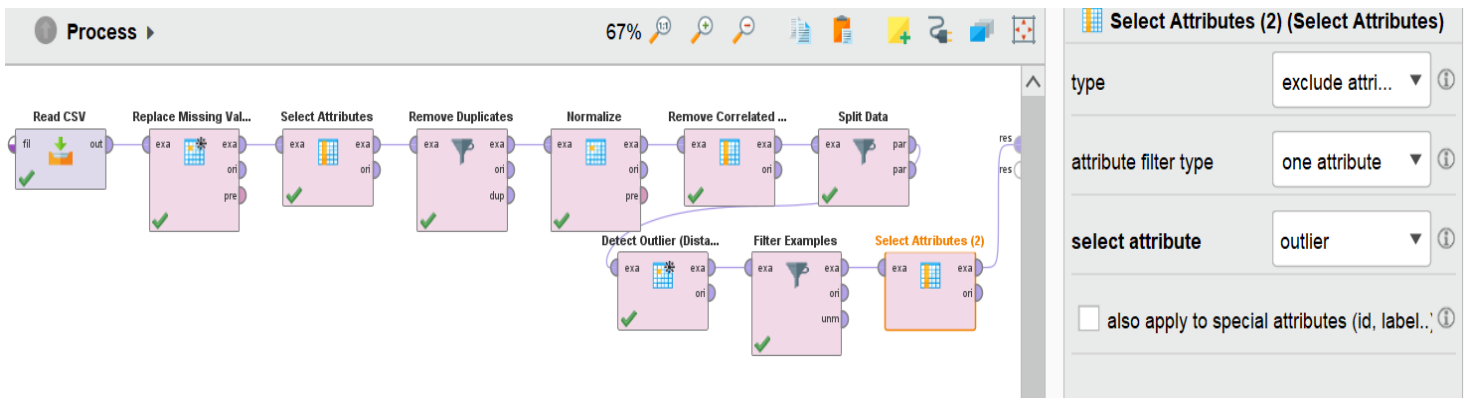
Open in Turbo Prep Auto Model Interactive Analysis

Filter (3,799 / 3,799 examples): all

Row No.	outlier	Choice	A_follower_...	A_following...	A_listed_co...	A_mentions...	A_mentions...	A_retweets...	A...
1	false	-1.021	-0.321	-0.252	-0.343	-0.092	-0.621	-0.529	-0
2	false	-1.021	-0.310	-0.234	-0.330	-0.088	-0.032	-0.002	-0
3	false	-1.021	-0.317	-0.234	-0.337	-0.091	-0.069	-0.273	-0
4	false	-1.021	-0.321	-0.258	-0.343	-0.091	-0.492	-0.529	-0
5	false	-1.021	-0.180	5.356	-0.147	-0.091	-0.621	-0.529	-0
6	false	0.979	-0.316	-0.258	-0.331	-0.090	-0.593	-0.529	-0
7	false	0.979	0.796	-0.166	0.343	0.140	0.728	0.874	3.
8	false	-1.021	-0.231	-0.230	-0.123	-0.087	1.840	-0.529	0.
9	false	-1.021	-0.319	-0.229	-0.333	-0.091	-0.621	-0.529	-0
10	false	0.979	-0.309	-0.248	-0.295	-0.088	0.037	-0.143	0.
11	false	-1.021	1.802	-0.253	1.538	0.150	-0.411	-0.405	0.
12	false	-1.021	-0.313	-0.245	-0.300	-0.090	0.098	-0.403	-0
13	false	0.979	-0.190	-0.254	-0.311	-0.070	-0.058	-0.529	-0

ExampleSet (3,799 examples: 1 special attribute 28 regular attributes)

- Visualization After Data Preparation



As we didn't need the outlier attribute, we've added the select attributes again to remove it.

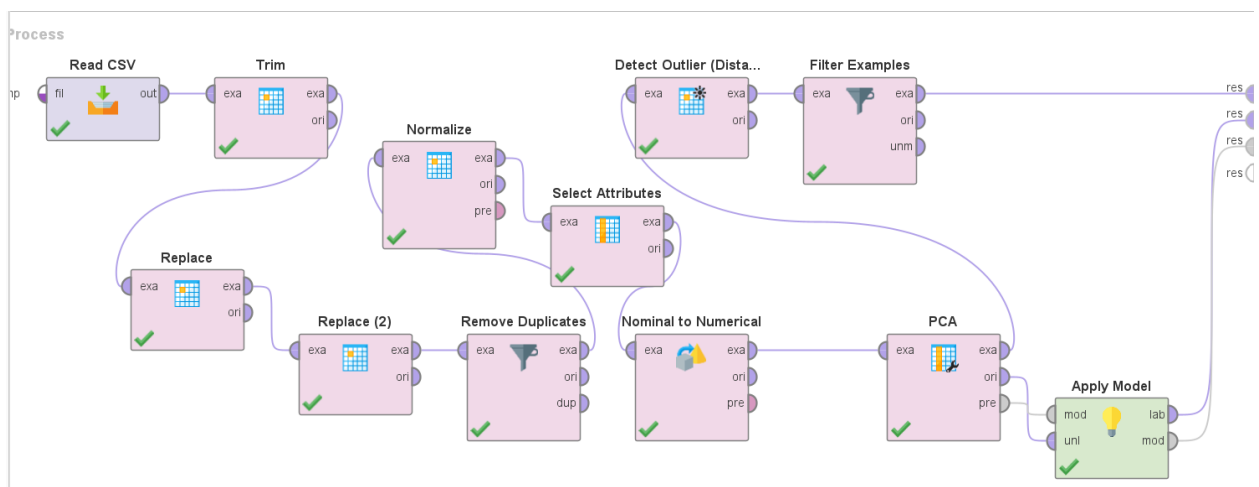
ExampleSet (Select Attributes (2))									
Open in		Turbo Prep		Auto Model		Interactive Analysis		Filter (3,799 / 3,799 examples): all	
Row No.	Choice	A_follower_...	A_following...	A_listed_co...	A_mentions...	A_mentions...	A_retweets...	A_posts	A...
10	0.979	-0.309	-0.248	-0.295	-0.088	0.037	-0.143	0.040	-0...
11	-1.021	1.802	-0.253	1.538	0.150	-0.411	-0.405	0.061	0.0...
12	-1.021	-0.313	-0.245	-0.300	-0.090	0.098	-0.403	-0.141	-0...
13	0.979	-0.190	-0.254	-0.311	-0.070	-0.058	-0.529	-0.087	-0...
14	0.979	0.035	-0.252	0.074	-0.034	-0.256	0.388	-0.142	-0...
15	0.979	-0.145	-0.216	-0.182	-0.091	-0.406	-0.266	-0.406	-0...
16	-1.021	-0.321	-0.258	-0.343	-0.092	-0.621	-0.529	-0.476	-0...
17	-1.021	-0.307	0.245	-0.335	-0.091	-0.520	-0.029	1.817	-0...
18	-1.021	-0.300	-0.257	-0.302	-0.090	-0.150	0.246	0.095	-0...
19	-1.021	1.042	0.155	0.567	0.082	-0.546	-0.405	-0.436	0.0...
20	0.979	-0.272	-0.210	-0.021	-0.078	0.246	2.074	0.092	-0...
21	-1.021	-0.320	-0.239	-0.341	-0.092	-0.512	-0.529	-0.406	-0...
22	-1.021	-0.298	-0.241	-0.181	-0.086	2.295	1.157	0.804	-0...

ExampleSet (3,799 examples, 1 special attribute, 24 regular attributes)

After pre-processing, the final data is shown here.

- Principal component analysis (PCA)

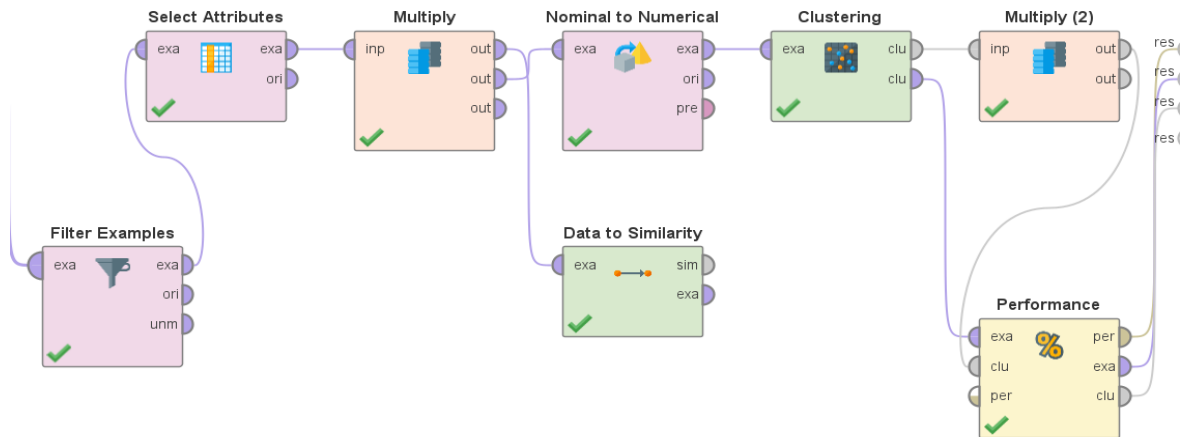
The Principal Component Analysis pre-processing of the data sets is shown in this section. This is applicable to sizable data sets with a high number of attributes that may result in some redundancy within the data sets.



The model performed the PCA using several procedures, as seen in the graphic Using the Select Attributes operator, it selected all the attributes except DATE because the PCA required that the data be converted from nominal to numerical. Second, the model uses the PCA method to remove a

significant portion of the data sets. Lastly, using Apply model operator to display the outcomes of running this model.

- Clustering



To determine whether approach to attribute dimensionality reduction is more effective, a clustering model has been used on the set of data. Select attributes is the initial operator used to choose numerical attributes. Next comes the multiply operator, followed by clustering using the K mean, and finally the performance operator.

As it is more effective to restrict the attributes in our data collection, we have chosen to preprocess the data using principal component analysis.

Task 4 - Building Data Mining Models

- Models

Building a Classification Model

Classification is the process of dividing different attributes into subgroups in order to determine their relationships. The main reason we classify is to expect the behavior of people based on Data set results. Additionally, this technique is helpful when it comes to making a decision, for instance, if an insurance company has a classified data set it can predict if the customer will renew the policy or not and based on that it can decide whether to give him promotions or not. The main purpose of classification in our model is to predict which platform people prefer to utilize for each category based on predictor attributes like platform, category, number of posts, number of tweets, number of followers and following.

First Classification Model:

- Random forest:

A random forest is one of the most vital tools in the data mining field in which the predictions of many simple models called decision trees are combined. They make the overall prediction by focusing on different parts of the data.

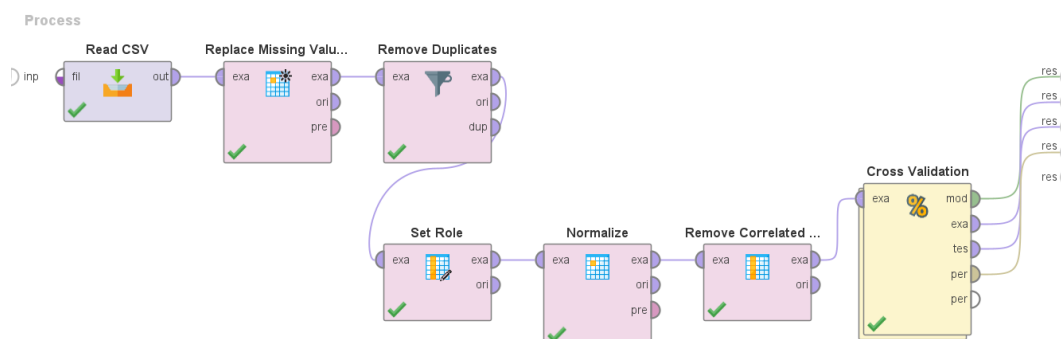
As we can see in the model, we first started off with the operator of replacing the missing values in order to address the missing entries of the data, which would affect the accuracy of the dataset. Next, we removed duplicates to achieve unique rows, for the same purpose.

We then used the 'category' attribute as the primary 'label' in the data by using the operator 'Set Role'. By this we indicate that "category" is the primary variable in our dataset.

After setting the role, we normalized the data with the attribute filter type 'Subset' and selecting all attributes. The purpose of normalizing is ensuring the consistency of the features and standardizing them to appear in the same range or scale. In order to ensure the effectiveness of the model, we first removed any associated attributes after normalizing the data and eliminating duplicates. This process focuses on separate features for the prediction tasks, which helped in enhancing the model's performance.

In order to address the correlated variables in the model, we utilized the operator 'Remove Correlated Attributes' with 0.95 as a parameter to ensure we were being selective and removing the most redundant features.

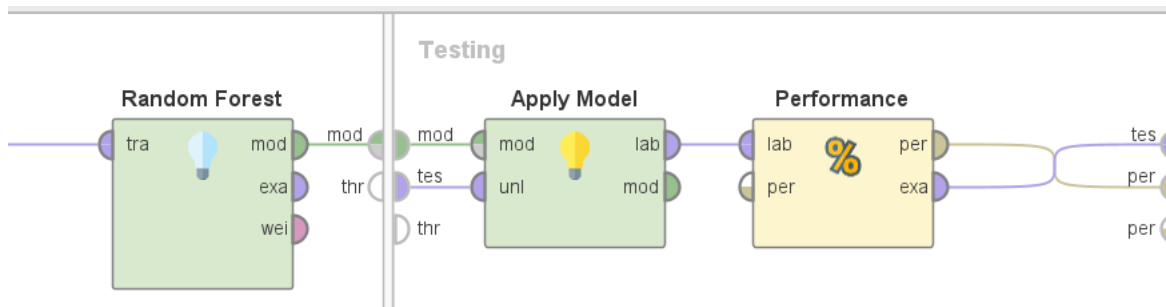
The next step was performing Cross Validation for the purpose of assessing the random forest model's predictive performance. Since it is 10 folds, the data was divided into 10 equal subsets, training on 9 folds and testing on one.



We then utilized the refined data to train the random forest algorithm. This method combines information from numerous individual decision trees for the purpose of increasing prediction accuracy and generating results suitable for forecasting and decision-making tasks.

After training the model, we applied it to new data using the apply model operator. This allowed us to observe the patterns that were discovered and produce the related predictions.

Finally, we conducted an evaluation of the model's performance. The purpose was to adjust parameters and enhance its predictive capability across a range of applications.



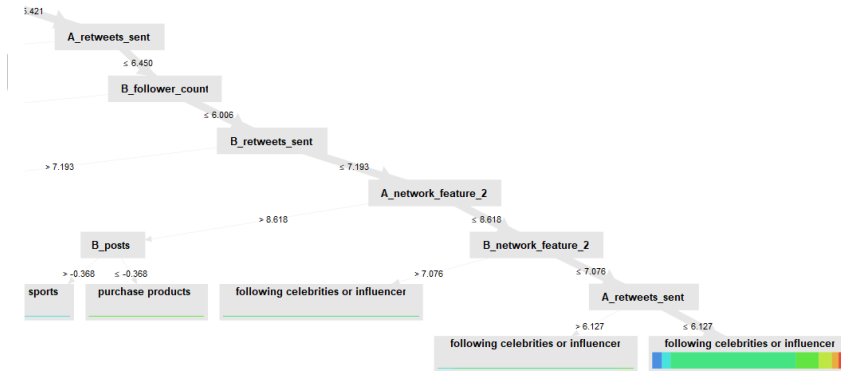
As demonstrated, these are the outcome of running the model.

▪ Cross validation Example set.



Here we notice that following celebrities and influencers is the most prominent category and the highest among all of the rest categories.

Random forest model



Evidently, the random forest model results show that the “following celebrities or influencers” category has been identified as the most critical factor in affecting whether a user is influential or not. This means that users who create posts or engage in activities in this category have more possibility in being an influential part in a Social Media society.

Remove correlated attributes

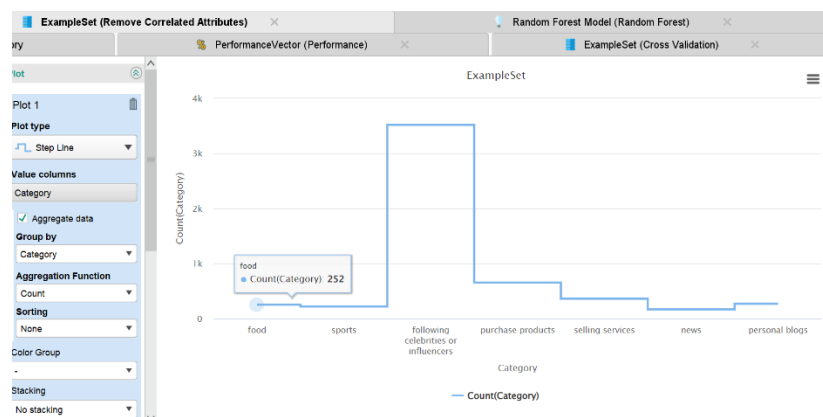
ExampleSet (Remove Correlated Attributes) | Random Forest Model (Random Forest)

Result History | PerformanceVector (Performance) | ExampleSet (Cross Validation)

Open in: Turbo Prep | Auto Model | Interactive Analysis | Filter (5,442 / 5,442 examples): all

Row No.	Category	Choice	A_follower...	A_following...	A_listed_co...	A_mentions...	A_mentions...	A_retweets...	A_posts	A_network...
3	following cel...	-1.021	-0.317	-0.234	-0.337	-0.091	-0.069	-0.273	-0.299	-0.179
4	purchase pro...	-1.021	-0.321	-0.258	-0.343	-0.091	-0.492	-0.529	-0.341	-0.182
5	selling serv...	0.979	-0.298	-0.241	-0.191	-0.086	2.295	1.157	0.804	-0.163
6	news	-1.021	-0.180	5.356	-0.147	-0.091	-0.621	-0.529	-0.489	-0.180
7	personal blogs	-1.021	-0.180	5.356	-0.147	-0.091	-0.621	-0.529	-0.489	-0.180
8	following cel...	0.979	-0.316	-0.258	-0.331	-0.090	-0.593	-0.529	-0.124	-0.175
9	following cel...	0.979	0.796	-0.196	0.343	0.140	0.728	0.674	3.109	0.563
10	following cel...	-1.021	-0.231	-0.230	-0.123	-0.087	1.840	-0.529	0.373	-0.162
11	following cel...	-1.021	-0.319	-0.192	-0.334	-0.092	-0.593	-0.529	-0.475	-0.182
12	following cel...	-1.021	-0.319	-0.229	-0.333	-0.091	-0.621	-0.529	-0.489	-0.182
13	following cel...	0.979	-0.309	-0.248	-0.295	-0.088	0.037	-0.143	0.040	-0.166
14	following cel...	1.071	1.805	-0.163	1.538	0.160	-0.411	-0.406	-0.061	-0.688

ExampleSet (5,442 examples, 1 special attribute, 23 regular attributes)

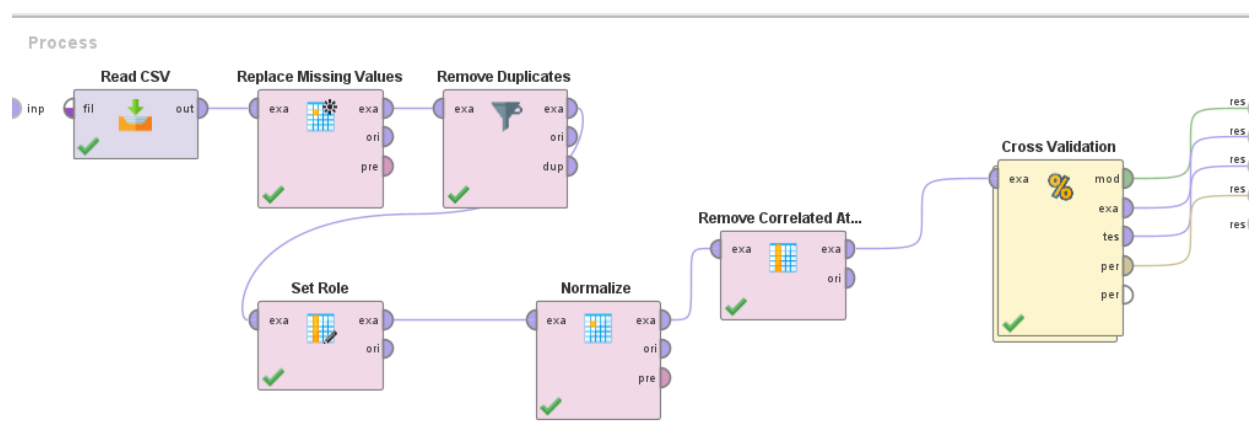


The analysis of the plot reveals a significant discrepancy in the count for the "celebrities and influencers" category compared to other categories. This suggests that the random forest model has identified a strong correlation between this category and user influence.

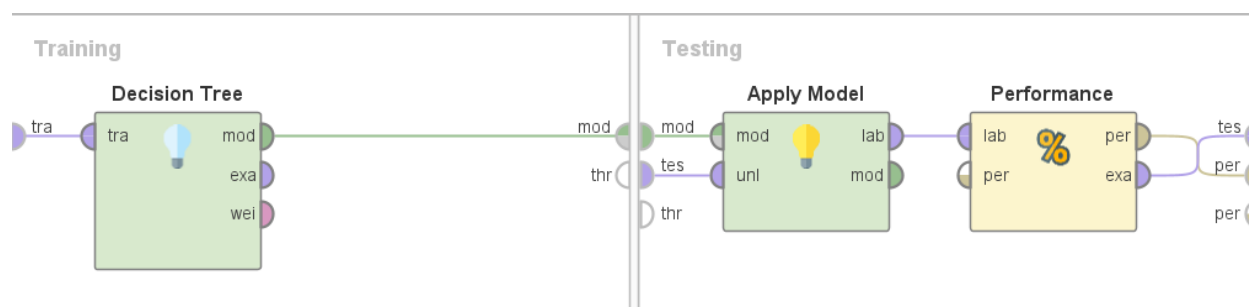
Second Classification Model:

- Decision Tree:

A decision tree is a tree-structured categorization model that shows the data set. By finding the greatest Entropy, it determines the data's root.

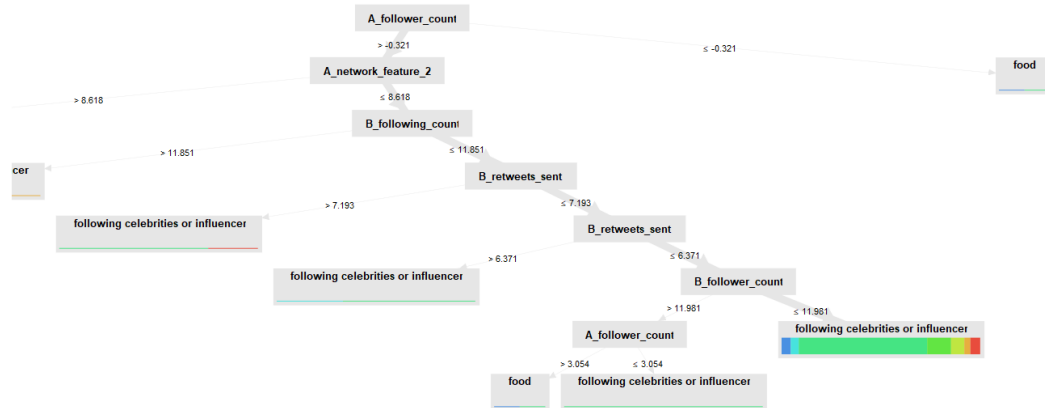


We used the refined dataset to train our final decision tree model after the cross-validation process. Next, the 'apply model' operator was used for this trained model to be applied to new data. This allowed us to use the model to predict and observe the way the stored decision rules would function when tested on actual cases, which is an essential step.

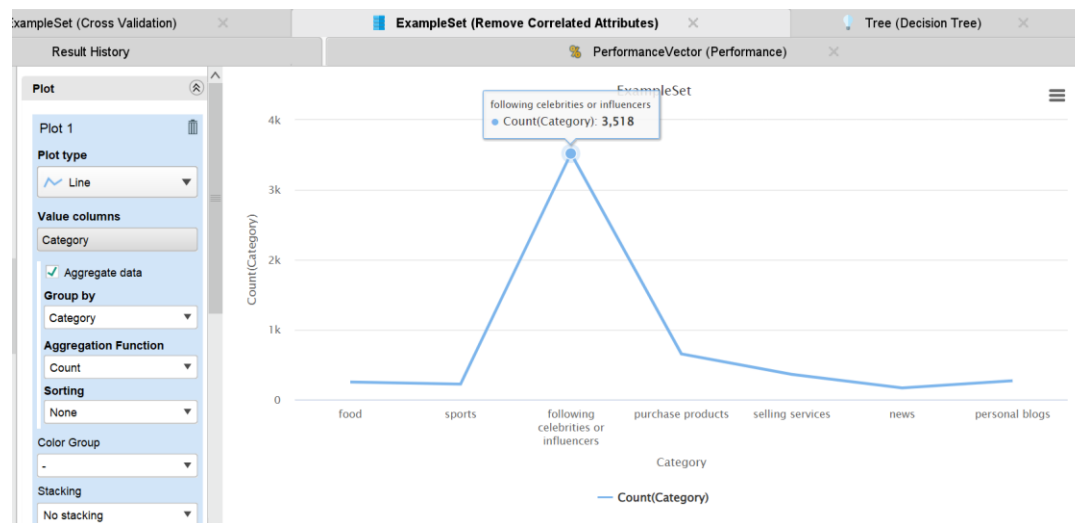


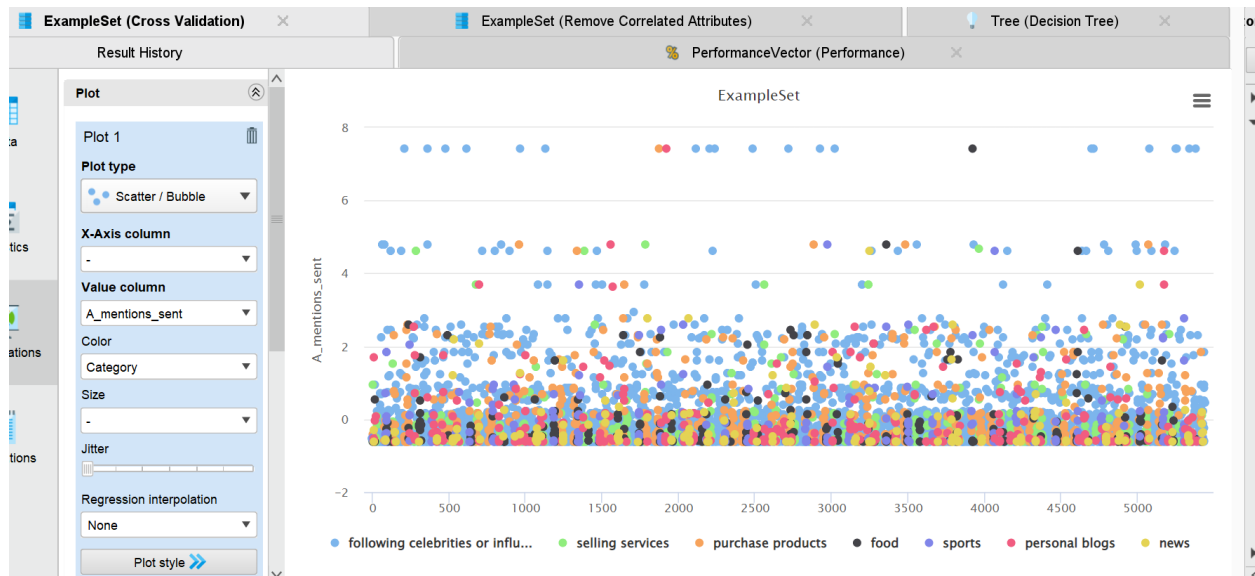
Decision tree advantages include:

- Easy model comprehension.
- Easy prediction of results following classification.



This decision tree model results confirm the Random Forest results that recognize the high significance of the “following celebrities and influencers” category as the most crucial one. Which leads us to believe that any users who participate in it are likely to be more influential than those who don’t.

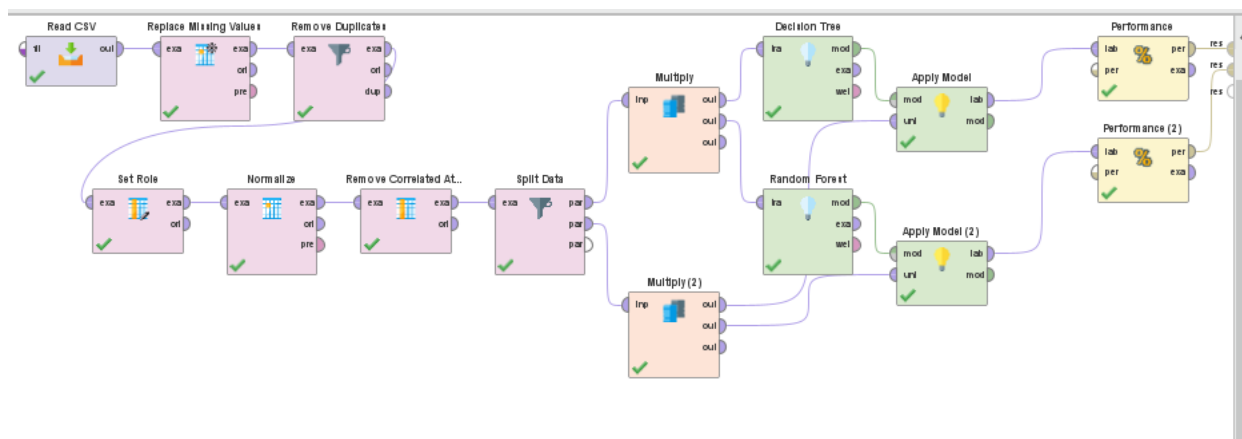




As clearly shown above, both the Decision Tree and Random Forest models agree on the importance of the category as it has the highest count. This finding has affects influencer marketing, social media strategy, identifying influencers, and user segmentation.

Third Classification Model:

- Ensemble method:
In Ensemble methods, several models on the same data are trained. the predictions of these models are averaged as a result of combining them in a specific way, as well as retrieving the provided most frequent answer. These models are combined for the purpose of integrating the strengths of each model and reducing their weaknesses.



Task 5 - Evaluating Data Mining Models

- Models Evaluation
- **Evaluation of the first model**

history

Criterion

accuracy

Table View Plot View

accuracy: 64.57% +/- 0.16% (micro average: 64.57%)

	true food	true sports	true following ...	true purchase...	true selling se...	true news	true personal ...	class precision
pred. food	0	0	0	0	0	0	0	0.00%
pred. sports	0	0	0	1	0	0	0	0.00%
pred. followin...	252	220	3514	653	361	167	269	64.64%
pred. purchas...	0	1	1	0	0	0	0	0.00%
pred. selling s...	0	0	0	0	0	0	0	0.00%
pred. news	0	0	2	0	0	0	0	0.00%
pred. persona...	0	0	1	0	0	0	0	0.00%
class recall	0.00%	0.00%	99.89%	0.00%	0.00%	0.00%	0.00%	

As you can see, the table above illustrates the performance of the decision tree model using the confusion matrix. Among the table's key metrics are:

Overall Accuracy: Using the same micro average, the model has an overall accuracy of 64.57%.

Prediction Counts: The total number of predictions made by the model as well as the number of accurate forecasts is presented in the table for each task.

Class Precision: Each task's class precision is shown in the final row. The accuracy of observing cases related to "following celebrities/influencers" is 64.64%.

Class Recall: The table indicates great opportunity for improvement with a recall rate of 0.00% across all tests except true following celebrities influencer which is 99.89%.

The rest of the requirements did not show for us, justification provided in Task 6. [Issues faced during the project.](#)

- **Evaluation of the second model**

Result History

Criterion

accuracy

classification error

kappa

Table View Plot View

accuracy: 64.63% +/- 0.09% (micro average: 64.63%)

	true food	true sports	true following ...	true purchase...	true selling se...	true news	true personal ...	class precision
pred. food	0	0	0	0	0	0	0	0.00%
pred. sports	0	0	1	0	0	0	1	0.00%
pred. followin...	252	221	3517	654	361	167	268	64.65%
pred. purchas...	0	0	0	0	0	0	0	0.00%
pred. selling s...	0	0	0	0	0	0	0	0.00%
pred. news	0	0	0	0	0	0	0	0.00%
pred. persona...	0	0	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	99.97%	0.00%	0.00%	0.00%	0.00%	

As you can see, the table above illustrates the performance of the decision tree model using the confusion matrix. Among the table's key metrics are:

Overall Accuracy: Using the same micro average, the model has an overall accuracy of 64.63%.

Prediction Counts: The total number of predictions made by the model as well as the number of accurate forecasts is presented in the table for each task.

Class Precision: Each task's class precision is shown in the final row. The accuracy of observing cases related to "following celebrities/influencers" is 64.65%.

Class Recall: The table indicates great opportunity for improvement with a recall rate of 0.00% across all tests except true following celebrities influencer which is 99.97%.

The rest of the requirements did not show for us justification provided in Task 6. [Issues faced during the project](#)

- **Evaluation of ensemble model**

This evaluation compares two performances of the ensemble method. By evaluating these tables, we can see the progression of the model over time for the purpose of identifying the areas of improvement or decline.

PerformanceVector (Performance (2)) PerformanceVector (Performance)

Criterion
accuracy

Table View Plot View

accuracy: 64.64%

	true food	true sports	true following ...	true purchase...	true selling se...	true news	true personal ...	class precision
pred. food	0	0	0	0	0	0	0	0.00%
pred. sports	0	0	0	0	0	0	0	0.00%
pred. followin...	76	65	1055	196	108	50	80	64.72%
pred. purchas...	0	1	0	0	0	0	1	0.00%
pred. selling s...	0	0	0	0	0	0	0	0.00%
pred. news	0	0	0	0	0	0	0	0.00%
pred. persona...	0	0	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	

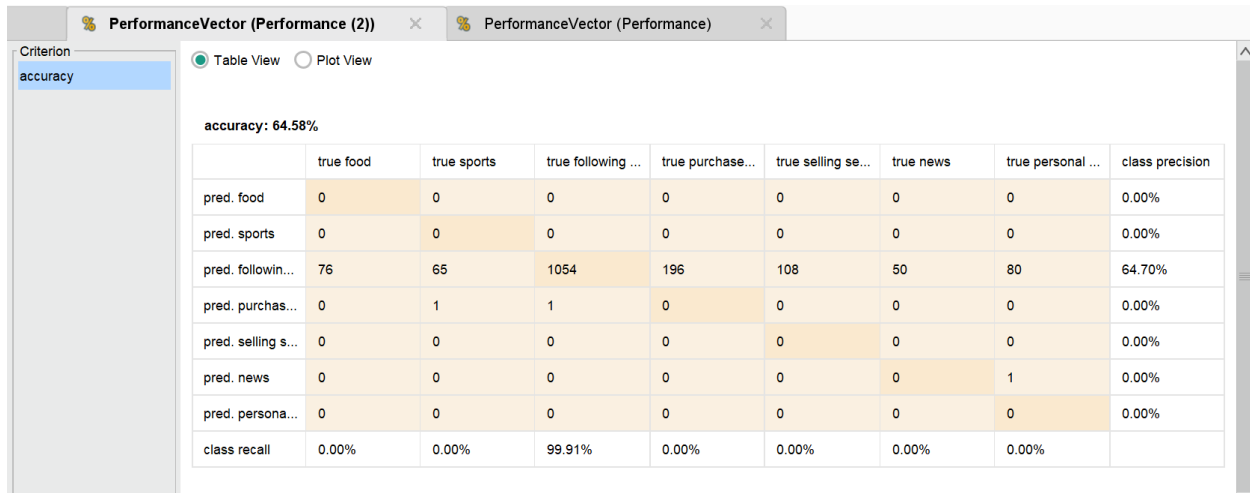
As you can see, the table above illustrates the performance of the ensemble method using the confusion matrix. Among the table's key metrics are:

Accuracy: The model's overall accuracy has increased slightly to 64.64%.

Prediction Counts: The total number of predictions made by the model as well as the number of accurate forecasts is presented in the table for each task.

Class Precision: The precision for identifying "true following ..." (like following celebrities and influencers) is now 64.72%, which demonstrates better accuracy in this area.

Class Recall: The recall for "true following ..." has improved significantly, reaching 100.00%, which shows that the model is effective at identifying instances in this category.



The screenshot shows a software window titled "PerformanceVector (Performance (2))" with a "Table View" selected. On the left, a sidebar lists "Criterion" with "accuracy" selected. The main area displays a table with the following data:

	true food	true sports	true following ...	true purchase...	true selling se...	true news	true personal ...	class precision
pred. food	0	0	0	0	0	0	0	0.00%
pred. sports	0	0	0	0	0	0	0	0.00%
pred. followin...	76	65	1054	196	108	50	80	64.70%
pred. purchas...	0	1	1	0	0	0	0	0.00%
pred. selling s...	0	0	0	0	0	0	0	0.00%
pred. news	0	0	0	0	0	0	1	0.00%
pred. persona...	0	0	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	99.91%	0.00%	0.00%	0.00%	0.00%	

As you can see, the table above illustrates another performance of the ensemble method using the confusion matrix. Among the table's key metrics are:

Accuracy: The model's overall accuracy has decreased slightly to 64.58%.

Prediction Counts: The total number of predictions made by the model as well as the number of accurate forecasts is presented in the table for each task.

Class Precision: The precision for identifying "true following ..." (like following celebrities and influencers) is now 64.70%, which demonstrates better accuracy in this area.

Class Recall: The recall for "true following ..." has decreased slightly to 99.91%.

The rest of the requirements did not show for us justification provided in Task 6. [Issues faced during the project.](#)

Task 6 - Inferences, Recommendation and Reflection

- Inferences

According to Task 5's performance measures, the **Ensemble model** seems to be the top performer. Compared to the other models, it has the best accuracy (64.64), meaning that it predicts the target variable more frequently and accurately.

Along with its good accuracy (64.63), the **decision tree** model also demonstrates respectable precision and recall. On all these criteria still, the Ensemble model performs better than it does.

Of the three models, the **Random Forest** model has the lowest accuracy (64.57) even though its accuracy and recall are comparable.

The Ensemble model is suggested as the best option based on these findings. It performs better as seen by its greater recall, accuracy, and precision. Despite this, before making a final decision, it is crucial to take several additional considerations including interpretability, and computational complexity.

- Recommendation

Based on the tasks applied throughout this project, a few recommendations can be made:

- The category in which a user interacts is the most essential determinant of whether they become influential in the social media community. This report found that following and engaging in activities about celebrities or other known influencers can greatly increase a user's potential in the online community. This is due to the significant correlation between this category and the most influential users compared in this dataset.
- Expanding the data to other platforms can allow exploration of a bigger data collection and more diverse user demographics that can aid the Data mining techniques in identifying what makes a user influential in social media overall and therefore help recognize current trends and future trends as well.

- Reflection

Data analysis and data cleaning were two of the lessons we learned throughout this project. We will be talking about sampling data, its applicability, the insights it may provide, and how it affects our comprehension of the significance of data pre-processing in this reflection.

One of the most important data analysis techniques is data sampling. Making meaning of the data is very hard when dealing with massive databases. Working with the entire dataset from the broader population would be very costly, time-consuming, and a huge waste of our resources. The ability to choose a subset of data points to represent characteristics is provided by samples. In this case, the sample enabled us to determine the real impact of following celebrities and influencers on social media sites, particularly Instagram and Twitter. We were able to obtain insightful knowledge by sampling this data while maintaining its correctness and integrity.

We may deal with a more obtainable dataset when we sample the data. It lessens the possibility of biased data and enables us to clean the data more effectively. It also improves the accuracy of our statistical conclusions. We decided to organize our dataset according to categories.

Accurate insights, well-informed decision-making, effective problem-solving and procedures, and dependable forecasts are all made possible by clean and accurate data. When you want to solve complicated problems or make smarter judgments, this is usually the most important stage. If the data set is not clean, when we make those crucial data-driven decisions, biases, outliers, missing numbers, and inconsistencies will all be considered and lead to mistakes. Our judgments become inaccurate and unreliable as a result.

Our understanding of the significance of data has fundamentally increased as a result of this project. As indicated and demonstrated in earlier sections, it increased our awareness of how important it is to pre-process data. Many of the procedures we performed, like clustering, outlier identification, random forest, decision tree, ensemble method would not have worked if we hadn't normalized our data. It wouldn't be correct even if it did.

- Issues faced during the project

The screenshot displays the Orange3 data mining environment. The main workflow area shows a sequence of operators: 'Random Forest' (model), 'Apply Model' (testing), and 'Performance' (evaluation). A red error box with a large 'X' is overlaid on the workflow, indicating a 'Missing label' error: 'Input ExampleSet does not have a label attribute.' This error is likely caused by the 'Set Role' operator in the background workflow not having a target attribute assigned. An 'Edit Parameter List: set roles' dialog is open, showing a table with 'Category' as the attribute name and 'label' as the target role. A 'Parameters' panel for the 'Performance (Performance (Binominal Classification))' operator is also visible, showing various performance metrics that can be selected for evaluation.

Cross Validation

Testing

Random Forest

Apply Model

Performance

Missing label
Input ExampleSet does not have a label attribute.

Edit Parameter List: set roles

attribute name	target role
Category	label

Parameters

Performance (Performance (Binominal Classification))

- ☐ AUC (pessimistic)
- ☒ precision
- ☒ recall
- ☐ lift
- ☐ fallout
- ☒ f measure
- ☐ false positive

We are facing an error in both models random forest and decision tree. The set role operator is already selected but still it shows an error that require a target attribute need to be

declared as label. Because of what went wrong, we were unable to evaluate all the metrics required like recall, precision, f-measure.

Extra Project Details

- Log files

Project Name: . Social Media Sector : Social Network Influencer Prediction				
Group No 3				
Student1 Name: Hind Busandal				ID 202002219
Student2 Name: Fatema Ali				ID 202000187
Student3 Name: Zainab Abbas				ID 202001012
Data Mining IT8416 Project				
Date	Week No	Task Name	Work done During the week (Bullet points /Description)	Issues Experienced if any
14-Mar-24	4	Choose Topic	Chosen the topic of Social Network Influencer Prediction	none
17-Mar-24	5	Task 1	Read the project document and get started on Task 1 to write the motivation, objectives, project statement, and expectations	none
28-Mar-24	6	Data Set	Explore the Data topic and analyze the article posted, download the available data set	could not download the data set due to restrictions on the website.
3-Apr-24	7	Downloading Resources	Download Rtools, Rstudio and Rapid Minor to prepare for the practical work	Fatema had trouble with downloading Rapid Minor due to an issue with registration and the Rapid minor website- took 4 days before the issue resolved
25-Apr-24	10	Task 2	Working on the dataset	none
5-May-24	12	Task 3	Preparing the data by applying the required techniques which are PCA and clustering	Took longer than expected but was done
12-May-24	13	Task 4 – Task 5	Building the models Random Forest and decision tree. Evaluating them	Had errors that affected the outcome
19-May-24	14 – Submission week	Task 6 and final edits	Wrote task 6 the inferences, recommendation, and reflection. Recorded the Video Demonstration	none

- Video Demonstration

Link: [Data Mining Video Demonstration.mp4](#)

- GitHub

Repository URL: <https://github.com/hindeb/DM-project.git>