SENG 372 Term Project

Final Report

Contents

1. Descriptions of the ETL Operations

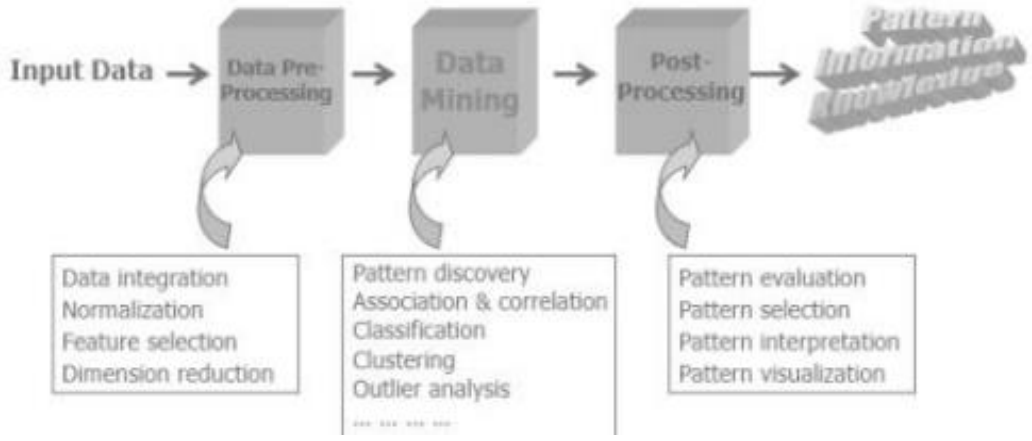
For this research, we will use these ETL operations:

Extract: The effectiveness of the subsequent procedures is determined by the extraction step, which is a vital part of the ETL process. In this step, raw data is exported to a staging area after being copied from its source location. We carried out extensive study to identify a trustworthy dataset in order to guarantee a correct extraction. We chose our datasets from Kaggle, a reliable and easily accessible source that offers information on mental health conditions and global happiness ratings for the data analysis. We can generate our own ideas on the subject using these datasets.

Transform: To make the extracted data compatible with the target system, the ETL process' transformation stage involves cleaning, enriching, and other changes. Data cleansing, mapping, enrichment, and validation are common sub steps that make up this process. It could also entail doing computations or aggregations, combining data from different sources, and applying business rules to the data. We executed data transformation by deleting superfluous columns and carrying out data cleansing in order to reach the desired outcome of correct answers from queries intended to analyze how the world happiness rates affect people's lives. We were able to correctly format and store the data for effective querying and analysis by following this procedure.

Load: Loading the transformed data into the target data warehouse requires moving it from the staging area at the ETL process's last stage. By doing this, we can make sure that the data loading phase interacts with the data in a way that satisfies the data quality performance requirements specified in the data table. This step is essential to ensure that the changed data is put into the target system appropriately and is accessible for queries and analysis to yield significant insights.

2. KDD Process



3. Brief Information About the Input Data

(# of samples, # of attributes, what is your class attribute? etc.)

In the data preprocessing section, we outlined the steps taken to preprocess and combine the Mental Health Disorders and the Crude Suicide Rates datasets using Orange Data Mining.

Data Import: We began by importing both the Mental Health Disorder dataset and the Crude Suicide Rates dataset using the 'File' widget in Orange.

Filter Columns: To simplify the analysis, we removed the 'Male' and 'Female' columns from the Suicide Rates dataset using the 'Select Columns' widget. This left us with the 'Both Sexes' column, which contains the average of the two sexes.

Merge Datasets: Next, we merged the two datasets by their common country attribute. We used the 'Merge Data' widget to accomplish this and set the merging option to include only countries present in both datasets. This step reduced the number of instances in the combined dataset.

Final Features: After preprocessing, we were left with the following features from the combined dataset:

From the Suicide Rates Dataset: Age Groups (10 to 19, 20 to 29, ..., 80 and above), Countries

From the Mental Health Dataset: Depression (%), Alcohol Usage Disorder (%)

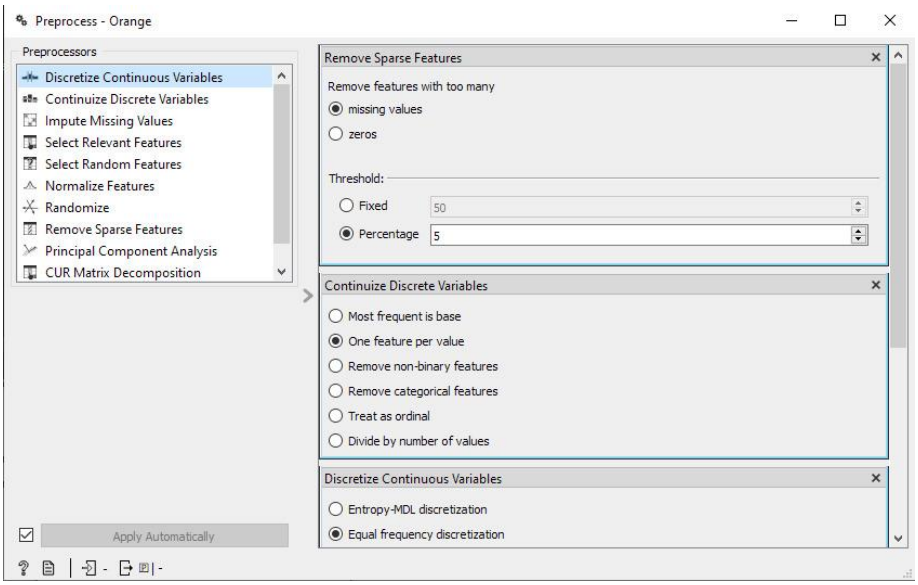
These features will now be used for further analysis, visualization, and modeling using Orange Data Mining tools.

The next actions involve exploring the dataset, visualizing relationships between variables, and building predictive models. We will:

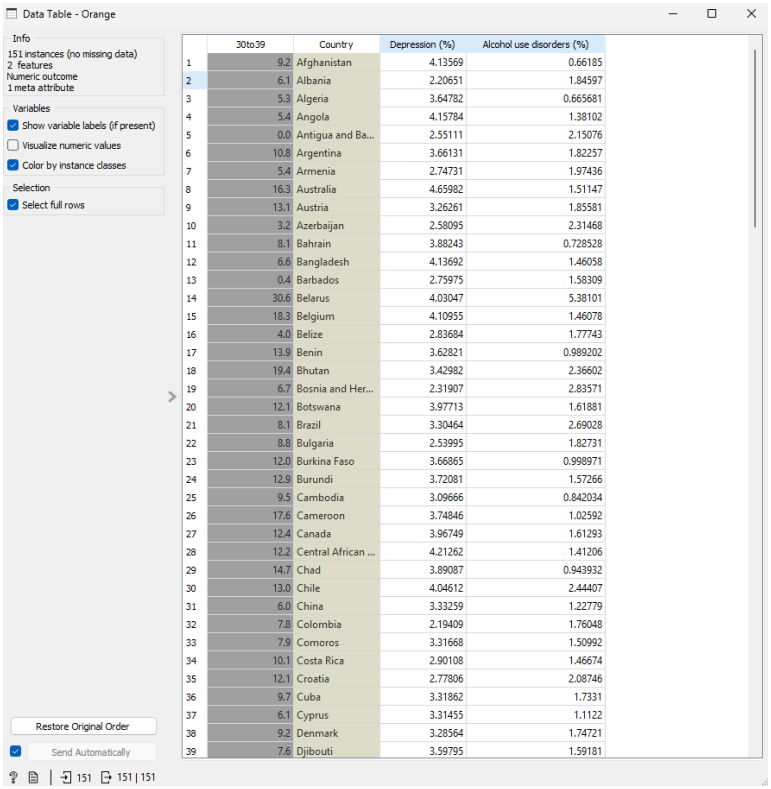
- Explore the dataset to understand its structure and identify any patterns or trends.
- Create visualizations such as scatter plots or box plots to examine the relationships between variables, like happiness and suicide rates.
- Select the most relevant features to improve the efficiency and effectiveness of the analysis.
- Apply clustering techniques to group similar instances together and reveal patterns related to happiness and suicide rates.
- Use supervised learning algorithms to build predictive models for happiness or suicide rates based on the selected features.
- Evaluate the performance of different models to choose the most suitable one for predicting happiness or suicide rates.

4. Data Pre-Processing Method

It is crucial to preprocess the data in order to confirm its quality and relevance for the research before beginning data mining. Feature selection, which involves choosing a subset of the most pertinent features for the analysis while removing redundant or unnecessary ones, is one of the most used data preparation techniques. When there are missing values in the data, imputation techniques like mean imputation or regression imputation can be used to fill in the gaps. Rows with missing values can also be erased, but doing so might mean losing important information. Additionally, if the data contains duplicate values, the duplicates can be eliminated using deduplication techniques like record linkage or hashing. Scaling and normalization are two more preprocessing techniques to make sure the data is uniform and normalized for analysis. In general, the type of preprocessing method chosen relies on the data's nature, the research issue, and the analysis's particular needs.



Feature scaling is a technique used to normalize the variety of independent variables or features in a dataset. This process is typically carried out during the data preprocessing step and is sometimes referred to as data normalization. Normalizing a dataset is crucial to minimizing data duplication and ensuring that only relevant data is stored in each table. There are three primary reasons for normalizing a dataset: to minimize duplicate data, to minimize or avoid data modification problems, and to simplify queries. We opted to use the Standardize to $\mu=0$, $\sigma 2 = 1$ option provided by Orange for normalization. This technique is also known as Standardization, which involves making the mean value 0 and the standard deviation 1, causing the distribution to approach normal. To achieve this, we subtract the mean value from the given value and divide it by the variance value, using the following formula.



	30to39	Country	Depression (%)	Alcohol use disorders (%)
1	9.2	Afghanistan	4.13569	0.66185
2	6.1	Albania	2.20651	1.84597
3	5.3	Algeria	3.64782	0.665681
4	5.4	Angola	4.15784	1.38102
5	0.0	Antigua and Ba...	2.55111	2.15076
6	10.8	Argentina	3.66131	1.82257
7	5.4	Armenia	2.74731	1.97436
8	16.3	Australia	4.65982	1.51147
9	13.1	Austria	3.26261	1.85581
10	3.2	Azerbaijan	2.58095	2.31468
11	8.1	Bahrain	3.88243	0.728528
12	6.6	Bangladesh	4.13692	1.46058
13	0.4	Barbados	2.75975	1.58309
14	30.6	Belarus	4.03047	5.38101
15	18.3	Belgium	4.10955	1.46078
16	4.0	Belize	2.83684	1.77743
17	13.9	Benin	3.62821	0.989202
18	19.4	Bhutan	3.42982	2.36602
19	6.7	Bosnia and Her...	2.31907	2.83571
20	12.1	Botswana	3.97713	1.61881
21	8.1	Brazil	3.30464	2.69028
22	8.8	Bulgaria	2.53995	1.82731
23	12.0	Burkina Faso	3.66865	0.998971
24	12.9	Burundi	3.72081	1.57266
25	9.5	Cambodia	3.09666	0.842034
26	17.6	Cameroon	3.74846	1.02592
27	12.4	Canada	3.96749	1.61293
28	12.2	Central African ...	4.21262	1.41206
29	14.7	Chad	3.89087	0.943932
30	13.0	Chile	4.04612	2.44407
31	6.0	China	3.33259	1.22779
32	7.8	Colombia	2.19409	1.76048
33	7.9	Comoros	3.31668	1.50992
34	10.1	Costa Rica	2.90108	1.46674
35	12.1	Croatia	2.77806	2.08746
36	9.7	Cuba	3.31862	1.7331
37	6.1	Cyprus	3.31455	1.1122
38	9.2	Denmark	3.28564	1.74721
39	7.6	Djibouti	3.59795	1.59181

Figure 1: Used Data Table

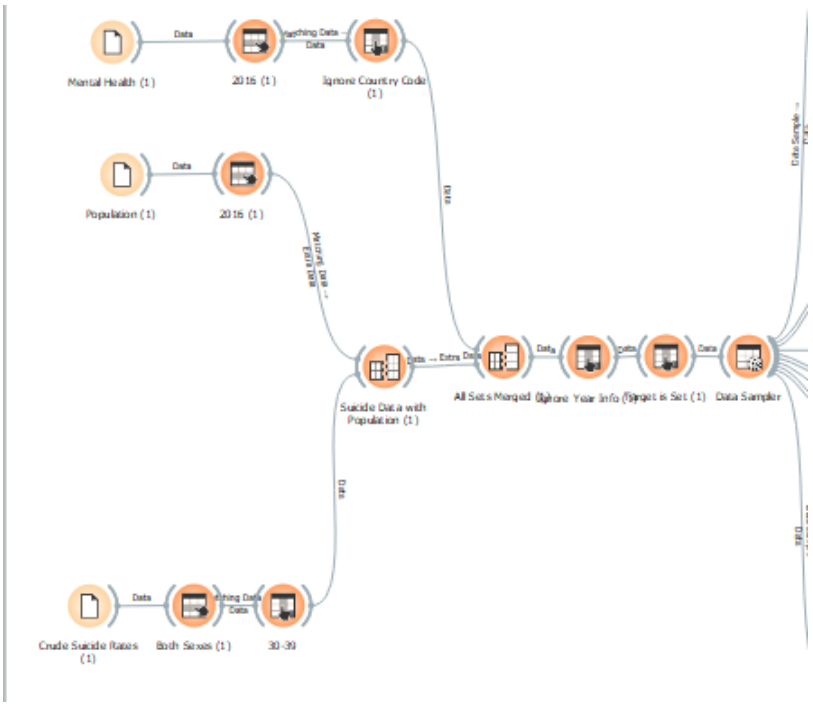


Figure 2: Preprocessing Steps

The image displayed above depicts the most recent edition of our preprocessing phase using Orange. Following this stage, we intend to utilize some algorithms that were taught in our class.

5. Choosing Mining Method

The data mining process involves the systematic look and examination of tremendous databases in arrange to distinguish already obscure designs, relationships, and experiences. In this study, we worked over Orange, and we used different methods to handle the data mining process.

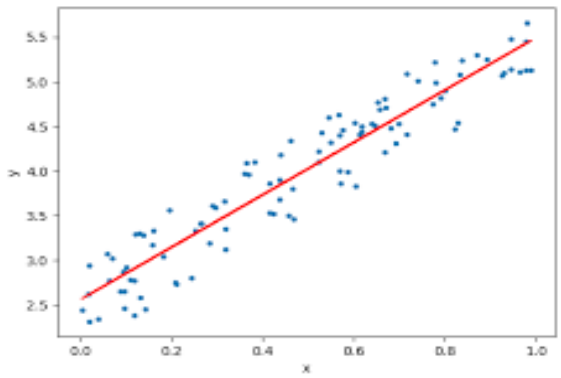
There are many countries and many people in the world. Since our datasets have some data about countries and people's ages, there could be so many predictions. That's why we decided to reduce our research area. We applied data mining techniques in various data, but we mainly focused on 2 subjects. After the preprocessing stages, we decided to use the people's ages (30-39) and the suicide rates around the world. We used some prediction algorithms in Orange.

We used Linear Regression and Random Forest to see if there are correlations present, and then used our data to train various models including RF. Finally, we used k-fold cross-validation to pick the model that fit our dataset the best.

6.Brief Description the Algorithm that We Use

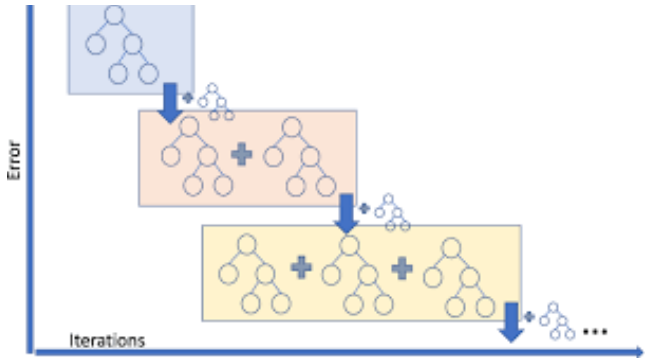
We created different and useful models with these algorithms. Here is the definition of these algorithms:

Linear Regression: A statistical modeling method called linear regression is used to examine the correlation between a dependent variable and one or more independent variables. It assumes that variables have a linear connection, with the objective of fitting the data to a straight line as closely as possible. The coefficients that minimize the sum of the squares of the difference between the predicted value and the actual value are estimated to produce this curve. In many different domains, linear regression is frequently used for prediction and inference tasks because it can reveal the nature and magnitude of correlations between variables.

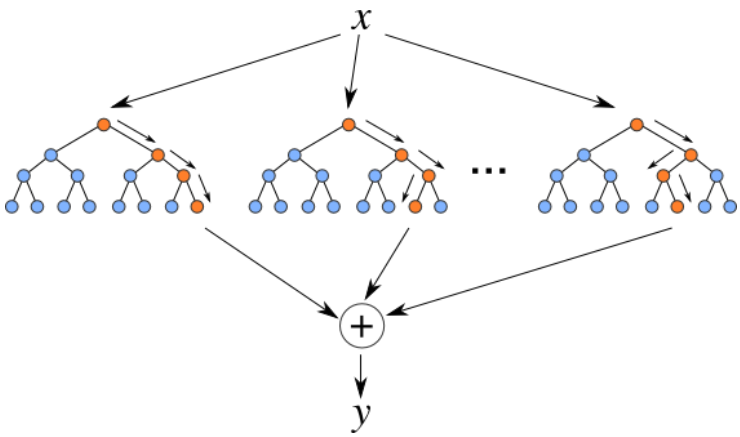


Gradient Boosting: Gradient Boosting is a machine learning technique that combines a number of weak prediction models, typically decision trees, to generate a powerful predictive model. It

works by repeatedly building new models that focus on capturing the errors or residuals of previous models, and then integrating these forecasts to get the final prediction. Each new model is trained to minimize the loss function using the errors of the previous models. Gradient Boosting is a popular and powerful machine learning method that can handle complex relationships and numerous types of data while producing high projected accuracy.



Random Forest: Powerful learning techniques like Random Forest are employed for both classification and regression applications. It mixes several decision trees, each of which was constructed using a randomly chosen subset of the training data and a randomly chosen subset of the features. The final prediction, either by majority vote (for classification) or by mean (for regression), is obtained by adding the predictions from all the trees. A well-liked and effective method in machine learning, Random Forest is noted for its capacity to handle multidimensional data sets, manage non-linear correlations, and over-smooth.



When we looked at the predictions, we found some results, but these results are not 100% correct. There are some errors. These errors are: MSE, RMSE, MAE, R2.

Mean Squared Error (MSE): Mean square error (MSE) is a frequently employed statistic to assess how well regression models work. Between the expected and actual numbers, it calculates the root mean squared difference between them. The MSE is calculated by first averaging the squared discrepancies between each projected value and its matching actual value. Better model performance is shown by a lower MSE, with values around 0 indicating a model that is more accurate. The reason MSE is so popular is that it may penalize greater errors more severely, giving a complete evaluation of the model's prediction accuracy.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Root Mean Squared Error (RMSE): A common evaluation metric for regression models is root mean square error (RMSE). It is calculated by squaring the difference between the expected and actual numbers and taking the square root of the mean. In the same units as the target variable, the RMSE offers a measurement of the average magnitude of the prediction mistakes. Similar to the MSE, a lower RMSE denotes better model performance, with values around zero denoting a model that is more accurate. When attempting to comprehend the normal error made by the model when predicting the target variable, RMSE is particularly helpful.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Mean Absolute Error (MAE): Mean Absolute Error (MAE) is a frequently used statistic to assess how well regression models perform. The mean absolute difference between expected and observed values is measured. The absolute difference between each predicted value and its matching actual value is measured in order to determine the MAE, which is subsequently averaged. Regardless of the direction of the errors, MAE gives a gauge of the typical magnitude of prediction errors. Better model performance is shown by a lower MAE, with values around zero indicating a model that is more accurate. MAE offers a straightforward interpretation of average error and is resistant to outliers.

$$\mathbf{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

R-squared (R2): The statistical measure R-squared (R2) expresses the percentage of the dependent variable's variance that can be accounted for by the independent variables in the regression model. It might be anywhere between 0 and 1, with 0 meaning that the model does not explain any variance and 1 meaning that it perfectly explains every variance. Higher values of R-squared indicate a better fit between the regression model and the data, which is used as an indication. R-squared should be used in conjunction with other measurements because it has

limitations, such as the inability to discriminate between an excellent model and a well-suited model.

$$R^2 = 1 - \frac{RSS}{TSS}$$

According to the algorithms we use for seeing the connection between depression and alcohol usage's effects on suicide rates between ages 30-39, [MSE] Mean Squared Error, Root Mean Squared Error (RMSE), R Square and Mean Absolute Error (MAE) values are as in the image below:

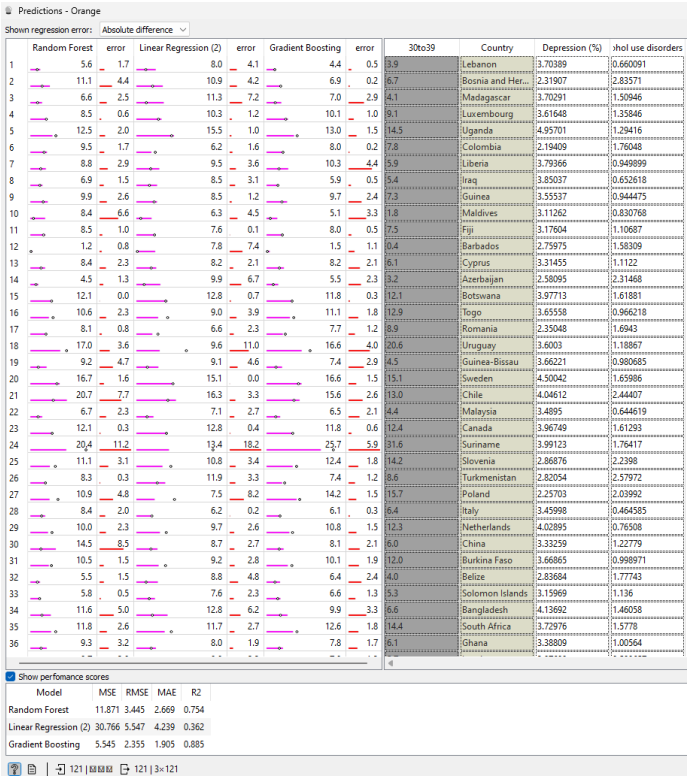
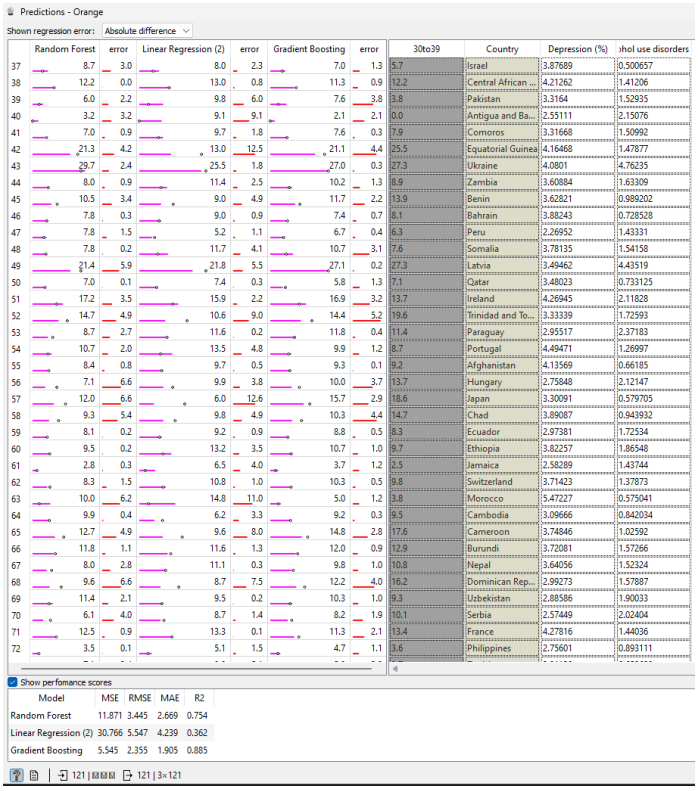


Figure 3: Prediction Table (line 1-36)



Predictions - Orange

Shown regression error: Absolute difference

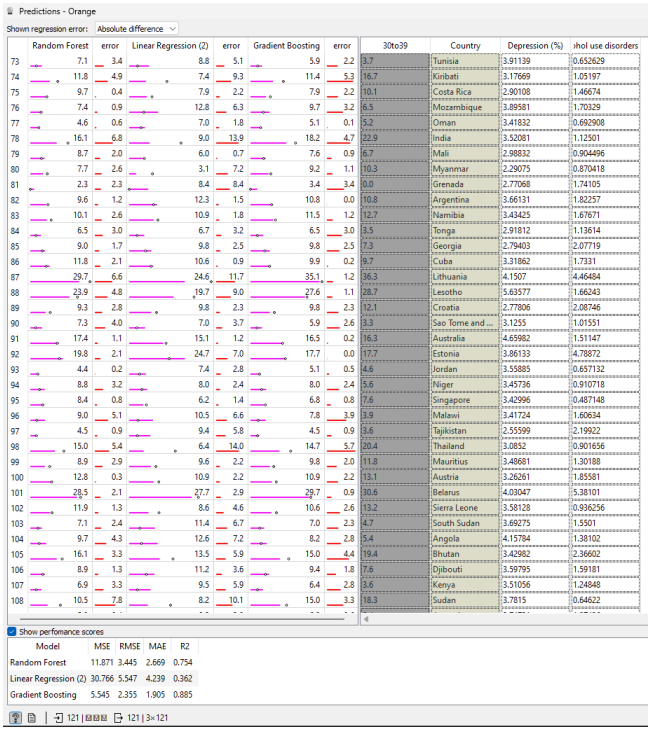
	Random Forest	error	Linear Regression (2)	error	Gradient Boosting	error	30to39	Country	Depression (%)	hol use disorders
37	8.7	3.0	8.0	2.3	7.0	1.3	5.7	Israel	3.87689	0.500657
38	12.2	0.0	13.0	0.8	11.3	0.9	12.2	Central African ...	4.21262	1.41206
39	6.0	2.2	9.8	6.0	7.6	3.8	3.8	Pakistan	3.3164	1.52935
40	3.2	3.2	9.1	9.1	2.1	2.1	0.0	Antigua and Ba...	2.55111	2.15076
41	7.0	0.9	9.7	1.8	7.6	0.3	7.9	Comoros	3.31668	1.50992
42	21.3	4.2	13.0	12.5	21.1	4.4	25.5	Equatorial Guinea	4.16468	1.47877
43	29.7	2.4	25.5	1.8	27.0	0.3	27.3	Ukraine	4.0801	4.76235
44	8.0	0.9	11.4	2.5	10.2	1.3	8.9	Zambia	3.60884	1.63309
45	10.5	3.4	9.0	4.9	11.7	2.2	13.9	Benin	3.62821	0.989202
46	7.8	0.3	9.0	0.9	7.4	0.7	8.1	Bahrain	3.88243	0.728528
47	7.8	1.5	5.2	1.1	6.7	0.4	6.3	Peru	2.26952	1.43331
48	7.8	0.2	11.7	4.1	10.7	3.1	7.6	Somalia	3.78135	1.54158
49	21.4	5.9	21.8	5.5	27.1	0.2	27.3	Latvia	3.49462	4.43519
50	7.0	0.1	7.4	0.3	5.8	1.3	7.1	Qatar	3.48023	0.733125
51	17.2	3.5	15.9	2.2	16.9	3.2	13.7	Ireland	4.26945	2.11828
52	14.7	4.9	10.6	9.0	14.4	5.2	19.6	Trinidad and To...	3.33339	1.72593
53	8.7	2.7	11.6	0.2	11.8	0.4	11.4	Paraguay	2.95517	2.37183
54	10.7	2.0	13.5	4.8	9.9	1.2	8.7	Portugal	4.49471	1.26997
55	8.4	0.8	9.7	0.5	9.3	0.1	9.2	Afghanistan	4.13569	0.66185
56	7.1	6.6	9.9	3.8	10.0	3.7	13.7	Hungary	2.75848	2.12147
57	12.0	6.6	6.0	12.6	15.7	2.9	18.6	Japan	3.30091	0.579705
58	9.3	5.4	9.8	4.9	10.3	4.4	14.7	Chad	3.89087	0.943932
59	8.1	0.2	9.2	0.9	8.8	0.5	8.3	Ecuador	2.97381	1.72534
60	9.5	0.2	13.2	3.5	10.7	1.0	9.7	Ethiopia	3.82257	1.86548
61	2.8	0.3	6.5	4.0	3.7	1.2	2.5	Jamaica	2.58289	1.43744
62	8.3	1.5	10.8	1.0	10.3	0.5	9.8	Switzerland	3.71423	1.37873
63	10.0	6.2	14.8	11.0	5.0	1.2	3.8	Morocco	5.47227	0.575041
64	9.9	0.4	6.2	3.3	9.2	0.3	9.5	Cambodia	3.09666	0.842034
65	12.7	4.9	9.6	8.0	14.8	2.8	17.6	Cameroon	3.74846	1.02592
66	11.8	1.1	11.6	1.3	12.0	0.9	12.9	Burundi	3.72081	1.57266
67	8.0	2.8	11.1	0.3	9.8	1.0	10.8	Nepal	3.64056	1.52324
68	9.6	6.6	8.7	7.5	12.2	4.0	16.2	Dominican Rep...	2.99273	1.57887
69	11.4	2.1	9.5	0.2	10.3	1.0	9.3	Uzbekistan	2.88586	1.90033
70	6.1	4.0	8.7	1.4	8.2	1.9	10.1	Serbia	2.57449	2.02404
71	12.5	0.9	13.3	0.1	11.3	2.1	13.4	France	4.27816	1.44036
72	3.5	0.1	5.1	1.5	4.7	1.1	3.6	Philippines	2.75601	0.893111

Show perfomance scores

Model	MSE	RMSE	MAE	R2
Random Forest	11.871	3.445	2.669	0.754
Linear Regression (2)	30.766	5.547	4.239	0.362
Gradient Boosting	5.545	2.355	1.905	0.885

121 | 121 | 3×121

Figure 4: Prediction Table (line 37-72)



Predictions - Orange

Shown regression error: Absolute difference

	Random Forest	error	Linear Regression (2)	error	Gradient Boosting	error	30to39	Country	Depression (%)	hol use disorders
73	7.1	3.4	8.8	5.1	5.9	2.2	3.7	Tunisia	3.91139	0.652629
74	11.8	4.9	7.4	9.3	11.4	5.3	16.7	Kiribati	3.17669	1.05197
75	9.7	0.4	7.9	2.2	7.9	2.2	10.1	Costa Rica	2.90108	1.46674
76	7.4	0.9	12.8	6.3	9.7	3.2	6.5	Mozambique	3.89581	1.70329
77	4.6	0.6	7.0	1.8	5.1	0.1	5.2	Oman	3.41832	0.692908
78	16.1	6.8	9.0	13.9	18.2	4.7	22.9	India	3.52081	1.12501
79	8.7	2.0	6.0	0.7	7.6	0.9	6.7	Mali	2.98832	0.904496
80	7.7	2.6	3.1	7.2	9.2	1.1	10.3	Myanmar	2.29075	0.870418
81	2.3	2.3	8.4	8.4	3.4	3.4	0.0	Grenada	2.77068	1.74105
82	9.6	1.2	12.3	1.5	10.8	0.0	10.8	Argentina	3.66131	1.82257
83	10.1	2.6	10.9	1.8	11.5	1.2	12.7	Namibia	3.43425	1.67671
84	6.5	3.0	6.7	3.2	6.5	3.0	3.5	Tonga	2.91812	1.13614
85	9.0	1.7	9.8	2.5	9.8	2.5	7.3	Georgia	2.79403	2.07719
86	11.8	2.1	10.6	0.9	9.9	0.2	9.7	Cuba	3.31862	1.7331
87	29.7	6.6	24.6	11.7	35.1	1.2	36.3	Lithuania	4.1507	4.46484
88	23.9	4.8	19.7	9.0	27.6	1.1	28.7	Lesotho	5.63577	1.66243
89	9.3	2.8	9.8	2.3	9.8	2.3	12.1	Croatia	2.77806	2.08746
90	7.3	4.0	7.0	3.7	5.9	2.6	3.3	Sao Tome and ...	3.1255	1.01551
91	17.4	1.1	15.1	1.2	16.5	0.2	16.3	Australia	4.65982	1.51147
92	19.8	2.1	24.7	7.0	17.7	0.0	17.7	Estonia	3.86133	4.78872
93	4.4	0.2	7.4	2.8	5.1	0.5	4.6	Jordan	3.55885	0.657132
94	8.8	3.2	8.0	2.4	8.0	2.4	5.6	Niger	3.45736	0.910718
95	8.4	0.8	6.2	1.4	6.8	0.8	7.6	Singapore	3.42996	0.487148
96	9.0	5.1	10.5	6.6	7.8	3.9	3.9	Malawi	3.41724	1.60634
97	4.5	0.9	9.4	5.8	4.5	0.9	3.6	Tajikistan	2.55599	2.19922
98	15.0	5.4	6.4	14.0	14.7	5.7	20.4	Thailand	3.0852	0.901656
99	8.9	2.9	9.6	2.2	9.8	2.0	11.8	Mauritius	3.48681	1.30188
100	12.8	0.3	10.9	2.2	10.9	2.2	13.1	Austria	3.26261	1.85581
101	28.5	2.1	27.7	2.9	29.7	0.9	30.6	Belarus	4.03047	5.38101
102	11.9	1.3	8.6	4.6	10.6	2.6	13.2	Sierra Leone	3.58128	0.936256
103	7.1	2.4	11.4	6.7	7.0	2.3	4.7	South Sudan	3.69275	1.5501
104	9.7	4.3	12.6	7.2	8.2	2.8	5.4	Angola	4.15784	1.38102
105	16.1	3.3	13.5	5.9	15.0	4.4	19.4	Bhutan	3.42982	2.36602
106	8.9	1.3	11.2	3.6	9.4	1.8	7.6	Djibouti	3.59795	1.59181
107	6.9	3.3	9.5	5.9	6.4	2.8	3.6	Kenya	3.51056	1.24848
108	10.5	7.8	8.2	10.1	15.0	3.3	18.3	Sudan	3.7815	0.64622

Show perfomance scores

Model	MSE	RMSE	MAE	R2
Random Forest	11.871	3.445	2.669	0.754
Linear Regression (2)	30.766	5.547	4.239	0.362
Gradient Boosting	5.545	2.355	1.905	0.885

121 | 121 | 3×121

Figure 5: Prediction Table (line 73-108)

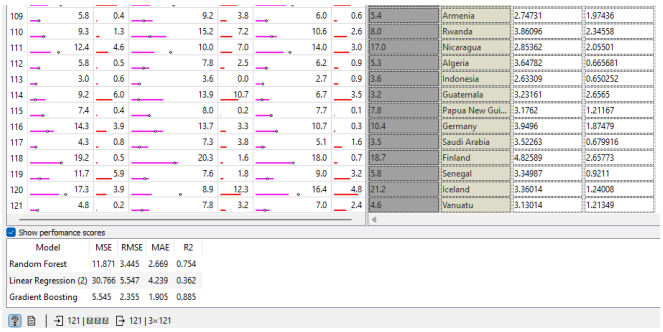


Figure 6: Prediction Table (line 109-121)

When we examined the prediction values shown in pink, we observed that there are less errors in gradient boosting than in any other model.

7. Prediction

Prediction is the practice of estimating or anticipating future occurrences or outcomes using statistical or machine learning methodologies and previous data in the context of data mining. It is critical in data mining because it allows us to analyze the data and make decisions based on the patterns and relationships we discover. We can notice patterns, possible hazards, or opportunities, allocate resources efficiently, and take proactive action to enhance productivity or handle complex challenges by anticipating outcomes. Prediction allows us to make use of data mining's ability to transform raw data into usable information, allowing firms to make decisions based on solid facts and gain an advantage over competitors in a range of disciplines and industries.

For accuracy, we gather the results from Random Forest, Gradient Boosting and Linear Regression. The prediction was tested and scored according to one age group which is 30-39. Here are the test results:

<u>Gradient Boosting:</u>

MSE = 5.544696371367683

RMSE = 2.3547178963450555

MAE = 1.9051613783442365

R2 = 0.8850547588122256

<u>Random Forest:</u>

MSE = 12.255918166965207

RMSE = 3.5008453503354313

MAE = 2.7634636447115786

R2 = 0.7459266702223498

Linear Regression:

MSE = 30.766403859328392

RMSE = 5.5467471421841825

MAE = 4.239421637524827

R2 = 0.36219199840177085

Model	MSE	RMSE	MAE	R2
Random Forest	11.871	3.445	2.669	0.754
Linear Regression (2)	30.766	5.547	4.239	0.362
Gradient Boosting	5.545	2.355	1.905	0.885

Figure 7: Performance scores of Depression and Alcohol Consumption's Effects on Suicides

8. Interpret the Result

With this study, we were able to observe the effect of depression and alcohol use on suicide rates. As alcohol use increases, suicide rates also increase for people in the 30-39 age range. However, when we look at the ages of 70-79, we observe that the suicide rate decreases compared to alcohol use. Depression, on the other hand, has a constant effect on the suicide rate. As depression increases, suicide rates also increase regardless of age.

We can say that alcohol use should also be reduced to reduce suicide rates. In order to reduce alcohol use, we can raise people's awareness about this issue, or we can try to prevent alcohol consumption with rules and restrictions. At the same time, precautions can also be taken to reduce depression, as it will reduce the suicide rate. Increasing the number of mental health facilities may be one of the precautions that can be taken in this regard.

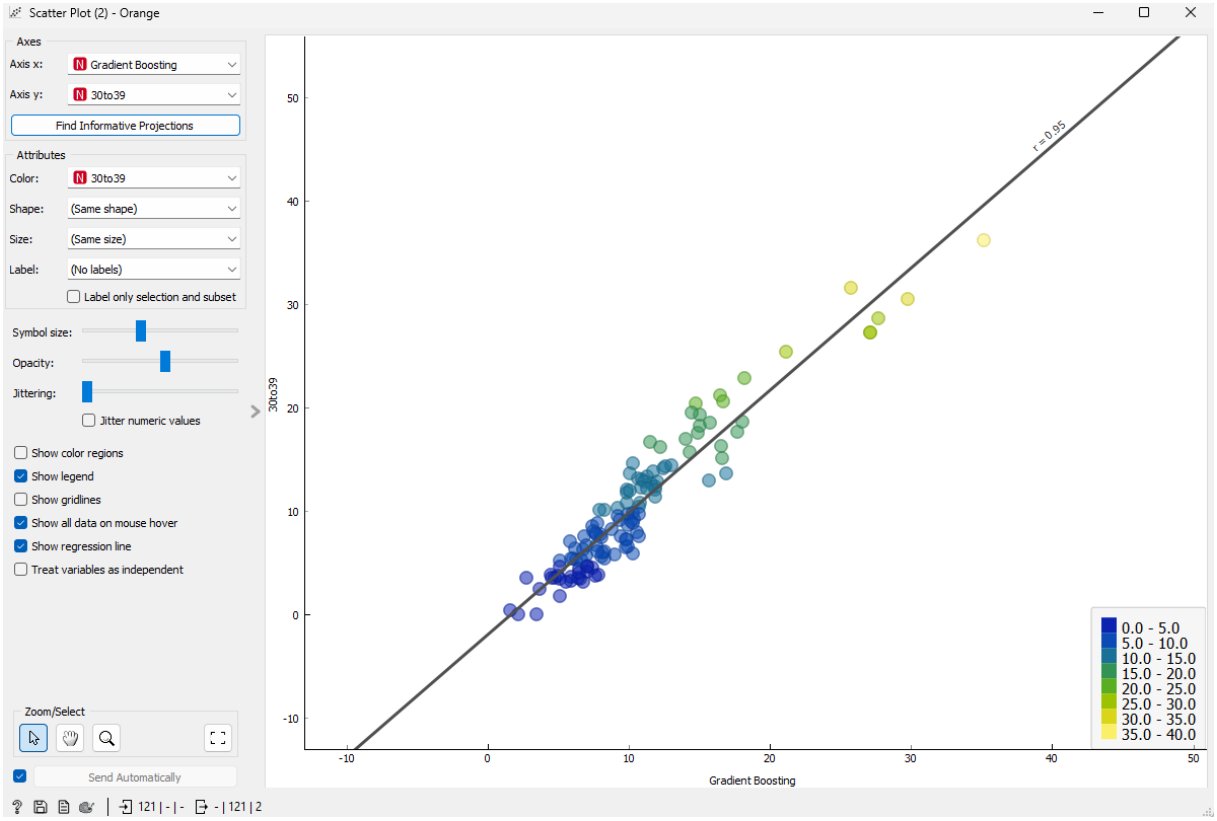


Figure 8: Gradient Boosting Performance Scatter Plot

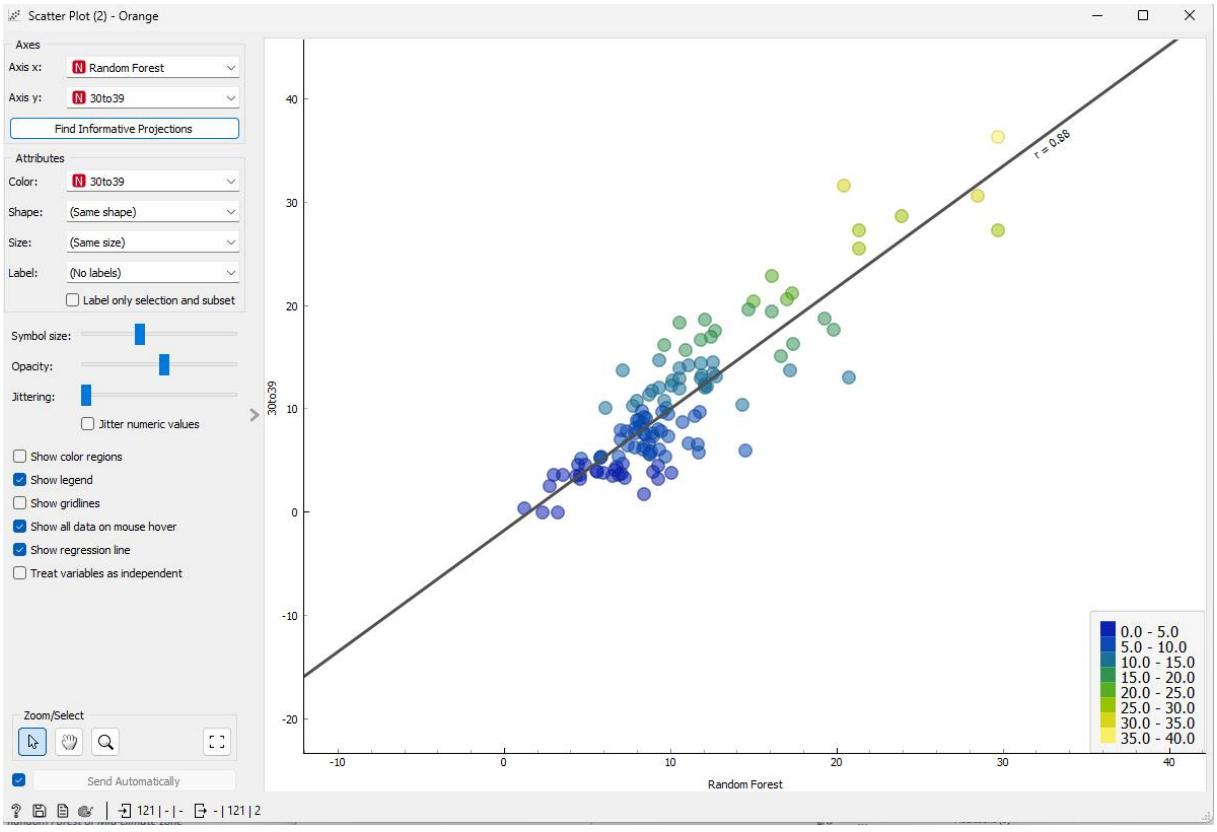


Figure 9: Random Forest Performance Scatter Plot

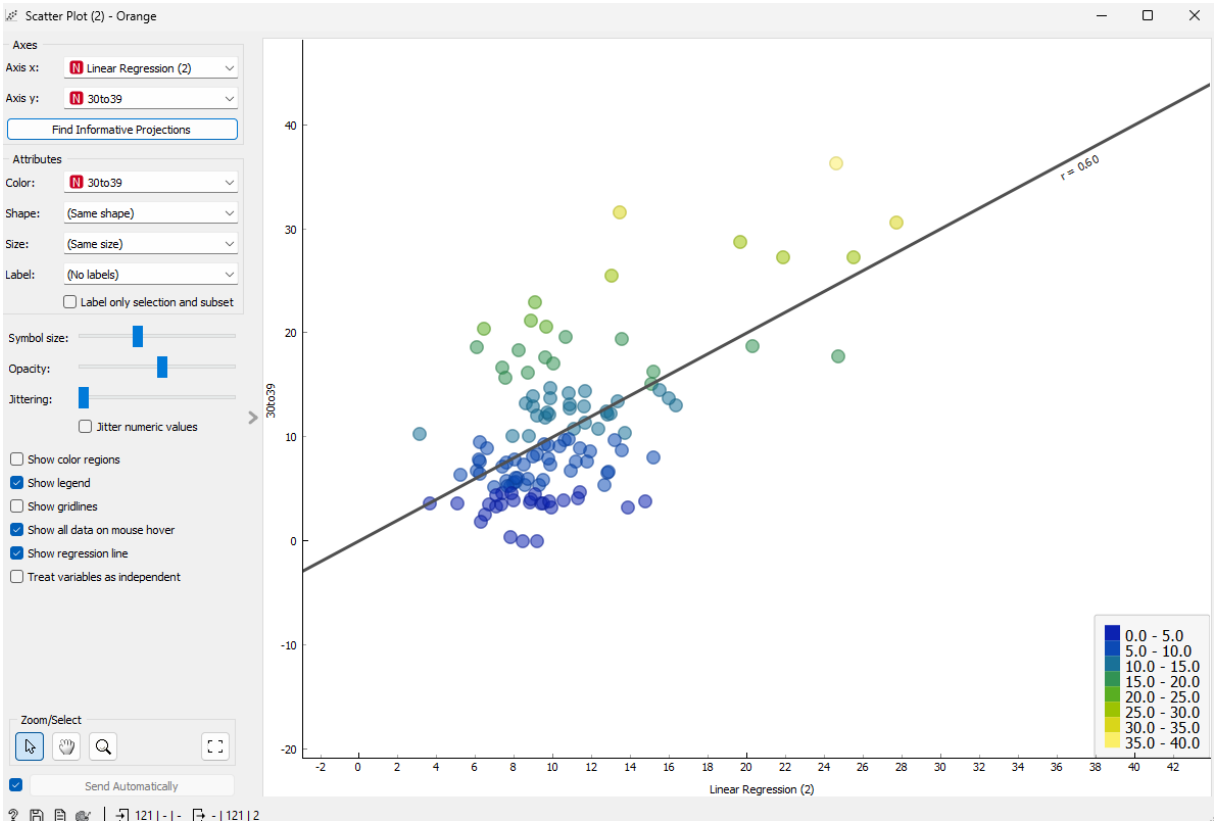


Figure 10: Linear Regression Performance Scatter Plot

9. References

https://www.geeksforgeeks.org/data-mining-process/

https://www.javatpoint.com/classification-and-predication-in-data-mining

https://www.marketingprofs.com/articles/2010/3567/the-nine-most-common-data-mining-techniques-used-in-predictive-analytics

https://www.geeksforgeeks.org/what-is-prediction-in-data-mining/

https://www.geeksforgeeks.org/ml-feature-scaling-part-2/

https://www.javatpoint.com/regression-in-data-mining

https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

https://towardsdatascience.com/quick-intro-to-random-forest-3cb5006868d8