MSIS – 5633 PREDICTIVE ANALYTICS TECHNOLOGIES

Homework Assignment #5

KNIME Data Mining II

Pre-processing and Analyzing Gambling Data Set Through the Implementation of Machine Learning Algorithms

Due Date

December 03, 2023

By

Rishitha Ganagoni

A20398497

Table of Contents

Executive Summary

The primary aim of this report is to assess the level of support for legalizing gambling in the United States. To achieve this goal, we utilized a dataset comprising 1200 rows and 31 columns. Employing machine learning techniques, we aimed to develop models capable of predicting the percentage of support for the legalization of gambling. The analysis of gambling conditions took into account various demographic, socio-economic factors. The report adopted the CRISP-DM methodology, a widely recognized standard in the field of Data Mining, to guide the analysis.

Within this report, we constructed three distinct models, namely Artificial Neural Network (ANN), Decision Tree (DT), and Random Forest (RF). Thorough consideration of multiple factors preceded the selection of the most effective model for identifying patterns. Metrics such as sensitivity, specificity, and ROC curve values played a crucial role in the final evaluation and selection of the best-performing model.

Additionally, sufficient time was allocated for data pre-processing. This involved the removal of certain columns and the replacement of missing values with the median, aiming to enhance the quality and reliability of the dataset for subsequent analysis.

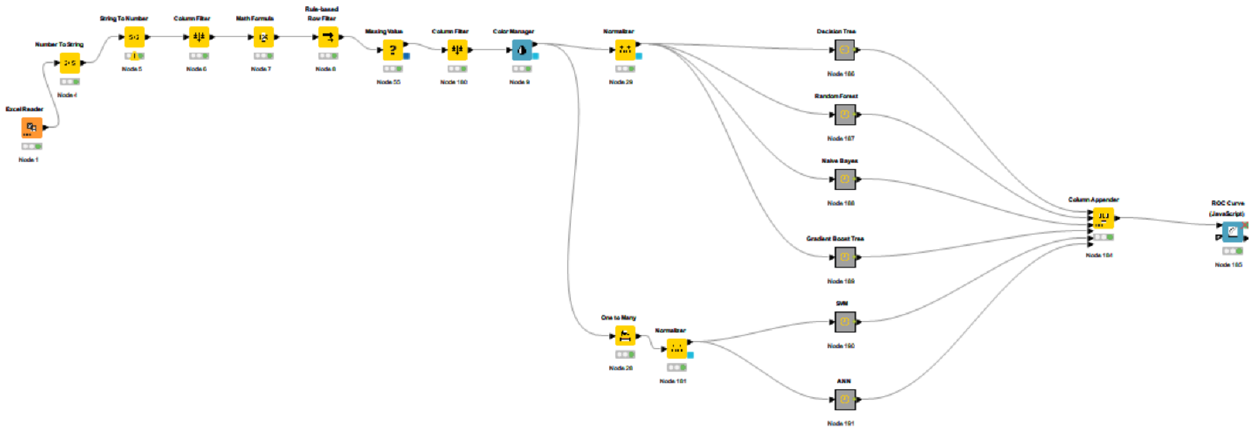

Figure 1: Overview of Workflow of all six models

CRISP-DM Methodology:

Debuted in 1999, it is known by its acronym, CRISP-DM, which stands for CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING, is widely respected and frequently employed as one of the most prevalent data mining methodologies. CRISP-DM serves as a process model that offers a comprehensive view of the data mining life cycle. CRISP-DM aims to address the following key stages in data mining:

Business Comprehension

Data Exploration

Data Cleansing and Preparatory Procedures

Application of Modeling Techniques

Model Assessment

Implementation

Business Comprehension:

There is no project without any business need. If there is no business requirement means no application of Data Mining techniques. For any project, understanding the customer requirements and drill downing according to it and fulfilling the customer requirement is the main motto. Firstly, We need to gather and analyze the data and according to that we have to take further steps. Within this project, our objective is to discern the stance of customers regarding the legalization of gambling. We have provided an avenue for customers to articulate their opinions through survey participation. In tandem with this, we have collected pertinent demographic, socio-economic variables to incorporate a comprehensive analysis into the study.

Data Exploration:

To do anything first we have to understand the given data which is the key step in any project. Removing the unwanted data and analyzing it, improves the richness of the Data.

Numeric Nominal Data Preview

Search:

Row ID	FOR	AGAINST	TOTAL CASTE	DEPENDENT VARIABLE	SIZE OF COUNTY	POPULATION DENSITY
Row0	42385	22992	65377	1	1192	223.290
Row3	957	856	1813	1	1349.400	3.990
Row4	942	1129	2071	0	2555.900	1.770
Row5	905	646	1551	1	1514	3.320
Row8	490	422	912	1	1781.500	1.340
Row9	1923	1140	3063	1	395.500	19.290
Row10	1063	1343	2406	0	1287.300	5.800
Row11	624	404	1028	1	1227	2.590
Row14	3645	3918	7563	0	1142.200	18.430
Row16	322	320	642	1	1067	1.410
Row17	11519	8261	19780	1	840.200	73.320
Row18	3006	2064	5070	1	1688	13.160

Figure 2: Cleaned data after Data Exploration

Data Cleansing and Preparatory Procedures:

Data pre-processing involves the data reduction, cleaning the data according to the requirements and transforming it. This step plays a critical role in analyzing the data because removing the unwanted data and analyzing it as the further steps are depended on this step itself. Here the data is given in the excel sheet. Importing the excel sheet into excel reader and exploring the data in the Data Exploartion. Here the missing values are replaced with the median values. Used the Column Filter where the State no and county no column has been removed because this is unique primary keys. Employed an equal-size sampler to address the issue of class imbalance. Utilized a column filter to eliminate specific columns such as the PCI, Medium Family Income, Percent White, Percent Black, Percent Other, No of Churches, No of Church Members, Ballot Type, Population. Employed a number-to-string filter to convert the target variables into string values.

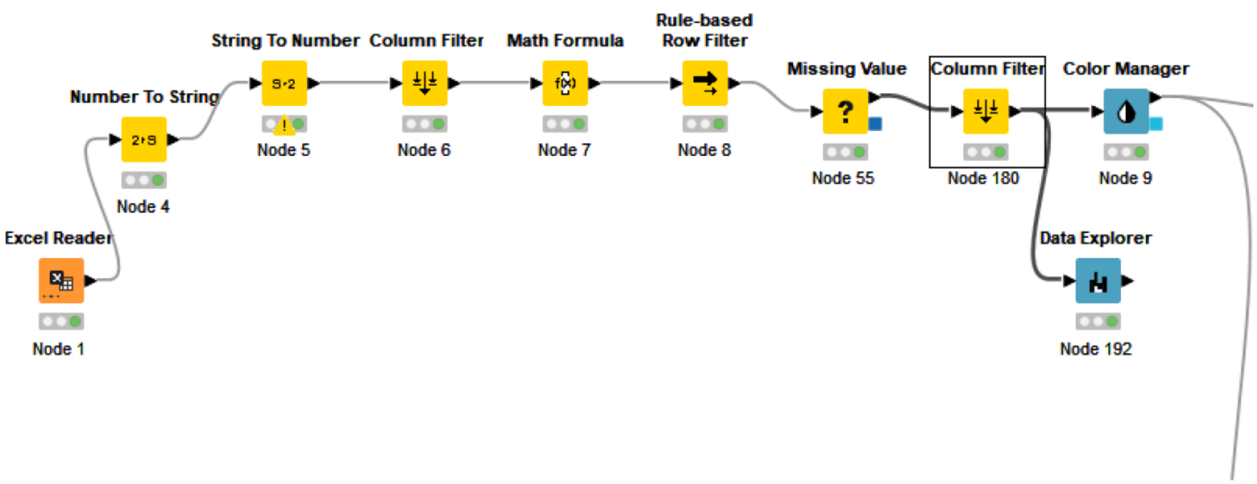

Figure 3: Data Pre-Processing done through Data Explorer, Column Filter and Color Manager nodes.

Table "default" - Rows: 935 Spec - Columns: 19 Properties Flow Variables

Row ID	D FOR	D AGAINST	D TOTAL ...	S DEPEN...	D SIZE O...	D POPUL...	D PERCE...	D PERCE...	D PERCE...	D POVER...	D UNEMP...	D AGE LE
Row0	42,385	22,992	65,377	1	1,192	223.29	0.496	0.504	0.094	10.4	5	80,600
Row3	957	856	1,813	1	1,349.4	3.99	0.51	0.49	0.214	16.9	5	1,647
Row4	942	1,129	2,071	0	2,555.9	1.77	0.495	0.505	0.217	19	1	1,192
Row5	905	646	1,551	1	1,514	3.32	0.52	0.48	0.159	20.4	3	1,401
Row8	490	422	912	1	1,781.5	1.34	0.509	0.491	0.207	11.6	1	775
Row9	1,923	1,140	3,063	1	395.5	19.29	0.519	0.481	0.042	9.5	3	2,058
Row10	1,063	1,343	2,406	0	1,287.3	5.8	0.497	0.503	0.021	33.9	11	2,694
Row11	624	404	1,028	1	1,227	2.59	0.5	0.5	0.003	34.6	10	945
Row14	3,645	3,918	7,563	0	1,142.2	18.43	0.493	0.507	0.192	17.8	6	5,393
Row16	322	320	642	1	1,067	1.41	0.508	0.492	0.251	14.5	4	431
Row17	11,519	8,261	19,780	1	840.2	73.32	0.502	0.498	0.094	3.2	3	19,411
Row18	3,006	2,064	5,070	1	1,688	13.16	0.527	0.473	0.053	7.5	3	5,825
Row19	2,179	1,205	3,384	1	1,850.9	5.27	0.501	0.499	0.09	10.4	4	3,083
Row20	58,502	38,892	97,394	1	2,126.7	186.81	0.502	0.498	0.172	6.9	7	115,224
Row22	4,646	3,744	8,390	1	2,947.5	10.29	0.509	0.491	0.114	9.3	4	8,678
Row24	2,350	1,440	3,790	1	1,849.8	4.33	0.531	0.469	0.077	9.3	3	2,112
Row26	236	177	413	1	1,117.8	0.42	0.527	0.473	0.195	13.9	2	87
Row28	431	269	700	1	1,613.3	0.99	0.533	0.467	0.045	10	2	429
Row29	86,173	64,270	150,443	1	772.2	569.43	0.493	0.507	0.098	5.8	3	121,829
Row30	472	413	885	1	1,771.1	0.95	0.488	0.512	0.265	13.8	2	501
Row31	1,552	1,464	3,016	1	2,161	3.29	0.495	0.505	0.387	15.2	1	2,169
Row32	1,063	631	1,694	1	376.9	15.98	0.516	0.484	0.074	15.7	6	1,788
Row33	5,386	4,210	9,596	1	1,692.1	19.18	0.504	0.496	0.13	12.3	5	8,859
Row35	2,405	1,348	3,753	1	4,773	2.88	0.487	0.513	0.086	26.2	8	3,725
Row36	1,023	776	1,799	1	2,586.3	1.75	0.49	0.51	0.293	17.9	2	1,249
Row37	4,102	2,533	6,635	1	1,838.6	9.52	0.488	0.512	0.268	14.9	3	4,984
Row39	264	160	424	1	875.8	0.63	0.514	0.486	0.243	13.1	6	129
Row40	1,886	1,320	3,206	1	4,742.5	2.4	0.506	0.494	0.18	11.1	5	3,835
Row41	2,363	2,534	4,897	0	2,036.9	9.19	0.487	0.513	0.165	20.2	7	6,093
Row42	4,110	3,978	8,088	1	2,240.7	10.92	0.488	0.512	0.2	14.2	6	7,002
Row43	3,832	2,779	6,611	1	1,285.5	17.06	0.49	0.51	0.316	16	4	6,766
Row45	529	603	1,132	0	542.1	4.25	0.505	0.495	0.127	9.6	9	590
Row47	976	954	1,930	1	687.7	6.09	0.472	0.528	0.45	14.1	1	1,141
Row49	2,280	1,731	4,011	1	1,640.5	8.12	0.488	0.512	0.307	21	5	4,375
Row50	26,991	14,514	41,505	1	2,388.8	51.51	0.484	0.516	0.135	20.2	7	34,270

Figure 4: Filtered Data after go through the Column Filter.

Application of Modeling Techniques:

This step also plays a vital role in Data Mining. Applying the Data model according to the data present is the key step in designing a model. Here, according to the data I am using six distinct data models namely ANN, DT, RF, NB, BT, and SVM.

Artificial Neural Networks(ANN):

ANN is a modeling technique inspired from Human Nervous system. It represents the data by a physical phenomenon or from a decision process. There is a unique feature for this modeling is to establish a relationship between the dependent variables and independent variables that extracts complex knowledge from the data sets.

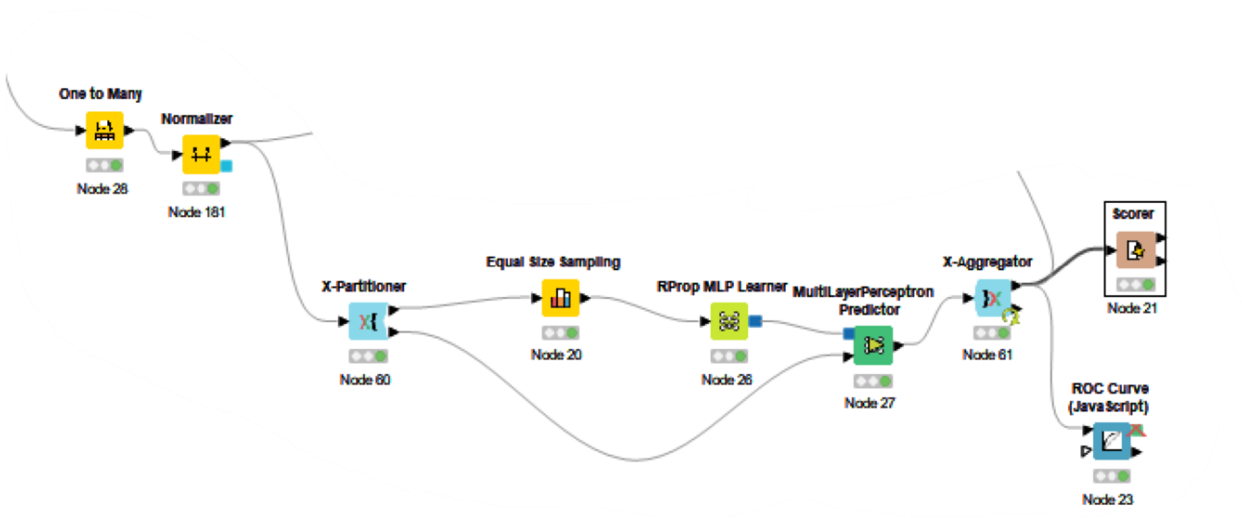

Figure 5: Designing Model for ANN

Support Vector Machine(SVM):

This extensively utilized machine learning algorithm partitions data points into distinct categories by creating a hyperplane or a series of hyperplanes within a multi-dimensional space. It is applicable to tasks involving both classification and regression.

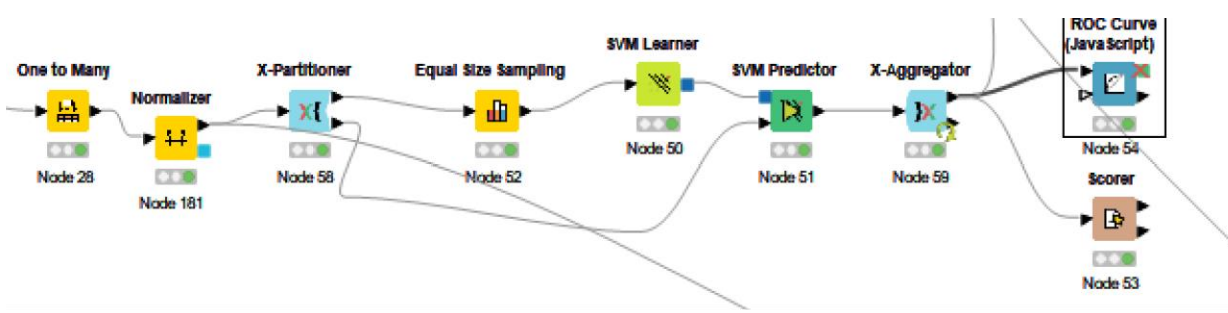

Figure 6: Designing Model for SVM

Decision Tree:

This algorithm is most widely used for classification and regression models. It splits the data into subsets based on the features of Data set and creates a tree-like structure of decision to predict the target variable. This process is repeated recursively for each subset until the criteria is met.

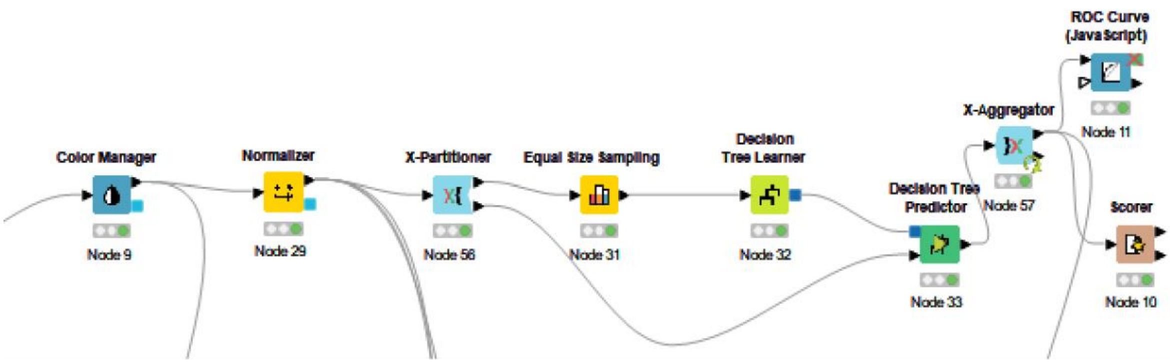

Figure 7: Designing Model for Decision Tree

Random Forests:

This approach involves combining multiple decision trees to enhance the accuracy and robustness of predictions, constituting an ensemble method. It is constructed on different subsets of Datasets. During the training process, each decision tree is trained independently with its own subset of the data set. Once all the trees are trained, they are combined to do the final prediction. This is made by taking the majority of all individual tree predictions.

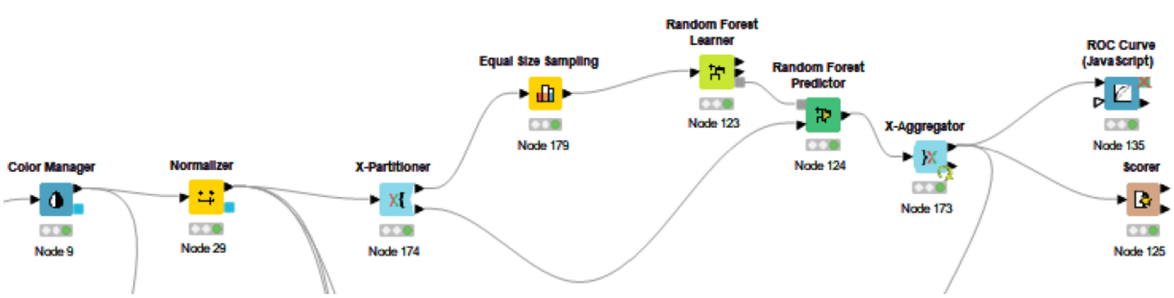

Figure 9: Designing Model of Random Forests

Boosted Trees:

Boosted trees are a technique that assembles numerous weaker models to form a robust model. The fundamental concept behind boosting trees is to progressively introduce new models to the ensemble, with each one aimed at rectifying the mistakes of the preceding models. The ultimate model is a weighted combination of all the individual models. This procedure is carried out iteratively, with each subsequent model concentrating on addressing the errors of the prior models.

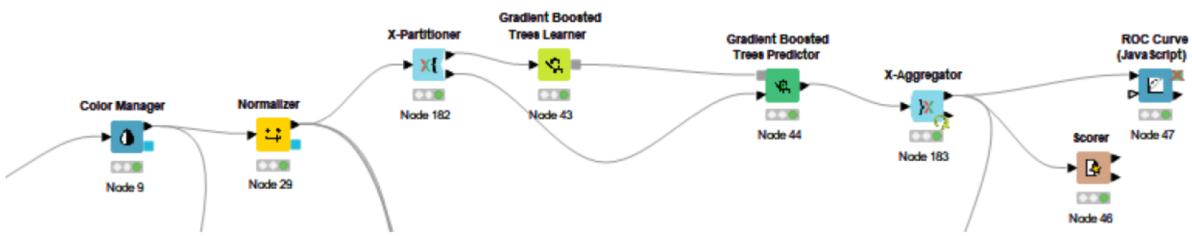

Figure 10: Designing model for Boosted Trees

Naive Bayes:

This classification approach aimed to create a model for how inputs are distributed in a specific class and assign them to instances of a problem. Unlike the models we used earlier, the naive Bayes classifier makes a strong assumption of attribute independence. This means it assumes that the value of a certain attribute doesn't depend on other attributes in the given context. This assumption makes it relatively easy to put into practice and computationally efficient. A significant advantage of the naive Bayes classifier is that it doesn't need a large amount of training data to estimate parameters and perform classification. This makes it particularly useful in situations where there's only a limited amount of available data.

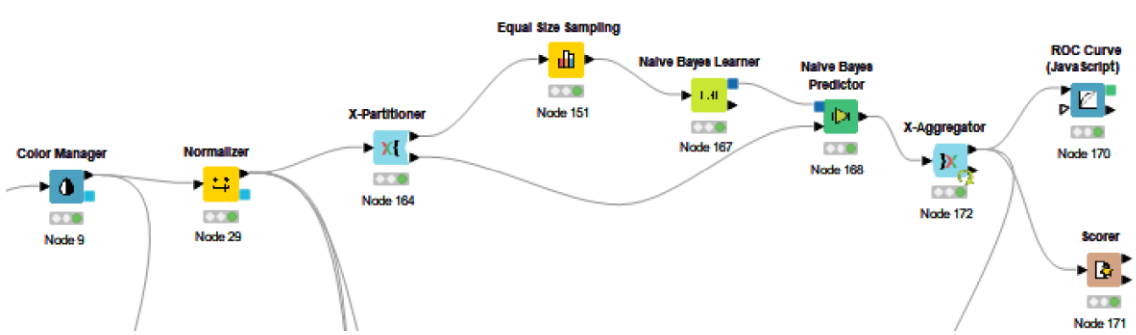

Figure 11: Designing model for Naïve Bayes

Evaluating the Models:

For every model we have some particular terms to evaluate the models they are accuracy, Sensitivity, specificity and roc on curve.

Model	Accuracy %	Sensitivity	Specificity	ROC on Curve
ANN	97.54	0.992	0.963	0.995
RF	86.84	0.879	0.861	0.949
DT	84.27	0.831	0.852	0.847
NB	60.53	0.891	0.902	0.814
SVM	69.30	0.708	0.829	0.758
BT	92.72	0.904	0.944	0.983

Table 1: Showing Accuracy, Sensitivity, Specificity and ROC Value.

Giving all the testing data to ROC Curve to know the best model among them. Combined all the testing data to a Column Appender and then connected to ROC Curve.

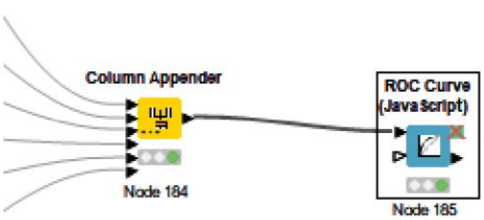

Figure 12: Inputs to Column Appender and ROC Curve.

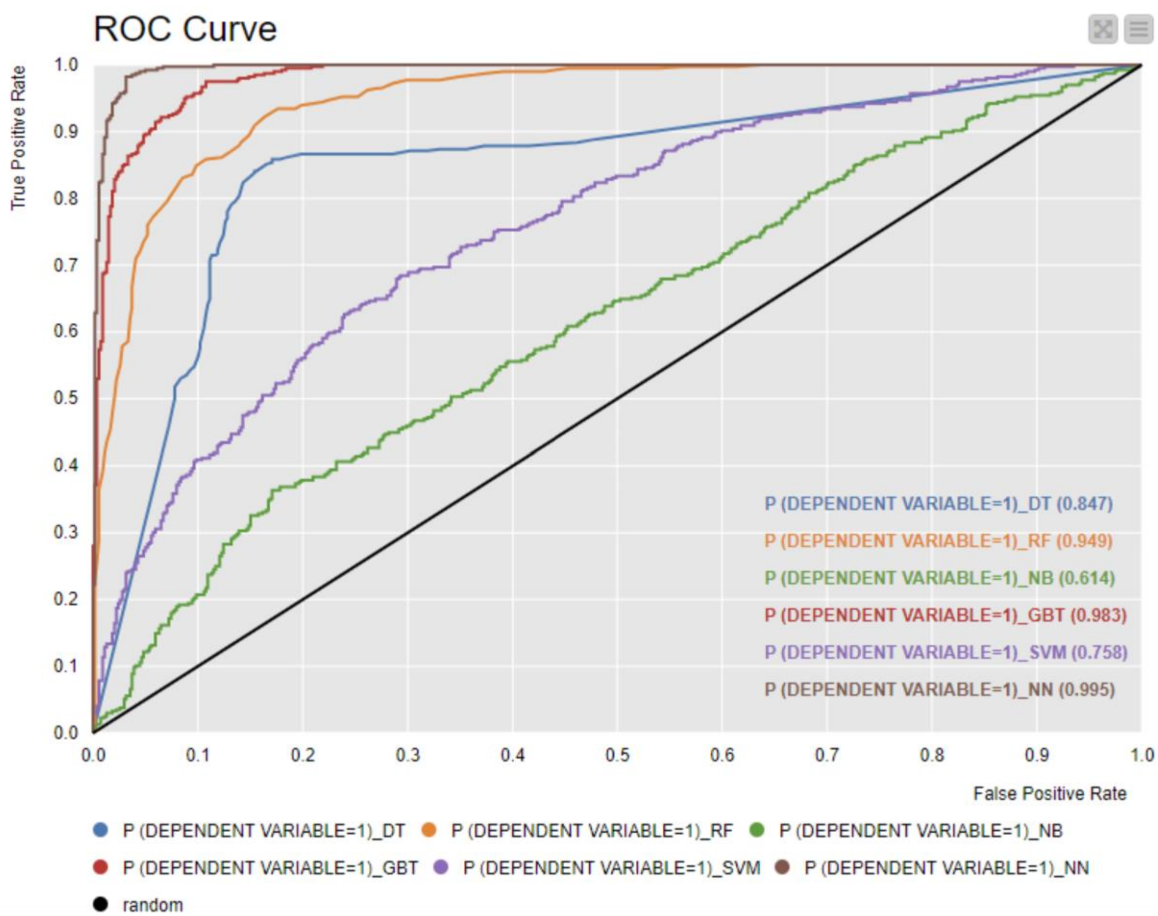

Figure 13: ROC Curve of all models

Table "spec_name" - Rows: 2 | Spec - Columns: 2 | Properties | Flow Variables

Row ID	1	0
1	393	3
0	20	519

Figure 14: Confusion Matrix of ANN

Table "spec_name" - Rows: 2 | Spec - Columns: 2 | Properties | Flow Variables

Row ID	1	0
1	348	48
0	75	464

Figure 15: Confusion Matrix of RF

Table "spec_name" - Rows: 2 | Spec - Columns: 2 | Properties | Flow Variables

Row ID	1	0
1	329	67
0	80	459

Figure 16: Confusion Matrix of DT

Table "spec_name" - Rows: 2 | Spec - Columns: 2 | Properties | Flow Variables

Row ID	1	0
1	80	316
0	53	486

Figure 17: Confusion Matrix of NB

File Edit Hilite Navigation View

Table "spec_name" - Rows: 2 | Spec - Columns: 2 | Properties | Flow Variables

Row ID	1	0
1	201	195
0	92	447

Figure 18: Confusion Matrix of SVM

Table "spec_name" - Rows: 2 | Spec - Columns: 2 | Properties | Flow Variables

Row ID	1	0
1	358	38
0	30	509

Figure 19: Confusion Matrix of GBT

Deployment

As of now we are not doing any deployment of the model into the real world.

Conclusion

To assess public sentiment on the legalization of gambling, I developed six models. The evaluation of these models involved the utilization of a confusion matrix and ROC curve values. Among all the constructed models, the Neural Network exhibited an impressive 97% accuracy, a sensitivity of 0.992, and a specificity of 0.963.

However, the accuracy of the Random Forest (RF) and Decision Tree (DT) models appeared to demonstrate overfitting tendencies. Generally, neural networks dynamically adjust their weights to minimize the error between input and output values. Similarly, in this project, the neural network exhibited signs of overfitting. Despite this, considering its commendable accuracy and robust specificity and sensitivity values, we have decided to proceed with the Neural Network model.

This model excels in predicting true negatives but may have limitations in predicting true positives. Nonetheless, these predictive capabilities are deemed sufficiently satisfactory for the purposes of this analysis.