

MSIS-4673-26637 – Predictive Analytics Technologies

Section – In - class

Homework Assignment #5

Gaming Ballot Prediction

KNIME Data Mining II

Due Date

April 21, 2024

By

Rhoda Alawiye

CONTENTS

EXECUTIVE SUMMARY	3
INTRODUCTION	3
BUSINESS UNDERSTANDING	4
DATA UNDERSTANDING	4
DATA PREPARATION.....	6
MODELING	9
EVALUATION.....	13
DEPLOYMENT	15
CONCLUSION.....	16

EXECUTIVE SUMMARY

This research paper delves into the voting behavior across various counties in the US, employing an in-depth analysis to evaluate the performance of five distinct prediction models to forecast the results of a gaming legalization vote. Employing the KNIME data analytics platform for the analysis, the provided dataset was preprocessed, and prediction models were constructed using decision trees, random forests, artificial neural networks, logistic regression, and Sector Vector Machines.

The results of the analysis showed that the neural network model surpassed the other models in accurately predicting the outcome of the gaming legalization vote. Furthermore, the study delves into the influential factors shaping voting behavior, encompassing demographic (PCI, population density and size of county) and economic (unemployment rate, poverty level and medium family income) elements among others. In the context of gaming legalization, the study identifies certain variables like percentage of white population and number of churches as pivotal in projecting ballot turnouts. Industries aiming for legalization should prioritize addressing these variables to bolster their prospects of success.

The research provides a comparative evaluation of the prediction models and based on their sensitivity performance, the neural network model emerging as the most effective, followed by the random forest and decision tree models. These insights offer valuable guidance to policymakers, stakeholders, and decision-makers in the gaming industry, shedding light on the interplay between different models and influencing factors in predicting voting outcomes. This study not only contributes a rigorous analysis of voting behavior and prediction model performance but also offers actionable insights that can empower policymakers and industry stakeholders to make informed decisions in the realm of gaming legalization.

INTRODUCTION

The debate over legalizing gaming has stirred controversy across many nations worldwide. This umbrella term encompasses casino gambling, sports betting, and lottery games, propelling it into a multi-billion-dollar industry on a global scale. Recognizing the potential for economic growth through increased tax revenue, tourism, and job opportunities, numerous countries have crafted legislation to legalize and oversee gaming activities.

The decision to legalize gaming is fraught with complexities, sparking debates over its social and economic ramifications. Critics often highlight concerns such as gambling addiction, money laundering, and links to organized crime. Balancing these apprehensions against the anticipated economic benefits can make the formulation and implementation of gaming legalization policies a challenging endeavor. While the potential negative social impacts of gaming legalization often dominate public discourse, proponents argue that the economic advantages cannot be ignored. The pivotal decision to legalize gaming typically hinges on the outcome of a vote. Therefore, understanding the underlying factors that shape voting behavior can equip policymakers and stakeholders with the insights needed to make well-informed choices.

In this study, the voting behavior across several counties in the US will be delved into, leveraging various prediction models to forecast the results of gaming legalization votes and comparison of the models' performance. Utilizing the data analytics platform KNIME for data preprocessing and model development, the analysis aims to shed light on influential voting factors and assess the predictive performance of different models in anticipating gaming legalization outcomes.

BUSINESS UNDERSTANDING

This paper aims to explore the complexities surrounding gaming legalization, recognizing that perspectives on this issue can be diverse and multifaceted. For businesses operating within the gaming sector, legalization often signifies a wealth of growth opportunities, expanding the market and fueling demand for gaming services. However, this potential growth comes with its own set of challenges, including heightened competition, stringent regulatory requirements, and the imperative to address social responsibility concerns. At a more localized level, subtle factors can exert significant influence over the outcome of legalization ballots. These may encompass the prevailing religious attitudes towards gaming, the demographic age distribution within the community, among other nuanced considerations.

From a business standpoint, grasping the intricacies outlined in this paper can provide stakeholders with valuable insights into steering counties towards a favorable vote on gaming legalization. Our analysis aims to shed light on the diverse factors that influence the prospects of gaming legalization across various counties, offering a comprehensive understanding that can inform strategic decision-making and advocacy efforts.


DATA UNDERSTANDING

The gaming ballot dataset was provided for this analysis and had a total of 1287 unique records and a total of 31 variables of which are categorical data variables and numeric variables that consist of several useful independent variables like information about the age groups, number of votes, number of churches in the counties and county population. The dataset comprised of demographic data, socio-economic data and other necessary data factors that were relevant to the analysis.

Understanding the various variables in the data would help provide insights into how to preprocess them and select them for model building. The dependent variable in the analysis indicates the voting outcome for gaming legalization, coded as 1 for “yes” and 0 for “no”. **Figure 1** and **Figure 2** depict the exploratory data analysis results for the provided numeric and nominal variables prior to data cleaning and preprocessing operations. **Figure 3** shows the correlation between variables before the data cleaning and preprocessing. This was done to help gain insights into variables that are needed for the analysis and variables to drop if they are not significant for the data analysis. The primary objective is to evaluate the performance of various classifiers in predicting these voting outcomes and to identify the most influential factors shaping these decisions.

Column	Column	Minimum	Maximum	Mean	Deviation	Variance	Skewness	Kurtosis	Sum	zeros	missings	NaN	+∞	-∞
State No	<input type="checkbox"/>	1	18	10.399	5.047	25.473	-0.260	-0.987	13384	0	0	0	0	0
County No	<input type="checkbox"/>	1	251	57.272	54.349	2953.825	1.652	2.307	73709	0	0	0	0	0
FOR	<input type="checkbox"/>	44	161415	6460.944	14766.946	218062690.440	5.195	34.671	8315235	0	0	0	0	0
AGAINST	<input type="checkbox"/>	15	121925	7330.511	15099.603	227998002.275	4.037	19.094	9434368	0	0	0	0	0
TOTAL CASTE	<input type="checkbox"/>	59	245523	13791.455	28431.805	808367551.443	4.103	19.848	17749603	0	0	0	0	0
DEPENDENT VARIABLE	<input type="checkbox"/>	0	1	0.426	0.495	0.245	0.300	-1.913	548	739	0	0	0	0
BALLOT TYPE	<input type="checkbox"/>	1	2	1.542	0.498	0.248	-0.170	-1.974	1985	0	0	0	0	0
POPULATION	<input type="checkbox"/>	327	1206243	52475.283	109737.626	12042346591.789	4.739	30.161	67535689	0	0	0	0	0
PCI	<input type="checkbox"/>	5720	44518	16652.772	3827.951	14653211.848	1.057	3.722	21432117	0	0	0	0	0
MEDIUM FAMILY INCOME	<input type="checkbox"/>	12225	55643	27398.726	6403.965	41010771.664	0.880	1.170	35262160	0	0	0	0	0
SIZE OF COUNTY	<input type="checkbox"/>	120.800	6347.800	1000.488	778.794	606520.375	2.721	10.619	1287627.800	0	0	0	0	0
POPULATION DENSITY	<input type="checkbox"/>	0.296	3404.480	81.573	221.759	49176.846	8.004	88.571	104985.066	0	0	0	0	0
PERCENT WHITE	<input type="checkbox"/>	0.026	0.998	0.804	0.189	0.036	-1.352	1.804	1035.251	0	0	0	0	0
PERCENT BLACK	<input type="checkbox"/>	0	0.798	0.081	0.134	0.018	2.138	4.249	103.902	160	0	0	0	0
PERCENT OTHER	<input type="checkbox"/>	0	0.973	0.115	0.167	0.028	2.581	7.253	147.816	1	0	0	0	0
PERCENT MALE	<input type="checkbox"/>	0.246	0.671	0.494	0.020	0.000	0.964	33.228	635.289	0	0	0	0	0

Figure 1a: Showing the Information about the numeric variables before preprocessing

Column	Exclude Column	No. missings	Unique values	All nominal values	Frequency Bar Chart
NO OF CHURCH MEMBERS	<input type="checkbox"/>	0	>1000	1752, 1593, 4871, 1239, 1244, [...], 1693, 9172, 4719, 2533, 5809	Not all nominal values calculated.
PERCENT CHURCH MEMBERS OF POPULATION	<input type="checkbox"/>	0	147	0.45, 0.48, 0.5, 0.46, 0.37, [...], 0.267820424948595, 0.115580448065173, 0.127114967462039, 0.9, 0.158346042682081	

Showing 1 to 2 of 2 entries

Figure 2a: Showing the Information about the nominal variables before preprocessing

DATA PREPARATION

Data preparation is essential for the accuracy and ease of the data analysis process. This step entails steps taken to collect, clean, and organize the data for analysis. The dataset was a real-world gaming ballot dataset encompassing information of voters across various categories over time for selected counties in the US.

The dataset required several data cleaning and manipulation procedures. The exploratory data analysis already showed immediately that variables “State No” and “County No” were not relevant because they had no information value needed for the prediction models but only served as column identifiers, hence were dropped. Data format consistency was checked for and two variables had the wrong format. Variables “No of church members” and “Percent church members of population” were represented as string nominal values but the number of unique variables each had showed that they were numeric variables, and they were both converted using the string manipulation node and the string to number node. **Figure 1** shows that the dependent variable, the ballot type and the MSA which were nominal variables given the values of their minimum, maximum and histogram, the rule engine node was used to transform them to nominal variables. The results of these changes are seen in **Figure 2b**. The search for missing values and incorrectly entered data points were investigated and there were two variables that had data points that were inconsistent with all other data points, the affected rows were dropped since they were few and would not significantly affect the data quality.




There were other variables that were dropped after carefully studying each of the variables and using the results from the correlation matrix and the exploratory data analysis for guidance. Variable “Population” was dropped because it was highly correlated with population density, PCI, number of church members and the Age distributions since population was used to derive these variables. Hence, another reason why variable population was dropped was because it was well represented among those other variables. The variables “For”, “Against” and “Total Caste” were also dropped because they were already represented as the dependent variable and using them would have negatively affected the model accuracy. “No of Older”, “No of younger”, and “Percent Minority” were dropped too since the information they have were already included in the age distribution variables, especially for the “Percent Minority” which is assumed to already be represented by the “Age less than 18” variable. At first, only “percent of church members of the population” was dropped and after seeing the performance of the models, variable “No of church members” was also dropped since their information is already in a way a subset of the “No of churches” variable.

Figure 1a and **Figure2a** show the distribution for the numeric and nominal variables before data cleaning and preprocessing while **Figure 1b** and **Figure 2b** shows the distribution after

preprocessing. **Figure 3a** and **Figure 3b** show the linear correlation matrix for the data variables before and after variable selection.

+ PCI	<input type="checkbox"/>	5720	44518	16652.772	3827.951	14653211.848	1.057	3.722	21432117	0	0	0	0	0
+ MEDIUM FAMILY INCOME	<input type="checkbox"/>	12225	55643	27398.726	6403.965	41010771.664	0.880	1.170	35262160	0	0	0	0	0
+ SIZE OF COUNTY	<input type="checkbox"/>	120.800	6347.800	1000.488	778.794	606520.375	2.721	10.619	1287627.800	0	0	0	0	0
+ POPULATION DENSITY	<input type="checkbox"/>	0.296	3404.480	81.573	221.759	49176.846	8.004	88.571	104985.066	0	0	0	0	0
+ PERCENT WHITE	<input type="checkbox"/>	0.026	0.998	0.804	0.189	0.036	-1.352	1.804	1035.251	0	0	0	0	0
+ PERCENT BLACK	<input type="checkbox"/>	0	0.798	0.081	0.134	0.018	2.138	4.249	103.902	160	0	0	0	0
+ PERCENT OTHER	<input type="checkbox"/>	0	0.973	0.115	0.167	0.028	2.581	7.253	147.816	1	0	0	0	0
+ PERCENT MALE	<input type="checkbox"/>	0.246	0.671	0.494	0.020	0.000	0.964	33.228	635.289	0	0	0	0	0
+ PERCENT FEMALE	<input type="checkbox"/>	0.335	0.754	0.507	0.020	0.000	-0.894	32.927	651.938	0	0	0	0	0
+ NO OF CHURCHES	<input type="checkbox"/>	1	738	58.703	68.467	4687.733	3.832	21.986	75551	0	0	0	0	0
+ POVERTY LEVEL	<input type="checkbox"/>	3.200	49.900	17.409	7.214	52.046	1.061	1.750	22405.600	0	0	0	0	0
+ UNEMPLOYMENT RATE	<input type="checkbox"/>	0.600	38.500	6.051	3.410	11.630	2.184	11.515	7787.000	0	0	0	0	0
+ AGE LESS THAN 18	<input type="checkbox"/>	87	353009	14419.316	30487.062	929462176.083	4.949	33.314	18557660	0	0	0	0	0
+ AGE24	<input type="checkbox"/>	16	137916	5204.805	11700.168	136893934.277	5.094	35.794	6698584	0	0	0	0	0
+ AGE44	<input type="checkbox"/>	108	437537	16497.123	37260.004	1388307877.086	5.049	33.835	21231797	0	0	0	0	0
+ AGE64	<input type="checkbox"/>	71	202133	10298.945	20948.274	438830197.433	4.448	24.833	13254742	0	0	0	0	0
+ AGE OLDER THAN 64	<input type="checkbox"/>	21	122335	6754.410	13508.404	182476983.264	4.439	23.454	8692926	0	0	0	0	0

Figure 2b: Showing the Information about the numeric variables after preprocessing

Column	Exclude Column	No. missings	Unique values	All nominal values	Frequency Bar Chart
Dependent Variable_Nom	<input type="checkbox"/>	0	2	No, Yes	
Ballot Type_Nom	<input type="checkbox"/>	0	2	Wagering, Gambling	
Metropolitan Stat Area	<input type="checkbox"/>	0	2	Yes, No	

Showing 1 to 3 of 3 entries

Figure 2b: Showing the Information about the nominal variables after preprocessing

Correlation Matrix - 3:56 - Linear Correlation

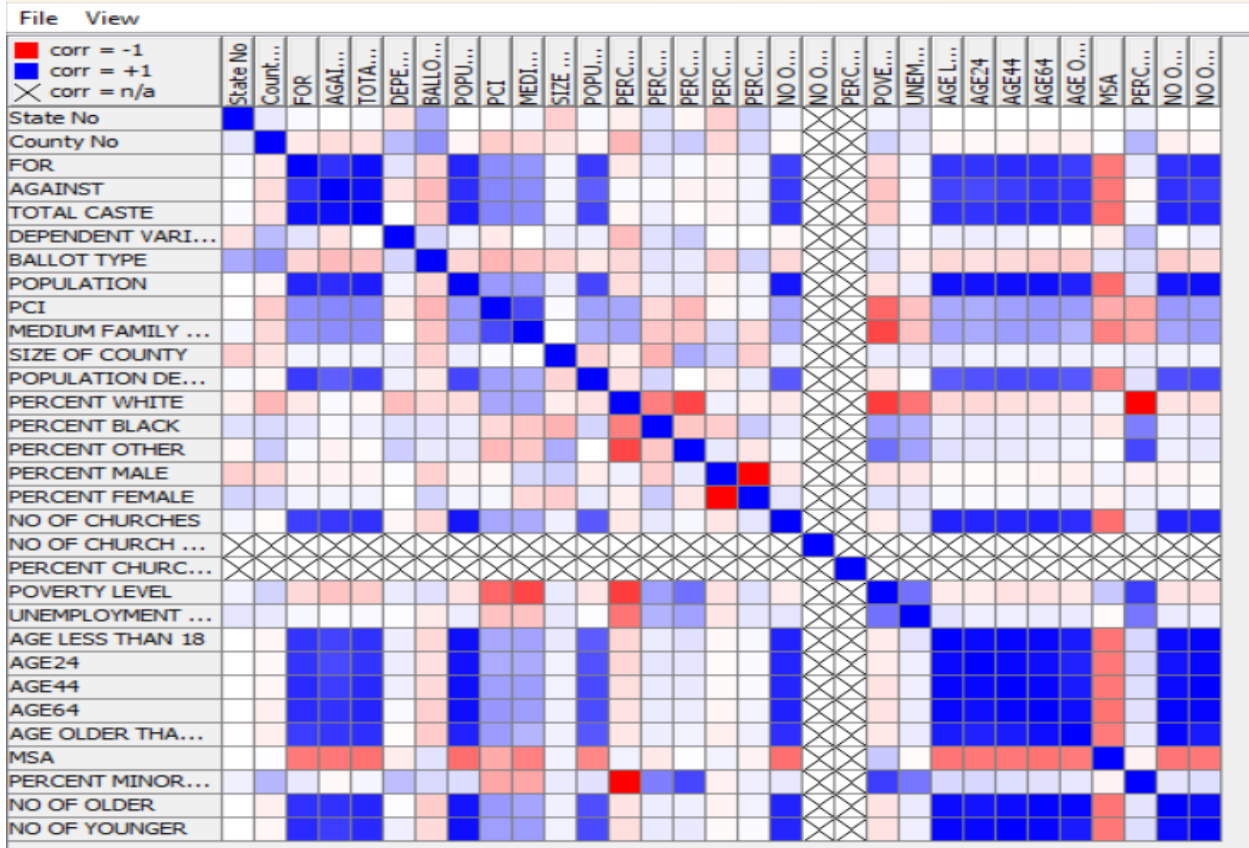


Figure 3a: Showing the linear correlation matrix before variable selection

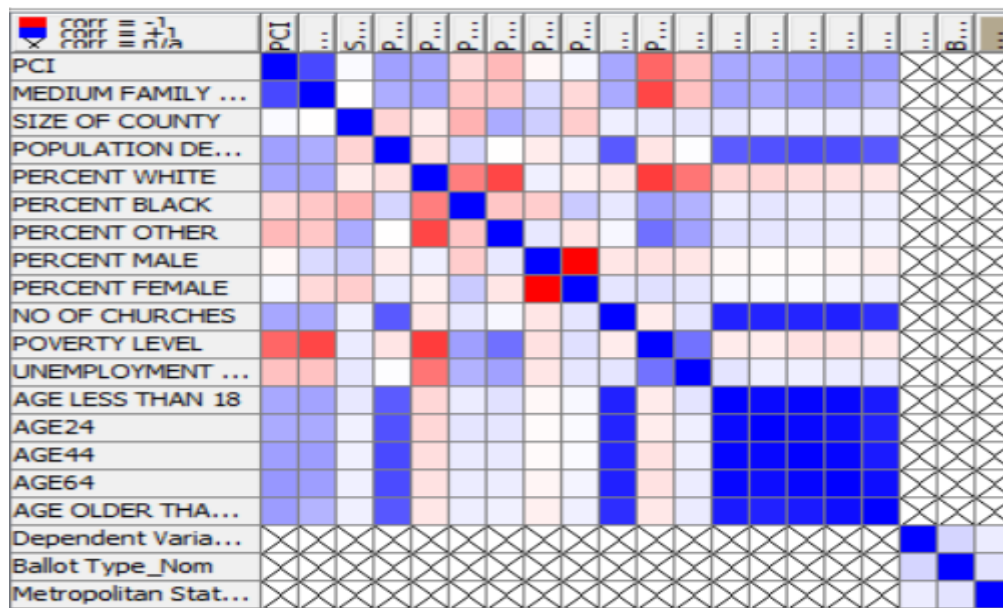


Figure 3b: Showing the linear correlation matrix after variable selection

MODELING

The data was first partitioned into the learning (training data) and the predictor (test data). Due to the likelihood of imbalance in the target variable as shown in **Figure 4**, the dataset was balanced using the “exact sampling” method in KNIME to ensure that the target variable was balanced, which was to ensure the assertiveness of the predictions. Data normalization task was conducted for gaussian distribution related models. Five machine learning algorithms were used for the analysis. The algorithms that were used are: Decision tree, Random Forest Classification, Artificial Neural Network, Logistic Regression and Sector Vector Machines. The first step of the data modeling after the data cleaning and preprocessing was to have the dataset partitioned with the 70/30 split for the training and validation for the decision tree and the random forest models. The data partitioning was also included in the other models, but other operation nodes were used prior to the use of the data partition node. The performance of the models was compiled and compared using performance metrics: Sensitivity, Specificity, Accuracy and AUC values.

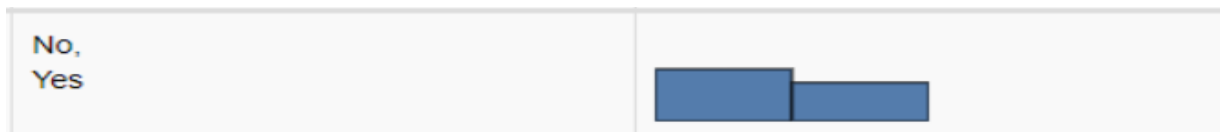


Figure 4: showing the distribution of the target variable “dependent variable”

Decision Tree

The decision tree classification of the target labels based on the values of other characteristics in the dataset, was utilized. It divides the dataset depending on the relevance of each variable, which is then utilized to produce predictions about the target variables. **Figure 7** depicts the results of the decision tree split.

Confusion Matrix - 3:20 - Scorer		
File	Hilite	
Dependent...	Yes	No
Yes	112	53
No	78	144
Correct classified: 256		
Wrong classified: 131		
Accuracy: 66.15%		
Error: 33.85%		
Cohen's kappa (κ): 0.321%		

Figure 5: showing the decision tree confusion matrix

#	RowID	TruePositives Number (integer)	FalsePositives Number (integer)	TrueNegatives Number (integer)	FalseNegatives Number (integer)	Recall Number (double)	Precision Number (double)	Sensitivity Number (double)	Specificity Number (double)	F-measure Number (double)	Accuracy Number (double)	Cohen's kappa Number (double)
1	Yes	112	78	144	53	0.679	0.589	0.679	0.649	0.631	0	0
2	No	144	53	112	78	0.649	0.731	0.649	0.679	0.687	0	0
3	Over...	0	0	0	0	0	0	0	0	0	0.661	0.321

Figure 6: showing the decision tree accuracy statistics table



Figure 7: Showing the Decision Tree Split

The decision tree initially split in **Figure 7** based on the percentage of white population, hinting at its potential importance in predicting voting outcomes within counties. Subsequently, the number of churches emerged as another influential variable. However, it's essential to approach these findings with caution. Statistically, white individuals often form the majority in many regions, raising concerns about classifying predictions based on race. Thus, while these variables showed significance, their impact on the voting outcome requires further scrutiny and validation. To further confirm this result, a variable of importance table was created using the random forest variable statistics and Excel for the visualization and the results shows “Unemployment Rate” as the most

important variable followed by “Age 24” age group, “Percent Other”, “Age 64” age group and “Percent White”, thereby making them the top 5 variables of importance that influences ballot outcomes as shown in **Figure 8**.

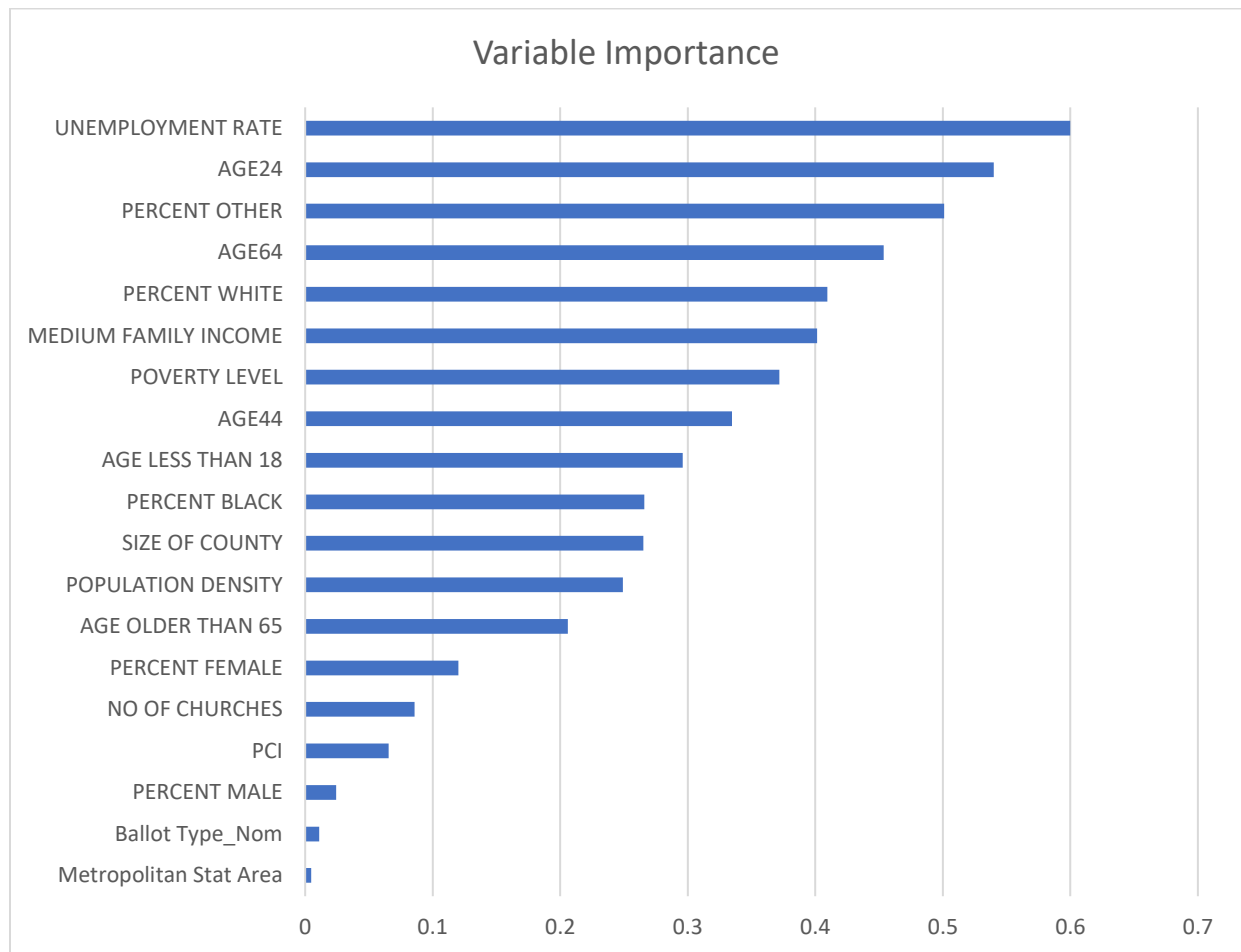


Figure 8: Showing the Variable of Importance

Random Forest

The random forest which is an ensemble learning method that uses multiple decision trees to make decisions, often provides improved predictive performance that are less prone to overfitting unlike the decision tree. **Figure 9** shows the accuracy statistics table for the random forest performance.

#	RowID	TruePositives Number (integer)	FalsePositives Number (integer)	TrueNegatives Number (integer)	FalseNegatives Number (integer)	Recall Number (double)	Precision Number (double)	Sensitivity Number (double)	Specificity Number (double)	F-measure Number (double)	Accuracy Number (double)	Cohen's kappa Number (double)
1	Yes	119	66	156	46	0.721	0.643	0.721	0.703	0.68	0.711	0.417
2	No	156	46	119	66	0.703	0.772	0.703	0.721	0.736	0.711	0.417
3	Over...	0	0	0	0	0	0	0	0	0	0.711	0.417

Figure 9: Showing the Random Forest Accuracy Statistics Table

Artificial Neural Network

An artificial neural network (ANN) is made up of several interconnected processing nodes known as neurons. ANN determines the correlations between the input data and the target label and then it uses the relationships to forecast new data. Neural networks are widely used to solve complex classification issues and may achieve extraordinarily high accuracy when it is trained on large datasets. **Figure 10** depicts the accuracy statistics table for artificial neural network performance.

#	RowID	TruePositives Number (integer)	FalsePositives Number (integer)	TrueNegatives Number (integer)	FalseNegatives Number (integer)	Recall Number (double)	Precision Number (double)	Sensitivity Number (double)	Specificity Number (double)	F-measure Number (double)	Accuracy Number (double)	Cohen's kappa Number (double)
1	Yes	122	62	160	43	0.739	0.663	0.739	0.721	0.699	0	0
2	No	160	43	122	62	0.721	0.788	0.721	0.739	0.753	0	0
3	Over...	0	0	0	0	0	0	0	0	0	0.729	0.453

Figure 10: Showing the ANN Accuracy Statistics Table

Logistic Regression

Logistic regression is used to predict binary outcome, like the customer churn dataset that we have. Linear regression finds the linear decision boundary most suitable for separating the two classes. **Figure 11** shows the accuracy statistics table for the logistic regression performance.

#	RowID	TruePositives Number (integer)	FalsePositives Number (integer)	TrueNegatives Number (integer)	FalseNegatives Number (integer)	Recall Number (double)	Precision Number (double)	Sensitivity Number (double)	Specificity Number (double)	F-measure Number (double)	Accuracy Number (double)	Cohen's kappa Number (double)
1	Yes	98	53	169	67	0.594	0.649	0.594	0.761	0.62	0	0
2	No	169	67	98	53	0.761	0.716	0.761	0.594	0.738	0	0
3	Over...	0	0	0	0	0	0	0	0	0	0.69	0.359

Figure 11: Showing the Logistic Regression Accuracy Statistics Table

Sector Vector Machines

Support Vector Machines (SVM) are employed to predict binary outcomes, like the gaming legalization voting dataset utilized in this study. SVM aims to find the optimal hyperplane that best separates the two classes within the dataset. This hyperplane is determined by maximizing the margin between the closest points of the two classes, known as support vectors. In the context of our analysis, SVM seeks to identify the decision boundary that effectively distinguishes counties that vote in favor of gaming legalization from those that do not. This classification approach is particularly valuable when dealing with datasets that may not be linearly separable, as SVM can utilize various kernel functions to transform the data into higher dimensions, making it easier to find a separating hyperplane. **Figure 12** presents the accuracy statistics table showcasing the

performance metrics of the SVM classifier in predicting the voting outcomes for gaming legalization.

#	RowID	TruePositives Number (integer)	FalsePositives Number (integer)	TrueNegatives Number (integer)	FalseNegatives Number (integer)	Recall Number (double)	Precision Number (double)	Sensitivity Number (double)	Specificity Number (double)	F-measure Number (double)	Accuracy Number (double)	Cohen's kappa Number (double)
1	Yes	91	46	176	74	0.552	0.664	0.552	0.793	0.603	0	0
2	No	176	74	91	46	0.793	0.704	0.793	0.552	0.746	0	0
3	Over...	0	0	0	0	0	0	0	0	0	0.69	0.352

Figure 12: Showing the SVM Accuracy Statistics Table

EVALUATION

The chosen models' performance was evaluated using the following metrics: sensitivity, specificity, accuracy, and the area under the receiver operating characteristic curve (AUC-ROC). Accuracy evaluates the fraction of properly identified cases, whereas AUC-ROC assesses the trade-off between sensitivity and specificity at various classification levels. **Table 1** displays the assessment findings for each model, whereas **Figure 13** depicts the ROC curve for all models. The neural network had the highest AUC-ROC value of 0.823, that is 82.3% while the Decision tree had the lowest AUC-ROC value of 0.667 that is 66.7%. Based on the accuracy measure the neural network had the highest accuracy score of 0.729 while the decision tree had the lowest accuracy score of 0.661. The AUC-ROC and Accuracy metric both gave the same performance evaluation results for the models. **Figure 14** below shows the final workflow for the analytical steps taken.

MODEL	SENSITIVITY	SPECIFICITY	ACCURACY	AUC-ROC
Decision Tree	0.679	0.649	0.661	0.667
Random Forest	0.721	0.703	0.711	0.792
Neural Network	0.739	0.721	0.729	0.823
Logistic Regression	0.594	0.761	0.69	0.756
SVM	0.552	0.793	0.69	0.733

Table 1: Showing the Model Performance Evaluation Metrics

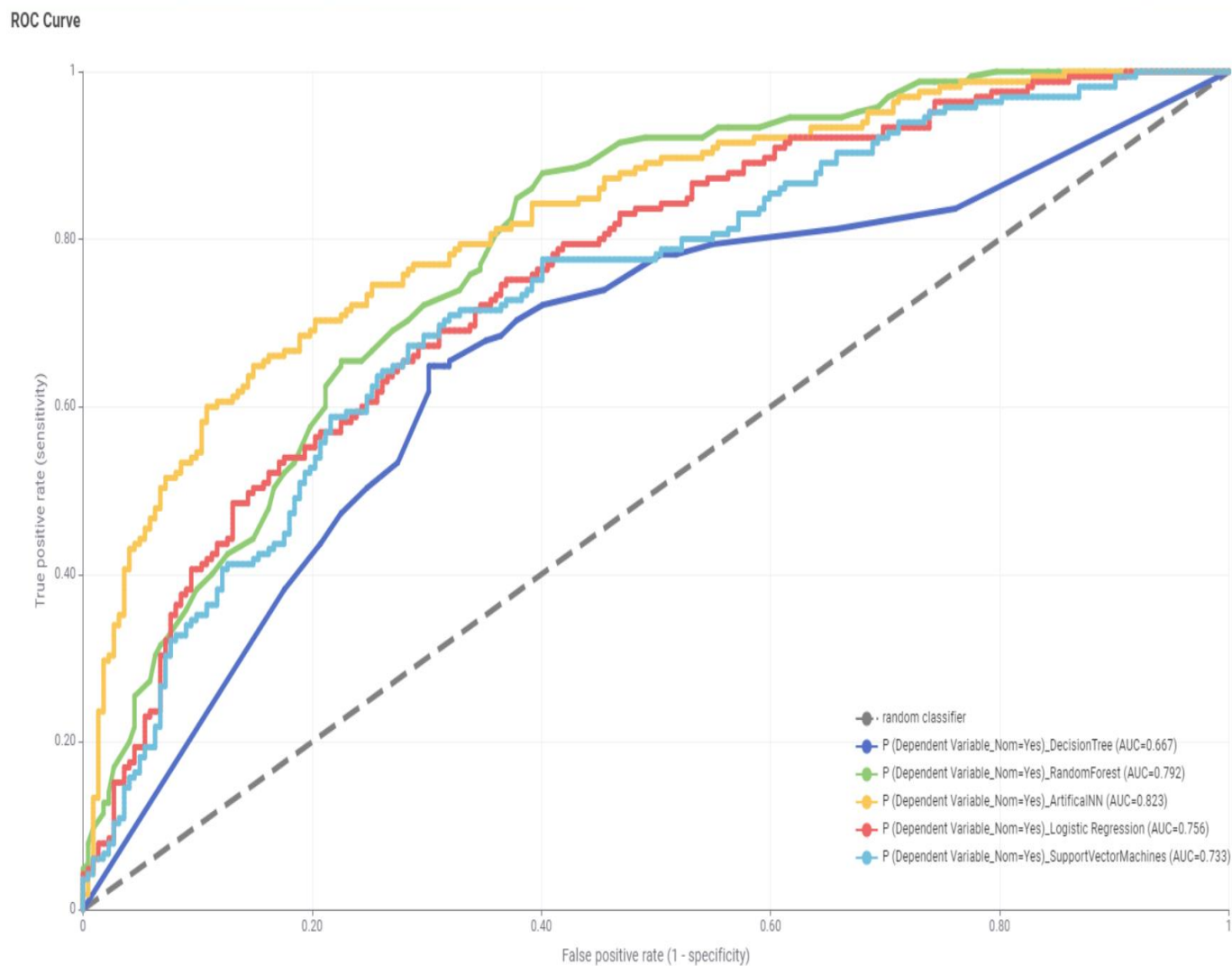


Figure 13: Showing the ROC curve for all the models.

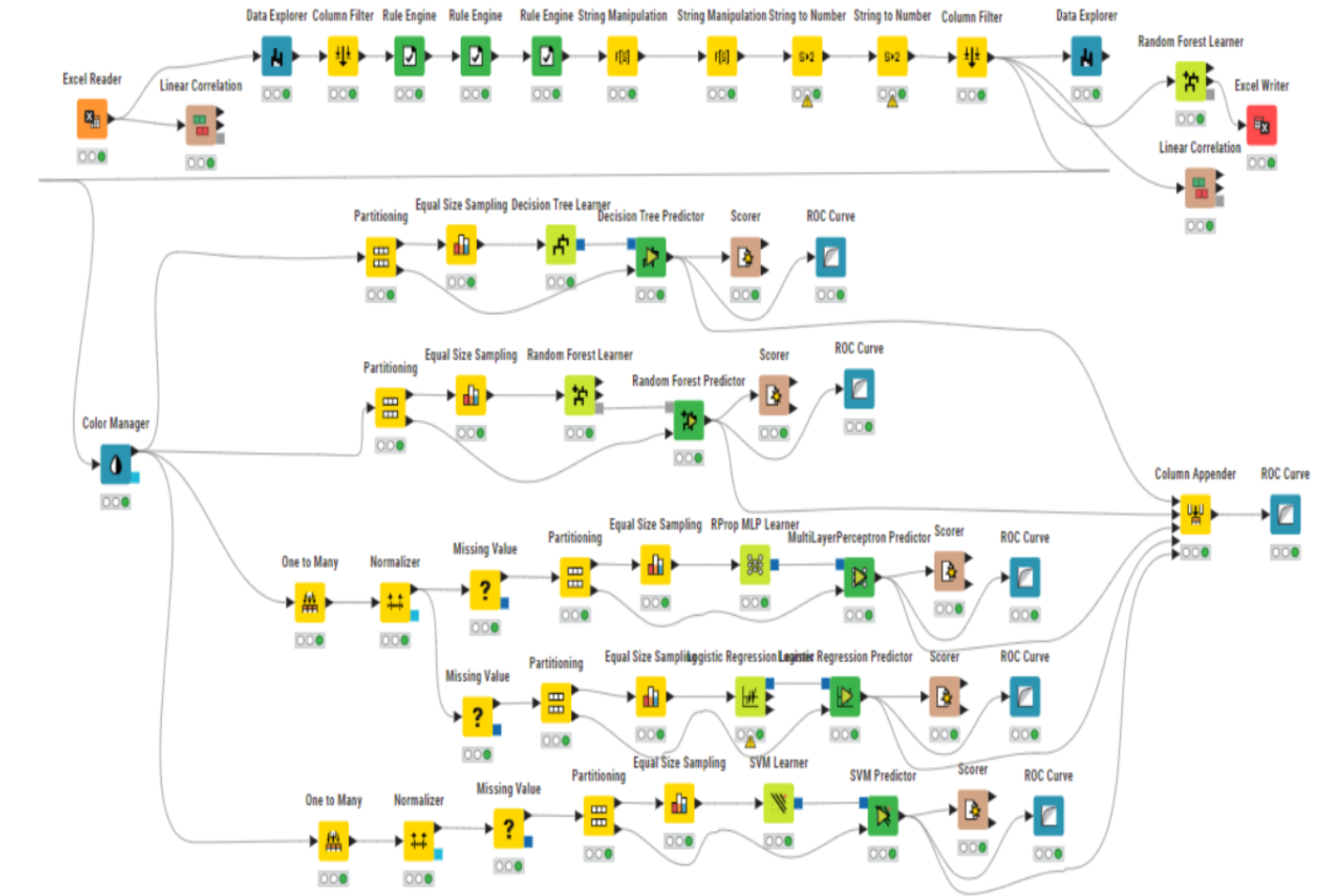


Figure 14: Showing the KNIME Workflow.

DEPLOYMENT

This research undertook a comprehensive assessment of multiple machine learning models to forecast the outcomes of gaming legalization votes, emphasizing both accuracy and AUC-ROC metrics. The results showcased the neural network model as the most adept in accurately predicting state-level voting outcomes within the study's scope. However, random forest and logistic regression models also displayed commendable predictive performance, each with its unique strengths and limitations. The analysis provides nuanced insights into the comparative advantages and drawbacks of employing different machine learning models for this prediction task, offering invaluable guidance for stakeholders in the gaming sector. Additionally, our research underscored the pivotal role played by specific variables, including the percentage of white population, number of churches, ballot type, proportion of church-affiliated individuals, and per capita income, in shaping ballot turnouts.

For industries seeking legalization in specific regions, a strategic approach may involve either addressing the influential factors that sway voting decisions or possibly redirecting efforts towards more receptive counties to optimize success rates.

CONCLUSION

In conclusion, my experience with KNIME has been largely positive, offering a comprehensive suite of pre-built machine learning models and data transformation nodes that facilitate the creation and customization of classification workflows. As familiarity with the platform grows, it is anticipated that its usability will improve, streamlining the process further. While there remains a certain comfort level with other tools like Excel and Python, the potential of KNIME as a robust tool for classification analysis is evident.

The research underscores the efficacy of employing various machine learning models, with the neural network model demonstrating superior predictive capabilities for gaming legalization voting outcomes. However, random forest and logistic regression models also showcased notable performance, each with distinct strengths and limitations. Furthermore, the analysis highlights the influential role of specific variables such as the percentage of white population, number of churches, and per capita income, in influencing voting outcomes. This insight provides valuable guidance for industries navigating the complex landscape of gaming legalization.

In summary, KNIME proves to be an invaluable asset for classification analysis, empowering users to construct accurate models and extract meaningful insights from their data, thereby contributing significantly to informed decision-making in the realm of gaming legalization.