Ballot Fortunes: Forecasting County Stance on Legalizing Gaming

By

Naveen Varma Patsamatla

Table of content

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 objective, we have utilized a dataset comprising 1287 rows and 31 columns. Our approach involves constructing machine learning models to predict the percentage of support for the legalization of gambling. Demographic, socio-economic, and other relevant factors are integral to our analysis of the conditions surrounding gambling.

In the pursuit of our objectives, we have diligently followed the CRISP-DM methodology, ensuring a robust and systematic analysis. To accomplish our goals, we explored a diverse range of machine learning models, such as Decision Tree, Random Forest, Artificial Neural Network - MLP, Logistic Regression, and SVM. A rigorous evaluation process, considering all pertinent factors, was employed to identify the model that excels in pattern recognition. Our model selection criteria were based on a comprehensive assessment of the predictive power and ROC curve values.

This report aims to offer actionable insights and recommendations, enabling businesses to make well-informed decisions regarding the support for legalizing gambling in the United States.

Phase 1: Business Understanding

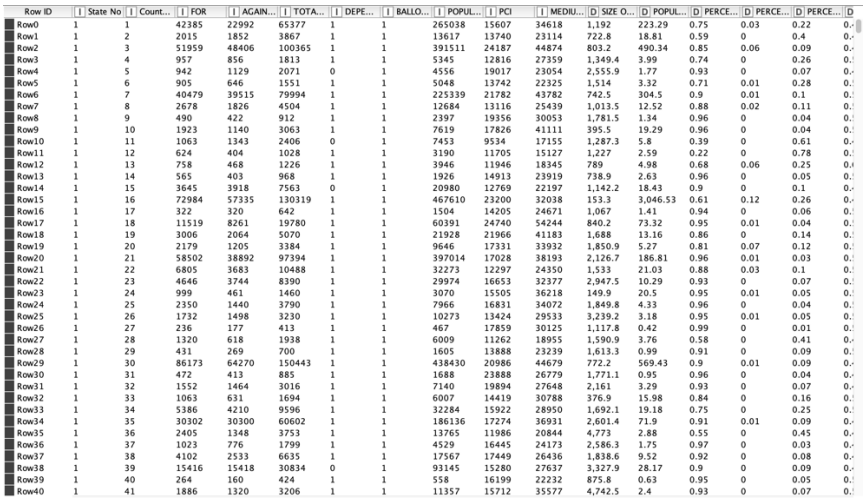
Understanding the business objectives and goals thoroughly is of utmost importance. In this context, our primary focus is on accurately predicting the stance of counties on legalizing gaming. From a data mining standpoint, the key to success lies in creating a predictive model that not only achieves high accuracy but also maintains a delicate balance between sensitivity and specificity. This equilibrium is crucial to align with the unique requirements and objectives of the business.
In addressing this challenge, we leverage the Knime platform, utilizing its capabilities to develop an effective solution.

Phase 2: Data Understanding

At the core of every data analysis endeavor, the phase of Data Understanding takes center stage. Here, the primary objective is to gather, depict, explore, and validate data, setting the foundation for a more nuanced and extensive analysis and modeling process.

Collecting Data:

Data is gathered and The data is organized by “State No” and “County No”. This is the starting point for our analysis.



Row ID	State No	Count...	FOR	AGAIN...	TOTA...	DEPE...	BALLO...	POPUL...	PCI	MEDIU...	SIZE O...	POPUL...	PERCE...	PERCE...	PERCE...	D
Row0	1	1	42385	22992	65377	1	1	265038	15607	34618	1,192	223.29	0.75	0.03	0.22	0.-
Row1	1	2	2015	1852	3867	1	1	13617	13740	23114	722.8	18.81	0.59	0	0.4	0.-
Row2	1	3	51959	48406	100365	1	1	391511	24187	44874	803.2	490.34	0.85	0.06	0.09	0.-
Row3	1	4	957	856	1813	1	1	5345	12816	27359	1,349.4	3.99	0.74	0	0.26	0.:
Row4	1	5	942	1129	2071	0	1	4556	19017	23054	2,555.9	1.77	0.93	0	0.07	0.-
Row5	1	6	905	646	1551	1	1	5048	13742	22325	1,514	3.32	0.71	0.01	0.28	0.:
Row6	1	7	40479	39515	79994	1	1	225339	21782	43782	742.5	304.5	0.9	0.01	0.1	0.:
Row7	1	8	2678	1826	4504	1	1	12684	13116	25439	1,013.5	12.52	0.88	0.02	0.11	0.:
Row8	1	9	490	422	912	1	1	2397	19356	30053	1,781.5	1.34	0.96	0	0.04	0.:
Row9	1	10	1923	1140	3063	1	1	7619	17826	41111	395.5	19.29	0.96	0	0.04	0.:
Row10	1	11	1063	1343	2406	0	1	7453	9534	17155	1,287.3	5.8	0.39	0	0.61	0.-
Row11	1	12	624	404	1028	1	1	3190	11705	15127	1,227	2.59	0.22	0	0.78	0.-
Row12	1	13	758	468	1226	1	1	3946	11946	18345	789	4.98	0.68	0.06	0.25	0.-
Row13	1	14	565	403	968	1	1	1926	14913	23919	738.9	2.63	0.96	0	0.05	0.:
Row14	1	15	3645	3918	7563	0	1	20980	12769	22197	1,142.2	18.43	0.9	0	0.1	0.-
Row15	1	16	72984	57335	130319	1	1	467610	23200	32038	153.3	3,046.53	0.61	0.12	0.26	0.-
Row16	1	17	322	320	642	1	1	1504	14205	24671	1,067	1.41	0.94	0	0.06	0.:
Row17	1	18	11519	8261	19780	1	1	60391	24740	54244	840.2	73.32	0.95	0.01	0.04	0.:
Row18	1	19	3006	2064	5070	1	1	21928	21966	41183	1,688	13.16	0.86	0	0.14	0.:
Row19	1	20	2179	1205	3384	1	1	9646	17331	33932	1,850.9	5.27	0.81	0.07	0.12	0.:
Row20	1	21	58502	38892	97394	1	1	397014	17028	38193	2,126.7	186.81	0.96	0.01	0.03	0.:
Row21	1	22	6805	3683	10488	1	1	32273	12297	24350	1,533	21.03	0.88	0.03	0.1	0.:
Row22	1	23	4646	3744	8390	1	1	29974	16653	32377	2,947.5	10.29	0.93	0	0.07	0.:
Row23	1	24	999	461	1460	1	1	3070	15505	36218	149.9	20.5	0.95	0.01	0.05	0.:
Row24	1	25	2350	1440	3790	1	1	7966	16831	34072	1,849.8	4.33	0.96	0	0.04	0.:
Row25	1	26	1732	1498	3230	1	1	10273	13424	29533	3,239.2	3.18	0.95	0.01	0.05	0.:
Row26	1	27	236	177	413	1	1	467	17859	30125	1,117.8	0.42	0.99	0	0.01	0.:
Row27	1	28	1320	618	1938	1	1	6009	11262	18955	1,590.9	3.76	0.58	0	0.41	0.-
Row28	1	29	431	269	700	1	1	1605	13888	23239	1,613.3	0.99	0.91	0	0.09	0.:
Row29	1	30	86173	64270	150443	1	1	438430	20986	44679	772.2	569.43	0.9	0.01	0.09	0.-
Row30	1	31	472	413	885	1	1	1688	23888	26779	1,771.1	0.95	0.96	0	0.04	0.-
Row31	1	32	1552	1464	3016	1	1	7140	19894	27648	2,161	3.29	0.93	0	0.07	0.-
Row32	1	33	1063	631	1694	1	1	6007	14419	30788	376.9	15.98	0.84	0	0.16	0.:
Row33	1	34	5386	4210	9596	1	1	32284	15922	28950	1,692.1	19.18	0.75	0	0.25	0.:
Row34	1	35	30302	30300	60602	1	1	186136	17274	36931	2,601.4	71.9	0.91	0.01	0.09	0.-
Row35	1	36	2405	1348	3753	1	1	13765	11986	20844	4,773	2.88	0.55	0	0.45	0.-
Row36	1	37	1023	776	1799	1	1	4529	16445	24173	2,586.3	1.75	0.97	0	0.03	0.-
Row37	1	38	4102	2533	6635	1	1	17567	17449	26436	1,838.6	9.52	0.92	0	0.08	0.-
Row38	1	39	15416	15418	30834	0	1	93145	15280	27637	3,327.9	28.17	0.9	0	0.09	0.-
Row39	1	40	264	160	424	1	1	558	16199	22232	875.8	0.63	0.95	0	0.05	0.:
Row40	1	41	1886	1320	3206	1	1	11357	15712	35577	4,742.5	2.4	0.93	0	0.07	0.:

Fig 1: snapshot of the data provided

Describing the Data:

For the second task, we offer an overview of the dataset, which encompasses a total of 1287 records and encompasses 31 features or attributes. These attributes can be classified into two categories: 28 are numeric features, and 3 are nominal, Binomial features. The dataset encompasses information pertaining to ballot type, state number, number of churches, county size, percentage of males, and more.

Exploring the Data:

Navigating into the dataset, we conduct an exploratory data analysis to unravel its intricacies. This involves delving into statistical concepts such as mean, median, and mode, coupled with the utilization of visual analysis tools. To deepen our understanding of feature distribution, we leverage measures like standard deviation, variance, and skewness. Noteworthy is the fact that we executed this comprehensive analysis utilizing the Data Explorer functionality within the Knime Analytics tool.

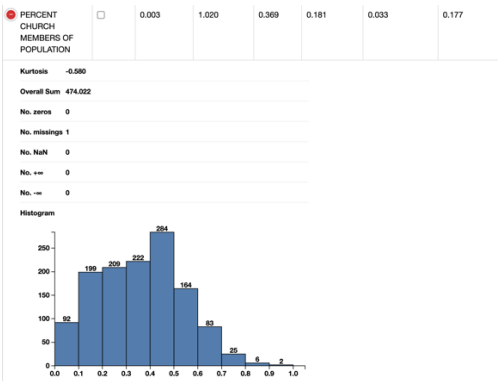


Fig 2: Image showing descriptive Statistics and Histogram of a feature

<u>Verifying Data Quality</u>:

In the fourth task, our focus shifts to evaluating the data's quality. Within the dataset, we uncover certain anomalies that warrant attention. Notably, the identification of outliers emerges as a crucial aspect, potentially influencing the robustness of our analysis. Additionally, a noticeable imbalance surfaces in relation to the target variable. Recognizing the importance of rectifying this imbalance to prevent bias towards the majority class, we must implement data balancing techniques. While the provided data appears to be of good quality briefly, it underscores the need for meticulous cleaning and refinement.

Fig 3: Image describing the data imbalance

Phase 3: Data Preparation

Within this phase, our primary emphasis is on preparing the data for the training of our model. This involves traversing through crucial steps, encompassing data cleaning, transformation, and reduction. The significance of this phase lies in its direct influence on the accuracy of our predictions, as both the quality and quantity of the data play pivotal roles. The adage holds true: the superior the quality of the data supplied to our model, the more accurate its predictions are poised to be.

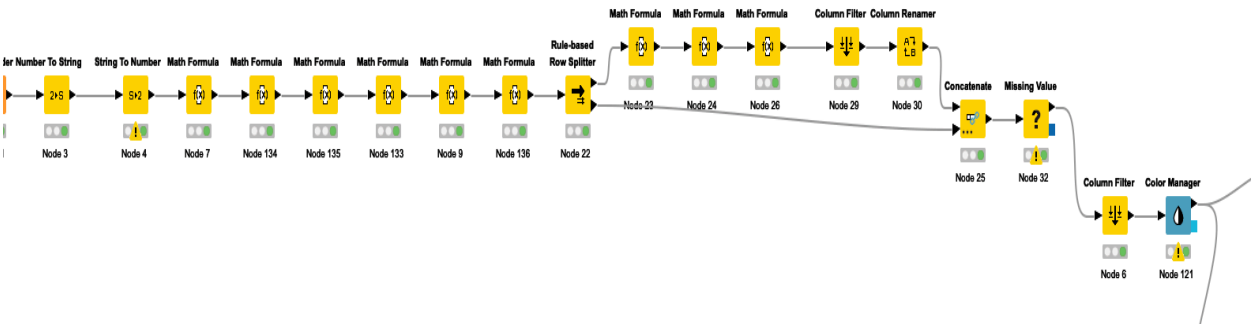


Fig 4 : Complete workflow of Data preprocessing

Selecting Data :

Ensuring the right selection of data (features or variables) is crucial. Including irrelevant or redundant information in our model training process can lead to overfitting, causing sensitivity to noise and potentially generating inaccurate outcomes. In our specific scenario, we meticulously assessed features like state number, county number, and unemployment rate, deeming them non-essential for revealing underlying parameter values. To address this, we utilized the Column Filter function in Knime, systematically eliminating these identified columns. This strategic filtering process resulted in retaining only 16 out of the initial 31 columns for training our models. This intentional feature curation aims to sharpen the model's focus on the most pertinent aspects of the data, fostering more accurate predictions.

Cleaning Data:

In this crucial step, we address multiple facets of data preprocessing. This includes handling missing data, managing outliers through either imputation or removal, addressing anomalies, and encoding the data in a format conducive to the model's requirements. Additionally, we perform data type conversion and meticulously double-verify derived features to ensure accuracy.

Furthermore, to mitigate bias and foster a balanced training environment, we employ an equal-sampling technique. This approach ensures an equal representation of "yes" and "no" instances, contributing to a more unbiased and robust model training process. This comprehensive data

preprocessing methodology is vital for enhancing the overall quality and reliability of our predictive models.

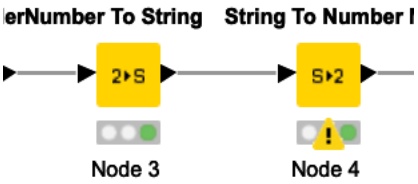


Fig 5 : Image showing using data type conversion nodes to retain the datatypes of columns that are changed due to knime.

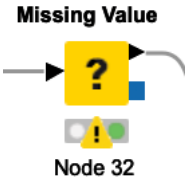


Fig 6 : Image showing using missing value node to handle the missing values

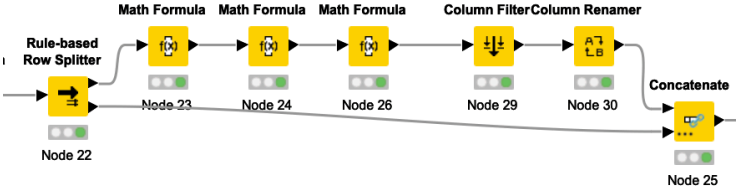


Fig 7 : Image showing manipulating rows of derived features percent white, percent black, percent others to meet the necessary requirements

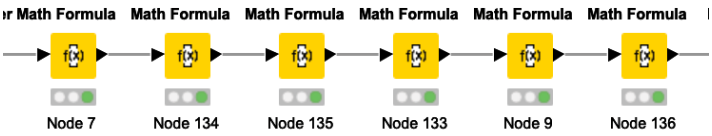


Fig 8 : Image showing replacing the outliers using math formula node

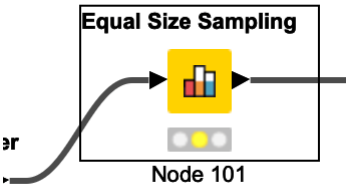


Fig 9 : Using equal-size sampling to balance the data

Formatting Data:

We've executed data encoding, a process that entails transforming relevant columns from nominal to numeric format. This conversion is essential to ensure compatibility with number-based models such as Artificial Neural Networks (ANN) and Logistic Regression. By making this adjustment, we enhance the model's ability to effectively process and derive insights from the encoded data.

We also used Normalizer using min-max normalization to smoothen the data into single range.

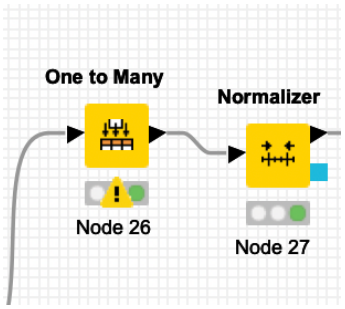


Fig 10: Image showing using of One-to-Many Node to perform Binary Encoding

Phase 4: Modeling

Modeling is important aspect as choosing appropriate model to the dataset is often hard and yields in better results.

Selecting Model techniques:

I have explored various models, encompassing both Set-based and Number-based approaches. The Set-based models include Decision Tree (DT) and Random Forest (RF), while the Number-based

models consist of Artificial Neural Networks (ANN)-MLP, Logistic Regression (LR), and Support Vector Machines (SVM). This diverse set of models allows for a comprehensive analysis, catering to different data patterns and structures.

Generate Test Design:

To ensure a robust evaluation of our models, we implemented the K-fold cross-validation methodology with K set to 10. This intricate process involves partitioning the dataset into 10 subsets and iteratively training and testing the models. During each iteration, 9 folds are dedicated to training, and the remaining fold is used for testing. The results from each testing phase are then averaged to provide a comprehensive performance assessment.

Despite the complexity of this approach, our primary focus remains on achieving both accuracy and a well-balanced model. To facilitate this, we utilized the X-partitioner and X-Aggregator nodes within the Knime platform. These nodes efficiently manage the partitioning and aggregation steps, streamlining the K-fold cross-validation procedure and contributing to the meticulous evaluation of our models.

Building Model:

As part of analysis, we've constructed several models, and in this discussion, we'll provide an overview of how they were developed.

Decision Tree (DT)

The Decision Tree is a type of non-parametric model used for tasks like classification and regression. It's designed to help us gain a deeper understanding of our data. In our case, we used Gini impurity as it suits many of our datasets. We didn't employ any pruning methods, and we chose to have binary nominal splits. Based on the decision tree splits longten ,employ variables has more influence on making final predictions of the model as they are close to the root node and highly important in decision making.

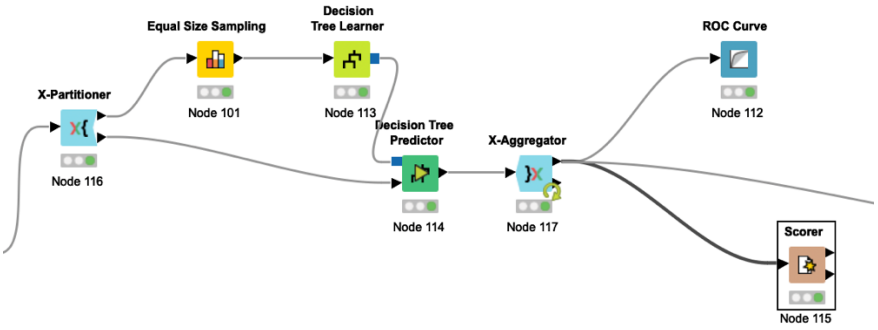


Fig 15: Complete workflow of DT model

Random Forest (RF)

Random Forest is a supervised ensemble model that combines multiple independent trees to predict the target variable. It makes decisions based on the majority vote of the trees. Additionally, we conducted Variable Importance analysis using 1000 trees to understand the significance of each variable.

Based on the Variable Importance Graph is clear that longten, cardten, longmon and few other variables are most used and must be included dataset used for training the model.As we calculated using the weight of each variable occurrence in all the levels.

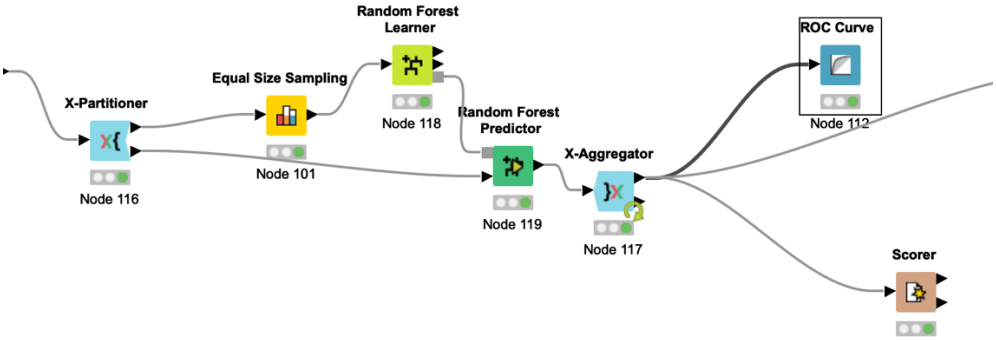


Fig 15: Complete workflow of RF model

Artificial Neural Network (ANN)

ANN is designed to mimic the biological neural networks in the human brain. Neurons in ANN calculate their outputs based on weights from various inputs or the previous layer, employing an activation function.

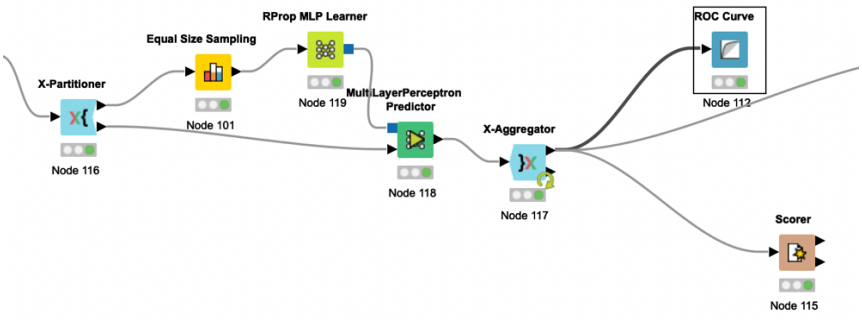


Fig 16: Complete workflow of ANN model

Logistic Regression (LR):

Logistic Regression is a commonly used model for classification tasks. It employs logistic or sigmoid functions to map values and make predictions.

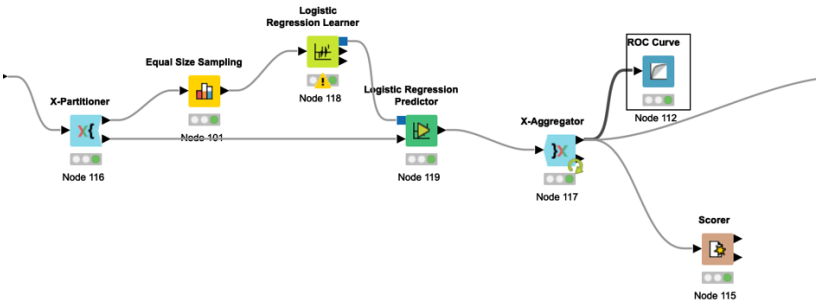


Fig 21: Complete workflow of LR model

Support Vector Machine(SVM):

Support Vector Machines (SVM) represent a supervised machine learning algorithm employed for tasks involving classification and regression. The goal of SVM is to identify a hyperplane

within an N-dimensional space with the explicit purpose of clearly delineating and classifying data into distinct categories.

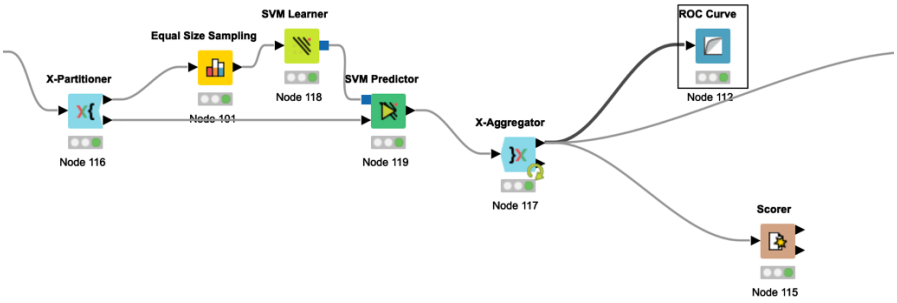


Fig 16: Complete workflow of SVM model

Phase 5: Evaluation

Evaluation is very important it is a measure of verifying the model whether the predictions are close to the actual values. Tabulated evaluations of each model have been listed below :

MODEL	AUC	ACCURACY(%)	SENSTIVITY(%)	SPECIFICTY(%)
DT	0.932	92.3	92.9	91.9
RF	0.958	88.6	89.6	87.8
ANN_MLP	0.998	98.1	99.1	97.3
SVM	0.735	67.6	56.3	75.9
LR	0.837	75.0	73.3	76.3

Table1 : Evaluation results for different models used.

Row ID	1	0
1	508	39
0	60	679

Fig 29: Confusion Matrix for DT

Row ID	1	0
1	490	57
0	90	649

Fig 29: Confusion Matrix for RF

Fig 29: Confusion Matrix for ANN_MLP

Row ID	1	0
1	542	5
0	20	719

Fig 29: Confusion Matrix for SVM

Row ID	1	0
1	308	239
0	178	561

Fig 29: Confusion Matrix for LR

Row ID	1	0
1	401	146
0	175	564

Comparative analysis of all Models:

Now we have multiple models built we need to compare the results to choose the best model. Here I used best models of each category(RF,DT,ANN,LR,SVM) and compared all of them using ROC graph.

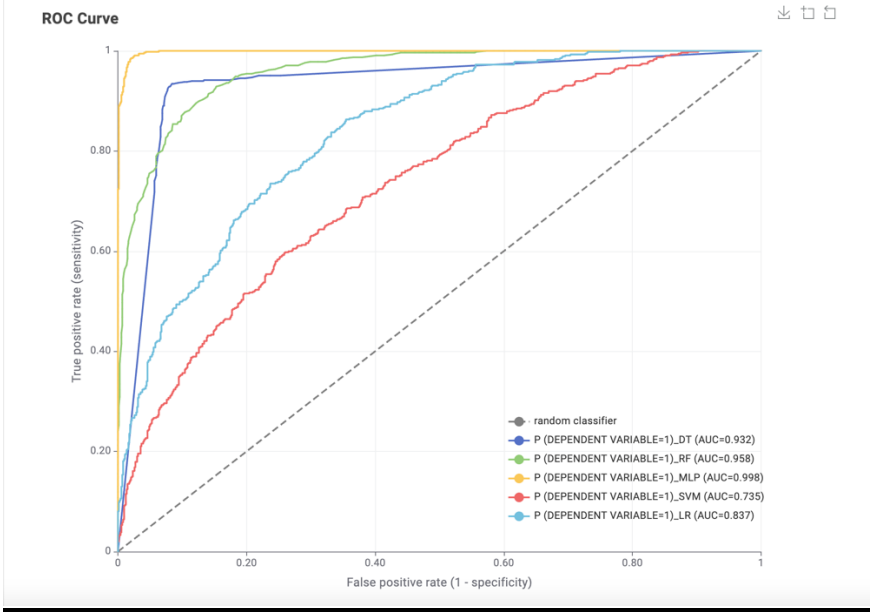


Fig 29: Combined ROC curve

Every model we constructed exhibits superior Area Under the Curve (AUC) values (AUC > 0.50), as evident from the Combined Receiver Operating Characteristic (ROC) graph. Upon careful evaluation of all outcomes, it is evident that Artificial Neural Network utilizing Multi-Layer Perceptron (ANN with MLP) consistently outperforms others, demonstrating the highest accuracy and a well-balanced power of prediction. This affirms the efficacy of ANN with MLP in our predictive modeling endeavors.

Overall final Workflow

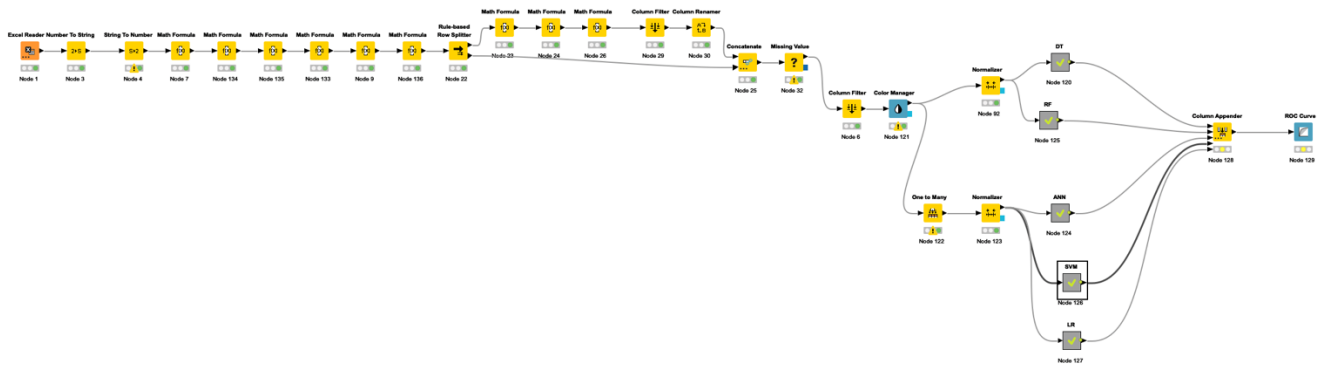


Fig 30: Overall final workflow

Phase 6: Deployment

Model remains undeployed at this stage, as we are still in the process of model development. But would recommend deploying ANN model to see the better results.

Conclusion

In conclusion, In our examination of public attitudes towards legalizing gambling, we crafted five distinct models. Evaluating their performance using metrics like the confusion matrix and ROC curve values, we observed a striking similarity in accuracy, sensitivity, and specificity between the Random Forest and Neural Network models. This suggests that both models are adept at capturing and discerning the intricacies of public sentiment on the legalization of gambling.

After a comprehensive analysis and evaluation of multiple models for predicting County Stance on Legalizing Gaming, the Artificial Neural Network - Multi-layer Perceptron emerged as the most promising model, boasting an impressive performance with an AUC of 0.998, Sensitivity of 99.1%, Specificity of 97.3%, and an Accuracy of 98.1%. This model demonstrates a high capability to accurately predict both true positives and true negatives.