

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | |
|  | | KNIME Data Mining II | | |  | |
|  |  | | | | |  |
|  | | |  |  | | |
|  | | | Moises MarinA20349918 mmarinm@okstate.edu |  | | |
|  | | | 4/20/2022—MSIS 5633Predictive Analytics Technologies |  | | |

**EXECUTIVE SUMMARY**

Gaming is an entertainment experience that can be enjoyed responsible, for this industry to comply with the standards of fairness and transparency, it needs the support of the different levels of government in the U.S. It is thru voting that populations shape the direction of the gaming industry in their states. In this report I follow the CRISP-DM methodology to build and evaluate three supervised learning models (Decision Tree, Support Vector Machine and Artificial Neural Network) to predict the results of gambling polls. The data set used contains a sample of results of elections in 1287 counties in 18 states of the U.S.

**1 - BUSINESS UNDERSTANDING**

The U.S. has 48 states and over 3,000 counties, a number that has vastly increased from the 292 counties we could find in the 1700s. Counties are a way to subdivide states to ease their administration and make the government get closer to the population of a region. The subdivision is not consistent, there are states with few counties like the 3 counties in Delaware and states with a couple hundred counties like the 254 we can find in Texas [1].

Elections in the USA are held at a local level, being the nation’s counties the organism in charge of supervising vote-casting. The county officials look after the conditions required to carry effective elections, like the accessibility and integrity of the polling process or the placement of the voting machines [2].

In addition to conducting elections, counties are also responsible for enforcing state laws, collecting taxes, assessing property, issuing licenses, providing parks, libraries, hospitals, public assistance, highways, police protection as well as urban planning [3]. This list of county functions gives us an idea of their importance in the government. Counties, as well as cities and towns, conform the lowest level of government in the U.S.: the local government. In a higher rank than local government we have the state government and above it is the federal government [4].

Among the many legislations that the three tiers of government oversee are those related to gambling. Gambling is defined as “*accepting, recording, or registering bets, or carrying on a policy game or any other lottery, or playing any game of chance, for money or other thing of value*” [5]. In the fiscal year 2020, around $30 billion were collected by State and local governments from different forms of gambling [6].

Nevada is of particular importance in the gambling scene in the U.S., more than 50% of all commercial casinos are in this western state, which legalized gambling in 1931 to attenuate the effects of the Great Depression.

The American Gaming Association, founded in 1994, promotes the gaming entertainment industry interests at the federal and state government levels. This association arranges gambling in 8 categories: card rooms, commercial casinos, charitable games, tribal casinos, sports betting, lotteries, parimutuel wagering and advance-deposit wagering. The legality of gambling, just like the subdivision of states into counties, is not consistent, not all forms of gambling are allowed in every state. In 2020 three more states voted in favor of authorizing legal sports betting: Maryland, South Dakota, and Louisiana.

What are the characteristics of the populations voting in favor of legalizing gambling? Are variables such as population size, family income, population density, racial and gender distribution, and religious identity enough to predict voting results? Predicting whether a county will vote in favor or against legalizing gambling is useful for the different levels of government, because it will help them prepare for changes required to effectively control, tax, and promote this remunerable, entertaining, and fast-growing industry.

**2 – DATA UNDERSTANDING**

The data set available for analysis is related to gaming ballots, the 1287 unique records are organized by state number and county number. There are no missing values, the 31 variables are listed in table 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **#** | **Variable name** | **Type** | **Description** | **Use in Model** |
| 1 | State No | Nominal | Primary Key Field | No |
| 2 | County No | Nominal | Primary Key Field | No |
| 3 | FOR | Numeric | Number of FOR votes | No |
| 4 | AGAINST | Numeric | Number of AGAINST votes | No |
| 5 | TOTAL CASTE | Numeric | Number of people voted | No |
| 6 | DEPENDENT VARIABLE | Binary Nominal | 1:Yes; 0:No | No |
| 7 | BALLOT TYPE | Binary Nominal | 1:Gambling; 2:Wagering | No |
| 8 | POPULATION | Numeric | Population of the County | Yes |
| 9 | PCI | Numeric | Per capita income | Yes |
| 10 | MEDIUM FAMILY INCOME | Numeric | Medium family income | Yes |
| 11 | SIZE OF COUNTY | Numeric | Size of the county (sq. mile) | No |
| 12 | POPULATION DENSITY | Numeric | Population density | Yes |
| 13 | PERCENT WHITE | Numeric | Racial distribution of the county | Yes |
| 14 | PERCENT BLACK | Numeric | Racial distribution of the county | Yes |
| 15 | PERCENT OTHER | Numeric | Racial distribution of the county | Yes |
| 16 | PERCENT MALE | Numeric | Gender distribution of the county | Yes |
| 17 | PERCENT FEMALE | Numeric | Gender distribution of the county | Yes |
| 18 | NO OF CHURCHES | Numeric | Religious identity of the county | Yes |
| 19 | NO OF CHURCH MEMBERS | Numeric | Religious identity of the county | No |
| 20 | PERCENT CHURCH MEMBERS OF POPULATION | Numeric | Religious identity of the county | No |
| 21 | POVERTY LEVEL | Numeric | Poverty level | Yes |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 22 | UNEMPLOYMENT RATE | Numeric | unemployment rate | No |
| 23 | AGE LESS THAN 18 | Numeric | age distribution of the county | Yes |
| 24 | AGE24 | Numeric | age distribution of the county | Yes |
| 25 | AGE44 | Numeric | age distribution of the county | Yes |
| 26 | AGE64 | Numeric | age distribution of the county | Yes |
| 27 | AGE OLDER THAN 65 | Numeric | age distribution of the county | Yes |
| 28 | MSA | Binary Nominal | Metropolitan statistical area-1:Yes; 0:No | Yes |
| 29 | PERCENT MINORITY | Numeric | percentage of minority | Yes |
| 30 | NO OF OLDER | Numeric | number of older population | No |
| 31 | NO OF YOUNGER | Numeric | number of younger population | No |

Table 1. List of variables

1 - State No

There are 18 state numbers in the data, the state is a nominal variable, it is a code of a state. The distribution of State No is not uniform, as expected some states have more counties than others. This variable won’t be selected for the classification model because I want to focus on other dimensions of the voters beyond their locality.

Chart, bar chart

Description automatically generated  
Figure 1. State No Histogram

2 - County No

Figure 1 indicates the number of counties in each state of the data. For example, state 9, most likely Texas, has 251 counties, followed by state 17, most likely Georgia, with 158 counties in the data set. This variable is a code for the county, like State No, this variable won’t be selected for the classification model.

3 – FOR

4 - AGAINST

5 - TOTAL CASTE

These variables are numeric, they refer to the number of votes counted in each county. They won’t be used in the classification model because they are not an intrinsic dimension of the voter.

6 - DEPENDENT VARIABLE

This is a binary nominal variable that indicates whether the county voted in favor of or against legalizing gaming at the ballot, it is a key variable that will help build and score the classification model. It has the value of 1 if the county had more FOR votes than AGAINST, it is 0 when AGAINST is greater than FOR. This variable is calculated correctly, I recalculated it using a FOR > AGAINST condition and obtained the exact same results.

7 - BALLOT TYPE

This variable will not be used for the model, it is the type of ballot with the two possible values of Gambling and Wagering. Examples of wagering are betting on sports events like football matches or horse races; examples of gambling are playing casino games like roulette, blackjack, or poker.

8 – POPULATION

This variable will be used in the model, it indicates the amount of people living in a county. This characteristic of a county could play a role in the outcome of the polls. Are bigger counties more open to legalizing gambling than smaller ones? And what’s the weight that the size play? Counties vary greatly in population size as we can see in figure 2.

Table

Description automatically generated

Figure 2. Descriptive statistics of county population in the data set.

9 - PCI

10 - MEDIUM FAMILY INCOME

Per Capita Income (PCI) is the total personal income divided by the total population, including children. The Median Family Income is the amount at the middle of the family incomes when they are sorted from lowest to highest. I decided to include both in the classification model because of the high correlation between PCI and the dependent variable and the information that Medium Family Income provides about income in a county.

Chart, bar chart, histogram

Description automatically generated

Figure 3. Medium Family Income Histogram

11 - SIZE OF COUNTY

12 - POPULATION DENSITY

Out of these two variables, I decided to include population density in the classification model because it is a good indicator of the county population, size in square miles could be relevant in terms of having space to build gambling spaces for example, but I don’t include it in this model.

13 - PERCENT WHITE

14 - PERCENT BLACK

15 - PERCENT OTHER

There is a small loss of precision with the racial distribution of the county, 2% of counties have the sum of racial distribution at 101% and another 2% at 99%. These variables are included in the model because they represent a dimension of the population worth analyzing. I am unable to make sure these derived variables are calculated correctly due to lack of detail in the data set, i.e., file is already aggregated to these percentages.

16 - PERCENT MALE

17 - PERCENT FEMALE

There are about 7% of records with a sum of male and female percentages over or under 100%, this could be due to rounding in percentages. These variables are included in the model because they represent a dimension of the population that is worth analyzing.

18 - NO OF CHURCHES

19 - NO OF CHURCH MEMBERS

20 - PERCENT CHURCH MEMBERS OF POPULATION

Out of these three variables only number of churches is included in the model because I think number of church members and percent church members of population are difficult to track with precision.

21 - POVERTY LEVEL

This variable is included in the model, it indicates percentage of population in the county living below the poverty line.

Table

Description automatically generated

Figure 4. Descriptive statistics of the poverty level

22 - UNEMPLOYMENT RATE

This variable is not included in the model because I already include two other income related variables in the model: Medium Family Income and Poverty Level.

23 - AGE LESS THAN 18

24 - AGE24

25 - AGE44

26 - AGE64

27 - AGE OLDER THAN 65

These variables are used because they provide information about the age of population in the county.

28 – MSA

This variable is used in the model, it indicates if the county has a relatively high population density and close economic ties in the area.

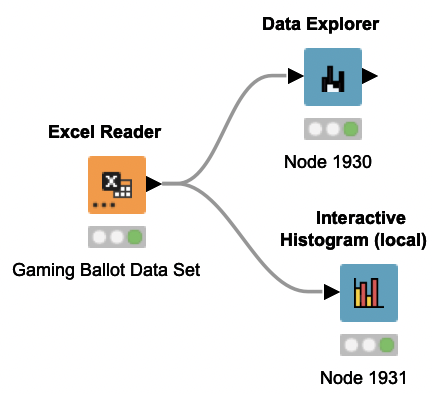
29 - PERCENT MINORITY

30 - NO OF OLDER

31 - NO OF YOUNGER

Percent minority is used due to its correlation with the dependent variable (0.25). Number of older and younger were not included in the model because I already have detailed information about age groups.

**3 – DATA PREPARATION**

Out of the original 31 variables 18 were selected for the model. There are several checks that were performed in the data set to validate that selected variables were ready to use. Checks were done using Knime’s Data explorer and Interactive Histogram nodes:

1. There are no missing values
2. There are no unexplained zero or not number values
3. There are no outliers

Figure 5. Knime’s nodes used to check selected variables

This is the list of variables used to build the models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **#** | **Variable name** | **Type** | **Description** | **Use in Model** |
| 1 | POPULATION | Numeric | Population of the County | Yes |
| 2 | PCI | Numeric | Per capita income | Yes |
| 3 | MEDIUM FAMILY INCOME | Numeric | Medium family income | Yes |
| 4 | POPULATION DENSITY | Numeric | Population density | Yes |
| 5 | PERCENT WHITE | Numeric | Racial distribution of the county | Yes |
| 6 | PERCENT BLACK | Numeric | Racial distribution of the county | Yes |
| 7 | PERCENT OTHER | Numeric | Racial distribution of the county | Yes |
| 8 | PERCENT MALE | Numeric | Gender distribution of the county | Yes |
| 9 | PERCENT FEMALE | Numeric | Gender distribution of the county | Yes |
| 10 | NO OF CHURCHES | Numeric | Religious identity of the county | Yes |
| 11 | POVERTY LEVEL | Numeric | Poverty level | Yes |
| 12 | AGE LESS THAN 18 | Numeric | age distribution of the county | Yes |
| 13 | AGE24 | Numeric | age distribution of the county | Yes |
| 14 | AGE44 | Numeric | age distribution of the county | Yes |
| 15 | AGE64 | Numeric | age distribution of the county | Yes |
| 16 | AGE OLDER THAN 65 | Numeric | age distribution of the county | Yes |
| 17 | MSA | Binary Nominal | Metropolitan statistical area-1:Yes; 0:No | Yes |
| 18 | PERCENT MINORITY | Numeric | percentage of minority | Yes |

Table 2. List of variables used in model

In terms of data transformation, normalization was done for the data set before feeding them into the SVM and ANN models. New attributes were not built based on the variables, correlation analysis was done to identify the variables that were more related to the dependent variable to not exclude them accidentally during analysis of variables. For the Decision Tree model, data type was changed from integer to string in two variables, the dependent variable (target) and the MSA. For the SVM and ANN models only the dependent variable (target) data type was changed to string.

**4 – MODEL BUILDING**

**Diagram

Description automatically generated**

Figure 6. Complete Classification Model

The complete classification model is showed in figure 6, it contains the data preparation and model generation nodes, along with their respective scorers and ROC curve nodes for evaluation.

The Decision Tree Learner uses the configuration depicted in figure 7, which was found to be optimal for this model. A color manager node was added for the Decision Tree to facilitate the analysis of its rules, green was assigned to voting in favor of gambling and red when voting against, as seen in figure 8.

Graphical user interface, application

Description automatically generated A picture containing text

Description automatically generated

Figure 7. Settings used in the Decision Tree Learner Node. Figure 8. Color coding used in Decision Tree.

**5 – TESTING AND EVALUATION**

Figures 9 to 14 show the confusion matrix and ROC curves of the models built. The Decision tree is the one that had the highest error rate with 32.56%, it was barely improved by the SVM model with a 32.17% error rate. The best model of the is the Artificial Neural Network (ANN) which had a 22.87% error rate and an accuracy of 77%.

Decision Tree

Table

Description automatically generated with medium confidence

Figure 9. Decision Tree’s confusion matrix

SVM

Table

Description automatically generated

Figure 11. SVM’s confusion matrix

ANN

Table

Description automatically generated

Figure 13. ANN’s confusion matrix

Chart, line chart

Description automatically generated

Figure 10. Decision Tree’s ROC Curve

Chart, line chart

Description automatically generated

Figure 12. SVM’s ROC Curve

Chart, line chart

Description automatically generated

Figure 14. ANN’s ROC Curve

**6 – DEPLOYMENT**

The goal of this project was to build the model and analyze their performance, there are no actions planned for their deployment.

**CONCLUSIONS**

KNIME is a very powerful tool that allows us to work with data sets and models easily. The learning curve is steep, I had to look for examples and read the documentation frequently. The nodes have many options that can be changed, unlike other data mining tools like some SAS offerings (Data Miner) that are limited in the settings, I didn’t find myself trying to change something in KNIME without success.

I spent a large amount of time playing with the different settings of the Decision Tree Learner node, changing the quality measure from Gain Index to Gini Ratio, changing the pruning method from no pruning to MDL, checking and unchecking the reduced error pruning and enabling/ disabling the Binary nominal splits option, all with the goal to find the highest accuracy of this model. The time expense was high and the improvement in accuracy was minimal. It made me think that having an option for the model to self-optimize its parameters would save some time and give an idea of the best result that could be obtained, all done programmatically without the time-consuming exercise of manually changing settings and executing.

I learned that running SVM without normalization is not a good idea, it takes the model a smaller amount of time to complete using normalized data.

I did not have to tweak many of the settings of the ANN, since the first execution it showed a better result than the other two models, I increased the number of hidden layers just to find that it decreased the accuracy compared to the previous execution, so I did not spend much time configuring the ANN settings.

Overall, this was a very interesting project assignment that showed the potential of ANN to predict polls results in counties related to gambling using 18 variables.

**REFERENCES**

[1]

Ang, C. (2020). Animated Map: The History of U.S. Counties. April 19, 2022, from visualcapitalist. Web site: <https://www.visualcapitalist.com/history-of-us-counties-animated-map/>

[2]

Hurley, E. (2018). All Elections Are Local, The county role in the elections process. April 19, 2022, from NACo. Web site: <https://www.naco.org/resources/featured/all-elections-are-local-county-role-elections-process>

[3]

HAMILTON COUNTY (2022). County Government Functions and Responsibilities. April 19, 2022, from Hamilton County. Web site:

<https://www.hamiltoncountyohio.gov/government/board_of_county_commissioners/county_government_functions_and_responsibilities#:~:text=As%20an%20agent%20of%20the,and%20officials%2C%20it%20provides%20parks%2C>

[4]

LWV (2020). Levels of Government: How Federal, State, and Local Government Work Together. April 19, 2022, from League of Women Voters. Web site:

<https://www.lwv-rva.org/levels-of-government-how-federal-state-and-local-government-work-together/#:~:text=State%20and%20local%20governments%20work,specific%20powers%20by%20the%20state>

[5]

US LEGAL (2022). Gambling Law and Legal Definition. April 19, 2022, from US LEGAL. Web site:

<https://definitions.uslegal.com/g/gambling/#:~:text=Gambling%20is%20accepting%2C%20recording%2C%20or,conduct%20an%20'illegal%20gambling%20business>.

[6]

URBAN INSTITUTE (2022). Lotteries, Casinos, Sports Betting, and Other Types of State-Sanctioned Gambling. April 19, 2022, from Urban Institute. Web site:

<https://www.urban.org/policy-centers/cross-center-initiatives/state-and-local-finance-initiative/state-and-local-backgrounders/lotteries-casinos-sports-betting-and-other-types-state-sanctioned-gambling>

[7]

History.com Editors (2010). Nevada legalizes gambling. April 19, 2022, from A&E Television Networks. Web site:

<https://www.history.com/this-day-in-history/nevada-legalizes-gambling#:~:text=In%20an%20attempt%20to%20lift,the%20Mexican%20War%20in%201848>.