Health Rankings: Illinois

COMP 300/400 Final Project
29th April 2020

Team 3
Mary Aldana
Kevin Criollo
Alison Lang
Brandi Letsche

Overview

I. The Problem Presenting to the Business/Organization

The original intended scope and plan of our project was to uncover the relationship between life expectancy and the rate of mortality of different diseases throughout the United States. In the initial stages of planning, it came to our attention that these two attributes were *too* similar in nature and, therefore, would not be able to provide meaningful results for analysis. Given this realization, we revamped our thought process to instead focus on creating a more clearer picture as to the overall, or average, level of health within a state. Furthermore, we narrowed our data farther to only include counties within the State of Illinois; this restricted our dataset from over 3,000 tuples down to a manageable 102 tuples for analysis.

Using two CSV files provided online by the A Robert Wood Johnson Foundation Program [1], we came to the conclusion that our goal of classification would be to "summarize" the overall level of health of each county in Illinois, given that county's average life expectancy. Additionally, we wished to analyze which additional regular attributes *correlated the strongest* in such countries that are positioned higher on the average life expectancy scale.

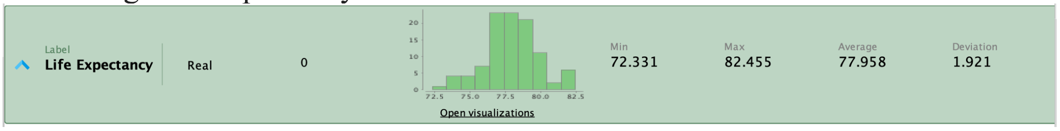
In summary, our final, cleaned dataset consisted of our label attribute of 'life expectancy,' an identifying attribute of 'county,' and seven regular attributes of '% Diabetic' (percentage of individuals living in a county who are diabetic), '# Food Insecure' (number of individuals living in a county who do not have food security), '% Insufficient Sleep' (percentage of individuals living in a county who are sleep deprived), '% Disconnected Youth' (percentage of individuals under the age of 18 who are considered to be "at-risk"), '# Firearm Fatalities' (number of fatalities due to firearm weapons in each county), '# Homeowners' (number of individuals in a county who own a home), and '% Female' (percentage of individuals in a county who identify as "female").

Data Exploration

I. Target Label/Class: Goal of Classification

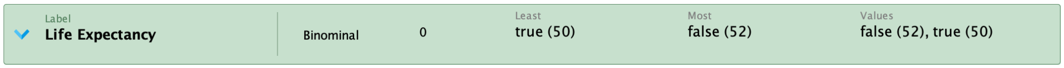
Life Expectancy

In order to summarize the overall health of a particular county in Illinois, we decided that the attribute of 'life expectancy' was most appropriate. In our analysis, however, we looked at counties in which their binomial attribute equivalent was "true" for being greater than the aggregated average life expectancy for all the countries combined.



Label
Life Expectancy
Real
0
Min 72.331
Max 82.455
Average 77.958
Deviation 1.921
Open visualizations

Fig. 1 *Label Attribute: Avg. Life Expectancy*

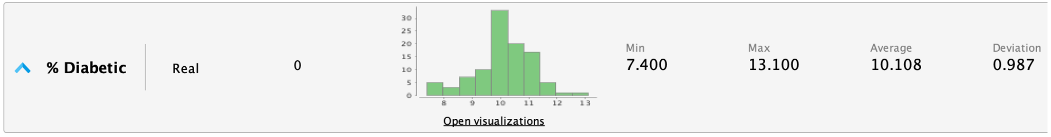


Label
Life Expectancy
Binominal
0
Least true (50)
Most false (52)
Values false (52), true (50)

Fig. 2 *Target Class - Above Average Life Expectancy*

II. Regular Attribute Descriptive Statistics

Percentage of Diabetics - *% Diabetic*



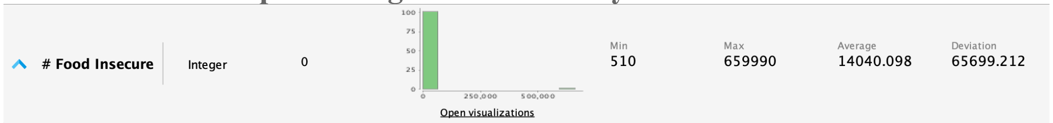
% Diabetic Real 0

Min 7.400 Max 13.100 Average 10.108 Deviation 0.987

Open visualizations

Fig. 3 *Diabetic Statistics*

Number of Individuals Experiencing Food Insecurity - *# Food Insecure*



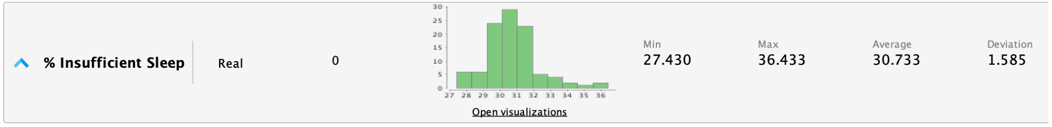
Food Insecure Integer 0

Min 510 Max 659990 Average 14040.098 Deviation 65699.212

Open visualizations

Fig. 4 *Food Insecurity Statistics*

Percentage of Individuals Sleep Deprived - *% Insufficient Sleep*



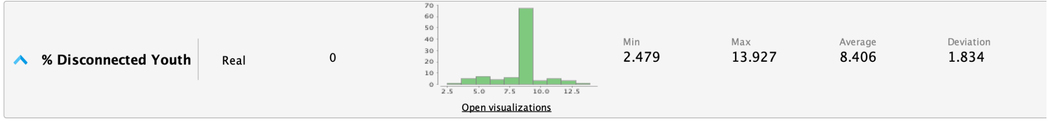
% Insufficient Sleep Real 0

Min 27.430 Max 36.433 Average 30.733 Deviation 1.585

Open visualizations

Fig. 5 *Insufficient Sleep Statistics*

Percentage of "At-Risk" Children - *% Disconnected Youth*



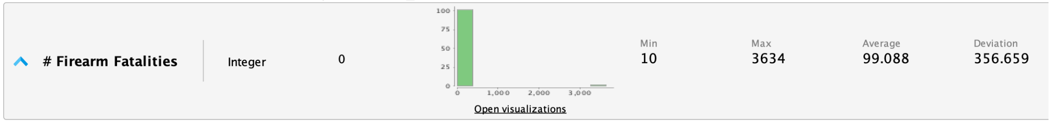
% Disconnected Youth Real 0

Min 2.479 Max 13.927 Average 8.406 Deviation 1.834

Open visualizations

Fig. 6 *Disconnected Youth Statistics*

Number of Fatalities Caused by Weapons - *# Firearm Fatalities*



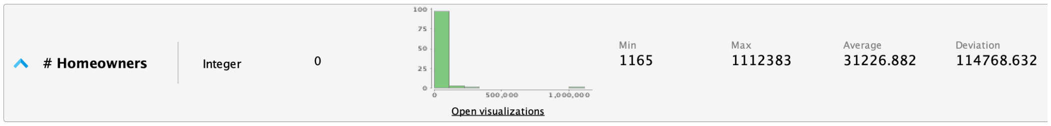
Firearm Fatalities Integer 0

Min 10 Max 3634 Average 99.088 Deviation 356.659

Open visualizations

Fig. 7 *Firearm Fatalities Statistics*

Number of Individuals who Own a Home - *# Homeowners*



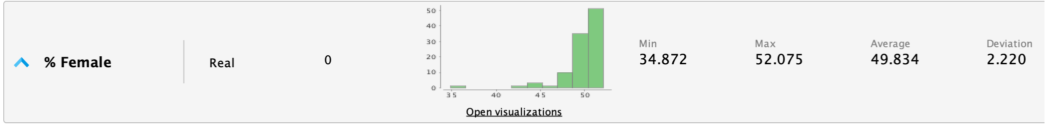
Homeowners Integer 0

Min 1165 Max 1112383 Average 31226.882 Deviation 114768.632

Open visualizations

Fig. 8 *Homeowner Statistics*

Percentage of Females/Women - *% Female*



% Female Real 0

Min 34.872 Max 52.075 Average 49.834 Deviation 2.220

Open visualizations

Fig. 9 *Female Demographics Statistics*

III. Missing Values

While the "Missing" statistic column in the figures for the above attributes specifies "0," our data did go through a process of accounting for missing attributes prior to being stored as our final, joined dataset in which we were to work from.

Using the operator, “Replace Missing Values,” we specified the process to evaluate *all* attributes and to then replace any missing values that existed with the average value of the attribute in question.

IV. Challenges

Many of the challenges we faced were in finding datasets with enough attribute variety, in addition to being able to brainstorm as to how multiple datasets may relate to one another in order to make meaningful predictions.

Furthermore, it was a matter of trial and error for us in regards to attribute/data reduction, normalization, and simply transforming our data into a form in which it would be “mendable” in order to be utilized for model instantiation and analysis. For example, our data was originally presented in numerical format, while many RapidMiner operators and performance tests require binomial and/or polynomial data. Therefore, we had some restructuring to do in terms of how our data was presented before it could be worked into a model and tested.

Data Visualization

I. Analysis of Variance (*Grouped ANOV*A)

One of our first visualization/modeling techniques was the *Grouped ANOVA* operator. This process allowed us to perform a ‘significance test’ for single regular attributes based upon the groups defined by the ‘life expectancy’ attribute.

The two figures shown below, *Fig. 10* and *Fig. 11*, show both the ANOVA process we constructed, in addition to a set of results given the attribute ‘*% Insufficient Sleep*.’

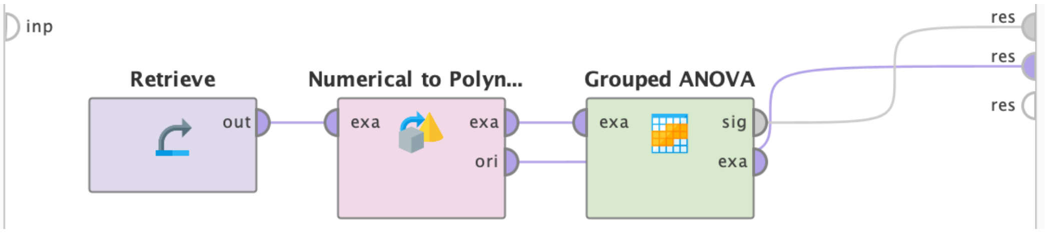
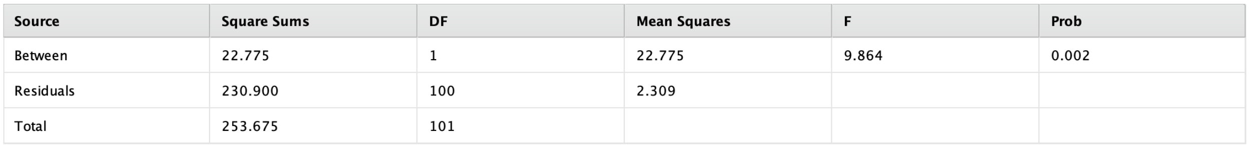


Fig. 10 *ANOVA RapidMiner Process*



Source	Square Sums	DF	Mean Squares	F	Prob
Between	22.775	1	22.775	9.864	0.002
Residuals	230.900	100	2.309		
Total	253.675	101			

Fig. 11 *ANOVA Test for Attribute ‘Insufficient Sleep’*

The above ANOVA results table contains a lot of information, which was useful in beginning to determine the significance of different attributes and where we should focus our attention going forward in our analysis. ‘Insufficient Sleep’ was determined to have the highest probability of significance among all of our attributes; (P-value) the difference between actual mean value is *most likely* significant, since 0.002 is less than our specified alpha significance of 0.05. In addition, ‘Insufficient Sleep’ resulted in a generally high F-statistic; tells us that the variance between the means of these two attribute groups is *most likely* significant.

The P-values for the remaining attributes were as follows:

Regular Attribute	**Prob. (P-value)**
% Diabetic	0.002
% Disconnected Youth	0.005
# Homeowners	0.137
# Food Insecure	0.267
% Female	0.343
# Firearm Fatalities	0.366

Fig. 12 *ANOVA Attribute P-values Listed Greatest Significance → Least Significant*

II. Correlation Matrix

We made use of the correlation matrix to find the correlation among multiple attributes. We chose the correlation matrix because the way in which it presents the attributes with the highest correlation is easy to read and understand. It helped with identifying patterns in the dataset and output different and interesting results.

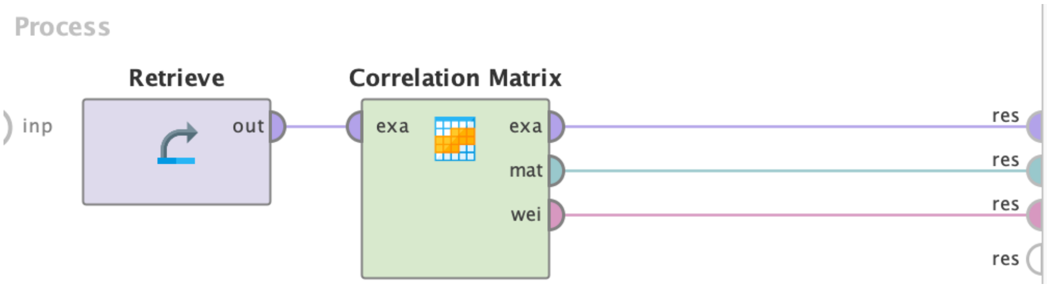
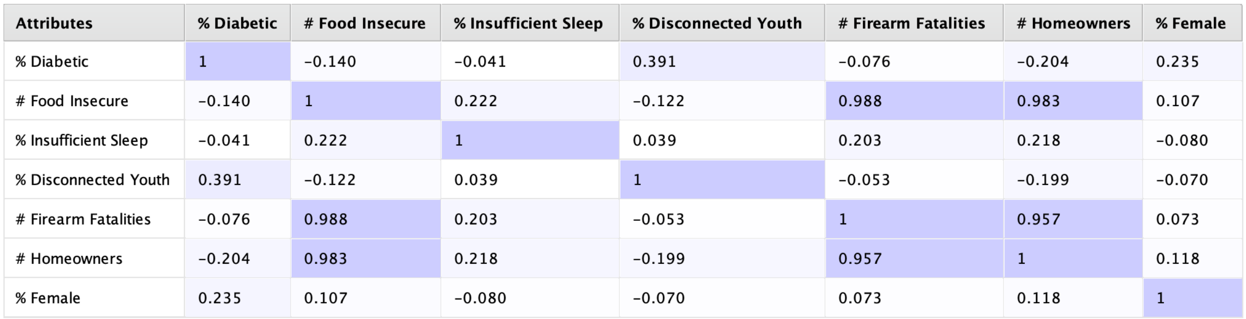


Fig. 13 *Correlation Matrix RapidMiner Process*



Attributes	% Diabetic	# Food Insecure	% Insufficient Sleep	% Disconnected Youth	# Firearm Fatalities	# Homeowners	% Female
% Diabetic	1	-0.140	-0.041	0.391	-0.076	-0.204	0.235
# Food Insecure	-0.140	1	0.222	-0.122	0.988	0.983	0.107
% Insufficient Sleep	-0.041	0.222	1	0.039	0.203	0.218	-0.080
% Disconnected Youth	0.391	-0.122	0.039	1	-0.053	-0.199	-0.070
# Firearm Fatalities	-0.076	0.988	0.203	-0.053	1	0.957	0.073
# Homeowners	-0.204	0.983	0.218	-0.199	0.957	1	0.118
% Female	0.235	0.107	-0.080	-0.070	0.073	0.118	1

Fig. 14 *Correlation Matrix Results Table*

For the purposes and scope of our project, we implemented the correlation matrix process in order to determine how our regular attributes relate to one another. In other words, we felt it may be useful to compare if certain attributes are more similar to each other, if those same attributes also tested 'true' in regards to being significant to our label attribute, 'life expectancy.'

A few key relationships can be seen here; one being that '% Disconnected Youth' and '% Diabetic' appear to be positively correlated. Meaning, that as one rises, the other one does as well. This is interesting as these are two attributes that the ANOVA results tables told us may be significant, based upon their P-value. Additionally, it seems as though '# Firearm Fatalities' and '# Food Insecure' are also positively correlated; it makes sense that as more people experience food insecurity, that the presence of firearms may rise.

III. Deployment Threshold: Decision Tree

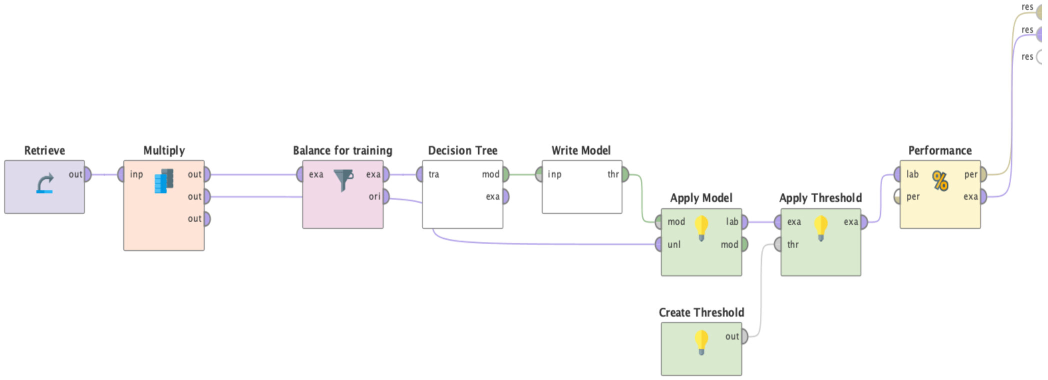
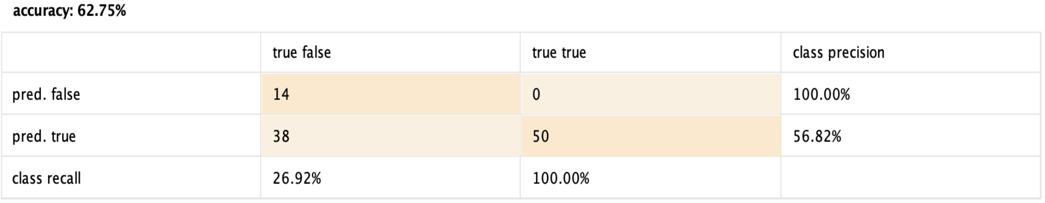


Fig. 15 *Deployment Thresholds RapidMiner Process*

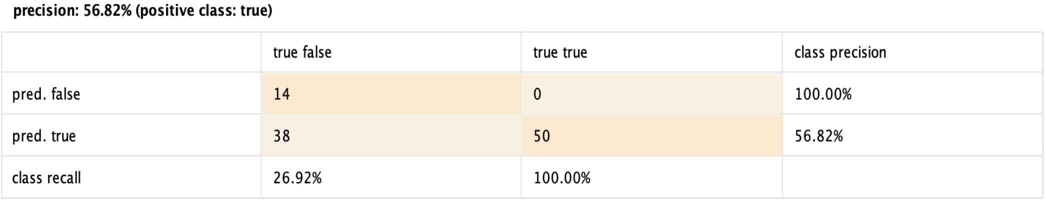
In creating this process, we first multiplied the data, training one set of data on a relative basis for modeling, and the other to act as our unlabeled, original data set. In doing this, we created and wrote a depreciated decision tree model (gini-index), then passed it along to an "apply model" and "apply threshold" operator; the threshold operator, set at 0.6, allows for greater clarity in classification results.



accuracy: 62.75%

	true false	true true	class precision
pred. false	14	0	100.00%
pred. true	38	50	56.82%
class recall	26.92%	100.00%	

Fig. 16 *Decision Tree Classification Accuracy*



precision: 56.82% (positive class: true)

	true false	true true	class precision
pred. false	14	0	100.00%
pred. true	38	50	56.82%
class recall	26.92%	100.00%	

Fig. 17 *Decision Tree Classification Precision*

We found that a decision tree using a gini-index provided the highest accuracy and precision for our model. Our decision tree and threshold process was able to achieve an approximate accuracy of 62.75% and a precision of approximately 56.82%. While we would have hoped for the precision measure to be a little bit higher, we felt as though our model was reliable based upon its accuracy; the main takeaway is that the decision tree accurately classified (based on our provided regular attributes) counties in which have a life expectancy higher than the overall average (class: "true").

IV. Neural Network

Our last model, Neural Net, we implemented in order to not only help make classifications on the basis of our target/class attribute (higher than average life expectancy), but to also uncover the strength of correlation among all of the attributes at the same time. .

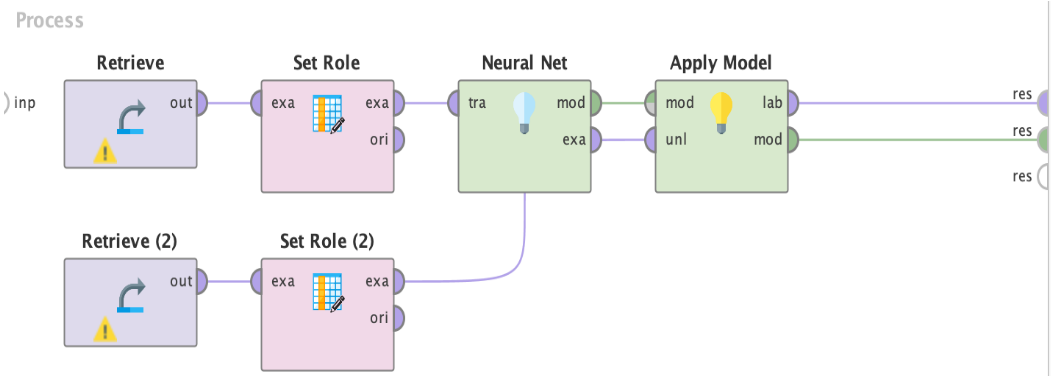


Fig. 18 *Neural Net RapidMiner Process*

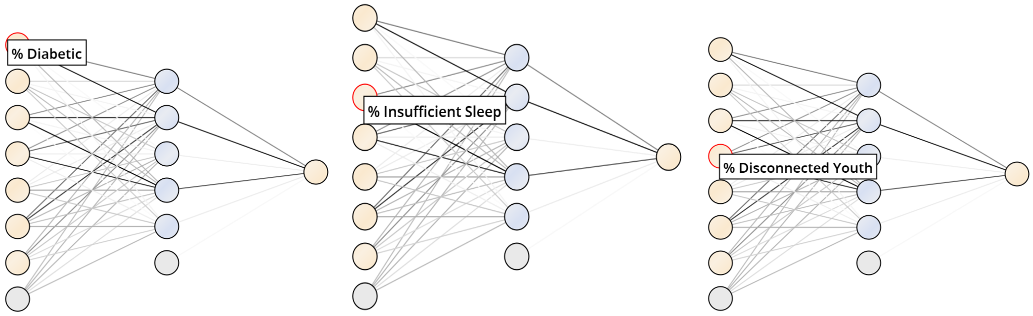
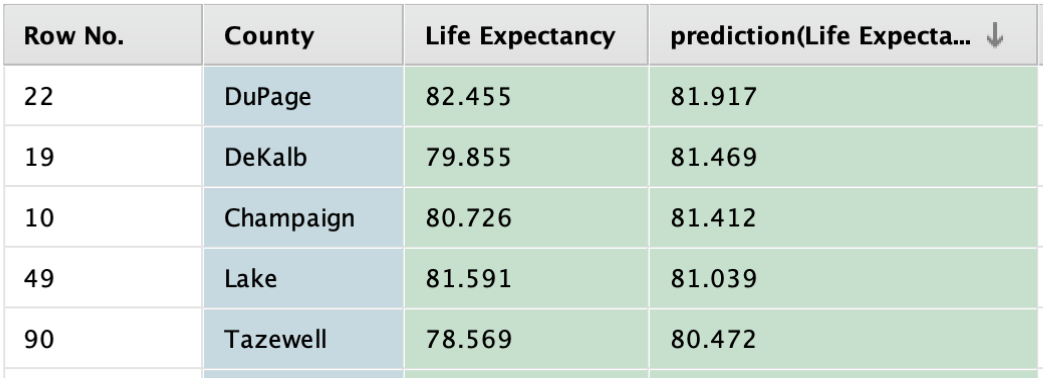


Fig. 19 *Neural Net Results*



Row No.	County	Life Expectancy	prediction(Life Expecta... ↓
22	DuPage	82.455	81.917
19	DeKalb	79.855	81.469
10	Champaign	80.726	81.412
49	Lake	81.591	81.039
90	Tazewell	78.569	80.472

Fig. 20 *Neural Net Model Applied*

Based upon the above two figures, *Fig. 19* and *Fig. 20*, we see that the attributes '% Diabetic,' '% Insufficient Sleep,' and '% Disconnected Youth' seem to be the most strongly correlated. With that in mind, the neural network model made predictions, based upon the training data supplied, of life expectancies for each county. Fig. 20 shows the top five highest-ranked counties in Illinois based upon the predicted life expectancy.

Summary

To summarize, our project data analysis concluded that the presence of diabetes and not getting enough sleep each night weigh most heavily on one's overall life expectancy. It may be interesting and beneficial, in further study and analysis, to look at some of the top-ranking counties in Illinois for life expectancy, and expand the range of attributes compared in order to get a more complete, robust picture.

It was most challenging to consider which attributes to choose to compare initially, however, due to the scope of this project, we were limited in terms of knowledge and ability to broaden the analyzation and prediction aspects on this project. Additional data mining could also include additional high-population states, such as New York, California, and Texas.

Resources

[1] "2020 County Health Rankings & Roadmaps." Internet: https://www.countyhealthrankings.org, 2020 [29th April 2020].