Predictive Analytics for County Health Rankings

Group Project – Team 1

CIS 9660 - Section PMWA

Phase #3 – 12/09/2019

County Health Rankings & Roadmaps

Building a Culture of Health, County by County

A Robert Wood Johnson Foundation program

Shobhit Ratan (shobhit.ratan@baruchmail.cuny.edu)
Zafirah Baksh (zafirah.baksh@baruchmail.cuny.edu)
Christian Cuvilly (Christian.cuvilly@baruchmail.cuny.edu)
Paul Jozefec (paul.jozefec@baruchmail.cuny.edu)
Agata Szawkalo (agata.wegrzyn@baruchmail.cuny.edu)
Bhavi Thakker (bhavimahesh.thakker@baruchmail.cuny.edu)
Alison Wen (alison.wen@baruchmail.cuny.edu)

Table of Contents

Description of Case Study

We have chosen to study county health data in an effort to determine which attributes are the most predictive of poor health. As a proxy for poor health we will be using the attribute ***'Years of Potential Lost Life'***. We will use a ***'Classification'*** approach and evaluate several models.

Business Problem

In 2018 Medicaid expenditures represented 29% of state budgets on average. This compares to just 20% in 2008. The burden is predicted to worsen as healthcare inflation of + 5% is well above most other categories. Unfortunately, the problem is not limited to just states and counties. Nearly 26% of the Federal budget represents spending on Medicare and Medicaid.

Although healthcare is a big concern, the government has limited resources and will typically implement just a few programs. We believe that our analysis would help steer their efforts towards the most beneficial projects.

Data Available

The dataset can be obtained from the following web link:
https://www.countyhealthrankings.org/sites/default/files/2019%20County%20Health%20Rankings%20Data%20-%20v2.xls

It contains **3,142** instances, which represent every US county. The dataset contains over 200 attributes, but many are redundant. Each recorded attribute is accompanied by additional information, such as 95% confidence intervals, quartiles, and other supplemental data. We would restrict our analysis to the **41** main measurements. The county health data set attributes are enclosed in Appendix 1.

Most instances have a complete set of data. We examined the percentage of missing values for each attribute and the greatest value was 8%. The average across all attributes was just 1.4%.

Our target variable will be '*Years of Potential Lost Life' per 100,000 citizens*, measured from 2015-2017. This variable represents the aggregate amount of life lost before the age of 75. For example, if an individual died at age 65, it would be recorded as 10. If an individual died at age 85, it would be recorded as 0.

We will make the target variable binary by using the median. Values above the median will be High Risk and values below will be Low Risk.

Exploratory data analysis

Frequency of the target variable

The target variable is '*years of lost life per 100,000 citizens*'. We converted the quantitative data to a categorical variable by using the median as the split point. The values above the median are classified as “High Risk” and the values below the median are classified as “Low Risk.” By using the median, this resulted in an equal amount of observations for each category. As a result, we are able to treat the analysis as a classification problem which continues to allow us to explore the data as a continuous variable.

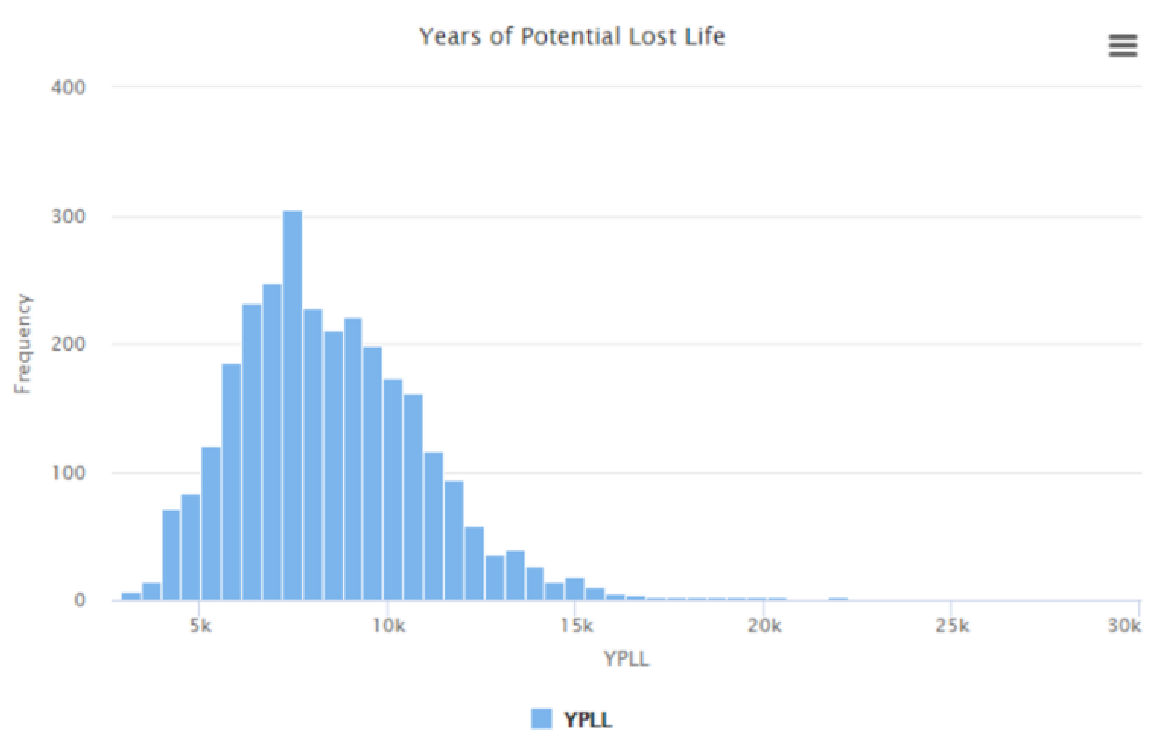


Figure 1: Distribution of the Target Variable: ‘Years of Potential Lost Life’

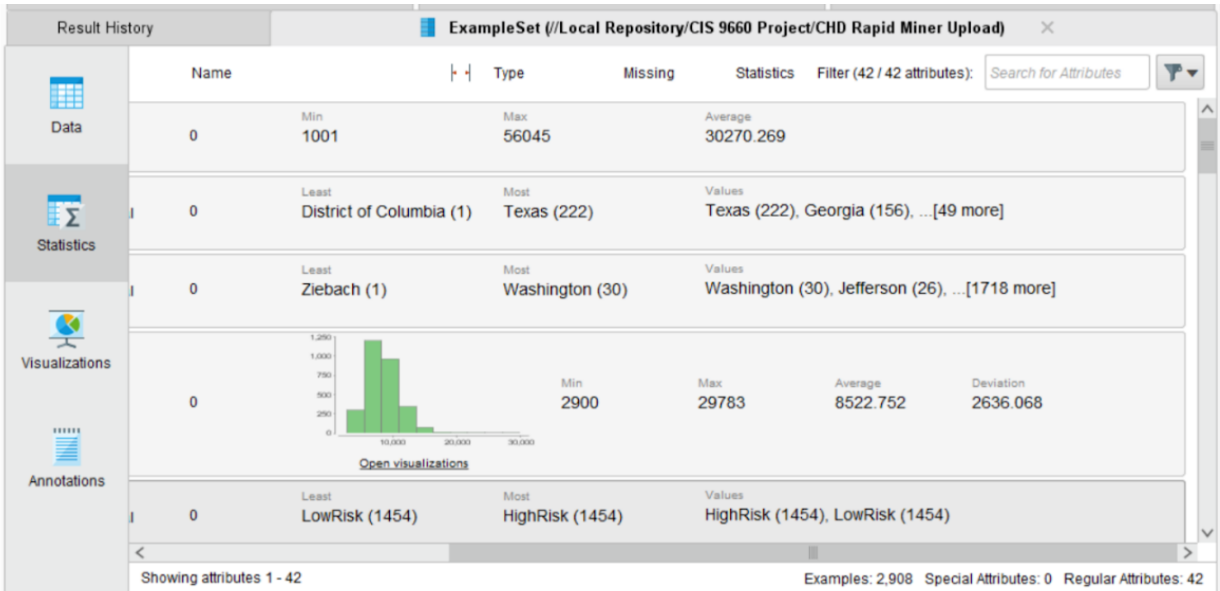


Figure 2: Years of Potential Lost Life Statistics

According to Figure 1 and 2, the data is positively skewed with a few extreme values. If the outliers are excluded, the variable becomes normally distributed.

Missing values, duplicates

SI	Attributes	Description	Missing Values
1	VCR	Violent Crime Rate Per 100,000 population	150
2	IDR	Injury Mortality Rate per 100,000 population	2
3	ADPM	Average daily amount of fine particulate matter in micrograms per cubic meter.	22
4	POV	County affected by a water violation: 1-Yes, 0-No	38
5	Dent	Dentists per 100,000 population	75
6	MHP	Mental Health Providers per 100,000 population	146
7	PHR	Discharges for Ambulatory Care Sensitive Conditions per 100,000 Medicare Enrollees	7
8	PS	Percentage of female Medicare enrollees having an annual mammogram (age 65 – 74)	5
9	PV	Percentage of annual Medicare enrollees having an annual flu vaccine	4

10	GR	Graduation rate	45
11	LBW	Percentage of births with low birth weight (<2500g)	6
12	FEI	Indicator of access to healthy foods – 0 is worst, 10 is best	19
13	WA	Percentage of the population for the places with actual physical activity	2
14	AI	Percentage of driving deaths with alcohol involvement	9
15	CR	Chlamydia cases per 100,000 population	24
16	TB	Births per 1,000 females ages 15-19	10
17	PCP	Primary Care Physicians per 100,000 population	91

Table 1: Attributes that had missing values

When we evaluated the data for completeness, we noticed that 21 out of 38 attributes had no missing values. According to Table 1, there were four attributes that had more than 50 missing values. The full dataset has 2,908 instances which decreases the significance of the 50 missing values. The dataset was mostly complete and took note that there were no duplicate attributes.

Relationship between variables

Firstly, we evaluated the relationship between variables by creating the correlation matrix.

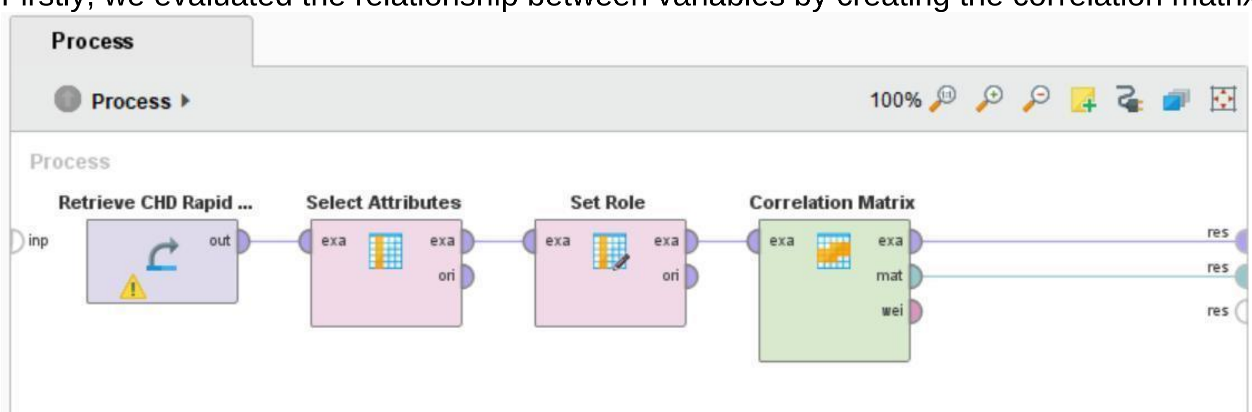


Figure 3: Correlation Matrix Process

We used RapidMiner Correlation Matrix operator to gauge the relationship strength between pairs of attributes and to help identify which variables could possibly serve as the best predictors in predicting our target variable. In Figure 3, one can see the operators used for this process. We decided to remove attributes that have little to no impact on the final result, which are County, State and Federal Information Processing Standard (FIPS).

According to Figure 4, the correlation heat map depicts many red-shaded and blue-shaded regions. The dark red boxes indicate a value that is closer to +1, whereas the dark blue boxes represent a value closer to -1. Lastly, the yellow boxes indicate a value closer to 0.

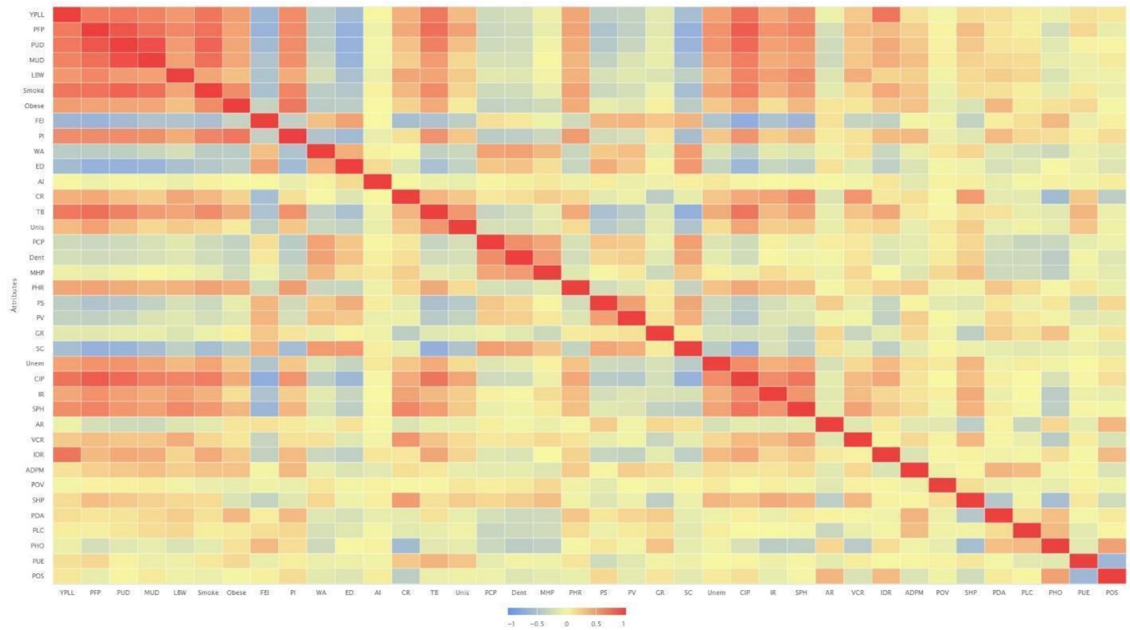


__Figure 4: Correlation Matrix Heat Map__

We examined the target variable of our dataset in relation to the other attributes. Looking at Figure 5 below, the darker shades of blue represents a higher correlation which is closer to 1 or -1. We noticed that in relation to YPLL, some attributes that had high correlation were Children In Poverty (CIP), Injury Death Rate (IDR), Smoke, and Teen Birth Rate (TB).

ExampleSet (Set Role) | Correlation Matrix (Correlation Matrix)

Attribut...	YPLL	PFP	PUD	MUD	LBW	Smoke	Obese	FEI	PI	WA
YPLL	1	0.678	0.687	0.627	0.514	0.703	0.493	-0.612	0.585	-0.413

ExampleSet (Set Role) | Correlation Matrix (Correlation Matrix)

ED	AI	CR	TB	Unis	PCP	Dent	MHP	PHR	PS	PV
-0.577	0.008	0.383	0.689	0.321	-0.275	-0.245	-0.065	0.442	-0.396	-0.353

ExampleSet (Set Role) | Correlation Matrix (Correlation Matrix)

Set Role.example set output → Correlation Matrix.example set

GR	SC	Unem	CIP	IR	SPH	AR	VCR	IDR	ADPM	POV
-0.137	-0.549	0.470	0.726	0.431	0.571	-0.060	0.244	0.718	0.115	-0.027

rix (Correlation Matrix)

SHP	PDA	PLC	PHO	PUE	POS
0.146	0.129	0.056	-0.076	0.119	0.088

__Figure 5: Target Variable vs. Other Attributes__

Top 10 Positively and Negatively Correlated Attributes (Relative to YPLL)

SI	Attribute	Correlation	Attribute	Correlation
1	% Children in Poverty	0.73	Food Environment Index	(0.61)
2	Injury Death Rate	0.72	% Excessive Drinking	(0.58)
3	% Smokers	0.70	% Some College	(0.55)
4	Teen Birth Rate	0.69	% W/Access to physical activity	(0.41)
5	Physically Unhealthy Days	0.69	% Mammogram Screened	(0.40)
6	% Fair/Poor Health	0.68	% Vaccinated	(0.35)
7	Mentally Unhealthy Days	0.63	Primary Care Docs/100,000	(0.28)
8	% Physically Inactive	0.59	Dentists/100,000	(0.25)
9	% Single Parent HH	0.57	Graduation Rate	(0.14)
10	% Low Birth Weight	0.51	% Homeowners	(0.08)

Table 2: Top 10 Positively and Negatively Correlated Attributes (Relative to YPLL)

Table 2 shows the top ten positively and negatively correlated attributes relative to our target variable, YPLL. The positively correlated attributes are mostly indicators of unhealthiness whereas negatively correlated attributes indicate healthy attributes. We noted that the socioeconomic status is also significant. Attributes like Children In Poverty (CIP), Teen Birth Rate (TB) and Percentage of Single Parent Households could indicate lower income counties. On the other hand, attributes like Percentage of Some College, Food Environment Index, and Percentage with Access to Physical Activity could be indicative of higher income counties. The Food Environment Index indicates whether people have access to healthy food and the Percentage with Access to Physical Activity indicates areas that people can be active in such as parks, track, or sports fields. Additionally, since lost life is measured as the difference between the age of death and 75, the younger an individual is at death the more heavily they would be weighted. While a death at the age of 65 is measured as 10, a death at the age of 15 would be measured as 60. Therefore, deaths of younger citizens due to gang violence, drug overdoses, or other non-health related measures would be of greater concern. This is likely captured in the socioeconomic attributes. Of note, the Percentage of Excessive Drinking is negatively correlated with our target variable, which was a surprising find during our research.

Figures 6 to 9: Scatter Plots – Most Positively Correlated

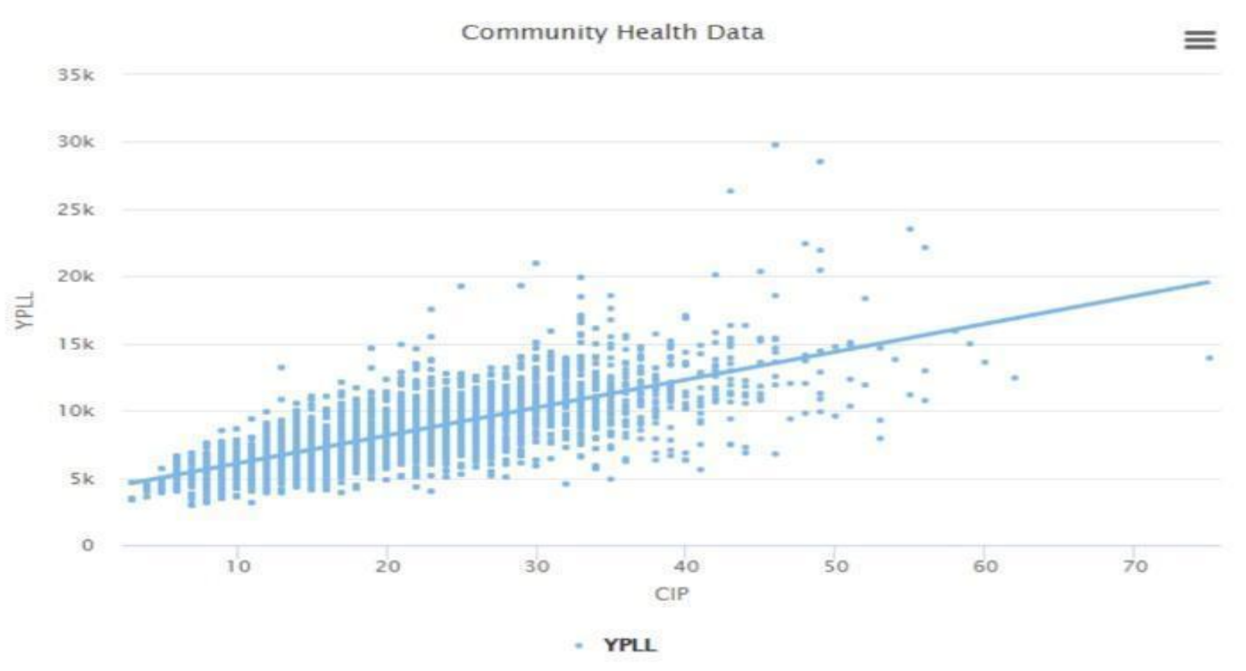


Figure 6: % Children in Poverty vs YPLL

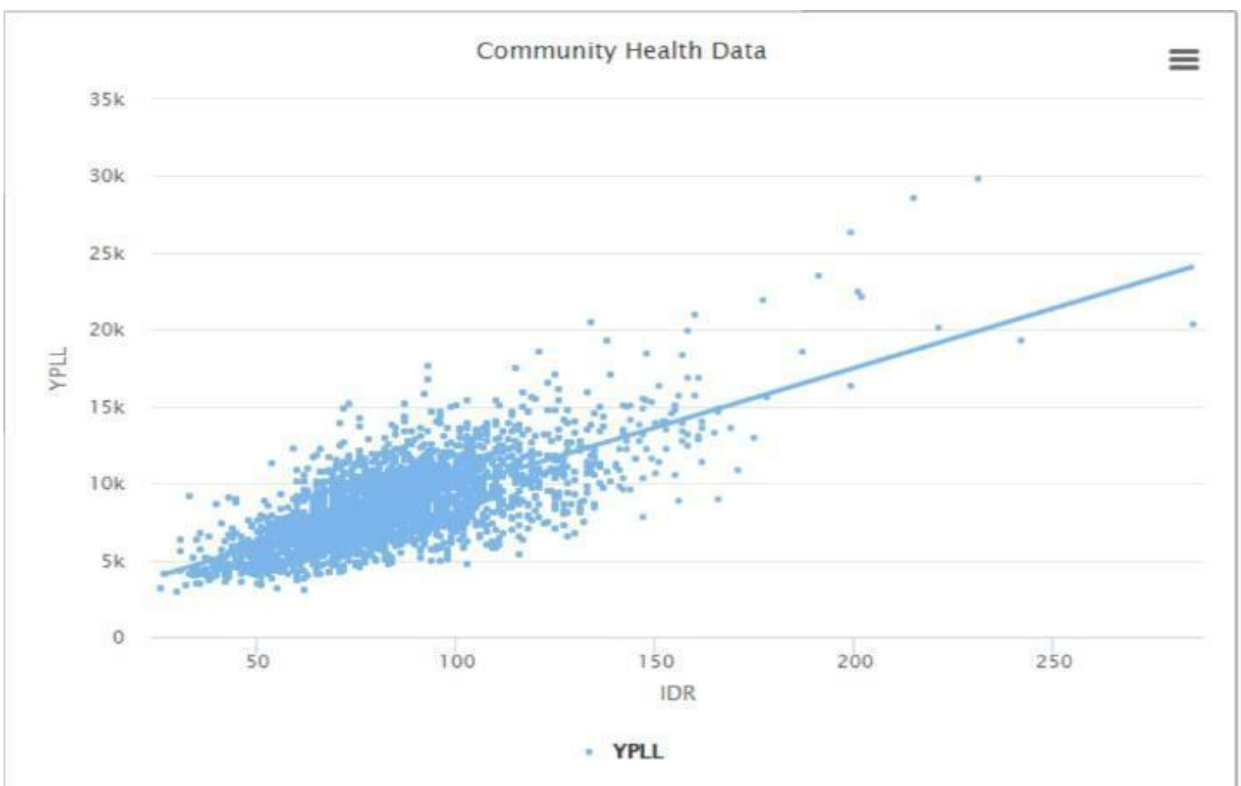


Figure 7: Injury Death Rate vs YPLL

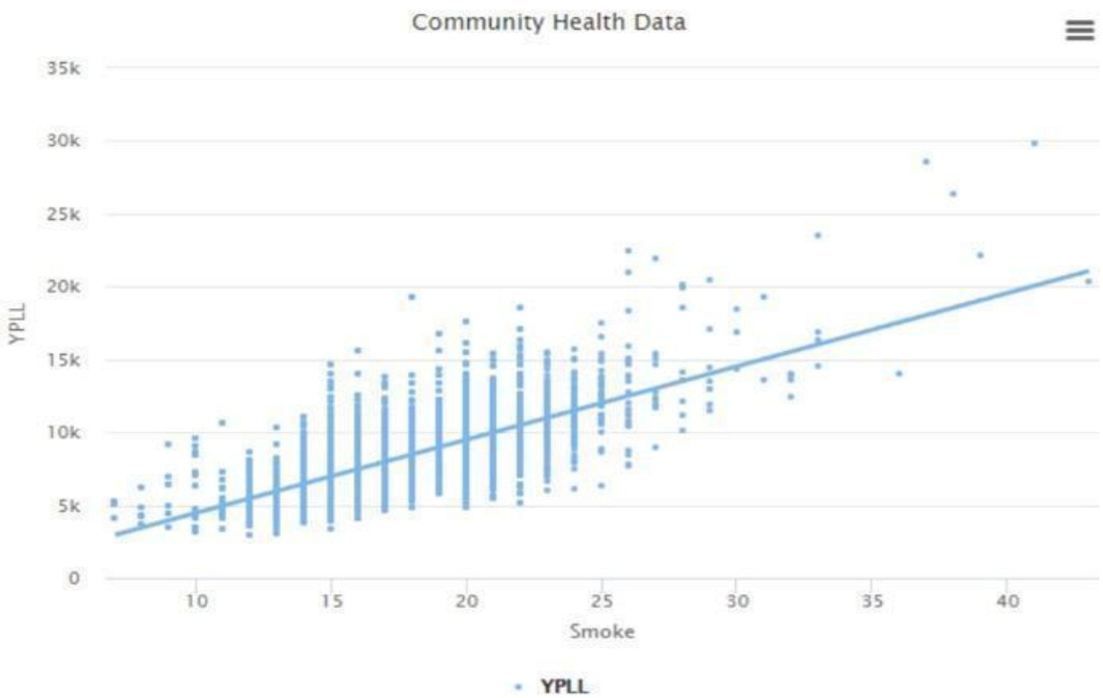


Figure 8: % Smokers vs YPLL

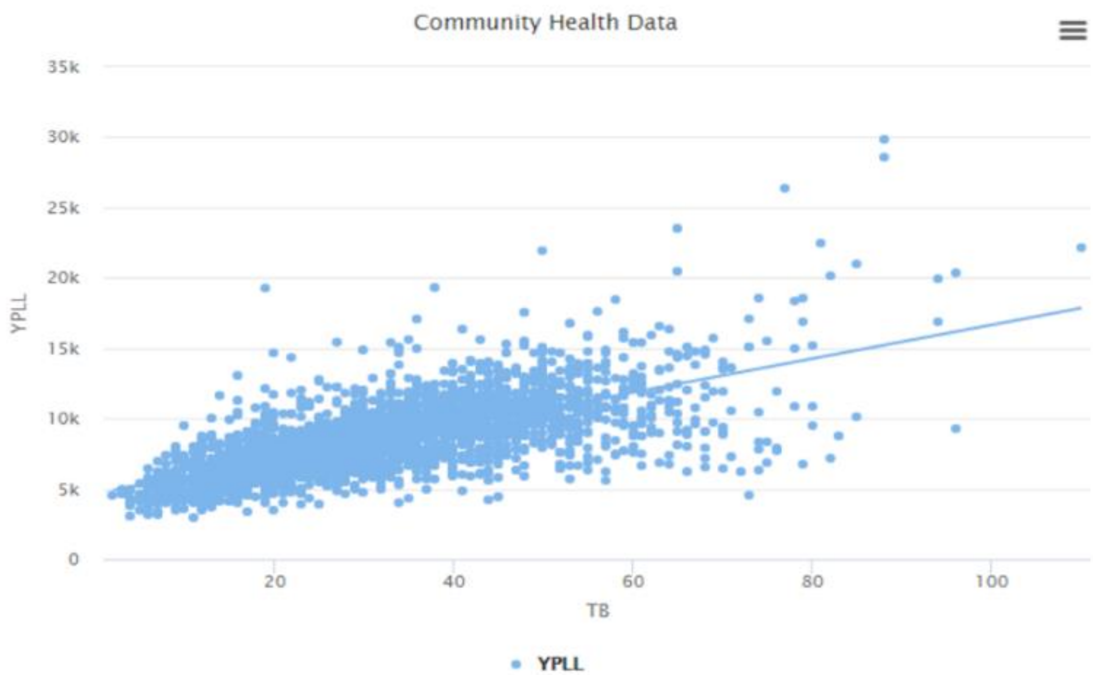


Figure 9: Teen Birth Rate vs YPLL

Figures 10 to 13: Scatter Plots – Most Negatively Correlated

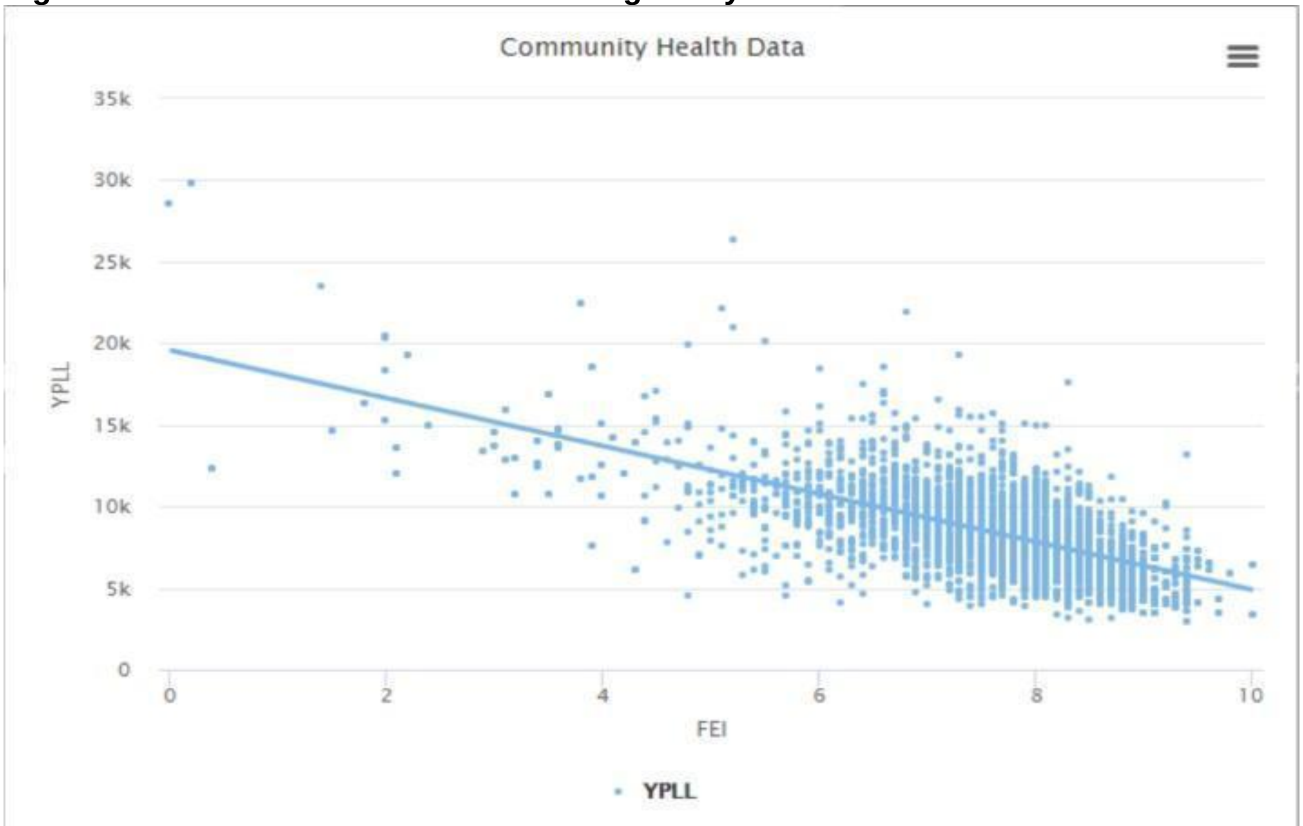


Figure 10: Food Environment Index vs. YPLL

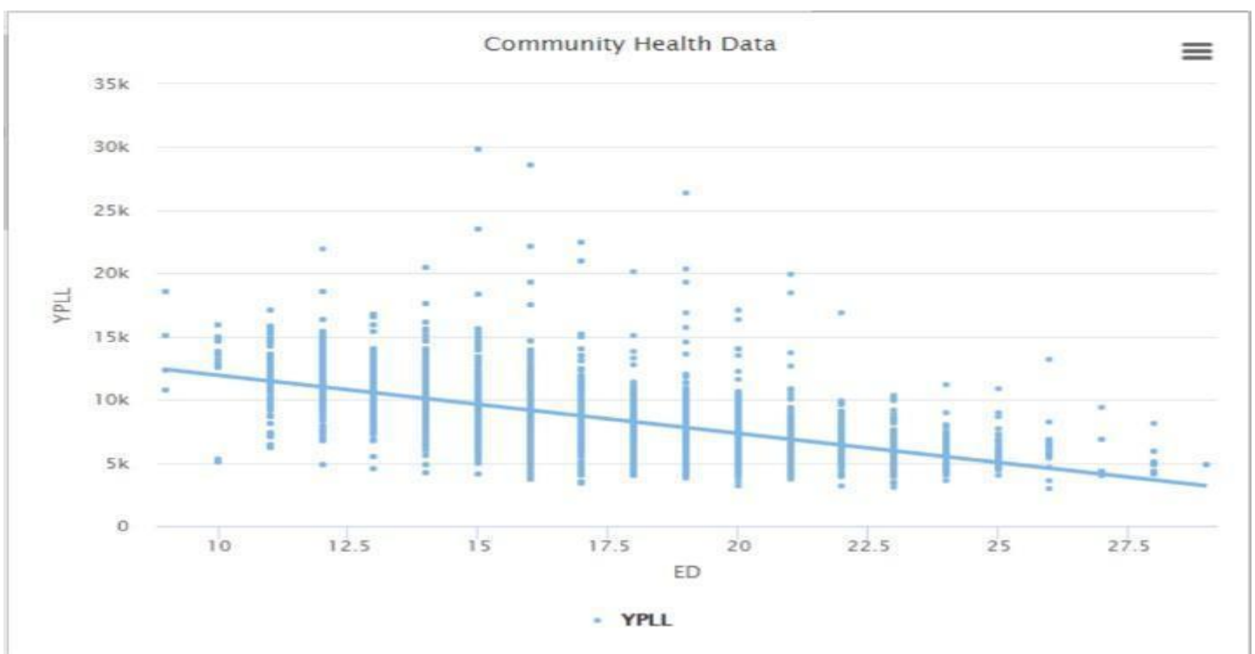


Figure 11: % Excessive Drinking vs YPLL

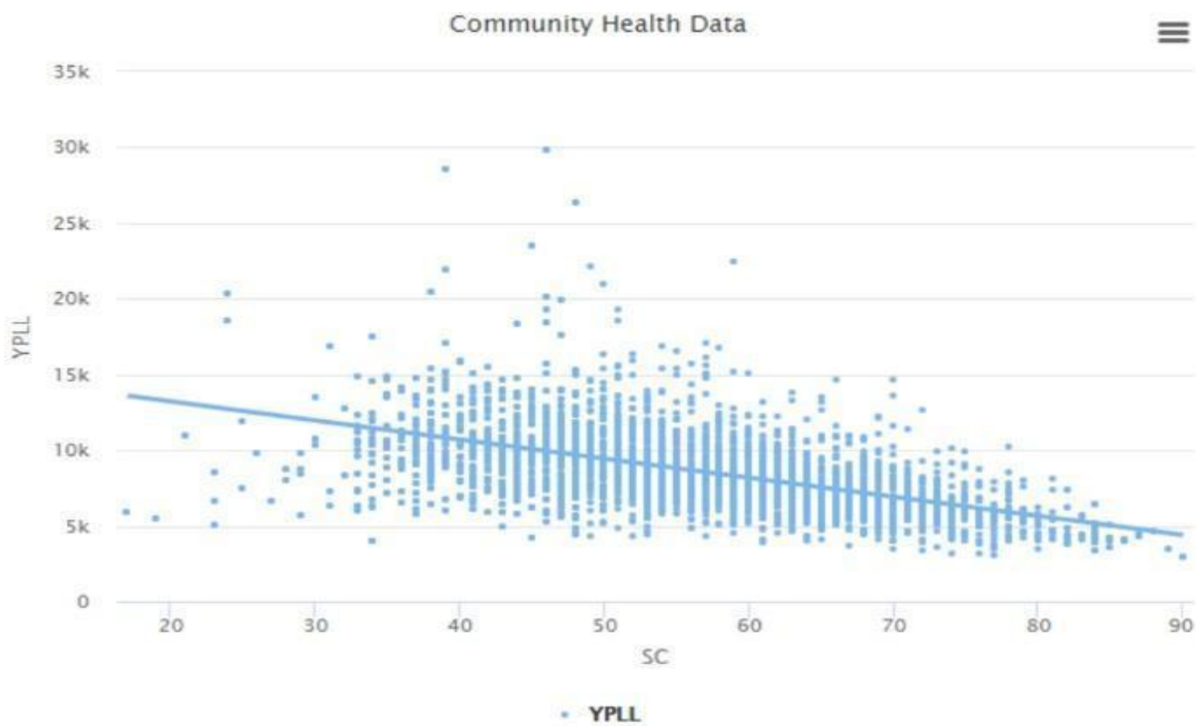


Figure 12: % Some College vs. YPLL

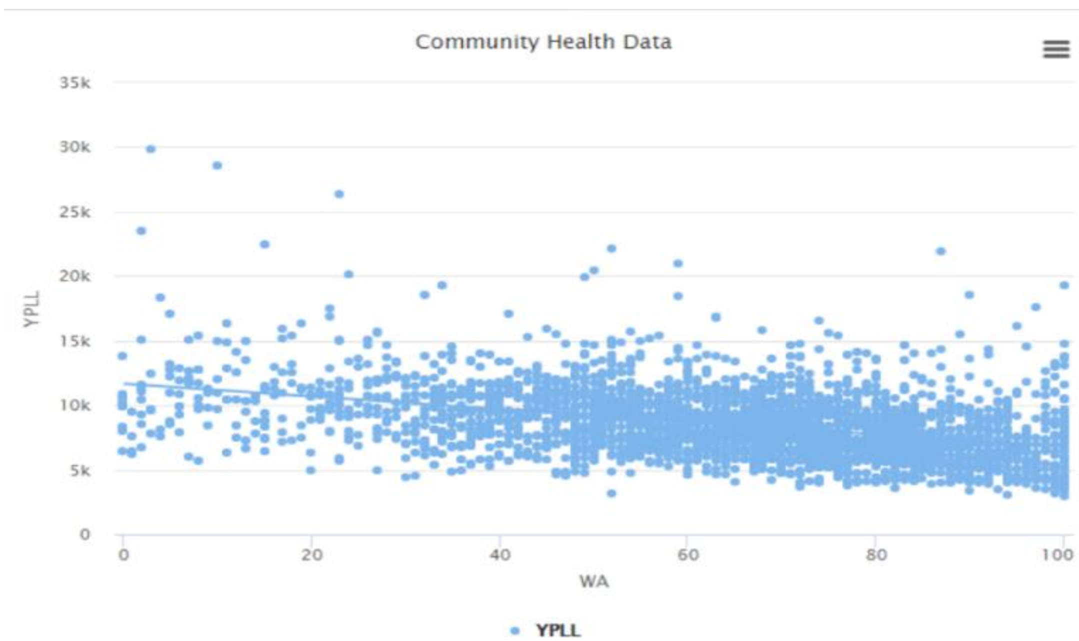


Figure 13: %Physically Active vs. YPLL

First Attribute	Second Attribute	Correlation
Physically Unhealthy Days	Mentally Unhealthy Days	0.91
% Fair/Poor Health	Physically Unhealthy Days	0.88
% Fair/Poor Health	% Children in Poverty	0.85
Physically Unhealthy Days	% Smokers	0.80
Physically Unhealthy Days	% Children in Poverty	0.77
% Fair/Poor Health	Teen Birth Rate	0.75
Mentally Unhealthy Days	% Smokers	0.74
% Fair/Poor Health	Mentally Unhealthy Days	0.74
% Fair/Poor Health	% Smokers	0.72
Teen Birth Rate	% Children in Poverty	0.72
% Children in Poverty	% Single Parent HH	0.71

Table 3: Pairwise Correlation

Looking at Table 3, you can see the results tabulated from RapidMiner's pairwise table. We limited it to correlations that are above 70% to eliminate some additional attributes from the final analysis. For example, Physically Unhealthy Days and Mentally Unhealthy Days are highly correlated to each other, which means that either should have a similar effect on our analysis and is why we can eliminate one of the attributes from each pair.

Outliers

Open in | Turbo Prep | Auto Model | Filter (2,908 / 2,908 examples): all

Row No.	FIPS	State	County	YPLL ↓	YPPLR	PFP	PUD	MUD
2227	46102	South Dakota	Oglala Lakota	29783	HighRisk	33	6.400	5.400
2201	46017	South Dakota	Buffalo	28531	HighRisk	31	5.800	4.800
1873	38085	North Dakota	Sioux	26337	HighRisk	32	5.700	4.900
2206	46031	South Dakota	Corson	23518	HighRisk	29	5.800	4.800
2216	46071	South Dakota	Jackson	22436	HighRisk	23	4.800	4
2232	46121	South Dakota	Todd	22124	HighRisk	30	5.900	4.900
1023	21189	Kentucky	Owsley	21923	HighRisk	26	5.600	5
1525	30003	Montana	Big Horn	20973	HighRisk	26	5.400	4.500
2224	46095	South Dakota	Mellette	20484	HighRisk	25	5.200	4.300
76	2158	Alaska	Kusilvak	20346	HighRisk	38	7.200	5.900
2210	46041	South Dakota	Dewey	20105	HighRisk	23	4.900	4.300

Figure 14: Outlier States

Figure 14 summarizes the outliers. There were several values that were over 20,000 that contributed to the positive skewed according to Figure 1. Of note, many of the values seemed to originate from North and South Dakota. The possible reason for this is because these states have some of the country's poorest and most rural counties. As such, healthcare may not be as accessible as it'd be in a wealthier suburban/urban county.

Baseline Model

Select Attributes → Remove (YPPL, County, FIPS, State)

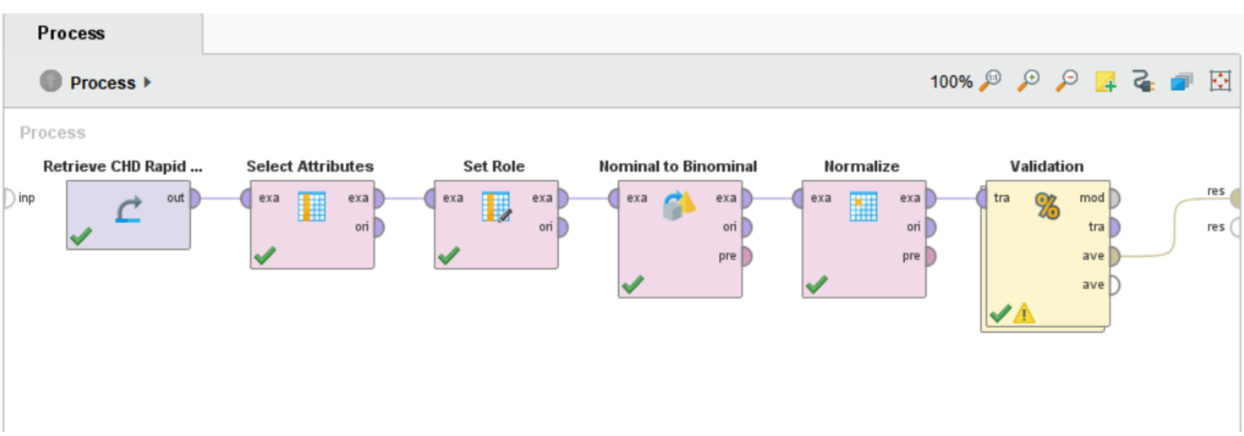


Figure 15: Baseline Model Process

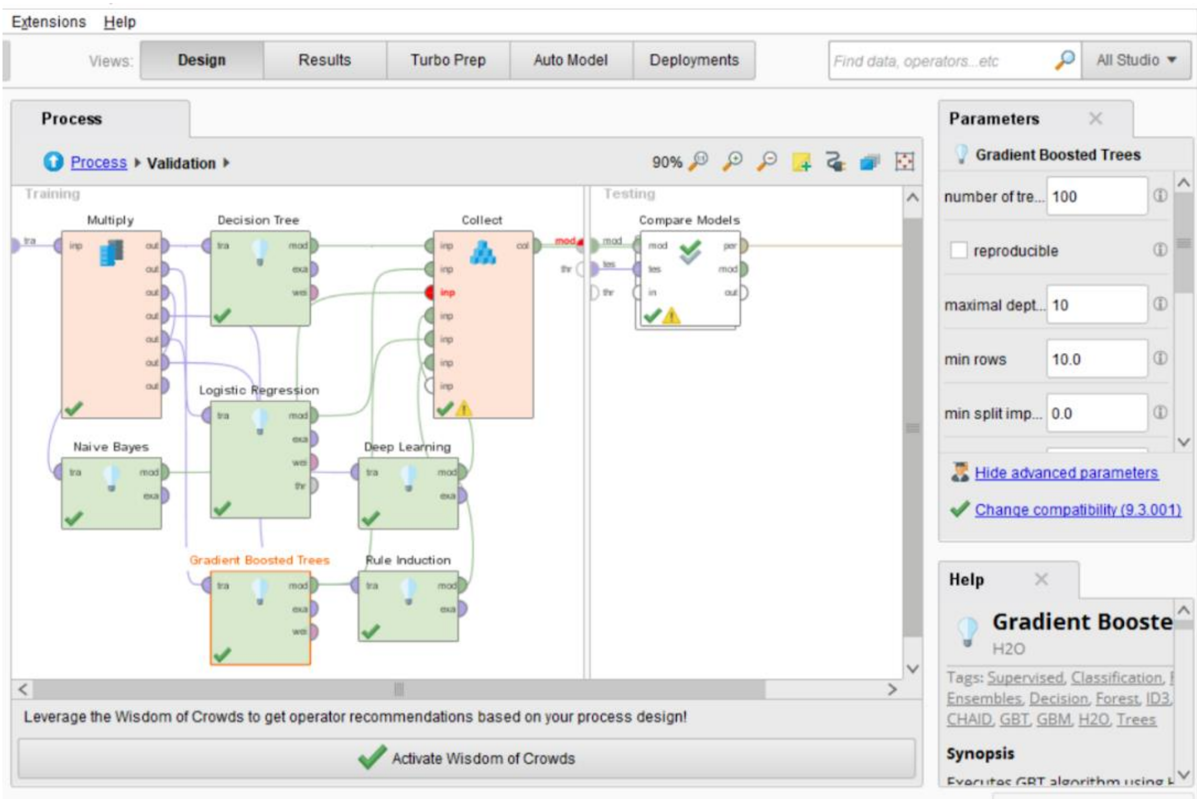


Figure 16: Baseline Model - Validation Operator

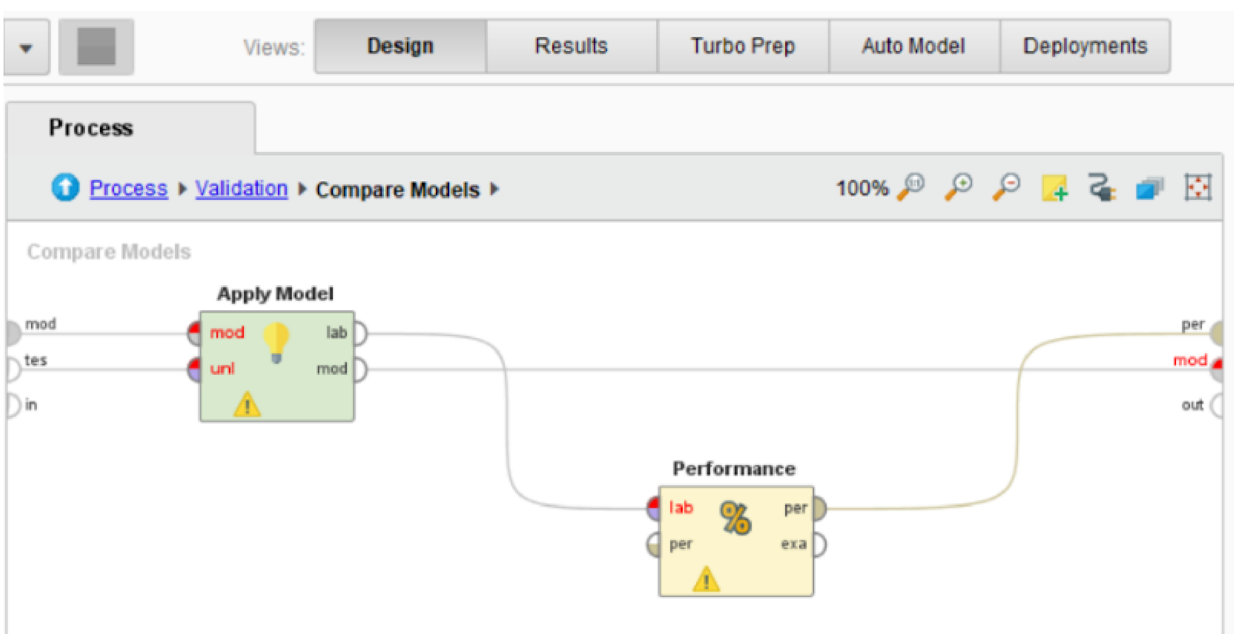


Figure 17: Baseline Model - Compare Models Operator

Decision Tree – Split Validation – Default Parameters

accuracy: 83.37%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	359	68	84.07%
pred. LowRisk	77	368	82.70%
class recall	82.34%	84.40%	

Figure 18: Decision Tree - Model Performance - Split Validation - Default Parameters

Accuracy	Precision	Recall	F-Measure
83.37%	82.70%	84.40%	83.54%

Logistic Regression – Split Validation – Default Parameters

accuracy: 90.25%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	394	43	90.16%
pred. LowRisk	42	393	90.34%
class recall	90.37%	90.14%	

Figure 19: Logistic Regression - Split Validation - Default Parameters

Accuracy	Precision	Recall	F-Measure
90.25%	90.34%	90.14%	90.24%

Naïve Bayes – Split Validation – Default Parameters

accuracy: 83.94%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	344	48	87.76%
pred. LowRisk	92	388	80.83%
class recall	78.90%	88.99%	

Figure 20: Naive Bayes - Split Validation - Default Parameters

Accuracy	Precision	Recall	F-Measure
83.94%	80.83%	88.99%	84.72%

Deep Learning – Split Validation – Default Parameters

accuracy: 89.79%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	391	44	89.89%
pred. LowRisk	45	392	89.70%
class recall	89.68%	89.91%	

Figure 21: Deep Learning - Split Validation - Default Parameters

Accuracy	Precision	Recall	F-Measure
89.79%	88.91%	90.14%	89.52%

Gradient Boosted Trees – Split Validation – Default Parameters

accuracy: 88.53%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	383	47	89.07%
pred. LowRisk	53	389	88.01%
class recall	87.84%	89.22%	

Figure 22: Gradient Boosted Trees - Split Validation - Default Parameters

Accuracy	Precision	Recall	F-Measure
88.53%	88.01%	89.22%	88.61%

Rule Induction – Split Validation – Default Parameters

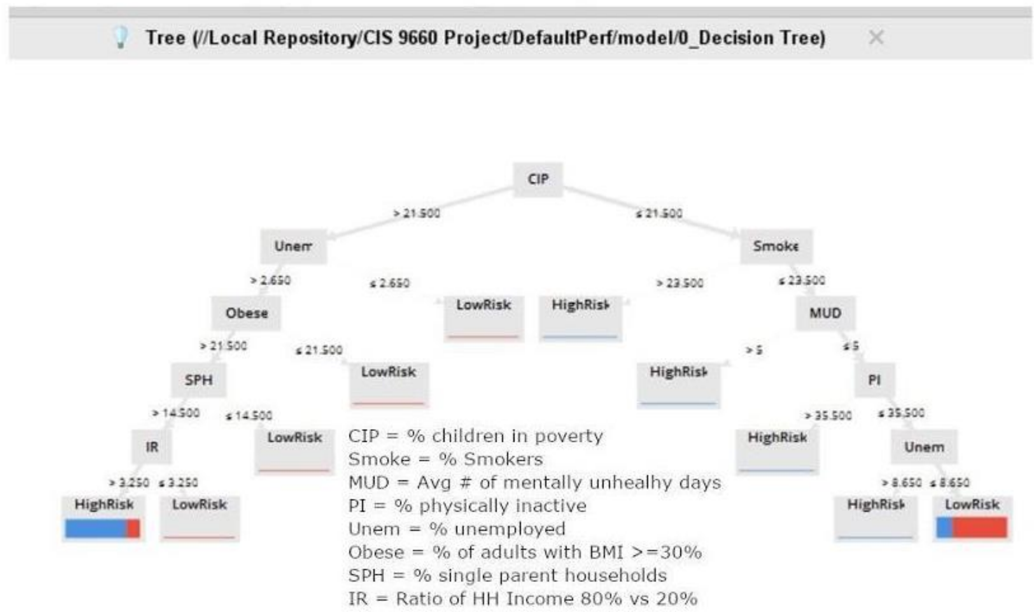
accuracy: 84.63%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	372	70	84.16%
pred. LowRisk	64	366	85.12%
class recall	85.32%	83.94%	

Figure 23: Rule Induction - Split Performance - Default Parameters

Accuracy	Precision	Recall	F-Measure
84.63%	85.12%	83.94%	84.53%

Default Decision Tree – Split Validation



Tree

```
CIP > 21.500
|   Unem > 2.650
|   |   Obese > 21.500
|   |   |   SPH > 14.500
|   |   |   |   IR > 3.250: HighRisk {HighRisk=770, LowRisk=173}
|   |   |   |   IR ≤ 3.250: LowRisk {HighRisk=0, LowRisk=2}
|   |   |   SPH ≤ 14.500: LowRisk {HighRisk=0, LowRisk=3}
|   |   Obese ≤ 21.500: LowRisk {HighRisk=0, LowRisk=3}
|   Unem ≤ 2.650: LowRisk {HighRisk=0, LowRisk=6}
CIP ≤ 21.500
|   Smoke > 23.500: HighRisk {HighRisk=4, LowRisk=0}
|   Smoke ≤ 23.500
|   |   MUD > 5: HighRisk {HighRisk=2, LowRisk=0}
|   |   MUD ≤ 5
|   |   |   PI > 35.500: HighRisk {HighRisk=2, LowRisk=0}
|   |   |   PI ≤ 35.500
|   |   |   |   Unem > 8.650: HighRisk {HighRisk=2, LowRisk=0}
|   |   |   |   Unem ≤ 8.650: LowRisk {HighRisk=238, LowRisk=831}
```

Figure 24: Default Decision Tree - Split Validation

According to Figure 24, this decision tree depicts that the percentage of children in poverty is the most important indicator of years of lost life. If the percentage of children in poverty is greater than 21.5% then the tree would move on to evaluate the percentage of unemployed. If the percentage of children in poverty is less than or equal to 21.5%, the tree would further evaluate the percentage of smokers. Although it is a mix of both health related attributes and socioeconomic indicators, it appears that county wealth is slightly more important to the model than county health.

Default Rule Model – Split Validation

RuleModel

```
if TB > 32.500 and MUD > 4.050 then HighRisk  (571 / 46)
if IDR ≤ 76.500 and CIP ≤ 18.500 then LowRisk  (15 / 483)
if CIP ≤ 21.500 and IDR ≤ 85.500 and PHR ≤ 4310 then LowRisk  (10 / 105)
if TB > 28.500 and IDR > 75.500 and CIP > 21.500 then HighRisk  (166 / 17)
if PI ≤ 23.500 and IDR ≤ 98.500 and Unem ≤ 4.450 then LowRisk  (8 / 72)
if PDA ≤ 79.500 and FEI > 7.450 and IDR ≤ 99.500 and CR ≤ 246.300 then LowRisk  (2 / 39)
if IDR ≤ 66.500 and CR ≤ 571.050 and AR ≤ 15.550 then LowRisk  (2 / 44)
if Smoke > 18.500 and IDR > 78.500 and PLC ≤ 39.500 then HighRisk  (50 / 5)
if PDA ≤ 83.500 and MHP ≤ 354 and PFP ≤ 16.500 and IDR ≤ 85.500 then LowRisk  (3 / 37)
if PDA > 82.500 and PV ≤ 44.500 and ED ≤ 20.500 and GR > 88.500 then HighRisk  (32 / 2)
if PDA ≤ 78.500 and FEI > 7.550 and SHP > 13.500 then LowRisk  (3 / 26)
if PI > 24.500 and FEI ≤ 7.250 and VCR > 423.500 then HighRisk  (17 / 0)
if AR ≤ 9.750 and IDR ≤ 93.500 and PUD ≤ 4.150 then LowRisk  (2 / 27)
if ADPM > 8.850 and PS ≤ 39.500 and VCR > 234 then HighRisk  (18 / 1)
if LBW > 6.500 and Obese > 32.500 and MHP > 114.500 and LBW > 7.500 then HighRisk  (18 / 1)
if Dent > 37.500 and IDR ≤ 99.500 and PUE ≤ 22.900 and PV ≤ 43.500 then LowRisk  (1 / 15)
if Unem > 4.650 and PLC > 43.500 then HighRisk  (16 / 0)
if IDR ≤ 95.500 and FEI > 7.750 and GR ≤ 89.500 then LowRisk  (2 / 20)
if POS ≤ 20.700 and PCP ≤ 62 and PUE > 21.250 and GR > 81.500 and CIP > 11.500 then HighRisk  (23 / 4)
if VCR > 176.500 and WA ≤ 80.500 and PHR ≤ 4855.500 then LowRisk  (4 / 24)
if GR ≤ 88.500 and ED > 17.500 then HighRisk  (13 / 1)
if SC > 66.500 and IR > 4 then LowRisk  (1 / 16)
if IDR > 97.500 and PI > 23.500 and VCR ≤ 162 then HighRisk  (17 / 1)
```

```
if IDR ≤ 76.500 and CIP ≤ 18.500 then LowRisk  (15 / 483)
if CIP ≤ 21.500 and IDR ≤ 85.500 and PHR ≤ 4310 then LowRisk  (10 / 105)
if TB > 28.500 and IDR > 75.500 and CIP > 21.500 then HighRisk  (166 / 17)
if PI ≤ 23.500 and IDR ≤ 98.500 and Unem ≤ 4.450 then LowRisk  (8 / 72)
if PDA ≤ 79.500 and FEI > 7.450 and IDR ≤ 99.500 and CR ≤ 246.300 then LowRisk  (2 / 39)
if IDR ≤ 66.500 and CR ≤ 571.050 and AR ≤ 15.550 then LowRisk  (2 / 44)
if Smoke > 18.500 and IDR > 78.500 and PLC ≤ 39.500 then HighRisk  (50 / 5)
if PDA ≤ 83.500 and MHP ≤ 354 and PFP ≤ 16.500 and IDR ≤ 85.500 then LowRisk  (3 / 37)
if PDA > 82.500 and PV ≤ 44.500 and ED ≤ 20.500 and GR > 88.500 then HighRisk  (32 / 2)
if PDA ≤ 78.500 and FEI > 7.550 and SHP > 13.500 then LowRisk  (3 / 26)
if PI > 24.500 and FEI ≤ 7.250 and VCR > 423.500 then HighRisk  (17 / 0)
if AR ≤ 9.750 and IDR ≤ 93.500 and PUD ≤ 4.150 then LowRisk  (2 / 27)
if ADPM > 8.850 and PS ≤ 39.500 and VCR > 234 then HighRisk  (18 / 1)
if LBW > 6.500 and Obese > 32.500 and MHP > 114.500 and LBW > 7.500 then HighRisk  (18 / 1)
if Dent > 37.500 and IDR ≤ 99.500 and PUE ≤ 22.900 and PV ≤ 43.500 then LowRisk  (1 / 15)
if Unem > 4.650 and PLC > 43.500 then HighRisk  (16 / 0)
if IDR ≤ 95.500 and FEI > 7.750 and GR ≤ 89.500 then LowRisk  (2 / 20)
if POS ≤ 20.700 and PCP ≤ 62 and PUE > 21.250 and GR > 81.500 and CIP > 11.500 then HighRisk  (23 / 4)
if VCR > 176.500 and WA ≤ 80.500 and PHR ≤ 4855.500 then LowRisk  (4 / 24)
if GR ≤ 88.500 and ED > 17.500 then HighRisk  (13 / 1)
if SC > 66.500 and IR > 4 then LowRisk  (1 / 16)
if IDR > 97.500 and PI > 23.500 and VCR ≤ 162 then HighRisk  (17 / 1)
else LowRisk  (21 / 28)

correct: 1877 out of 2029 training examples.
```

Figure 25: Default Rule Model – Split Validation

Decision Tree – Cross Validation – Default Parameters

accuracy: 75.60%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	108	34	76.06%
pred. LowRisk	37	112	75.17%
class recall	74.48%	76.71%	

Figure 26: Decision Tree - Cross Validation - Default Parameters

Accuracy	Precision	Recall	F-Measure
75.60%	75.17%	76.71%	75.93%

Logistic Regression – Cross Validation – Default Parameters

accuracy: 88.66%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	127	15	89.44%
pred. LowRisk	18	131	87.92%
class recall	87.59%	89.73%	

Figure 27: Logistic Regression - Cross Validation - Default Parameters

Accuracy	Precision	Recall	F-Measure
88.66%	87.92%	89.73%	88.81%

Naïve Bayes – Cross Validation – Default Parameters

accuracy: 80.41%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	110	22	83.33%
pred. LowRisk	35	124	77.99%
class recall	75.86%	84.93%	

Figure 28: Naive Bayes - Cross Validation - Default Parameters

Accuracy	Precision	Recall	F-Measure
81.79%	77.99%	84.93%	81.31%

Deep Learning – Cross Validation – Default Parameters

accuracy: 87.63%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	124	15	89.21%
pred. LowRisk	21	131	86.18%
class recall	85.52%	89.73%	

Figure 29: Deep Learning - Cross Validation - Default Parameters

Accuracy	Precision	Recall	F-Measure
87.63%	86.18%	89.73%	87.92%

Gradient Boosted Trees – Cross Validation – Default Parameters

Table View Plot View

accuracy: 88.66%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	128	15	89.51%
pred. LowRisk	18	130	87.84%
class recall	87.67%	89.66%	

Figure 30: Gradient Boosted Trees - Cross Validation - Default Parameters

Accuracy	Precision	Recall	F-Measure
88.66%	87.84%	89.66%	88.74%

Rule Induction – Cross Validation – Default Parameters

Table View Plot View

accuracy: 88.66%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	128	15	89.51%
pred. LowRisk	18	130	87.84%
class recall	87.67%	89.66%	

Figure 31: Rule Induction - Cross Validation - Default Parameters

Accuracy	Precision	Recall	F-Measure
88.66%	87.84%	89.66%	88.74%

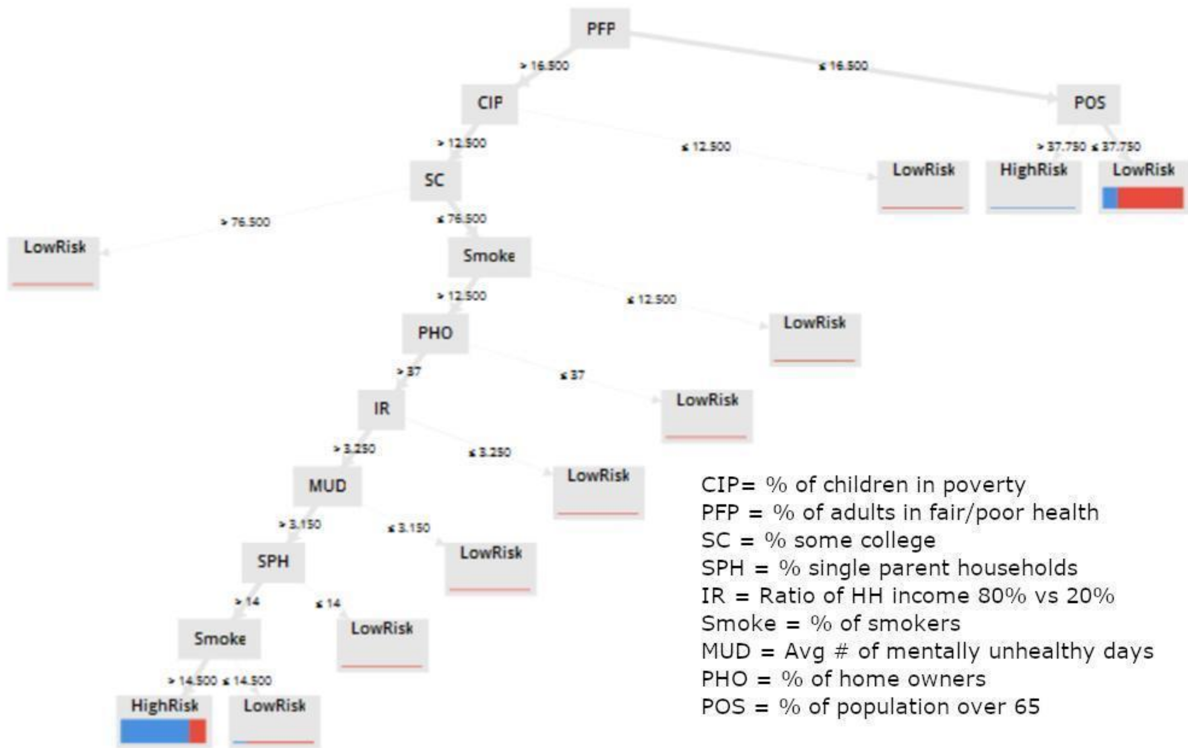
Overall, the accuracy was higher for Split Validation than Cross Validation for each model type.

Model	Split Validation	Cross Validation
Logistic Regression	90.25%	88.66%
Decision Tree	79.24%	75.60%
Naïve Bayes	83.94%	81.79%
Deep Learning	89.79%	87.63%
Gradient Boosted Trees	88.53%	88.66%
Rule Induction	84.63%	88.66%

Table 4: Summary of Split Validation and Cross Validation Models

After evaluating all the models, the accuracies were better for split validation in 4 out of 6 models. Thus, our base model would be built using split validation rather than cross validation.

Default Decision Tree – Cross Validation



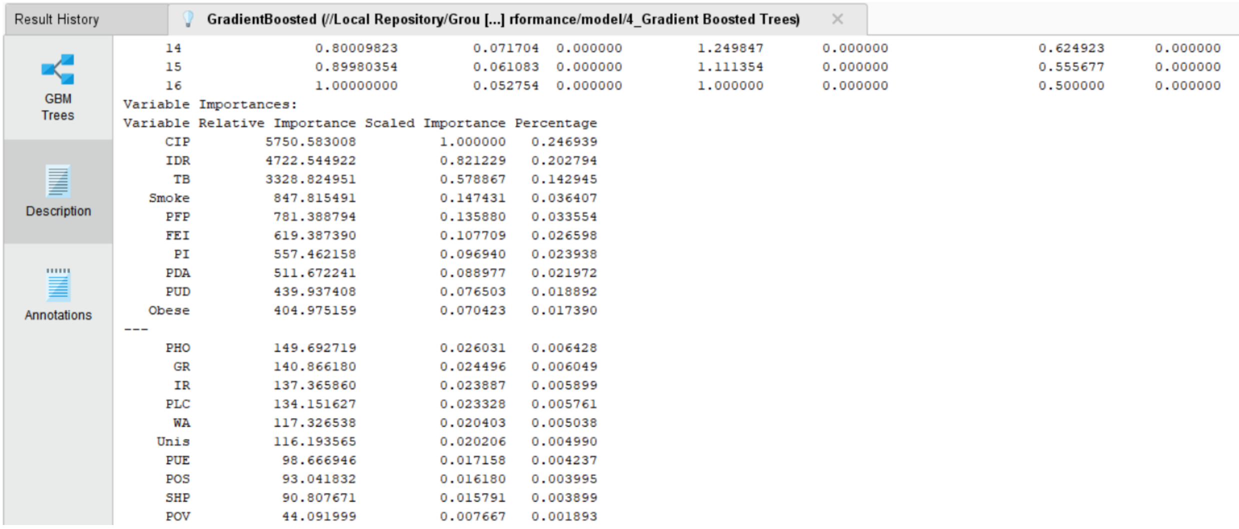
Tree

```
PFP > 16.500
|   CIP > 12.500
|   |   SC > 76.500: LowRisk {HighRisk=0, LowRisk=7}
|   |   SC ≤ 76.500
|   |   |   Smoke > 12.500
|   |   |   |   PHO > 37
|   |   |   |   |   IR > 3.250
|   |   |   |   |   |   MUD > 3.150
|   |   |   |   |   |   |   SPH > 14
|   |   |   |   |   |   |   |   Smoke > 14.500: HighRisk {HighRisk=1082, LowRisk=264}
|   |   |   |   |   |   |   |   Smoke ≤ 14.500: LowRisk {HighRisk=6, LowRisk=30}
|   |   |   |   |   |   |   SPH ≤ 14: LowRisk {HighRisk=0, LowRisk=2}
|   |   |   |   |   |   MUD ≤ 3.150: LowRisk {HighRisk=0, LowRisk=2}
|   |   |   |   |   IR ≤ 3.250: LowRisk {HighRisk=0, LowRisk=3}
|   |   |   |   PHO ≤ 37: LowRisk {HighRisk=0, LowRisk=5}
|   |   |   Smoke ≤ 12.500: LowRisk {HighRisk=0, LowRisk=6}
|   CIP ≤ 12.500: LowRisk {HighRisk=0, LowRisk=11}
PFP ≤ 16.500
|   POS > 37.750: HighRisk {HighRisk=2, LowRisk=0}
|   POS ≤ 37.750: LowRisk {HighRisk=219, LowRisk=978}
```

Figure 32: Default Decision Tree - Cross Validation

According to Figure 32, the decision tree that was produced using cross validation is very different from the decision tree that was produced by split validation, Figure 24. Although, the Percentage of Children in Poverty is high in the split validation decision tree, it is no longer at the top for the cross-validation decision tree and was replaced by Percentage of Adults in Fair or Poor Health, which is also not included in the Figure 24.

Most Important Features




```
     14          0.80009823       0.071704  0.000000       1.249847     0.000000          0.624923     0.000000
     15          0.89980354       0.061083  0.000000       1.111354     0.000000          0.555677     0.000000
     16          1.00000000       0.052754  0.000000       1.000000     0.000000          0.500000     0.000000
Variable Importances:
Variable Relative Importance Scaled Importance Percentage
     CIP         5750.583008          1.000000   0.246939
     IDR         4722.544922          0.821229   0.202794
      TB         3328.824951          0.578867   0.142945
   Smoke          847.815491          0.147431   0.036407
     PFP          781.388794          0.135880   0.033554
     FEI          619.387390          0.107709   0.026598
      PI          557.462158          0.096940   0.023938
     PDA          511.672241          0.088977   0.021972
     PUD          439.937408          0.076503   0.018892
   Obese          404.975159          0.070423   0.017390
---
     PHO          149.692719          0.026031   0.006428
      GR          140.866180          0.024496   0.006049
      IR          137.365860          0.023887   0.005899
     PLC          134.151627          0.023328   0.005761
      WA          117.326538          0.020403   0.005038
    Unis          116.193565          0.020206   0.004990
     PUE           98.666946          0.017158   0.004237
     POS           93.041832          0.016180   0.003995
     SHP           90.807671          0.015791   0.003899
     POV           44.091999          0.007667   0.001893
```

Figure 33: Important Features

SI	Most Important Variables	Description
1	CIP	%Children in Poverty
2	IDR	Injury Death Rate
3	TB	Teen Birth Rate
4	Smoke	%Smokers
5	FEI	Food Environment Index
6	PDA	%Drive Alone
7	PUD	Physically Unhealthy Days
8	PI	%Physically Inactive
9	PFP	%Fair / Poor
10	Obese	%Obese

Table 5: Description List of Important Features

Feature Engineering

Prior to doing any complex feature engineering, we needed to address missing values and normalize the dataset, as we were unable to perform PCA without handling the missing values. The dataset contains around 600 missing values which needed to be handled. We compared resolving it using two ways – replace missing value with average versus impute missing values using k-NN – results were similar so neither method was superior. We opted to replace missing values with average for simplicity and quicker processing time.

The majority of attributes were a percentage of relevant population, some attributes are a rate (# for every 100,000 people) and some are other types of numerical values. These are identified for each feature in the table below as "Feature Type".

We normalized values of the features with numerical values as the ranges and magnitudes varied. For example, MUD and PUD had possible ranges of 0 to 30, percentage features had a possible full range of 100, rate features of 100,000 where the section of the ranges with values in the feature varies.

Sl	Feature	Feature Type	Min Value	Max Value	Min after Norm	Max after Norm
1	Obese	Percentage	14	50	-3.9	3.8
2	Smoke	Percentage	7	43	-3.0	6.7
3	LBW	Percentage	3	19	-2.5	5.4
4	ED	Percentage	9	29	-2.5	3.5
5	WA	Percentage	0	100	-2.9	1.6
6	PI	Percentage	8	45	-3.4	3.6
7	PFP	Percentage	8	41	-2.0	4.9
8	Unis	Percentage	2	31	-1.8	4.1
9	POS	Percentage	4.8	56.9	-3.1	8.8
10	PUE	Percentage	7.2	41.2	-4.4	5.5
11	PHO	Percentage	20	90	-6.3	2.3
12	Unem	Percentage	1.6	20.1	-1.8	9.4
13	AI	Percentage	0	100	-2.1	5.3
14	PLC	Percentage	0	85	-2.5	4.3
15	PDA	Percentage	5	96	-11.1	2.3
16	SHP	Percentage	4	71	-2.2	12.3

17	PV	Percentage	4	65	-4.0	2.5
18	PS	Percentage	7	62	-4.5	3.0
19	SPH	Percentage	7	80	-2.5	4.6
20	CIP	Percentage	3	75	-2.0	5.7
21	SC	Percentage	17	90	-3.5	2.8
22	CR	Rate (per 100k)	40	2897	-1.3	10.0
23	IDR	Rate (per 100k)	26	285	-2.3	8.2
24	VCR	Rate (per 100k)	0	1820	-1.3	8.3
25	AR	Rate (per 100k)	0	48.9	-2.2	6.1
26	PHR	Rate (per 100k)	471	17731	-2.5	7.3
27	MHP	Rate (per 100k)	4	2003	-0.9	11.7
28	Dent	Rate (per 100k)	0	725	-1.5	23.0
29	PCP	Rate (per 100k)	2	447	-1.6	12.0
30	YPLL	Rate (per 100k)	2900	29783	-2.1	8.0
31	FEI	Numerical (Scale)	0	10	-6.8	2.2
32	MUD	Numerical (other)	2.5	6	-2.5	3.4
33	PUD	Numerical (other)	2.3	7.2	-2.3	4.5
34	TB	Rate (per 1000)	2	110	-1.9	5.1
35	POV	Binary	n/a	n/a	n/a	n/a
36	ADPM	Numerical (other)	3	19.7	-3.3	5.6
37	IR	Ratio	2.7	9.1	-2.5	6.2
38	GR	Rate (other)	36	100	-7.3	1.6

Table 6: List of Features before and after Normalization

[Section Change from Phase II to Phase III revision: Table 6 updated, Figure 34 removed, Table 7 & Table 8 removed and combined into Table 6, Text Updated]

Principal Component Analysis (PCA)

Then we performed PCA on this dataset as follows:

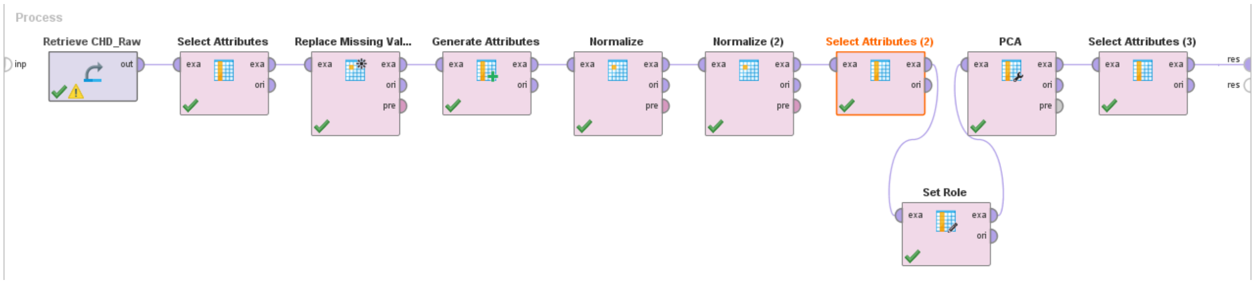


Figure 34: Creation of PCA process flow

This process flow includes the data pre-processing mentioned above, plus it sets the role of feature "YPLL" as the target variable.

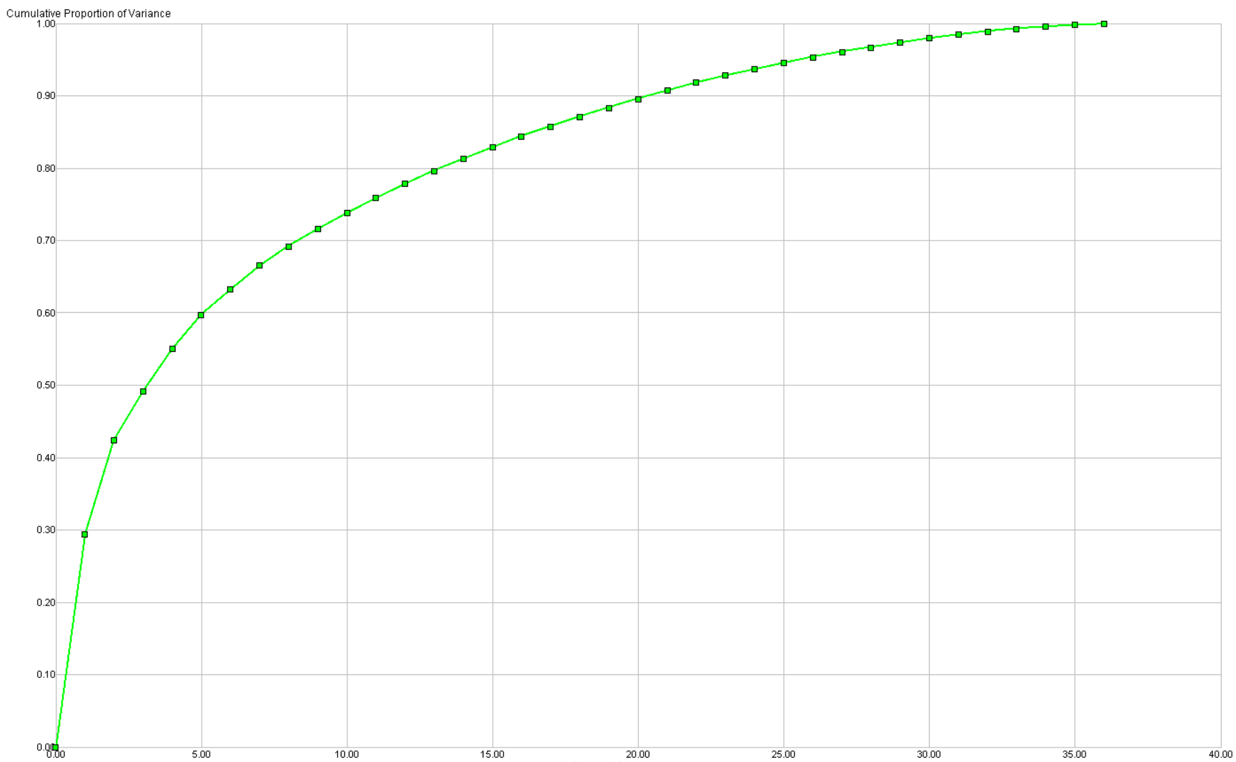


Figure 35: Creation of PCA Plot

This plot shows diminishing returns as the number of Principal Components (PC) calculated increases. From the plot we see that about 60% variance in the target variable (YPLL) is described by 5 Principal Components. From these results we decided not to add any of the PCs to the model, since the addition of the PCs adds to the complexity of interpretation and there isn't significant gain by adding it into the model. Instead, we used it to further explore the features to inform our understanding of their importance in the model.

We highlight the 5 PC (principle components) below that explain the 60% variance:

VARS	PC1	PC2	PC3	PC4	PC5	PC1.a	PC2.a	PC3.a	PC4.a	PC5.a	sum
PI	0.214	-0.193	-0.106	-0.037	-0.161	0.063	-0.025	-0.007	-0.002	-0.008	0.105
SPH	0.212	0.157	-0.081	-0.161	-0.094	0.063	0.020	-0.005	-0.009	-0.004	0.102
SC	-0.227	0.131	-0.118	-0.066	-0.089	-0.067	0.017	-0.008	-0.004	-0.004	0.100
CIP	0.276	0.043	0.084	-0.103	-0.014	0.081	0.006	0.006	-0.006	-0.001	0.099
CR-P	0.141	0.282	-0.122	0.071	-0.166	0.042	0.037	-0.008	0.004	-0.008	0.098
TB-P	0.245	-0.004	0.105	0.165	-0.151	0.072	-0.001	0.007	0.010	-0.007	0.097
Obese	0.173	-0.159	-0.157	-0.011	-0.285	0.051	-0.021	-0.011	-0.001	-0.013	0.096
FEI	-0.22	-0.133	-0.083	0.06	0.1	-0.065	-0.017	-0.006	0.004	0.005	0.096
MUD	0.24	-0.013	-0.093	-0.166	0.148	0.071	-0.002	-0.006	-0.010	0.007	0.095
PFP	0.284	0.036	-0.016	0.037	0.019	0.084	0.005	-0.001	0.002	0.001	0.093
PUD	0.27	0	-0.026	-0.083	0.112	0.080	0.000	-0.002	-0.005	0.005	0.092
Smoke	0.243	-0.028	-0.083	-0.065	-0.051	0.072	-0.004	-0.006	-0.004	-0.002	0.087
IR	0.172	0.179	-0.046	-0.13	0.039	0.051	0.023	-0.003	-0.008	0.002	0.087
LBW	0.203	0.055	-0.166	-0.113	-0.006	0.060	0.007	-0.011	-0.007	0.000	0.085
PS	-0.16	-0.009	-0.199	-0.267	-0.156	-0.047	-0.001	-0.013	-0.016	-0.007	0.085
PCP-P	-0.116	0.242	-0.023	-0.194	-0.126	-0.034	0.031	-0.002	-0.011	-0.006	0.085
WA	-0.154	0.218	-0.084	-0.07	0.014	-0.045	0.028	-0.006	-0.004	0.001	0.084
SHP	0.081	0.325	0.036	0.035	0.282	0.024	0.042	0.002	0.002	0.013	0.084
IDR-P	0.147	-0.06	0.297	-0.166	-0.04	0.043	-0.008	0.020	-0.010	-0.002	0.083
Unem	0.183	0.06	0.071	-0.125	0.184	0.054	0.008	0.005	-0.007	0.009	0.083
ED	-0.219	0.083	0.001	0.06	-0.034	-0.065	0.011	0.000	0.004	-0.002	0.081
PHO	-0.053	-0.369	0.098	-0.064	0.126	-0.016	-0.048	0.007	-0.004	0.006	0.080
AR-P	-0.064	-0.125	0.158	-0.173	-0.504	-0.019	-0.016	0.011	-0.010	-0.024	0.080
Dent-P	-0.107	0.273	-0.023	-0.106	-0.089	-0.032	0.035	-0.002	-0.006	-0.004	0.079
Unis	0.147	0.024	0.223	0.283	-0.002	0.043	0.003	0.015	0.017	0.000	0.078
PHR-P	0.163	-0.084	-0.143	0.029	-0.125	0.048	-0.011	-0.010	0.002	-0.006	0.076
PV	-0.133	0.03	-0.377	-0.101	0.012	-0.039	0.004	-0.025	-0.006	0.001	0.075
POS	-0.007	-0.2	0.256	-0.468	0.032	-0.002	-0.026	0.017	-0.028	0.002	0.074
VCR-P	0.103	0.21	-0.143	-0.026	-0.093	0.030	0.027	-0.010	-0.002	-0.004	0.073
PDA	0.042	-0.222	-0.314	-0.046	-0.169	0.012	-0.029	-0.021	-0.003	-0.008	0.073
PLC	0.031	-0.165	-0.211	0.026	0.487	0.009	-0.021	-0.014	0.002	0.023	0.069
ADPM	0.073	-0.06	-0.458	0.014	0.08	0.022	-0.008	-0.031	0.001	0.004	0.065
PUE	0.044	0.056	-0.002	0.544	-0.2	0.013	0.007	0.000	0.032	-0.009	0.062
MHP-P	-0.04	0.277	0.057	-0.156	-0.011	-0.012	0.036	0.004	-0.009	-0.001	0.061
GR	-0.058	-0.232	-0.114	0.049	-0.042	-0.017	-0.030	-0.008	0.003	-0.002	0.060
AI	-0.014	0.029	0.12	-0.099	0.058	-0.004	0.004	0.008	-0.006	0.003	0.025

Figure 36: PCA Results

From the PCA (principal component analysis) results, we extracted PC1 through PC5, which represent 60% of the variance in the target variable, YPLL. Then we multiplied this by the weights of each PC to generate the values for columns PC1.a to PC5. a. The PCs are weighted as follows, by proportion of variance it explains. Then we calculated the sum of proportion each feature explains, by summing the absolute value of columns PC1.a to PC5.a. The table above is sorted by the sum column so the features at the top add the most value. They are % Physically Inactive (PI), % Single-Parent Households (SPH), % Some College (SC), % Children in Poverty (CIP), Chlamydia Rate (CR), Teen Birth Rate (TB), Obese, Food Environment Index (FEI), Mentally Unhealthy Days (MUD), Percentage of adults that report fair or poor health (PFP) and Physically Unhealthy Days (PUD).

PC#	Proportion of Variance
PC 1	0.295
PC 2	0.13
PC 3	0.067
PC 4	0.059
PC 5	0.047

Clustering

First, we did an unsupervised clustering exercise to see if there were natural clusters in the data. We used clustering (k-means) with “measure types” set to “Numerical Measures” and “numerical measure” set to “Euclidean Distance”. We selected these options because we are using the dataset with the normalized values and generated features as described above. Since our goal is to classify each county as Low Risk or High Risk for high YPLL, we set the number of clusters equal to 2 to see how the dataset forms only 2 clusters.

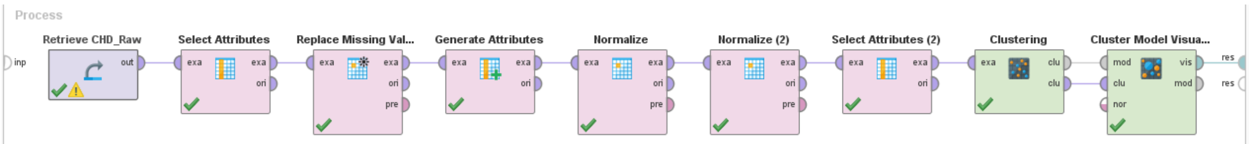


Figure 37: Creation of Clustering process flow

The results are as follows:

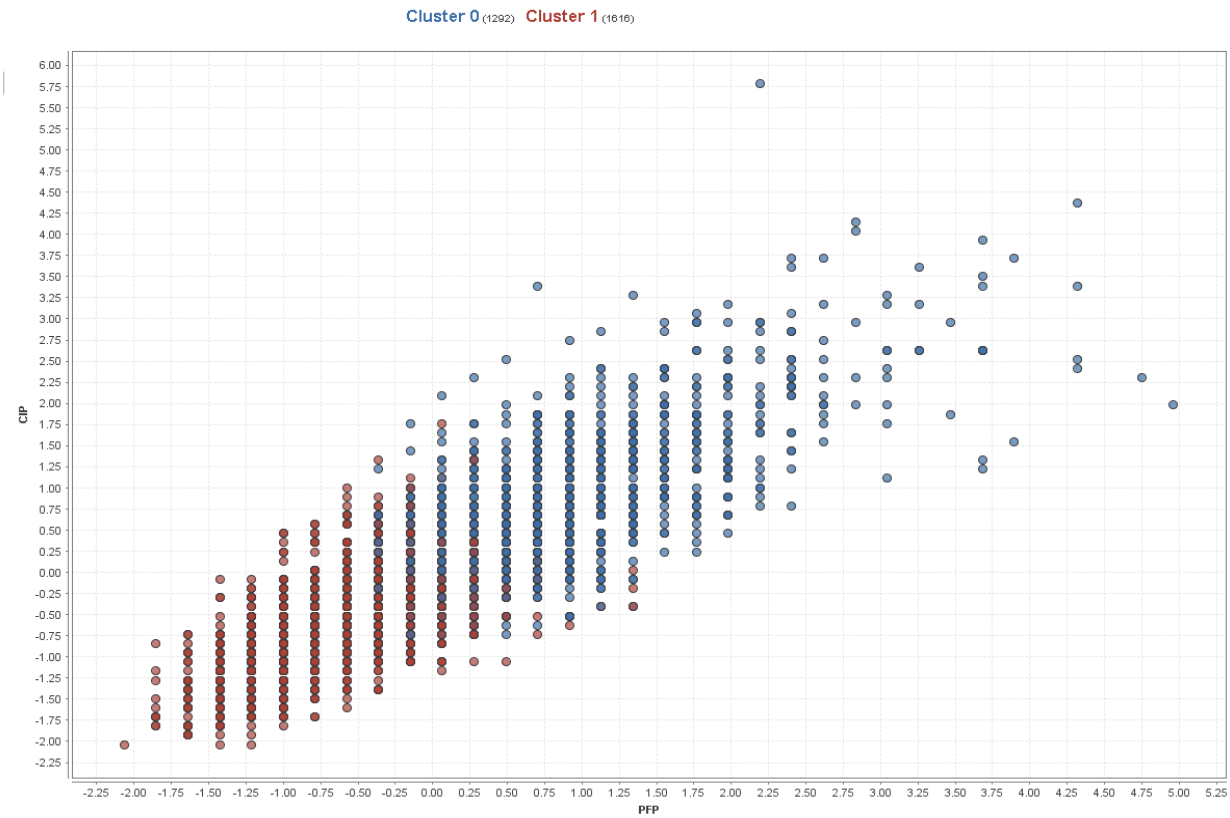


Figure 38: Clustering Graph

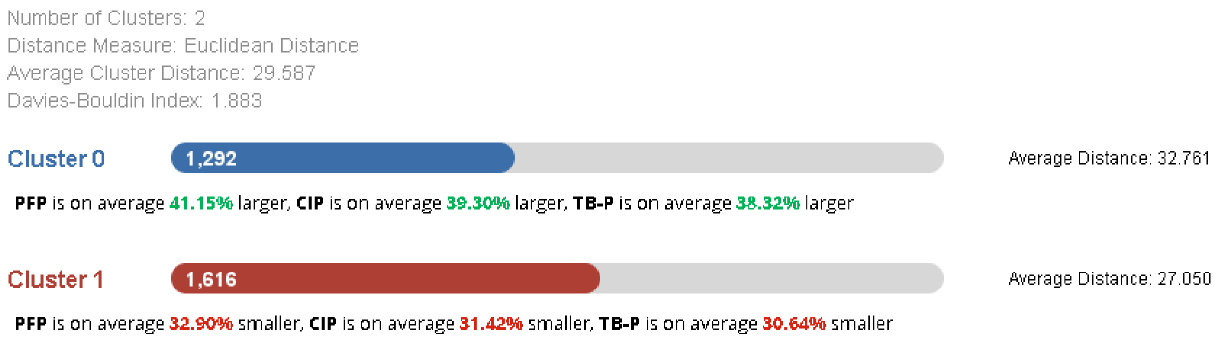


Figure 39: Clustering Results

From these results we see that PFP (% Fair/Poor Health), CIP (% Children in Poverty), and TB (Teen Birth Rate) were used to differentiate the clusters, and that on average these 3 values are at least 30% larger in Cluster 0, and around 30% smaller in Cluster 1.

Now, we do the clustering exercise again, adding the YPLL as a feature and get similar results:

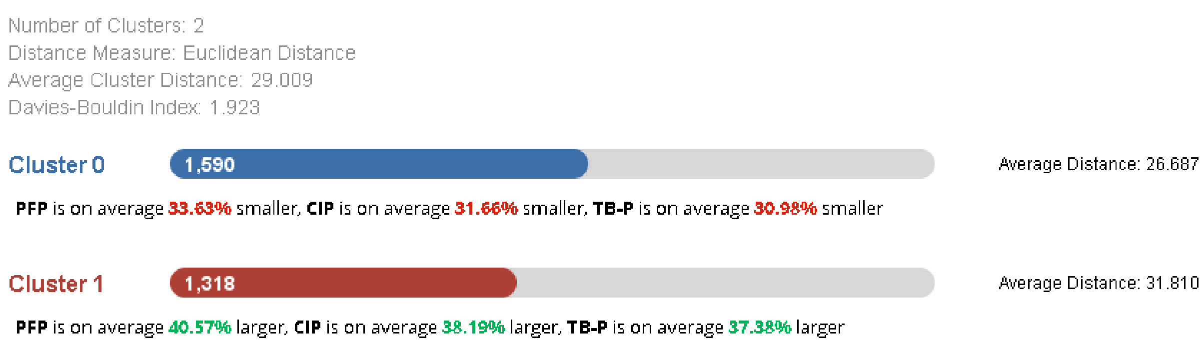


Figure 40: Clustering Results

We received consistent results in both Figure 38 and 39. Thus, confirming our results achieved from our Classification Analysis.

Parameter Optimization

Decision Tree – Best Model Performance – Accuracy

Sl	Parameters	Accuracy(%)	Ref.
1	Default Parameters (Figure 41)	83.37	Fig. 42
2	Disabled Apply Pre-pruning button with everything else remaining same	83.14	Fig. 43
3	Disabled Apply Pre-pruning button and increased the Maximal Depth to 15 and Confidence to 0.2	82.68	Fig. 44
4	Disabled Apply Pre-pruning button and increased the Maximal Depth to 20	82.91	Fig. 45
5	Disabled Apply Pre-pruning button and decreased the Maximal Depth to 7	84.86	Fig. 46
6	Disabled Apply Pre-pruning button and decreased the Maximal Depth to 5	84.17	Fig. 47

Table 7: Decision Tree - Best Model Performance - Accuracy

Decision Tree – Best Model Performance – F-Measure

Sl	Parameters	F-Measure(%)	Ref.
1	Default Parameters (Figure 41)	83.54	Fig. 42
2	Disabled Apply Pre-pruning button with everything else remaining same	83.24	Fig. 43
3	Disabled Apply Pre-pruning button and increased the Maximal Depth to 15 and Confidence to 0.2	82.82	Fig. 44
4	Disabled Apply Pre-pruning button and increased the Maximal Depth to 20	83.05	Fig. 45
5	Disabled Apply Pre-pruning button and decreased the Maximal Depth to 7	84.90	Fig. 46
6	Disabled Apply Pre-pruning button and decreased the Maximal Depth to 5	83.65	Fig. 47

Table 8: Decision Tree - Best Model Performance - F-Measure

Logistic Regression – Best Model Performance – Accuracy

Sl	Parameters	Accuracy(%)	Ref.
1	Default Parameters (Figure 48)	90.25	Fig. 49
2	With Reproducible enabled	90.25	Fig. 50
3	With maximum number of threads increased to 10	90.25	Fig. 51
4	With maximum number of threads increased to 15	90.25	Fig. 52

Table 9: Logistic Regression - Best Model Performance - Accuracy

Logistic Regression – Best Model Performance – F-Measure

Sl	Parameters	F-Measure(%)	Ref.
1	Default Parameters (Figure 48)	90.24	Fig. 49
2	With Reproducible enabled	90.24	Fig. 50
3	With maximum number of threads increased to 10	90.24	Fig. 51
4	With maximum number of threads increased to 15	90.24	Fig. 52

Table 10: Logistic Regression - Best Model Performance - F-Measure

Naïve Bayes – Best Model Performance – Accuracy

Sl	Parameters	Accuracy(%)	Ref.
1	Default Parameters (Figure 53)	83.94	Fig. 54
2	Displaced Laplace correction	83.94	Fig. 55

Table 11: Naive Bayes - Best Model Performance - Accuracy

Naïve Bayes – Best Model Performance – F-Measure

Sl	Parameters	F-Measure(%)	Ref.
1	Default Parameters (Figure 53)	84.72	Fig. 54
2	Displaced Laplace correction	84.72	Fig. 55

Table 12: Naive Bayes - Best Model Performance - F-Measure

Gradient Boosted Trees – Best Model Performance – Accuracy

SI	Parameters	Accuracy(%)	Ref.
1	Default Parameters (Figure 56)	88.53	Fig. 57
2	Increased the number of trees from 100 to 125	88.99	Fig. 58
3	Increased the number of trees from 125 to 150	89.22	Fig. 59
4	Increased the number of trees from 150 to 200	89.33	Fig. 60
5	Increased the number of trees from 200 to 225	89.22	Fig. 61

Table 13: Gradient Boosted Trees - Best Model Performance - Accuracy

Gradient Boosted Trees – Best Model Performance – F-Measure

SI	Parameters	F-Measure(%)	Ref.
1	Default Parameters (Figure 56)	88.61	Fig. 57
2	Increased the number of trees from 100 to 125	88.97	Fig. 58
3	Increased the number of trees from 125 to 150	89.17	Fig. 59
4	Increased the number of trees from 150 to 200	89.47	Fig. 60
5	Increased the number of trees from 200 to 225	89.39	Fig. 61

Table 14: Gradient Boosted Trees - Best Model Performance - F-Measure

Deep Learning – Best Model Performance – Accuracy

SI	Parameters	Accuracy(%)	Ref.
1	Default Parameters (Figure 62)	89.91	Fig. 63
2	Changed the Activation function from Rectifier to Tanh	89.68	Fig. 64
3	Changed the Activation function from Rectifier to Maxout	89.33	Fig. 65
4	Changed the Activation function from Maxout to ExpRectifier	90.14	Fig. 66
5	Changed the Epochs from 10 to 15	90.02	Fig. 67
6	Changed the Epochs from 10 to 7	89.11	Fig. 68

Table 15: Deep Learning - Best Model Performance - Accuracy

Deep Learning – Best Model Performance – F-Measure

SI	Parameters	F-Measure(%)	Ref.
1	Default Parameters (Figure 62)	89.86	Fig. 63
2	Changed the Activation function from Rectifier to Tanh	89.91	Fig. 64
3	Changed the Activation function from Rectifier to Maxout	89.30	Fig. 65
4	Changed the Activation function from Maxout to ExpRectifier	90.09	Fig. 66
5	Changed the Epochs from 10 to 15	90.01	Fig. 67
6	Changed the Epochs from 10 to 7	89.46	Fig. 68

Table 16: Gradient Boosted Trees - Best Model Performance - F-Measure

Rule Induction – Best Model Performance – Accuracy

SI	Parameters	Accuracy(%)	Ref.
1	Default Parameters (Figure 69)	84.63	Fig. 70
2	Changed the sample ratio from 0.9 to 0.95	83.94	Fig. 71
3	Changed the sample ratio from 0.9 to 0.85	85.89	Fig. 72
4	Changed the sample ratio from 0.9 to 0.8	86.01	Fig. 73
5	Changed the sample ratio from 0.9 to 0.75	84.40	Fig. 74
6	Changed the minimal prune benefit from 0.25 to 0.3	85.67	Fig. 75
7	Changed the minimal prune benefit from 0.25 to 0.2	84.40	Fig. 76

Table 17: Rule Induction - Best Model Performance - Accuracy

Rule Induction – Best Model Performance – F-Measure

SI	Parameters	F-Measure(%)	Ref.
1	Default Parameters (Figure 69)	84.53	Fig. 70
2	Changed the sample ratio from 0.9 to 0.95	83.72	Fig. 71
3	Changed the sample ratio from 0.9 to 0.85	85.91	Fig. 72
4	Changed the sample ratio from 0.9 to 0.8	85.98	Fig. 73
5	Changed the sample ratio from 0.9 to 0.75	83.69	Fig. 74
6	Changed the minimal prune benefit from 0.25 to 0.3	85.35	Fig. 75
7	Changed the minimal prune benefit from 0.25 to 0.2	83.69	Fig. 76

Table 18: Rule Induction - Best Model Performance - F-Measure

Model Building

Based on Tables 7-18, the best models are as follows:

SI	Models	Accuracy(%)	Ref.
1	Decision Tree	84.86	Fig. 46
2	Logistic Regression	90.25	Fig. 52
3	Naïve Bayes	83.94	Fig. 54
4	Gradient Boosted Trees	89.33	Fig. 60
5	Deep Learning	90.14	Fig. 66
6	Rule Induction	85.98	Fig. 73

Table 19: Comparison of Models - Accuracy

Sl	Models	F-Measure(%)	Ref.
1	Decision Tree	84.90	Fig. 46
2	Logistic Regression	90.24	Fig. 52
3	Naïve Bayes	84.72	Fig. 54
4	Gradient Boosted Trees	89.47	Fig. 60
5	Deep Learning	90.09	Fig. 66
6	Rule Induction	86.01	Fig. 73

Based on Table 19-20, the best models are Logistic Regression, Gradient Boosted Trees and Deep Learning.

Analysis and Recommendations

Death rates from dozens of causes have been rising over the past decade for young and middle-aged adults, driving down overall life expectancy in the United States. Our initial expectations led us to believe that YPLL would be significantly determined by mostly health and healthcare related attributes. We were surprised to learn that socioeconomic factors are just as determinant.

During our study of this dataset, there were certain parameters that recurred more often in that they appear to be of importance regardless of our modeling approach. Some of these are Food Environment Index (FEI), PFP (% Fair/Poor Health), CIP (% Children in Poverty), and TB (Teen Birth Rate). Most of these attributes are socio-economic issues that would presumably give way to actual health problems, all of which amalgamates to an alarming reversal of historical patterns in human longevity. Despite spending more on health care than any other country, the United States has seen increasing mortality and falling life expectancy for people age 25 to 64, which is in contrast to other “wealthy” nations.

People are less likely to live longer if they are poor, get little exercise and lack access to health care. The quality and availability of that health care has a significant effect on health outcomes. Smoking, physical inactivity, obesity, high blood pressure are all preventable risk factors that are not directly addressed in the way that the United States currently delivers it’s healthcare, which will have reverberations in future generations. The government and associated governing bodies need to rethink how we deliver medical care in this country, with a much greater investment in prevention and a more holistic approach to creating healthy communities that are free of preventable health related drivers like food deserts.

Our analysis unfortunately cannot take all things into account, such as whether there exists a causality between this trend and the ongoing opioid epidemic as well as other possible environmental drivers. Regardless, our analysis of this data points to a larger overall erosion of the health of Americans.

Appendix

County Health Attributes Description

SI	Data Elements	Code	Description
1	**FIPS**	**FIPS**	Federal Information Processing Standard
2	**State**	**State**	
3	**County**	**County**	
4	**Years of Potential Life Lost Rate**	**YPPL**	Age-adjusted YPLL rate per 100,000
5	**% Fair/Poor**	**PFP**	Percentage of adults that report fair or poor health
6	**Physically Unhealthy Days**	**PUD**	Average number of reported physically unhealthy days per month
7	**Mentally Unhealthy Days**	**MUD**	Average number of reported mentally unhealthy days per month
8	**% LBW**	**LBW**	Percentage of births with low birth weight (<2500g)
9	**% Smokers**	**Smoke**	Percentage of adults that reported currently smoking
10	**% Obese**	**Obese**	Percentage of adults that report BMI >= 30
11	**Food Environment Index**	**FEI**	Indicator of access to healthy foods - 0 is worst, 10 is best
12	**% Physically Inactive**	**PI**	Percentage of adults that report no leisure-time physical activity
13	**% With Access**	**WA**	Percentage of the population with access to places for physical activity
14	**% Excessive Drinking**	**ED**	Percentage of adults that report excessive drinking
15	**% Alcohol-Impaired**	**AI**	Percentage of driving deaths with alcohol involvement
16	**Chlamydia Rate**	**CR**	Chlamydia cases per 100,000 population
17	**Teen Birth Rate**	**TB**	Births per 1,000 females ages 15-19
18	**% Uninsured**	**Unis**	Percentage of people under age 65 without insurance
19	**PCP Rate**	**PCP**	Primary Care Physicians per 100,000 population
20	**Dentist Rate**	**Dent**	Dentists per 100,000 population
21	**MHP Rate**	**MHP**	Mental Health Providers per 100,000 population
22	**Preventable Hosp. Rate**	**PHR**	Discharges for Ambulatory Care Sensitive Conditions per 100,000 Medicare Enrollees
23	**% Screened**	**PS**	Percentage of female Medicare enrollees having an annual mammogram (age 65-74)

24	**% Vaccinated**	**PV**	Percentage of annual Medicare enrollees having an annual flu vaccination
25	**Graduation Rate**	**GR**	Graduation rate
26	**% Some College**	**SC**	Percentage of adults age 25-44 with some post-secondary education
27	**% Unemployed**	**Unem**	Percentage of population ages 16+ unemployed and looking for work
28	**% Children in Poverty**	**CIP**	Percentage of children (under age 18) living in poverty
29	**Income Ratio**	**IR**	Ratio of household income at the 80th percentile to income at the 20th percentile
30	**% Single-Parent Households**	**SPH**	Percentage of children that live in single-parent households
31	**Association Rate**	**AR**	Associations per 10,000 population
32	**Violent Crime Rate**	**VCR**	Violent crimes per 100,000 population
33	**Injury Death Rate**	**IDR**	Injury mortality rate per 100,000
34	**Average Daily PM2.5**	**ADPM**	Average daily amount of fine particulate matter in micrograms per cubic meter
35	**Presence of violation**	**POV**	County affected by a water violation: 1-Yes, 0-No
36	**% Severe Housing Problems**	**SHP**	Percentage of households with at least 1 of 4 housing problems: overcrowding, high housing costs, or lack of kitchen or plumbing facilities
37	**% Drive Alone**	**PDA**	Percentage of workers who drive alone to work
38	**% Long Commute - Drives Alone**	**PLC**	Among workers who commute in their car alone, the percentage that commute more than 30 minutes
39	**% Homeowners**	**PHO**	Percentage of population Home Owners
40	**% < 18**	**PUE**	Percentage of population Under 18
41	**% 65 and over**	**POS**	Percentage of population over 65

Note: Our target variable at Sl 4 above highlighted in yellow

Decision Tree - Default Parameters

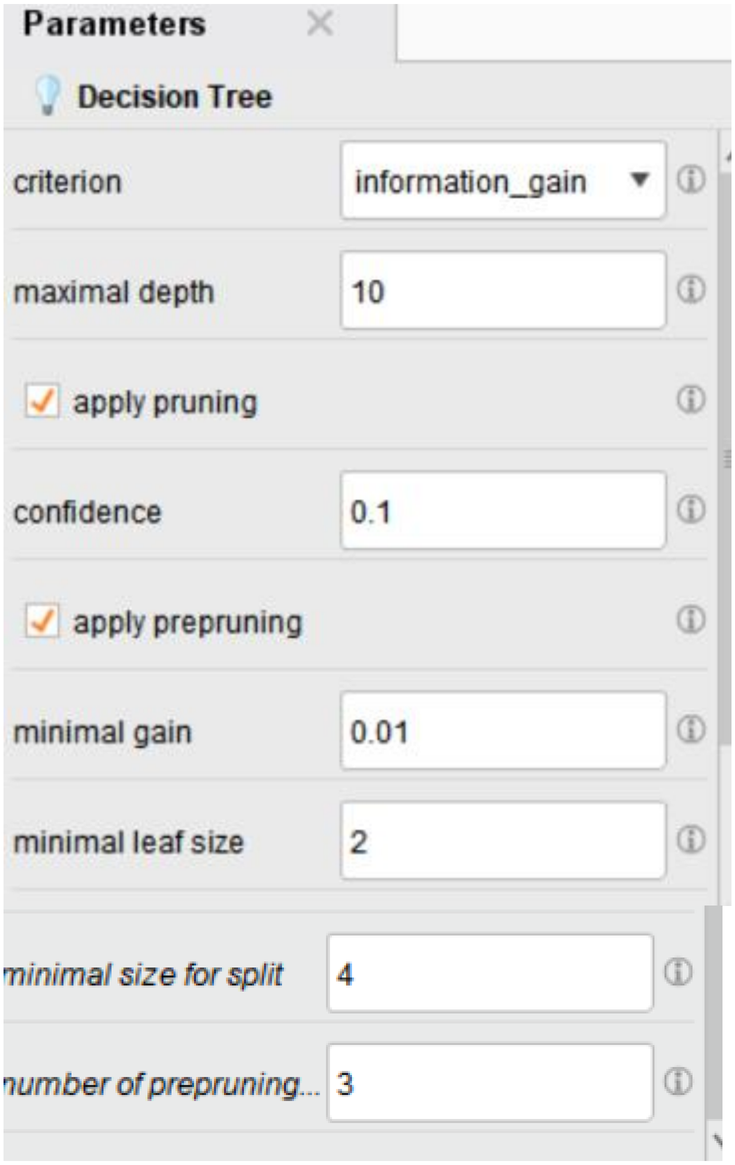


Figure 41: Decision Tree - Default Parameters

accuracy: 83.37%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	359	68	84.07%
pred. LowRisk	77	368	82.70%
class recall	82.34%	84.40%	

Figure 42: Decision Tree - Model Performance - Default Parameters

Accuracy	Precision	Recall	F-Measure
83.37%	82.70%	84.40%	83.54%

Decision Tree - Disabled Apply Pre-pruning button with everything else remaining same - Performance

accuracy: 83.14%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	360	71	83.53%
pred. LowRisk	76	365	82.77%
class recall	82.57%	83.72%	

Figure 43: Decision Tree - Disabled Apply Pre-pruning button with everything else remaining same - Performance

Accuracy	Precision	Recall	F-Measure
83.14%	82.77%	83.72%	83.24%

Decision Tree - Disabled Apply Pre-pruning button and increased the Maximal Depth to 15 and Confidence to 0.2 - Performance

accuracy: 82.68%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	357	72	83.22%
pred. LowRisk	79	364	82.17%
class recall	81.88%	83.49%	

Figure 44: Decision Tree - Disabled Apply Pre-pruning button and increased the Maximal Depth to 15 and Confidence to 0.2 – Performance

Accuracy	Precision	Recall	F-Measure
82.68%	82.17%	83.49%	82.82%

Decision Tree - Disabled Apply Pre-pruning button and increased the Maximal Depth to 20 - Performance

accuracy: 82.91%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	358	71	83.45%
pred. LowRisk	78	365	82.39%
class recall	82.11%	83.72%	

Figure 45: Decision Tree - Disabled Apply Pre-pruning button and increased the Maximal Depth to 20 - Performance

Accuracy	Precision	Recall	F-Measure
82.91%	82.39%	83.72%	83.05%

Decision Tree - Disabled Apply Pre-pruning button and decreased the Maximal Depth to 7 - Performance

accuracy: 84.86%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	369	65	85.02%
pred. LowRisk	67	371	84.70%
class recall	84.63%	85.09%	

Figure 46: Decision Tree - Disabled Apply Pre-pruning button and decreased the Maximal Depth to 7 - Performance

Accuracy	Precision	Recall	F-Measure
84.86%	84.70%	85.09%	84.90%

Decision Tree - Disabled Apply Pre-pruning button and decreased the Maximal Depth to 5 - Performance

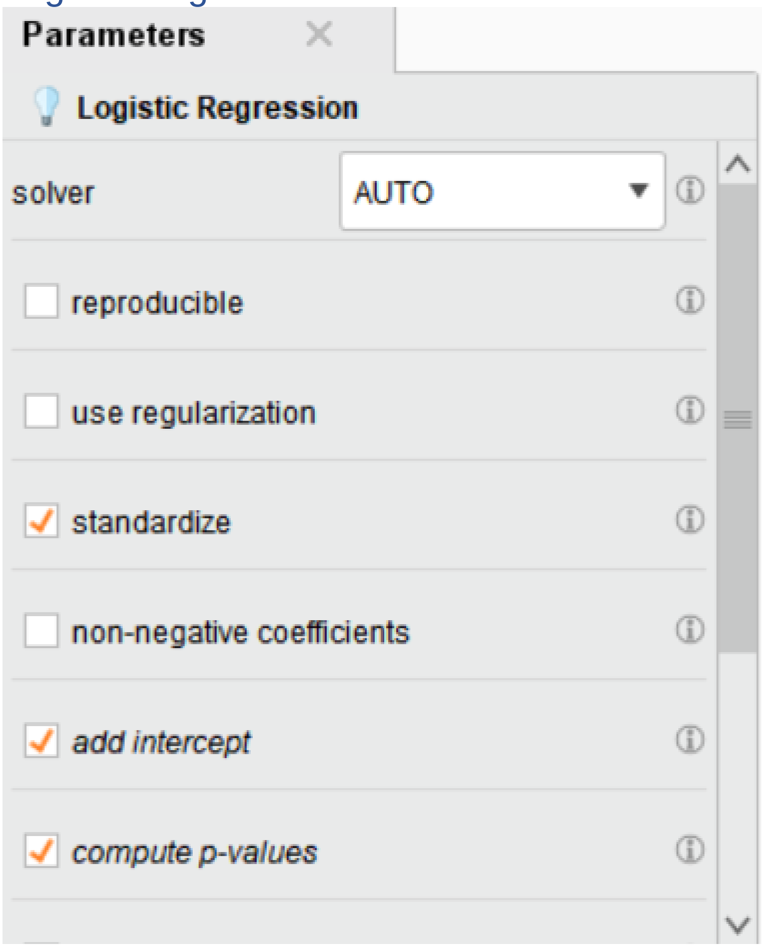
accuracy: 84.17%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	381	83	82.11%
pred. LowRisk	55	353	86.52%
class recall	87.39%	80.96%	

Figure 47: Decision Tree - Disabled Apply Pre-pruning button and decreased the Maximal Depth to 5 – Performance

Accuracy	Precision	Recall	F-Measure
84.17%	86.52%	80.96%	83.65%

Logistic Regression - Default Parameters



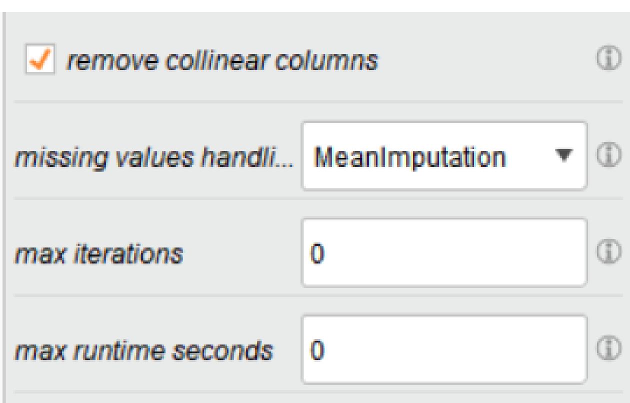


Figure 48: Logistic Regression - Default Parameters

accuracy: 90.25%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	394	43	90.16%
pred. LowRisk	42	393	90.34%
class recall	90.37%	90.14%	

Figure 49: Logistic Regression - Model Performance - Default Parameters

Accuracy	Precision	Recall	F-Measure
90.25%	90.34%	90.14%	90.24%

Logistic Regression - With Reproducible enabled

accuracy: 90.25%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	394	43	90.16%
pred. LowRisk	42	393	90.34%
class recall	90.37%	90.14%	

Figure 50: Logistic Regression - With Reproducible enabled

Accuracy	Precision	Recall	F-Measure
90.25%	90.34%	90.14%	90.24%

Logistic Regression - With maximum number of threads increased to 10

accuracy: 90.25%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	394	43	90.16%
pred. LowRisk	42	393	90.34%
class recall	90.37%	90.14%	

Figure 51: Logistic Regression - With maximum number of threads increased to 10

Accuracy	Precision	Recall	F-Measure
90.25%	90.34%	90.14%	90.24%

Logistic Regression - With maximum number of threads increased to 15

accuracy: 90.25%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	394	43	90.16%
pred. LowRisk	42	393	90.34%
class recall	90.37%	90.14%	

Figure 52: Logistic Regression - With maximum number of threads increased to 15

Accuracy	Precision	Recall	F-Measure
90.25%	90.34%	90.14%	90.24%

According to Figure 54, the optimum model should have a maximum number of threads of 15.

Naive Bayes - Default Parameters

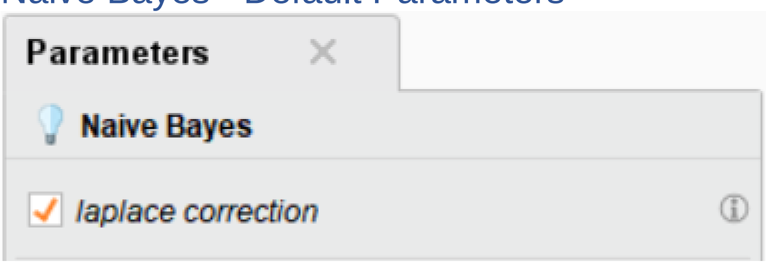


Figure 53: Naive Bayes - Default Parameters

accuracy: 83.94%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	344	48	87.76%
pred. LowRisk	92	388	80.83%
class recall	78.90%	88.99%	

Figure 54: Naive Bayes - Model Performance - Default Parameters

Accuracy	Precision	Recall	F-Measure
83.94%	80.83%	88.99%	84.72%

Naive Bayes - Disabled Laplace correction

accuracy: 83.94%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	344	48	87.76%
pred. LowRisk	92	388	80.83%
class recall	78.90%	88.99%	

Figure 55: Naive Bayes - Disabled Laplace correction

Accuracy	Precision	Recall	F-Measure
83.94%	80.83%	88.99%	84.72%

Gradient Boosted Trees - Default Parameters

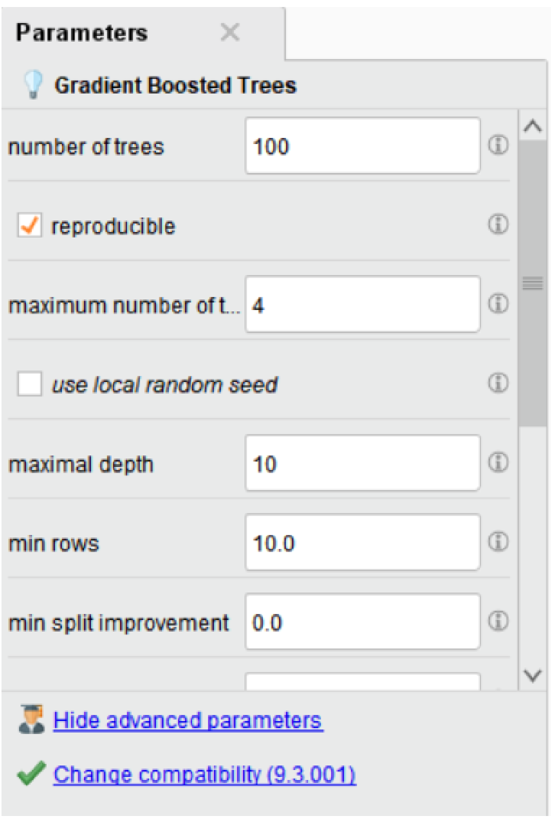


Figure 56: Gradient Boosted Trees - Default Parameters

accuracy: 88.53%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	383	47	89.07%
pred. LowRisk	53	389	88.01%
class recall	87.84%	89.22%	

Figure 57: Gradient Boosted Trees - Model Performance - Default Parameters

Accuracy	Precision	Recall	F-Measure
88.53%	88.01%	89.22%	88.61%

Gradient Boosted Trees - Increased the number of trees from 100 to 125

accuracy: 88.99%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	389	49	88.81%
pred. LowRisk	47	387	89.17%
class recall	89.22%	88.76%	

Figure 58: Gradient Boosted Trees - Increased the number of trees from 100 to 125

Accuracy	Precision	Recall	F-Measure
88.99%	89.17%	88.76%	88.97%

Gradient Boosted Trees - Increased the number of trees from 125 to 150

accuracy: 89.22%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	391	49	88.86%
pred. LowRisk	45	387	89.58%
class recall	89.68%	88.76%	

Figure 59: Gradient Boosted Trees - Increased the number of trees from 125 to 150

Accuracy	Precision	Recall	F-Measure
89.22%	89.58%	88.76%	89.17%

Gradient Boosted Trees - Increased the number of trees from 150 to 200

accuracy: 89.33%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	384	41	90.35%
pred. LowRisk	52	395	88.37%
class recall	88.07%	90.60%	

Figure 60: Gradient Boosted Trees - Increased the number of trees from 150 to 200

Accuracy	Precision	Recall	F-Measure
89.33%	88.37%	90.60%	89.47%

Gradient Boosted Trees - Increased the number of trees from 200 to 225

accuracy: 89.22%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	382	40	90.52%
pred. LowRisk	54	396	88.00%
class recall	87.61%	90.83%	

Figure 61: Gradient Boosted Trees - Increased the number of trees from 200 to 225

Accuracy	Precision	Recall	F-Measure
89.22%	88.00%	90.83%	89.39%

Deep Learning - Default Parameters

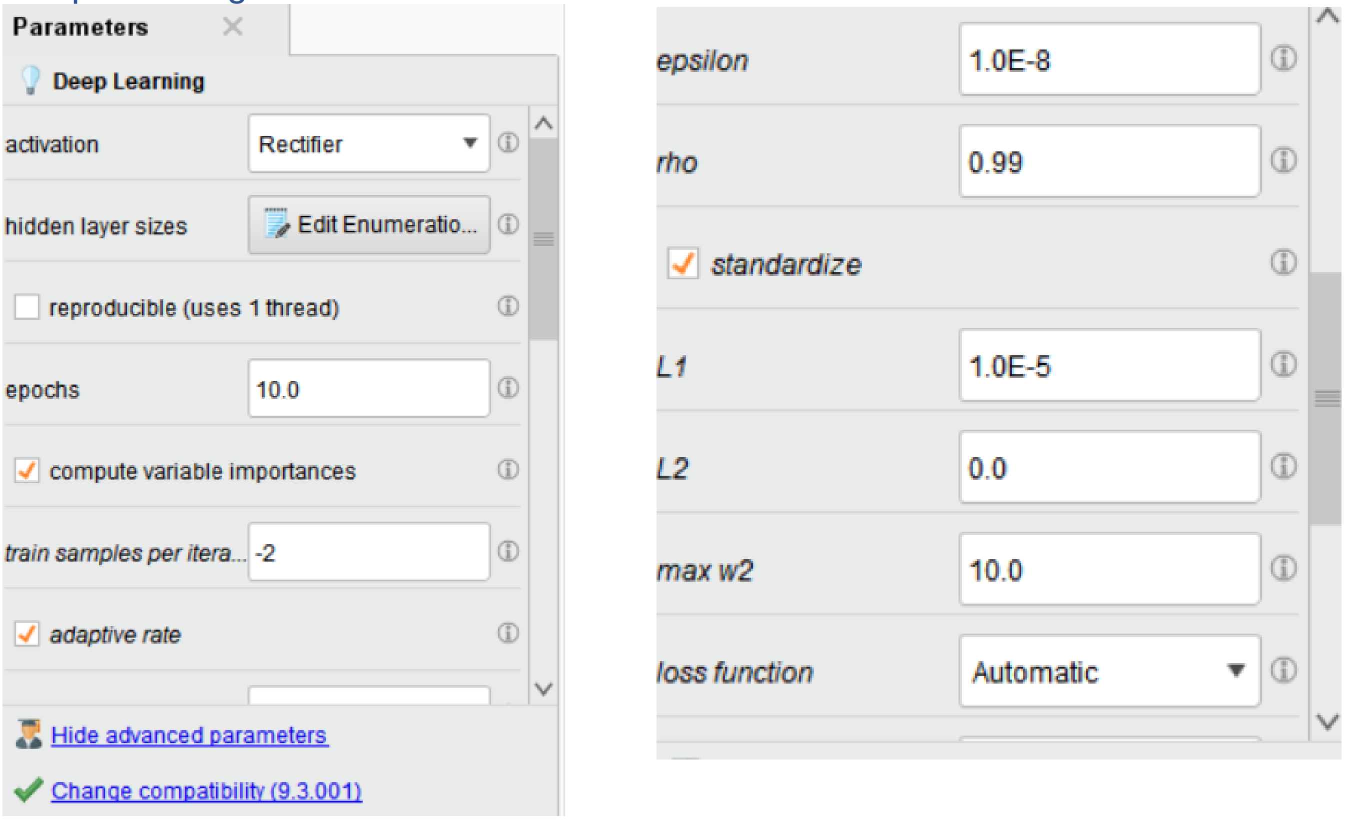
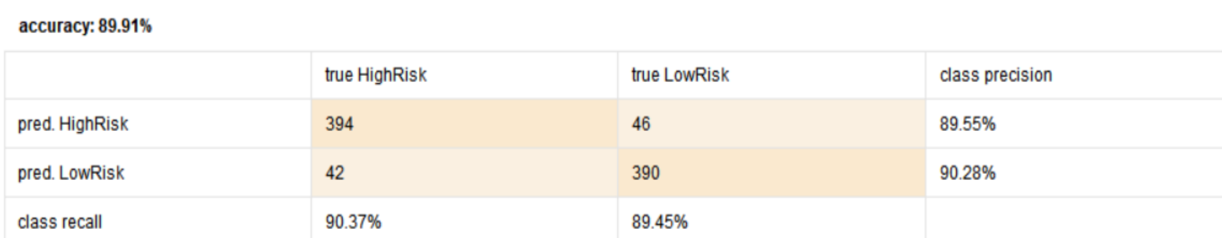


Figure 62: Deep Learning - Default Parameters


accuracy: 89.91%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	394	46	89.55%
pred. LowRisk	42	390	90.28%
class recall	90.37%	89.45%	

Figure 63: Deep Learning - Model Performance - Default Parameters

Accuracy	Precision	Recall	F-Measure
89.91%	90.28%	89.45%	89.86%

Deep Learning - Changed the Activation function from Rectifier to Tanh

accuracy: 89.68%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	381	35	91.59%
pred. LowRisk	55	401	87.94%
class recall	87.39%	91.97%	

Figure 64: Deep Learning - Changed the Activation function from Rectifier to Tanh

Accuracy	Precision	Recall	F-Measure
89.68%	87.94%	91.97%	89.91%

Deep Learning - Changed the Activation function from Rectifier to Maxout

accuracy: 89.33%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	391	48	89.07%
pred. LowRisk	45	388	89.61%
class recall	89.68%	88.99%	

Figure 65: Deep Learning - Changed the Activation function from Rectifier to Maxout

Accuracy	Precision	Recall	F-Measure
89.33%	89.61%	88.99%	89.30%

Deep Learning - Changed the Activation function from Maxout to ExpRectifier

accuracy: 90.14%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	395	45	89.77%
pred. LowRisk	41	391	90.51%
class recall	90.60%	89.68%	

Figure 66: Deep Learning - Changed the Activation function from Maxout to ExpRectifier

Accuracy	Precision	Recall	F-Measure
90.14%	90.51%	89.68%	90.09%

Deep Learning - Changed the Epochs from 10 to 15

accuracy: 90.02%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	393	44	89.93%
pred. LowRisk	43	392	90.11%
class recall	90.14%	89.91%	

Figure 67: Deep Learning - Changed the Epochs from 10 to 15

Accuracy	Precision	Recall	F-Measure
90.02%	90.11%	89.91%	90.01%

Deep Learning - Changed the Epochs from 10 to 7

accuracy: 89.11%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	374	33	91.89%
pred. LowRisk	62	403	86.67%
class recall	85.78%	92.43%	

Figure 68: Deep Learning - Changed the Epochs from 10 to 7

Accuracy	Precision	Recall	F-Measure
89.11%	86.67%	92.43%	89.46%

After reviewing Figures 68-70, we decided to use the Deep Learning Model with ExpRectifier and 10 Epochs because it has the highest accuracy and f-measure.

Rule Induction - Default Parameters

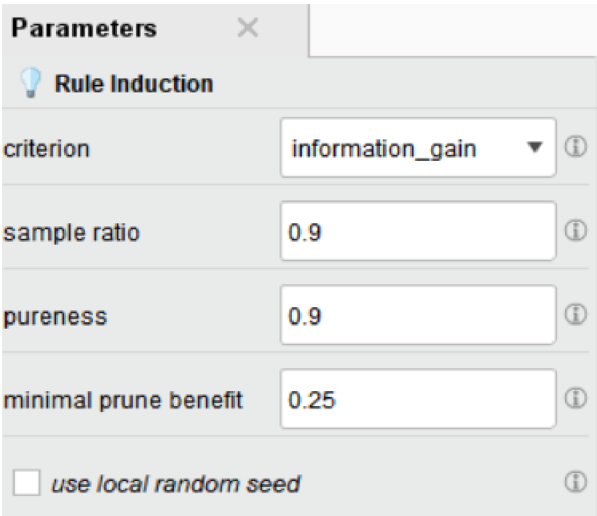


Figure 69: Rule Induction - Default Parameters

accuracy: 84.63%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	372	70	84.16%
pred. LowRisk	64	366	85.12%
class recall	85.32%	83.94%	

Figure 70: Rule Induction - Model Performance - Default Parameters

Accuracy	Precision	Recall	F-Measure
84.63%	85.12%	83.94%	84.53%

Rule Induction - Changed the sample ratio from 0.9 to 0.95

accuracy: 83.94%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	372	76	83.04%
pred. LowRisk	64	360	84.91%
class recall	85.32%	82.57%	

Figure 71: Rule Induction - Changed the sample ratio from 0.9 to 0.95

Accuracy	Precision	Recall	F-Measure
83.94%	84.91%	82.57%	83.72%

Rule Induction - Changed the sample ratio from 0.9 to 0.85

accuracy: 85.89%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	374	61	85.98%
pred. LowRisk	62	375	85.81%
class recall	85.78%	86.01%	

Figure 72: Rule Induction - Changed the sample ratio from 0.9 to 0.85

Accuracy	Precision	Recall	F-Measure
85.89%	85.81%	86.01%	85.91%

Rule Induction - Changed the sample ratio from 0.9 to 0.8

accuracy: 86.01%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	376	62	85.84%
pred. LowRisk	60	374	86.18%
class recall	86.24%	85.78%	

Figure 73: Rule Induction - Changed the sample ratio from 0.9 to 0.8

Accuracy	Precision	Recall	F-Measure
86.01%	86.18%	85.78%	85.98%

Rule Induction - Changed the sample ratio from 0.9 to 0.75

accuracy: 84.40%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	387	87	81.65%
pred. LowRisk	49	349	87.69%
class recall	88.76%	80.05%	

Figure 74: Rule Induction - Changed the sample ratio from 0.9 to 0.75

Accuracy	Precision	Recall	F-Measure
84.40%	87.69%	80.05%	83.69%

After reviewing Figures 70-74, we decided to use the sample ratio of 0.8 for further model optimization.

Rule Induction - Changed the minimal prune benefit from 0.25 to 0.3

accuracy: 85.67%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	383	72	84.18%
pred. LowRisk	53	364	87.29%
class recall	87.84%	83.49%	

Figure 75: Rule Induction - Changed the minimal prune benefit from 0.25 to 0.3

Accuracy	Precision	Recall	F-Measure
85.67%	87.29%	83.49%	85.35%

Rule Induction - Changed the minimal prune benefit from 0.25 to 0.2

accuracy: 84.40%

	true HighRisk	true LowRisk	class precision
pred. HighRisk	387	87	81.65%
pred. LowRisk	49	349	87.69%
class recall	88.76%	80.05%	

Figure 76: Rule Induction - Changed the minimal prune benefit from 0.25 to 0.2

Accuracy	Precision	Recall	F-Measure
84.40%	87.69%	80.05%	83.69%

After reviewing figures 72-78, we decided to use the Rule Induction model with a sample ratio of 0.8 and minimal prune benefit of 0.25 because it has the highest accuracy and f-measure.