CPE 695 Final Project Report

Xiangyang Xu
ECE Department
Stevens Institute of Technology
Hoboken, USA
xxu46@stevens.edu

Zifan Cao
EE Department
Stevens Institute of Technology
Jersey City, USA
zcao23@stevens.edu

Qiuyang Tang
CPE Department
Stevens Institute of Technology
Jersey City, USA
qtang7@stevens.edu

I. Introduction

A. Goal of Project

The goal of this project is to use machine learning method to predict American people's life expectancy based on the datasets provided on BCHC (Big Cities Health Coalition).

The reason this topic is picked is because BCHC provide many categories of health issue that will directly or indirectly influence the American people's life expectancy. For example, based on the diabetes mortality rate in 2010 can directly influence the prediction of American people''s life expectancy in the 2011 and even in the next few years. And some health issue has indirect influence, such as adult smoking rate or obesity rate. Also, with the improvement of people's life style and the development of the medical equipment, Longer life expectancy becomes hot topic in today's world. People in America care about how long they could live, and what factors could affect their expectancy live years.

B. Task Allocation:

Each team member should pick an area to apply the proper machine learning and find his answer. At the end, the team will compare the advantage and disadvantage between all the methods (accuracy, computational cost etc.) and find a general prediction of the American people's life expectancy.

- General study of Life expectation - Xiangyang Xu
- Cancer & Life expectation study - Xiangyang Xu
- Chronic Disease - Zifan Cao
- Injury & Violence - Qiuyang Tang

II. Related Work

Similar works have been made by other people, we collect two related work as reference

Kwetishe Joro Danjuma [1] has done a machine learning research in post-operative life expectancy in the lung Cancer patients. Danjuma used multilayer Perceptron, J48, and the Naive Bayes algorithms to train and test models on Thoracic Surgery datasets obtained from the University of California Irvine machine learning repository. Finally, he got result that multilayer perceptron performed best with classification accuracy of 82.3%, J48 came out second with classification accuracy of 81.8%, and Naive Bayes came out the worst with classification accuracy of 74.4%. The quality and outcome of the chosen machine learning algorithms depends on the ingenuity of the clinical miner.

Maciej Zieba team presents boosted SVM dedicated to solve imbalanced data problems. Proposed solution combines the benefits of using ensemble classifiers for uneven data together with cost-sensitive support vectors machines. Further, we present oracle-based approach for extracting decision rules from the boosted SVM. In the next step we examine the quality of the proposed method by comparing the performance with other algorithms which deal with imbalanced data. Finally, boosted SVM is used for medical application of predicting post-operative life expectancy in the lung cancer patients.

III. Solution

A. Data Processing.

We collect Big Cities Health Inventory Data from BCHI. We select Life expectancy, Suicide Rate, Homicide Rate, Pneumonia and Influenza, Firearm, Vehicle, Heart Disease Diabetes HIV and Cancer By using Pandas Dataframe and Numpy Array. If two data have same place, year, sex and race feature, we think that these two data can be merged. We merge all data into one csv as our database.

Then team members select the data they need to complete individual tasks.

	Place	Year	Sex	Race	Life_expectancy	Suicide_Rate	Homicide_Rate	Pneumonia_and_Influenza	Firearm	Vehicle	Heart_Disease	Diabetes	HIV
0	Baltimore, MD	2013	Both	All	73.9	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	Boston, MA	2012	Both	All	80.1	5.4	NaN	18.3	4.4	5	131.8	19.6	879.3
2	Boston, MA	2012	Both	Asian/PI	87.2	NaN	NaN	NaN	NaN	NaN	45	NaN	NaN
3	Boston, MA	2012	Both	Black	77	3.1	19.9	NaN	13.7	NaN	156.2	39.6	1585.2
4	Boston, MA	2012	Both	Hispanic	86.4	NaN	NaN	NaN	NaN	NaN	80.9	NaN	899.5

Fig. 1. Part of database

B. Xiangyang Xu Work Part

1) Train Data and Test Data

I treat all people are the same, regardless of their city, year, race and gender. Therefore, I drop Place, Year, Race and Sex feature.

Because data is not enough, I think it's unwise to drop more data. Therefore I randomly set missing data from mean – std to mean + std, which don't make data change a lot. As is

shown in following figure, the describe of the feature of old data and new data are very close.

	Life_expectancy	Suicide_Rate	Homicide_Rate	Pneumonia_and_Influenza	Firearm	Vehicle	Heart_Disease	Diabetes	HIV	Cancer
count	467.000000	341.000000	316.000000	368.000000	332.000000	321.000000	436.000000	416.000000	399.000000	434.000000
mean	79.734475	11.584164	13.003797	16.029348	15.010843	7.759190	169.012385	23.384375	752.536842	165.176959
std	4.540366	6.719269	15.366981	7.755649	15.016814	4.258057	67.957100	10.548302	605.582888	44.020602
min	68.000000	0.000000	0.400000	0.000000	0.400000	1.100000	39.600000	0.000000	25.100000	27.600000
25%	76.800000	6.300000	3.200000	10.300000	4.800000	4.300000	124.950000	16.375000	302.950000	136.500000
50%	79.600000	10.000000	6.400000	14.350000	9.950000	6.800000	159.100000	21.650000	599.400000	163.750000
75%	82.400000	16.100000	16.825000	20.200000	18.600000	10.500000	205.250000	27.500000	1052.000000	192.775000
max	97.600000	40.000000	86.900000	55.800000	89.600000	20.000000	468.300000	80.800000	4125.600000	398.400000

Fig. 2. Description of Old Data

	Life_expectancy	Suicide_Rate	Homicide_Rate	Pneumonia_and_Influenza	Firearm	Vehicle	Heart_Disease	Diabetes	HIV	Cancer
count	480.000000	480.000000	480.000000	480.000000	480.000000	480.000000	480.000000	480.000000	480.000000	480.000000
mean	79.725000	11.335833	13.452500	15.905833	14.874167	7.509792	169.521667	23.343542	764.867083	165.168333
std	4.501259	6.067254	13.499642	7.108633	13.354059	3.800671	66.110007	10.013507	569.608472	42.526504
min	68.000000	0.000000	0.000000	0.000000	0.000000	1.100000	39.600000	0.000000	25.100000	27.600000
25%	76.800000	6.550000	3.900000	10.900000	5.375000	4.500000	125.675000	16.875000	353.100000	138.000000
50%	79.600000	10.150000	8.000000	14.800000	10.850000	7.000000	160.500000	22.000000	688.000000	163.450000
75%	82.400000	15.450000	19.925000	19.900000	19.850000	9.925000	206.175000	27.600000	1054.775000	192.625000
max	97.600000	40.000000	86.900000	55.800000	89.600000	20.000000	468.300000	80.800000	4125.600000	398.400000

Fig. 3. Description of New Data

In order to test the accuracy, overfit for my model I split my data into test data and train data. I randomly set about 70% data as train data and about 30% as test data.

2) Basic Work – Life Expectancy vs Cancer

I select life expectancy and cancer data from database. Label (Trainy) is Life expectancy and feature (TrainX) is Cancer number.

Because the feature is a one-dimension data, I choose linear regression as its algorithm. I try linear model linear equation, quadratic equation and cubic equation. I get solution as following figures show.

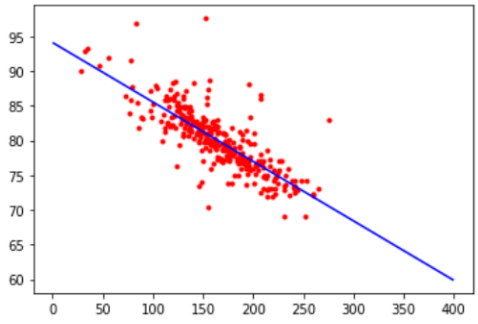

Fig. 4. Linear Equation Model

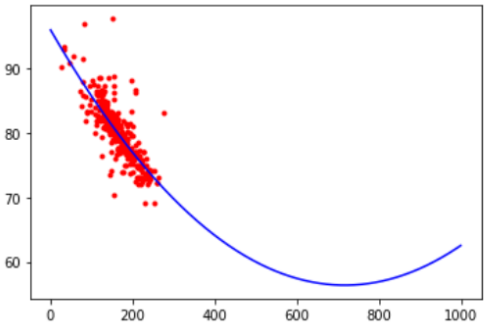

Fig. 5. Quadratic Equation Model

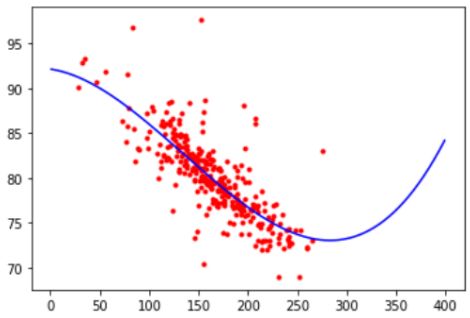

Fig. 6. Cubic Equation Model

As is shown in figures, cubic equation model has shown its overfitting, because it's unreasonable that when cancer number is more than 300, the life expectancy will increase with the increase of cancer number. Quadratic equation model will show the same overfitting if the cancer number become more as its mathematical feature.

Finally, I choose linear equation as model, and get the model y = -0.086x + 94.12. Then I input train data to test its accuracy. The describe result of error is mean error = 1.793, min error = 0.010, max error = 16.587, mean error = 1.793, variance = 4.056; I input test data to test its accuracy and get following solution. The describe result of error is mean error = 2.156, min error = 0.0001, max error = 9.411, variance = 4.755. All error in describe is absolute value of original error, because I don't want positive error and negative error to offset each other. The average life expectancy is 79.82.

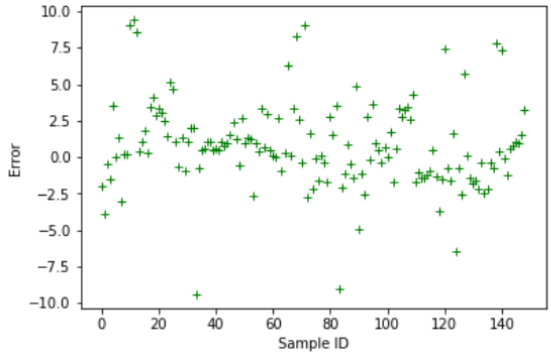

Fig. 7. Test Data Error of Linear Equation Model

3) Additional Work – General Study of Life Expectancy.

In the beginning of teamwork, I still lack knowledge about this project. After I learn PCA and other new algorithm, I think it's time to make an additional work.

a) Data Processing

I select Life Expectancy as label and all feature in database as feature. Then use PCA method to make dimensional reduction.

PCA (principal components analysis) is a commonly used data reduction technique. This method seeks to find linear combinations of the predictors, known as principal components (PCs), which capture the most possible variance. The first PC is defined as the linear combination

of the predictors that captures the most variability of all possible linear combinations. Then, subsequent PCs are derived such that these linear combinations capture the most remaining variability while also being uncorrelated with all previous PCs.

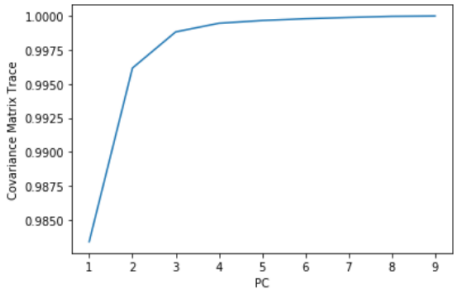

Fig. 8. PC vs Covariance Matrix Trace

In this project, after PCA, we can get that PC1 + PC2 can account for more than 99% covariance matrix trace, then I set PC1 and PC2 as feature.

b) *Linear Regression Method*

This method based on PCA method, using dimension reduction data.

Because current data just has feature PC1 and PC2, I choose linear regression method. Also, linear regression is a very fast algorithm. PC1 has more than 98% covariance matrix trace, and PC1, PC2 has more than 99%. Therefore, we can assume that all original data are similar in mathematic. As the experience I get from life expectancy vs cancer. I think this model should be a flat, as linear in 2D, set y = $u_1 \times x_1 + u_2 \times x_2$, and I get following model Expectancy Life = -0.0029 × PC1 + -0.0508×PC2 + 80.0353.

```
Red - Original Points, Blue - Predict Plat
```

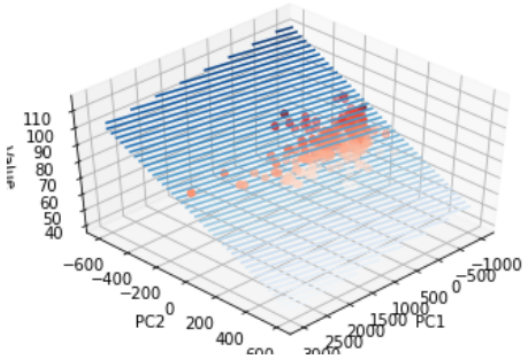

Fig. 9. Linear Regression Model

Then I input train data to test its accuracy. The describe result of error is mean error = 1.730, min error = 0.0031, max error = 19.5410, variance = 3.4397; I input test data to test its accuracy and get following solution. The describe result of error is mean error = 2.1360, min error = 0.0133, max error = 11.1013, variance = 3.8991.

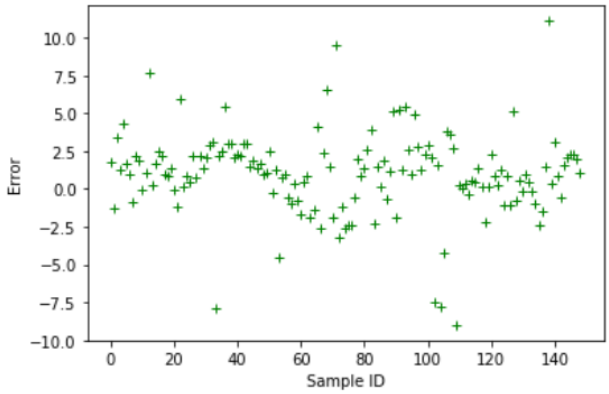

Fig. 10. Test Data Error of Linear Regression

c) *KNN Method*

This method based on PCA method, using dimension reduction data. KNN is one kind of clustering algorithms. Clustering algorithm has advantage as follows, relatively simple to implement, guarantees convergence, easily adapts to new examples and generalizes to clusters of different shapes and sizes, such as elliptical clusters.

I try different n_neighbors 2,3,4 and 5, and different KNN algorithm auto, brute, KD_tree and ball_tree. I choose the method with min mean error by test data and get n_neighbor = 2; due to the error of 4 algorithm is very close, I can't choose the method by their mean error. Because auto is a powerful algorithm which can adjust automatically, I choose algorithm = auto.

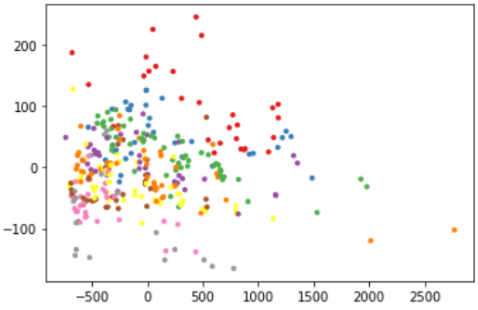

Fig. 11. KNN Model with Train Data

Then I input train data to test its accuracy. The describe result of error is mean error = 1.0843, min error = 0, max error = 18.7000, variance = 5.0117; I input test data to test its accuracy and get following solution. The describe result of error is mean error = 1.8329, min error = 0.0, max error = 10.5, variance = 4.1064.

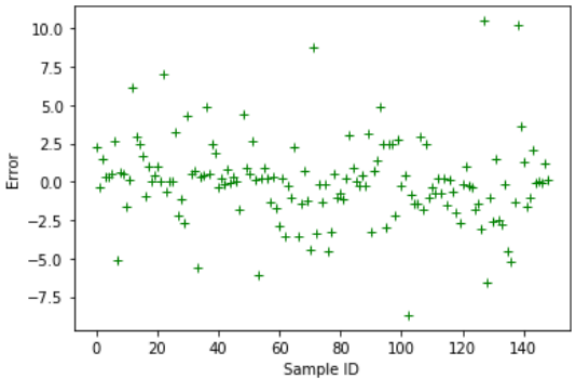

Fig. 12. Test Data Error of KNN

d) Decision Tree Classification Method'

This method based on original database, and more feature may help better classifier. I set all different life expectancy as different classifications. Decision tree is comprehensive and specific. It forces the consideration of all possible outcomes of a decision and traces each path to a conclusion. It assigns specific values to each problem, decision path and outcome.

In order to find the best parameter to prune the tree, I use cross-validation method. I find depth = 9 is the best parameter for this tree.

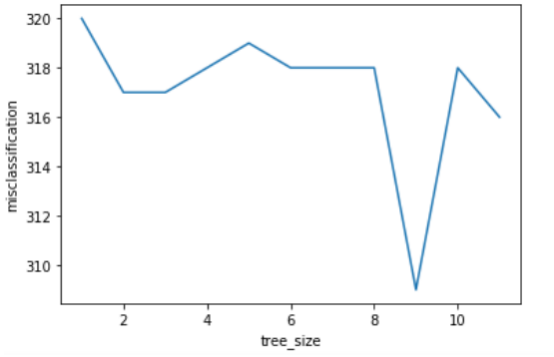

Fig. 13. Cross validation

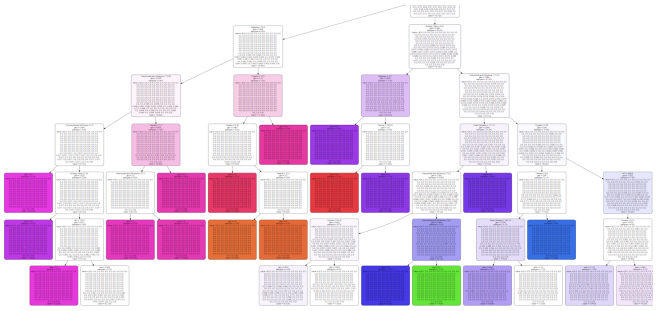

Fig. 14. Part of Pruned tree

Then I input train data to test its accuracy. The describe result of error is mean error = 1.0547, min error = 0, max error = 20.9000, variance = 5.2272; I input test data to test its accuracy and get following solution. The describe result of error is mean error = 2.0181, min error = 0. max error = 13.4000, variance = 5.5669.

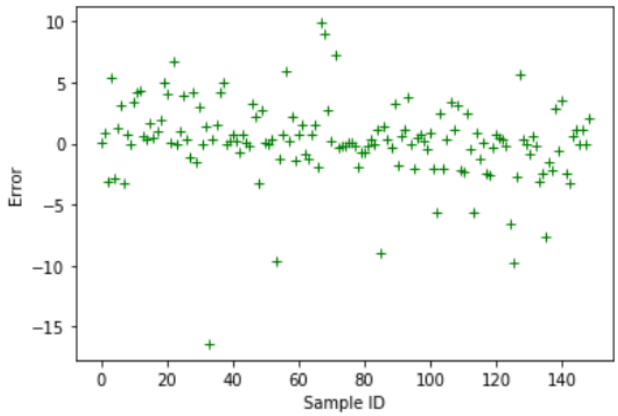

Fig. 15. Test Data Error of Decision Tree Classification Method

e) Ensemble

I use linear method to ensemble above result of life prediction, $x = u_1x_1 + u_2x_2 + u_3x_3 + u_4$. I use linear regression method get parameter of each method $u_1 = 0.2149$, $u_2 = 0.4386$, $u_3 = 0.4286$ and $u_4 = -5.8991$.

By ensemble three models I get better prediction result. The error result as follows. I input train data to test its accuracy. The describe result of error is mean error = 1.1462, min error = 0.0045, max error = 19.1515, variance = 2.4905; I input test data to test its accuracy and get following solution. The describe result of error is mean error = 1.9443, min error = 0.0097, max error = 9.4474, variance = 3.1676.

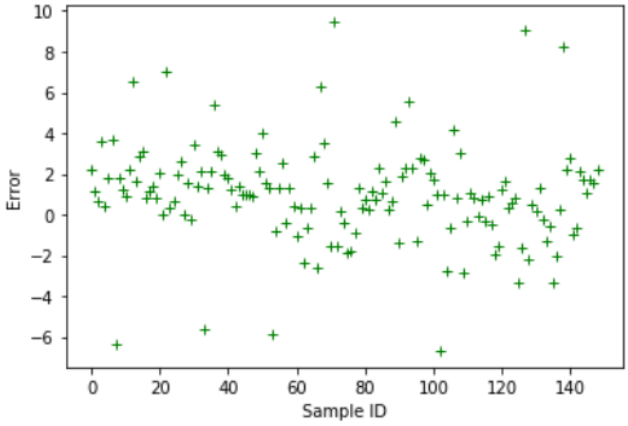

Fig. 16. Test Data Error of Ensemble

C. Zifan Cao's Part Chronic Disease influence the life expectancy

1) description of dataset:

In the big city health inventory data, there are nine categories under chronic disease, which are asthma emergency department visit rate, diabetes mortality rate, heart disease mortality rate, teen obesity, adult smoking, teen smoking, adult obesity, teen physical activity levels and adult physical activity levels.And those factors are all considered to be reasons affect American people's life expectancy. In each category dataset, there are no more than

one thousand samples (relatively small dataset to analysis), and each of them will include the Indicator, year, sex, race/ethnicity, value, place, ninety percent confidence low, ninety percent confidence high, ,ninety five percent confidence low,ninety five percent confidence high. So in this project, the features of “sex”, “race/ethnicity”, “value”, “place”, “95% confidence level - low”, “95% confidence level - high” are defined as inputs, and the value (mortality rate or the rate of certain issue happens) is the target value.

2) machine learning method and reason:

The artificial neural network (ANN) machine learning method is used to analyze the relationship between those features and the target. Since there are many categories of input and some have missing data, the tolerance of ANN is very important for implementing this project. Even though there are some corruption of one or more cells of ANN does not prevent it from generating output. This means, if none missing data are filled, the method can still generate some value output. but the more complete the dataset is, the more accurate value appears. In this project, the average value of the column is chosen to fill the missing or NAN. Even though I cannot say the filling is perfect, but the ANN can give back a fairly better output. besides, the ANN has advantages of storing information on the entire network, parallel processing ability and ability to train machine.

3) implement process:

Filt and fill the data: as described before, chronic disease can be caused by nice categories, and each category can directly or indirectly cause the influence on life expectancy. In the ANN analysis of chronic disease, the information of “sex”, “race/ethnicity”, “value”, “place”, “95% confidence level - low”, “95% confidence level - high” are chosen as the input and “value” is chosen as output.

In the analysis process, the missing data of “value”, “95% confidence level - low”, “95% confidence level - high” are found. As mentioned before, a more complete input gives a more accurate output. So the missing place are replaced with the average value from that column.

Then, the “encode”processing is used to convert the categories of words into numerical value (int), so that the machine learning system could better “understanding”.

Take asthma emergency department visit rate data as an example:

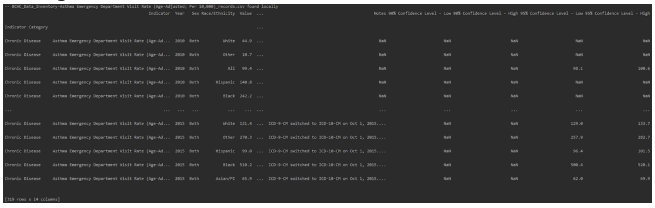

Fig. 17. before filleting and filling process

Indicator Category	Sex	Race/Ethnicity	Value	Place	95% Confidence Level - Low	95% Confidence Level - High
Chronic Disease	0	6	82	0	125.090361	125.090361
Chronic Disease	0	5	7	0	125.090361	125.090361
Chronic Disease	0	0	178	1	98.100000	100.600000
Chronic Disease	0	4	248	0	125.090361	125.090361
Chronic Disease	0	3	274	0	125.090361	125.090361
...	...	...	...	...	...	...
Chronic Disease	0	6	243	5	129.000000	133.700000
Chronic Disease	0	5	277	5	257.900000	282.700000
Chronic Disease	0	4	176	5	96.400000	101.500000
Chronic Disease	0	3	284	5	500.400000	520.100000
Chronic Disease	0	2	142	5	62.000000	69.900000

[319 rows x 6 columns]

Fig. 18. after filleting and filling process

Before processing the data, the asthma emergency department visit rate data contains three hundred nineteen rows, fourteen columns and many NAN values. After the processing, the dataset contains three hundred and nineteen rows, six columns, and all the missing data are filled.

ANN implement:

Below is the unmodified ANN implementation:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=27)
clf = MLPClassifier(hidden_layer_sizes=(100, 100, 100), max_iter=1000, alpha=0.0001,
                    solver='sgd', verbose=10, random_state=21, tol=0.000000001)
clf.fit(X_train, y_train)
```

Fig. 19. implement of ANN

X is the features set (input), and y is the target set (output). X and y are divided into test and training set by the ratio of three to seven. In the ANN there are three layers and each layer contains one hundred nodes. stochastic gradient descent is used. Alpha or the learning rate is one of ten thousand, and tol (momentum) is set to one of billion. the accuracy is based one the difference between the predicted value and the actual value.

Below is an example of predicted value based on the asthma emergency department visit rate data and accuracy for all the categories without optimization:

```
y_train_pred is
 [ 51.  101.2  51.   51.   51.  101.2  51.   51.   51.  101.2 101.2  51.
 101.2  51.   51.  101.2  51.   51.   51.   51.   51.  101.2  51.   51.
  51.   51.   51.   51.  101.4  51.   51.   51.  101.2  51.  101.2 101.2
  51.   51.   51.   51.   51.   29.9  51.   51.   51.   51.   51.   51.
 101.2  51.   51.   51.   51.   51.   51.   51.  109.   51.   51.  101.2
  51.   51.   51.  109.  109.   51.   51.   51.   51.   51.   51.   51.
  51.  109.  101.2  51.   51.   51.   51.   51.   51.   51.   51.  101.2
  51.   51.   51.   51.  37.6  51.   51.  101.2 101.2  51.   51.   51.
  51.   51.  101.4  51.   51.   51.   51.   51.   51.   51.   51.   51.
  51.   51.   51.   51.   51.   51.   51.  101.2  51.  37.6  51.   51.
  51.   51.   51.   51.   29.9  51.   51.   51.  58.6  51.   51.   29.9
  51.   51.   51.  37.6  51.   51.   51.  101.2 109.  101.2  51.   51.
  51.   51.  101.2  51.   51.   51.  101.2  51.   51.   51.   51.   51.
  51.   51.   51.  101.2  51.   51.  101.2  51.   51.   51.   51.   51.
  51.   29.9  51.   51.  101.2  51.   51.   51.   29.9  51.   51.   51.
  51.  101.2 101.2  51.   51.   51.   51.  195.  101.2  58.6  51.   51.
  51.  109.   51.   51.   51.   51.  101.2  51.   51.   51.   51.  109.
  51.   51.   51.   51.   51.   51.  101.2  51.   51.  109.   51.  109.
  37.6  51.   51.   51.  101.2  51.   51. ]
y_test_pred is
 [ 51.   51.   51.   51.   51.   51.   51.   51.   51.   51.   58.6
  51.   51.   51.   51.   51.   51.   51.   51.   51.   51.   51.
  51.   51.  101.2  51.   51.   51.  101.2  51.   51.   51.   29.9  51.
  51.   58.6 101.2  51.   51.   51.   51.   51.  101.2  51.   51.   51.
  51.   51.   51.   51.   51.   51.  109.   51.   51.   51.   51.   51.
  51.   51.   51.   51.   51.   51.   51.  101.2 109.   51.   51.   51.
  51.   51.   51.   51.   51.   51.   51.   51.   51.   51.  109.   51.
  51.   51.   51.   51.   51.   51.   51.   51.   51.   51.   51.   51. ]
```

Fig. 20. example train prediction and test prediction

Category	In Sample Accuracy	Out of Sample Accuracy
Asthma Emergency Department Visit Rate (Age-Adjusted; Per 10,000)_records	0.4190309497	0.3957228885
Diabetes Mortality Rate (Age-Adjusted; Per 100,000 people)_records	0.4654991436	0.40121273
Heart Disease Mortality Rate (Age-Adjusted; Per 100,000 people)_records	0.5538845314	0.5920331858
Percent of Adults Who Are Obese_records	0.6832404329	0.6930433512
Percent of Adults Who Currently Smoke_records	0.6387265654	0.6571209455
Percent of Adults Who Meet CDC-Recommended Physical Activity Levels_records	0.6146180011	0.649143299
Percent of High School Students Who Are Obese_records	0.5504328268	0.5362386626
Percent of High School Students Who Currently Smoke_records	0.5135948877	0.4428819292
Percent of High School Students Who Meet CDC-Recommended Physical Activity Levels_records	0.6256864643	0.6734903494

Fig. 21. accuracy without optimization

Optimization:
In all, there are three factors to adjust to optimize the accuracy. they are tree size, alpha value, and tol value. The strategy used here is to test keep the other two values at the same time, make changes to one variable. find a solution with maximum accuracy output, then plug the solution back, change another variable and keep testing.

Following graphs shows the relationship between changes and the output based on percent of adults who are obese records data:

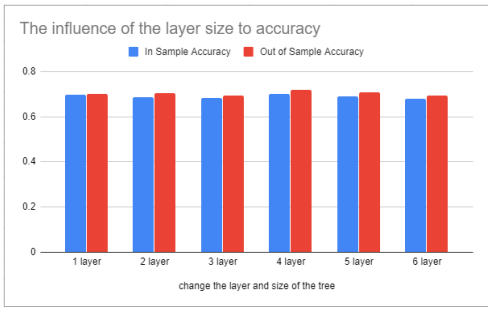

Fig. 22. change layer size

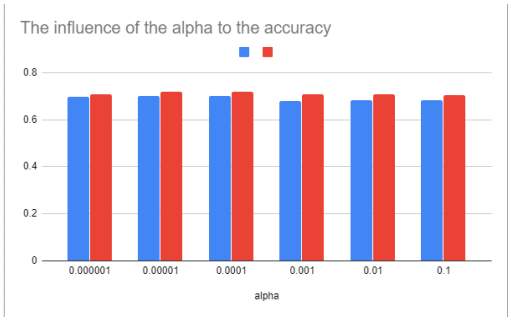

Fig. 23. change alpha size

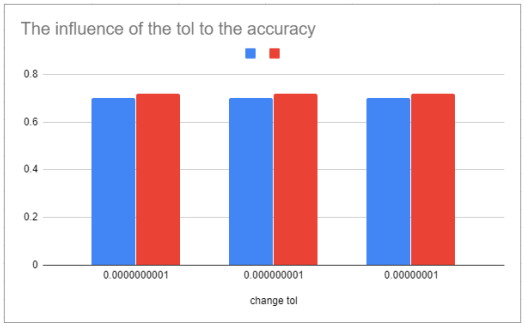

Fig. 24. change tol size

From the graph, the in-sample and out-sample accuracy reaches maximum when there are four layers and the alpha is 0.0001. The change of tol does not make any difference. However, the accuracy does not have big change due to the change of these factors.

Following is the table and graph of all categories' after optimized accuracy vs. original accuracy

Category	In Sample Accuracy	Out of Sample Accuracy	In Sample Accuracy (after)	Out of Sample Accuracy (after)
Asthma Emergency Department Visit Rate (Age-Adjusted; Per 10,000)_records	0.4190309497	0.3957228885	0.4190309497	0.3957228885
Diabetes Mortality Rate (Age-Adjusted; Per 100,000 people)_records	0.4654991436	0.40121273	0.5198296404	0.4568306697
Heart Disease Mortality Rate (Age-Adjusted; Per 100,000 people)_records	0.5538845314	0.5920331858	0.6487887551	0.6890250162
Percent of Adults Who Are Obese_records	0.6832404329	0.6930433512	0.7015620149	0.7190149583
Percent of Adults Who Currently Smoke_records	0.6387265654	0.6571209455	0.6503790293	0.6682021378
Percent of Adults Who Meet CDC-Recommended Physical Activity Levels_records	0.6146160011	0.649143299	0.6164437656	0.6493089073
Percent of High School Students Who Are Obese_records	0.5504328268	0.5362386626	0.5453368801	0.5473986458
Percent of High School Students Who Currently Smoke_records	0.5135948877	0.4428819292	0.5181382492	0.4428819292
Percent of High School Students Who Meet CDC-Recommended Physical Activity Levels_records	0.6256864643	0.6734903494	0.6250118725	0.6720019086

Fig. 25. table of optimized accuracy vs. original accuracy

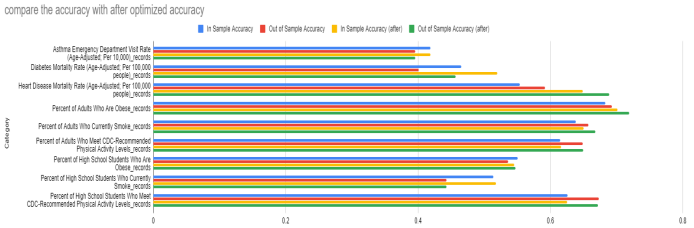

Fig. 26. graph of optimized accuracy vs. original accuracy

As graph and table shows, after changing the size of tree and alpha, the accuracy of some category increases.

Calculate the American people's life expectancy:
Assume there is a negative linear relationship between life expectancy and the number of deaths caused by chronic disease. Then use the accounted life expectancy minus the number of deaths over ten thousand (since the value are all counted from ten thousand people), and the result is the predicted American people's life expectancy based on chronic disease research.

average of train prediction	
za	60.54663677
zb	25.54205083
zc	168.6447846
zd	27.33122739
ze	19.68631995
zf	50.90365012
zg	14.17871551
zh	10.47666253
zi	24.397724
ztotal = sum/10000	0.03282206219
2015 american life expectancy	78.69
	78.65717794

Fig. 27. chronic disease life expectancy

D. Qiuyang Tang Work Part

1) Task Description

One important factor that has effects on life expectancy is violent death, including homicide, suicide, firearm-related mortality and motor vehicle mortality. My task is to predict life-expectancy of the United State based on the injury/violence category of Big City Health Inventory Data.

2) Data Analysis

Original data set gives different sexes' life expectancies, homicide rates, suicide rates, firearm-related mortality rates and motor vehicle mortality rates of several big cities in the United States from 2010 to 2016. Each city's data can be considered as examples that picked uniformly from the whole country, regardless of the race but consider the sex.

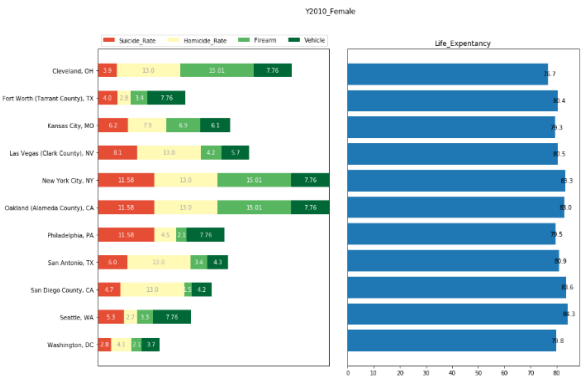

Fig. 28. Different Cities' Data of Male in 2010

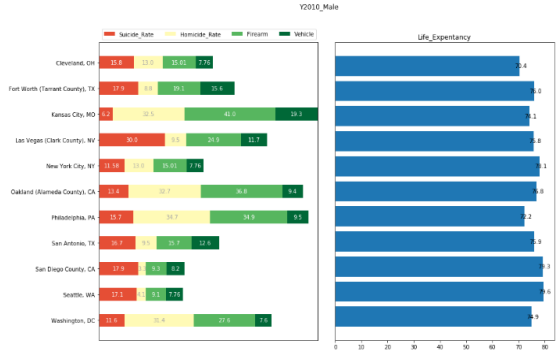

Fig. 29. Different Cities Data of Female in 2010

3) *Build Model*

To build the model, I chose the decision tree. Decision tree has several advantages to make it a good choice for this task and the most important one is being able to handle both numerical and categorical data[t1]. Origin data has both numerical data such as suicide rate and categorical data such as sex. Specifically, regression tree is used to get continuous value prediction. Instead of associating a class label to every node, a real value or a functional dependency of some of the inputs is used[t2].

4) *Regression Tree*

Before using grid search to find the best parameters for generating the tree, generate some trees to find the approximate range of parameters.

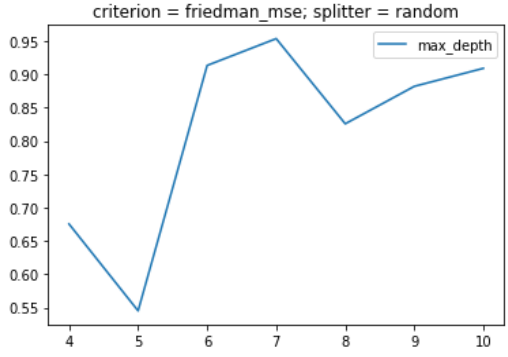

Fig. 30. Score Curve by Adjusting the Depth of Tree

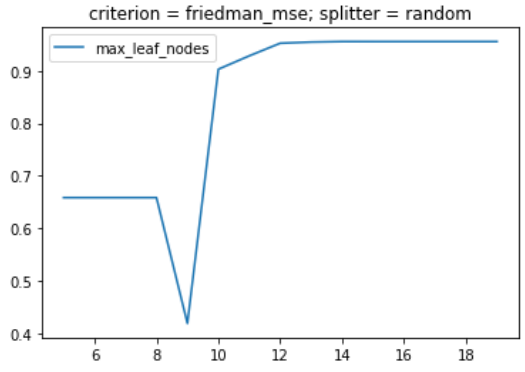

Fig. 31. Score Curve by Adjusting the Number of Nodes

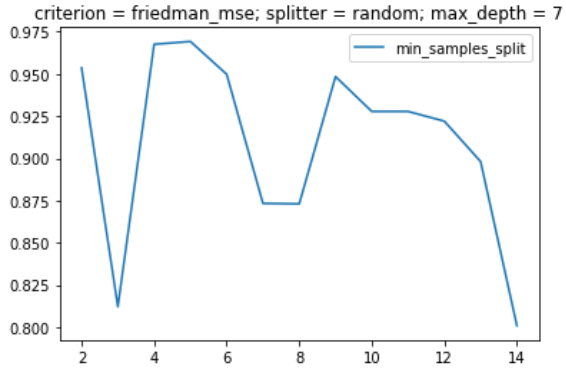

Fig. 32. Score Curve by Adjusting the Minimum Number of Samples Required to Split an Internal Node

Now we have an approximate estimation of the parameters for generating the best tree. Then, use grid search to find best parameters combination. The range of some parameters is listed below:

TABLE I. GRID SEARCH RESULT

Grid Search	
criterion	"mse", "friedman_mse"
max_depth	4,5,6,7,8,9,10
min_samples_split	2,4,6,8,10
max_leaf_nodes	10,12,14,16

The measure of impurity is mean square error(MSE)[t3]. It is equal to variance reduction as feature selection criterion and minimizes the L2 loss using the mean of each terminal node. There is another way using mean square error with Friedman's improvement score(friedman_mse) for potential splits[t4].

The best parameter combination found is using friedman mse, max_depth =7, max_leaf_node = 14, min_samples_split =2.

5) Random Forest

Use random forest to build another model. In contrast to the original publication[t5], the combination function implemented by averaging their probabilistic prediction instead of letting each regressor voting for results.

Still use grid search to find the best parameter combination and the result found is n_estimators = 100. For each single tree, parameters are criterion = 'friedman_mse', max_depth = 5, max_leaf_nodes = 16, min_samples_leaf = 2, min_samples_split = 8.

6) Evaluate

The function used to evaluate each model is R Square regression score function[t6]. Best possible score is 1.0.

TABLE II. TEST SCORE OF THREE METHODS

Model	Unpruned Regression Tree	Pruned Regression Tree	Random Forest
training score	0.99998	0.63130	0.79838
Testing score	0.88292	0.95645	0.97482

According to testing scores, random forest has the best performance for generalization.

Predicted values using above models are listed below:

TABLE III. PREDICT VALUE OF THREE METHOD

Model	Predicted value
Unpruned Regression Tree	79.32
Pruned Regression Tree	79.37
Random Forest	79.38

Consider each model's score, the result is 79.38 years for the United States in 2015.

IV. COMPARISON

A. Team Part

Original data set only provides life expectancy of U.S. total from 2010 to 2014. Predict the life expectancy of U.S. total in 2015 based on models built and each city's data of different categories in that year. True value is from[t7]. It gives life expectancy at birth of each country in different years. This source shows U.S. total's life expectancy is 78.74 in 2015. The predicted values of different categories is shown below:

TABLE IV. PREDICT OF THREE DIFFERENT FEATURES

Predicted Life Expectancy in 2015			
Indicator Category	Cancer	Chronic Disease	Injury/Violence
Result	79.82	78.66	79.38

Cancer and Injury/Violence categories all have errors of approximate +1 and Chronic Disease is very close to the true value.

B. General Part by Xiangyang

If not consider ensemble result, KNN has minimum mean error, linear regression has minimum variance. With ensemble result, ensemble result has very close mean error to KNN, and ensemble result has much better variance than other 3 method.

TABLE V. RESULT ERROR: LIFE EXPECTANCY VS GENERAL DATA

Method		Linear Regression	KNN	Decision Tree	Ensemble
Train error	Mean	1.730	1.0843	1.0547	1.1462
	Min	0.0031	0	0	0.0045
	Max	19.5410	18.7000	20.9000	19.1515
	Var	3.4397	5.0117	5.2272	2.4905
Test error	Mean	2.1360	1.8329	2.0181	1.9443
	Min	0.0133	0	0	0.0097
	Max	11.1013	10.5000	13.4000	9.4474
	Var	3.8991	4.1064	5.5669	3.1676

V. FUTURE RESEARCH DIRECTIONS

There are many future research our team could work on. Based on the research now, a roughly prediction on American's life expectancy could be told, that is because there is not enough dataset to do further learning. However, if we could combine all the dataset to one, and use some deep learning method, we could get a more valuable answer. Not only the data set on the BCHC could be used, but also from other reliable platform, and from more perspectives. For now, the effect of cancer, chronic disease, injury and violence are considered, but there are more effect that affects human's life expectancy, for instance, food safety, environment, the medical support and so on.

VI. CONCLUSION

Based on different indicator categories, we all get acceptable results to predict the life expectancy of U.S. total in 2015. The idea of doing the task separately is learnt from ensemble learning, which combines multiple hypotheses to form a better hypothesis. We chose several factors that may

affect life expectancy most. From the results we can find that those factors do affect life expectancy a lot. The prediction using chronic disease is the closest to true value may indicate chronic disease has much more effects on life expectancy than we thought. Health damaging behaviors- particularly tobacco use, lack of physical activity, and poor eating habits- are major contributors to the leading chronic diseases. This conclusion reminds us to pay more attention to chronic disease and take actions to prevent it, such as quitting smoking, doing more exercise and eating healthily. There are certainly other factors that also have effects on life expectancy, that's why our models always have some errors.

We can use the prediction of several years to see the trend of the life expectancy in the United States and that may help us to do better in the future.

References

[1] Kwetishe Joro Danjuma, "Performance Evaluation of Machine Learning Algorithms in Post operative Life Expectancy in the Lung Cancer Patients"

[2] Maciej Zieba∗, Jakub M. Tomczak, Marek Lubicz, Jerzy "Boosted SVM for extracting rules from imbalanced data inapplication to prediction of the post-operative life expectancy in the lung cancer patients,"

[3] Tom M. Mitchell 'Machine Learning'

[4] https://intellipaat.com/community/21886/advantages-and-disadvantages-of-neural-networks

[5] https://bchi.bigcitieshealth.org/indicators/1837/searches/34461

[6] https://www.python-course.eu/neural_networks_with_python_numpy.php

[7] https://jakevdp.github.io/PythonDataScienceHandbook/03.04-missing-values.html

[8] Gareth, James; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert. "An Introduction to Statistical Learning". New York: Springer, 2015, pp. 315.

[9] Dobra, A., "Classification and Regression tree construction". Thesis proposal. Department of Computer Science, Cornell University, Ithaca NY, 2002, pp.5

[10] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. "Classification and Regression Trees". Wadsworth, Belmont, 1984.

[11] Friedman, J.H.."Greedy Function Approximation: a Gradient Boosting Machine". Technical report, Dept. of Statistics, Stanford University.

[12] Breiman, "Random Forest", Machine Learning, 45(1), 5-32, 2001

[13] https://en.wikipedia.org/wiki/Coefficient_of_determination

[14] https://www.populationpyramid.net/hnp/life-expectancy-at-birth-total-years/2015/united-states-of-america/