Comparison of Three Data Mining Models for Health Index Prediction by risk factors

QE57001-M. Tech Project

By
Arka Shankar Pradhan
16QE30003

Under The supervision of
Professor J. Maiti
Department of Industrial and Systems Engineering

Indian Institute of Technology, Kharagpur

November 2019

ACKNOWLEDGEMENT

Throughout the duration of the project, I have received a great deal of guidance and support. I would first like to thank my supervisor, Prof. J Maiti Department of Industrial and Systems Engineering for supervising me on this project. His constant guidance ensured continuous progress. His expertise and mentoring were invaluable.

I would like to thank Asish Garg Sir for constant support throughout the project.

Date - 14 November 2020
Place: IIT Kharagpur

Arka Shankar Pradhan
16QE30003

Contents

Abstract

The study compares different models namely logistic regression, artificial neural networks, and random forest for the prediction of Health index of the person. We created three prediction models utilizing 21 input variables and an output variable which is multiclass in nature. The model is evaluated using accuracy, sensitivity, and specificity. This report is divided into four sections of which contains the introduction about the project second, we discuss data, methodology and at last, we discuss the result obtained. We will discuss the multiclass prediction methods and accuracy based on the risk factors and we will try to find the variable which affects the prediction results the most for each method using sensitivity analysis. We will analyze about the 533 people and will try to find out which model gives the best results and which model gives the lowest accuracy.

1. Introduction

1.1 Motivation

Health index is the characteristic of the population which is the measure of health of the population. The survey method is used to accumulate information about certain people in different regions where people have different lifestyles, different cultures and food and use the statistics for generalizing the information collected to the entire population.

1.2 Project Objective

Data mining is the method to select and explore different models on a huge amount of data and to see if there is any pattern or trend in the data. This helps us to discover unknown patterns or relationships that help us to predict very accurate results. Data mining methods are very useful in the medical field as per the study and much research is conducted using many data mining methods. The main purpose of this project is to compare different models to predict the health index of the person using classification models which include logistic regression, artificial neural network (ANNs) and decision tree.

1.3 About data mining methods

1.3.1 Logistic regression

Multinomial logistic regression is a method that uses the generalization of logistic regression to multiclass problems i.e. two or more possible outcomes. This models used to predict the probabilities of different outcomes of a categorical distributed dependent variable, given a set of the independent variable. To determine the multinomial logistic regression for K possible outcomes, K-1 independent binary outcomes are regressed using a “pivot” point.

$$\ln \frac{\Pr(Y_i = 1)}{\Pr(Y_i = K)} = \boldsymbol{\beta}_1 \cdot \mathbf{X}_i$$
$$\ln \frac{\Pr(Y_i = 2)}{\Pr(Y_i = K)} = \boldsymbol{\beta}_2 \cdot \mathbf{X}_i$$
$$\cdots\cdots$$
$$\ln \frac{\Pr(Y_i = K-1)}{\Pr(Y_i = K)} = \boldsymbol{\beta}_{K-1} \cdot \mathbf{X}_i$$

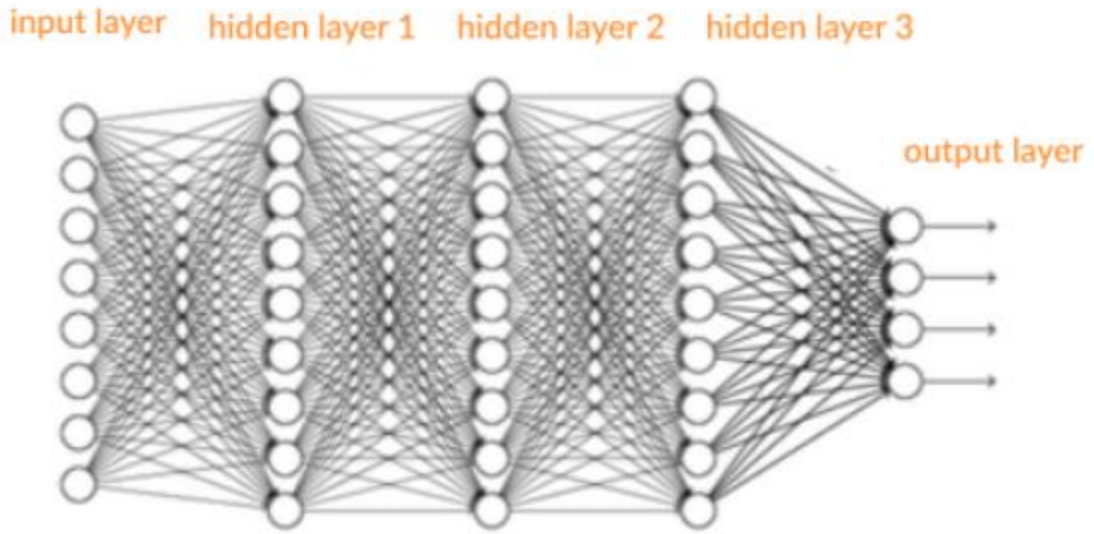
Where Xi is the observation and βk is the vector of weights or regression coefficient corresponding to outcome k.
Using mathematics, we derive the final probabilities.

$$\Pr(Y_i = 1) = \frac{e^{\boldsymbol{\beta}_1 \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}_k \cdot \mathbf{X}_i}}$$

$$\Pr(Y_i = 2) = \frac{e^{\boldsymbol{\beta}_2 \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}_k \cdot \mathbf{X}_i}}$$
$$\cdots\cdots$$
$$\Pr(Y_i = K-1) = \frac{e^{\boldsymbol{\beta}_{K-1} \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}_k \cdot \mathbf{X}_i}}$$

1.3.2 Artificial Neural Network

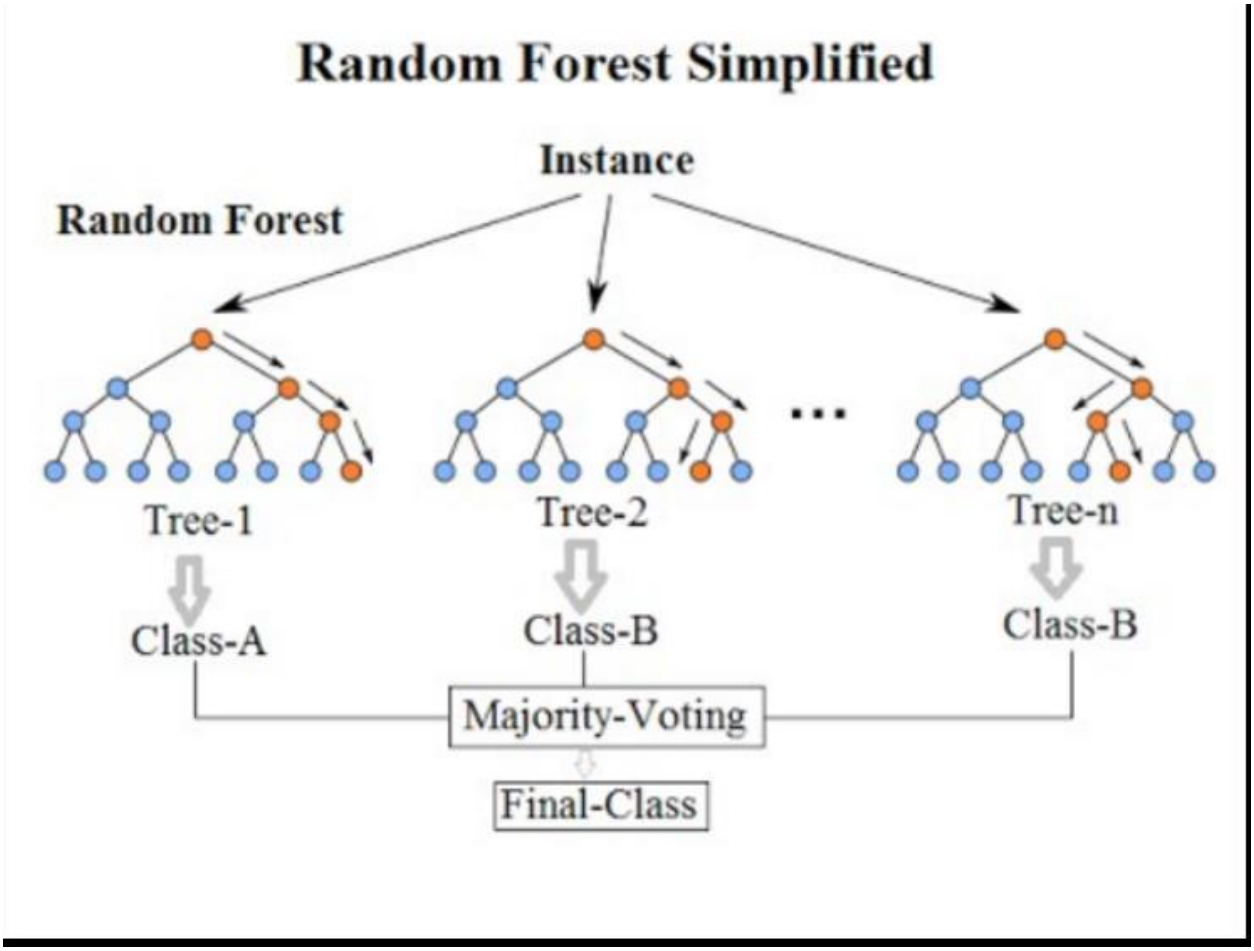
An artificial neural network is the model which consists of nodes and neuron like structures which connects them to the node of another layer. ANNs learns from the data and that enable them to produce accurate results as more data becomes available.



1.3.3 Random Forest

Random forest is a supervised learning algorithm. Random forest consists of multiple decision trees which have very low correlations and count votes from all the decision trees to get the final answer. It is useful for searching for the best features.

The low correlation between models operating as a committee will outperform any specific model. The reason for this is that the tree protects each other from their individual errors.



2. Data Collection

In this project, we will analyze health index data and in data, we have age, food habits, sleep habit and daily habits etc. We have in total 533 entries and 21 different attributes. We will discuss the attribution in the table below and values.

Sr.No	Attributes	values
1	Exp	4 <= EXP <= 57
2	Age	[above 46 years,36-45 years,up to 25 years,26-35 years]
3	I eat a variety of fruits and vegetable	[Sometimes,Always,Often,Seldom,Never]
4	I try not to eat too many foods and drinks that are salty, fatty, and sugary . I avoid sugary drinks and fatty or salty food.	[Sometimes,Always,Often,Seldom,Never]
5	Whether I prepare my own meals or eat out, I look for healthier choices to order.	[Sometimes,Always,Often,Seldom,Never]
6	I spend no more than 2 hours a day on recreational screen time such as watching TV, gaming, or on the internet.	[Sometimes,Always,Often,Seldom,Never]
7	When eating, mindful of my food intake by watching my portion sizes and taking time to savour the flavours, smells and textures of my meals.	[Sometimes,Always,Often,Seldom,Never]
8	I enjoy doing at least 60 minutes of moderate physical activities every day (Walking, Gardening, Cycling, Swimming etc.)	[<6 hours, 6-8 hours,>8hours]
9	I look for ways to include activity in my daily life (Taking stairs, Talk & Walk, Fetch water, Walking to market etc.)	[Yes, No]
10	I do some kind of stretching and strength activities & Yoga etc. at least 3 times a week.	[Yes, No]
11	How many hours of sleep do you get everyday	[Sometimes, Always, Often, Seldom,Never]
12	Do you suffer from any lifestyle disease like : Diabetes / Hypertension / High Cholesterol / Obesity / Osteoporosis / Thyroid / Arthritis / Any other	[Sometimes,Always,Often,Seldom,Never]
13	Have you undergone Annual Health check-up regularly	[Sometimes,Always,Often,Seldom,Never]
14	When you receive advice and/or medication from a physician, do you follow the advice and take the medication as prescribed.	[Yes,No]
15	Do you smoke or use any Tobacco product like Khaini, Jarda, Pan masala etc	[Sometimes,Always,Often,Seldom,Never]
16	Do you have symptoms like Discomfort or pain in the shoulders, neck, or upper or lower back / Eye Strain / Tingling & Numbness in the hands or fingers / Swelling, or joint stiffness / General feeling of muscle tightness, cramping, or discomfort / Clumsiness or loss of coordination / Any other	[Sometimes,Always,Often,Seldom,Never]
17	I spent leisure time with my family for travel, daily meals, recreation etc.	[Sometimes,Always,Often,Seldom,Never]
18	Blood Pressure	[0,2,4]
19	Blood sugar	[0,2,4]
20	Cholesterol (CHO)	[0,2,4]
21	Body Mass Index (BMI)	[0,2,4]

Finally, the last column is the health index which has 6 values which include 8, 10, 12, 14, 16. So we will design the model to get the prediction for the multiclass classification.

2.1 Data Preprocessing

In our data, several attributes are integer in nature while several attributes are string in nature. So, we need to convert the data to categorical data. We can do it with the help of label encoder using sklearn.preprocessing and importing LabelEncoder and then using one label encoder we can split the column according to the categories. One hot encoding is also imported from sklearn.preprocessing.
For example, the Age attribute has 4 categories namely ‘above 46 years’, ‘36-45’, ‘up to 25 years’ and ‘26-35 years’ which are categorized as 0,1,2 and 3 respectively, then One hot encoding is used.

Below is the table that describes the way variables are encoded using label encoder.

Index	0	1	2	3	4	5
0	1	0	None	None	None	nan
1	above 46 years	36-45 years	up to 25 years	26-35 years	None	nan
2	Sometimes	Always	Often	Seldom	Never	nan
3	Always	Sometimes	Never	Seldom	Often	nan
4	Sometimes	Always	Seldom	Often	Never	nan
5	Always	Often	Sometimes	Never	Seldom	nan
6	Sometimes	Often	Seldom	Always	Never	nan
7	< 6 hours	6 - 8 hours	> 8 hours	None	None	nan
8	NO	YES	None	None	None	nan
9	YES	NO	None	None	None	nan
10	Sometimes	Often	Always	Seldom	Never	nan
11	Sometimes	Often	Always	Never	Seldom	nan
12	Sometimes	Always	Often	Seldom	Never	nan
13	YES	NO	None	None	None	nan
14	Sometimes	Always	Never	None	None	nan
15	Sometimes	Never	Always	None	None	nan
16	Sometimes	Always	Never	None	None	nan
17	2	4	0	None	None	nan
18	4	2	0	None	None	nan
19	4	2	0	None	None	nan
20	4	2	0	None	None	nan
21	14	12	10	8	6	16

3.Methodology

3.1 Statistical Analyses

Scipy in windows is used for statistical analyses. We used the Anaconda navigator for our code. We have done statistical analysis for all the attributes. This test was carried out to find out differences between proportions with a significance value of 0.05. The dataset was divided into training and testing sets. We trained the data on the multinomial regression model, ANNs and random forest models and tested on the testing set.
The training set consists of 70% of the total dataset and 30% is the testing set.

The output variable was multiclass in nature consisting of 6 classes, where each class represents the health index of the individual.

3.1.1Parameter settings

3.1.1.1Multinomial Logistic Regression

We used the sklearn library and imported the linear_model, using the library we trained our data for the multinomial logistic regression model.

3.1.1.2 ANN

For implementing this model we used Keras library which includes classes such as models, layers, wrappers, util using which we imported Sequential, Dense, KerasClassifier and np_utils respectively.
We used 10 input nodes and the input dimension of 21 and used activation function ‘relu’. We used several hidden layers and obtained different accuracy and finally, we have 6 nodes for output, and we used ‘softmax’ as the activation function. And finally, compile the model using loss as ‘sparse_categorical_crossentropy’, optimizer as ‘adam’ and metrics.
The estimator of the model used baseline_model and epochs are 200 and batch_size is 5 and verbose=0.

3.1.1.3 Random Forest

For implementing this we used the sklearn library and using ensemble we imported the RandomForestClassifier. The parameters for this model are n_estimator is 10, criterion as entropy and the random state as 42.

3.1.2 Pearson Chi-Square Test

Sr.No	Attributes	values	HI=14	HI=12	Hi=10	HI=16	HI=8	HI=6	Pearson Chi-square test	p-value
1	Exp	>16 <16	512 21	1618	1784	863	615	60	4.4028	0.4925
2	**Age** above 46 years 26-35 years 36-45 years up to 25 years	 199 162 152 20	 69 49 48 3	 67 59 54 2	 45 15 27 2	 3 36 15 12	 11 3 6 1	 4 0 2 0	92.0338	<0.001
3	**I eat a variety of fruits and vegetable** Sometimes Always Often Seldom Never	 185 143 143 39 23	 52 53 47 12 5	 67 46 48 14 7	 32 26 22 4 5	 25 11 19 7 4	 6 6 6 2 1	 3 1 1 0 1	13.193	0.8689
4	**I try not to eat too many foods and drinks th at are salty, fatty, and sugary.** **I avoid sugary drinks and fatty or salty food** Always Often Sometimes Seldom Never	 215 154 109 37 18	 68 46 40 11 4	 79 56 31 10 6	 33 25 21 7 3	 25 24 11 6 0	 7 3 5 3 3	 3 0 1 0 2	38.341	0.0080
5	**Whether I prepare my own meals or eat out, I look for healthier choices to order.** Always Often Sometimes Seldom Never	 250 166 80 20 17	 81 47 33 4 4	 95 49 21 11 6	 35 36 10 4 4	 27 29 10 0 0	 10 4 4 1 2	 2 1 2 0 1	33.415	0.0303
6	**I spend no more than 2 hours a day on recre ational screen time such as watching TV, ga ming, or** **on the internet.** Always	 227 120	 75 45	 81 38	 39 18	 23 18	 6 1	 3 0	24.28	0.2303

	Often Sometimes Seldom Never	109 39 38	32 8 9	34 15 14	19 8 5	13 6 6	8 2 4	3 0 0		
7	**When eating, mindful of my food intake by w atching my portion sizes and taking time to savour the flavours, smell s and textures of my meals.** Always Often Sometimes Never Seldom	217 162 107 29 18	68 44 43 9 5	74 55 34 11 8	40 28 14 6 1	25 26 10 1 4	6 8 5 2 0	4 1 1 0 0	17.8 792	0.5953
8	**I enjoy doing at least 60 minutes of moderate physical activities every day (Walking, Gardening, Cycling, Swimmin g etc.)** 6 - 8 hours < 6 hours > 8 hours	414 92 27	142 20 7	131 41 10	72 13 4	49 13 4	16 3 2	4 2 0	11.0 24	0.3555
9	**I look for ways to include activity in my daily life (Taking stairs, Talk & Walk, Fetch water, Walking to market etc.)** NO YES	377 156	129 40	127 55	57 32	50 10	6 15	2 4	31.0 14	<0.001
10	**I do some kind of stretching and strength act ivities & Yoga etc. at least 3 times a week** YES NO	512 21	161 8	178 4	86 3	61 5	20 1	6 0	4.40 28	0.4925
11	**How many hours of sleep do you get everyday** Always Sometimes Often Seldom Never	202 135 109 47 40	71 43 33 14 8	73 45 42 10 12	32 28 14 9 6	16 12 16 11 11	6 6 4 3 2	4 1 0 0 1	30.9 49	0.0558
12	**Do you suffer from any lifestyle disease like : Diabetes / Hypertension / High Cholesterol / Obesity / Osteoporosis / Thyroid / Arthritis / Any other** Always Often Sometimes Seldom Never	262 148 88 24 11	93 48 22 5 1	94 47 34 3 4	32 29 18 8 2	28 17 12 7 2	11 5 2 1 2	4 2 0 0 0	32.9 199	0.0344
13	**Have you undergone Annual Health check-up regularly** Always Sometimes Often Never Seldom	179 139 100 68 47	62 38 34 24 11	70 45 27 20 20	21 32 21 8 7	16 14 17 12 7	7 8 1 4 1	3 2 0 0 1	28.3 906	0.1004

14	**When you receive advice and/or medication f rom a physician, do you follow the advice and take the medication as prescribed.** YES NO	498 35	163 6	172 10	81 8	57 9	19 2	6 0	9.79 57	0.0812
15	**Do you smoke or use any Tobacco product like Khaini, Jarda, Pan masala etc** Never Sometimes Always	411 93 29	131 32 6	145 27 10	64 17 8	56 8 2	11 7 3	4 2 0	15.9 979	0.0996
16	**Do you have symptoms like Discomfort or p ain in the shoulders, neck, or upper or lower back / Eye Strain / Tingling & Numbness in the hands or fingers / Swelling, or joint stiffness / General feeling of muscle tightness, cramping, or discomfort / Clumsiness or loss of coordination / Any other** Never Sometimes Always	253 251 29	82 78 9	89 86 7	43 39 7	25 37 4	9 10 2	5 1 0	8.50 57	0.5795
17	**I spent leisure time with my family for travel, daily meals, recreation etc.** Always Sometimes Never	264 246 23	98 66 5	82 92 8	45 41 3	26 36 4	10 9 2	3 2 1	13.2 739	0.2087
18	**Blood Pressure** 2 4 0	372 138 23	116 53 0	160 16 6	76 3 10	0 66 0	15 0 6	5 0 1	296. 0694	<0.001
19	**Blood sugar** 4 2 0	480 44 9	165 4 0	166 16 0	67 18 4	66 0 0	16 3 2	0 3 3	153. 2454	<0.001
20	**Cholesterol (CHO)** 4 2 0	438 85 10	147 22 0	154 28 0	62 22 5	66 0 0	7 9 5	2 4 0	116. 0956	<0.001
21	**Body Mass Index (BMI)** 4 2 0	257 224 52	143 26 0	37 142 3	11 43 35	66 0 0	0 11 10	0 2 4	428. 8479	<0.001

From the table we can conclude that many factors or attribute had no significance statistically, e.g. spending leisure time with family for travel, daily meals etc., consumption of tobacco, hours of sleep etc. While others show significant statistical significance.

4. RESULTS

We used a confusion matrix to see the performance of all three models for all 6 classes of health index. We calculated the confusion matrix and accuracy for each method.
Confusion Matrix for Multinomial Logistics regression

	6	8	10	12	14	16
6	0	2	0	0	0	0
8	0	3	5	0	0	0
10	1	1	23	0	0	0
12	0	0	0	57	0	0
14	0	0	0	0	47	1
16	0	0	0	0	2	18

Confusion matrix for ANN

	6	8	10	12	14	16
6	0	1	1	0	0	0
8	0	2	6	0	0	0
10	0	2	19	3	1	0
12	0	0	5	43	9	0
14	0	0	0	4	42	2
16	0	0	0	0	4	16

Confusion matrix for Random Forest

	6	8	10	12	14	16
6	0	0	2	0	0	0
8	0	2	5	1	0	0
10	0	0	12	13	0	0
12	0	0	3	47	7	0
14	0	0	0	4	44	0
16	0	0	0	0	3	17

As logistic regression is most used for classification problems, we achieved an accuracy of 92%. While using Ann we achieved an accuracy of 76.25% and random forest model achieved an accuracy of 75.6%. Hence, we can conclude that multinomial logistic regression is performing best for this model. But using a dense layer in Ann it outperforms the multinomial logistic regression e.g. using 2 dense layers of 23 and 23 nodes we get an accuracy of 94%.

Our main task is to find the risk factors that affect the most in our prediction models.

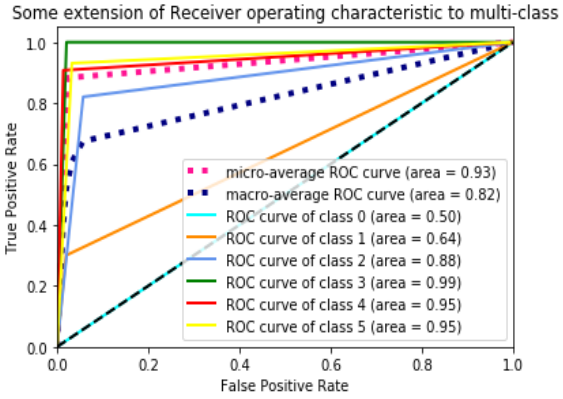


Fig: ROC for multinomial linear regression

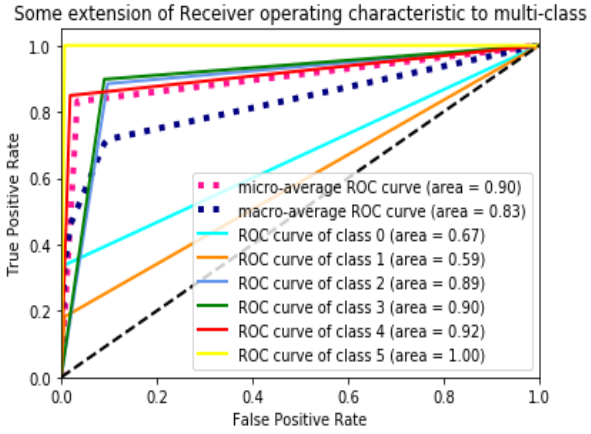


Fig: ROC for ANN

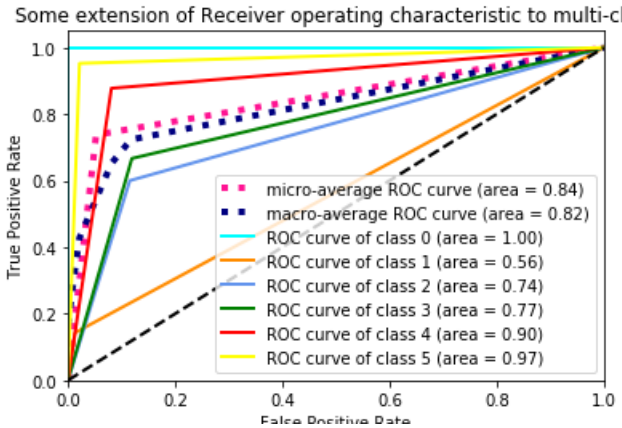


Fig: ROC for Random Forest

The Roc curve of the multinomial linear regression is given above it shows that steepness of each class true positive rate. As the steepness increases the better is true positive rate.The graph shows the micro average of the prediction rate. Micro average is the method to draw ROC curve considering each element of the label indicator matrix as binary prediction. Here class 0 is 6, class 1 is 8, class 2 is 10, class 3 is 12, class 4 is 14 and finally class 5 is 16.

As we can compare by seeing the graph that area under the curve of the classes of the ROC curve we can conclude that ANN and multinomial linear regression perform better than Random forest and we can also compare the steepness of the ROC curves of classes we can conclude that the ANN and logistic regression performs much better than random forest. We also have done the ranking of each attribute (how they affect the output) for multinomial linear regression

Logistic Regression

Sr.No	Attributes	Rank
1	Exp	5
2	Age	6
3	I eat a variety of fruits and vegetable	12
4	I try not to eat too many foods and drinks that are salty, fatty, and sugary . I avoid sugary drinks and fatty or salty food	14
5	Whether I prepare my own meals or eat out, I look for healthier choices to order.	13
6	I spend no more than 2 hours a day on recreational screen time such as watching TV, gaming, or on the internet.	19
7	When eating, mindful of my food intake by watching my portion sizes and taking time to savour the flavours, smells and textures of my meals	18
8	I enjoy doing at least 60 minutes of moderate physical activities every day (Walking, Gardening, Cycling, Swimming etc.)	11
9	I look for ways to include activity in my daily life (Taking stairs, Talk & Walk, Fetch water, Walking to market etc.)	8
10	I do some kind of stretching and strength activities & Yoga etc. at least 3 times a week	4
11	How many hours of sleep do you get everyday	7
12	Do you suffer from any lifestyle disease like : Diabetes / Hypertension / High Cholesterol / Obesity / Osteoporosis / Thyroid / Arthritis / Any other	15
13	Have you undergone Annual Health check-up regularly	21
14	When you receive advice and/or medication from a physician, do you follow the advice and take the medication as prescribed.	9
15	Do you smoke or use any Tobacco product like Khaini, Jarda, Pan masala etc	10
16	Do you have symptoms like Discomfort or pain in the shoulders, neck, or upper or lower back / Eye Strain / Tingling & Numbness in the hands or fingers / Swelling, or joint stiffness / General feeling of muscle tightness, cramping, or discomfort / Clumsiness or loss of coordination / Any other	16
17	I spent leisure time with my family for travel, daily meals, recreation etc.	17
18	Blood Pressure	20
19	Blood sugar	2
20	Cholesterol (CHO)	1
21	Body Mass Index (BMI)	3

5. Conclusion and Future Work

Now it is very important to understand data in the medical field also. We want some results that depend upon some trend or relation. This project discussed how multiclass data can be predicted and compared models to get very accurate predictions. Data analysis was also done on the risk factors and after conducting chi square test we got the attributes which are more involved in getting the class. Also, this project helps us to understand what are the attributes that affect the output results.

We need to do more test analysis on the ANN method and Random Forest model as it is very important to find what factors govern the output from each model and rank them accordingly.

References

[1] Pan XR, Yang WY, Li GW, Liu J. Prevalence of diabetes and its risk factors in China, 1994. Diabetes Care 1997:1664–9.

2. Chang CD, Wang CC, Jiang BC. Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors. Expert Syst Appl 2011; 38:5507–13.

3. https://www.geeksforgeeks.org/
4. https://machinelearningmastery.com/multi-class-classification-tutorial-keras-deep-learning-library/
5. **Comparison of three data mining models for predicting diabetes or prediabetes by risk factors**

   Xue-Hui Meng [a], Yi-Xiang Huang [a], Dong-Ping Rao [b], Qiu Zhang [a], Qing Liu [b,*]