Comparison of three data mining models for predicting Health Index by risk factors

MTP Presentation

Arka Shankar Pradhan

16QE30003

MTP Guide – Prof. J Maiti

IIT Kharagpur

Introduction

- Motivation-Health index is the characteristic of the population which is evidence for describing the health of the population. Health index is the characteristic of the population which is evidence for describing the health of the population.
- Objective - 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, along with a 10-fold cross-validation technique.

Multinomial 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 model is 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.

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

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

$$\cdots\cdots$$

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

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

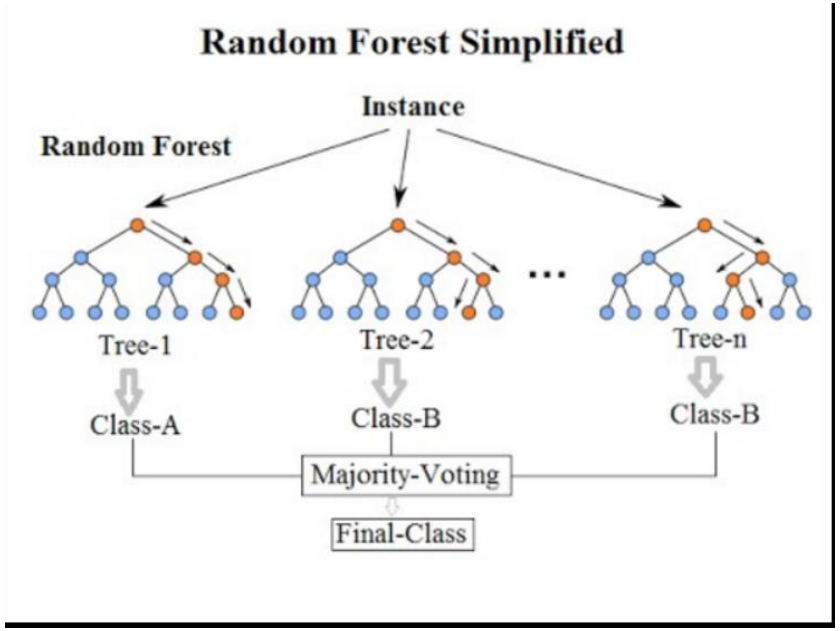
$$\ln \frac{\Pr(Y_i = 2)}{\Pr(Y_i = K)} = \boldsymbol{\beta}_2 \cdot \mathbf{X}_i$$

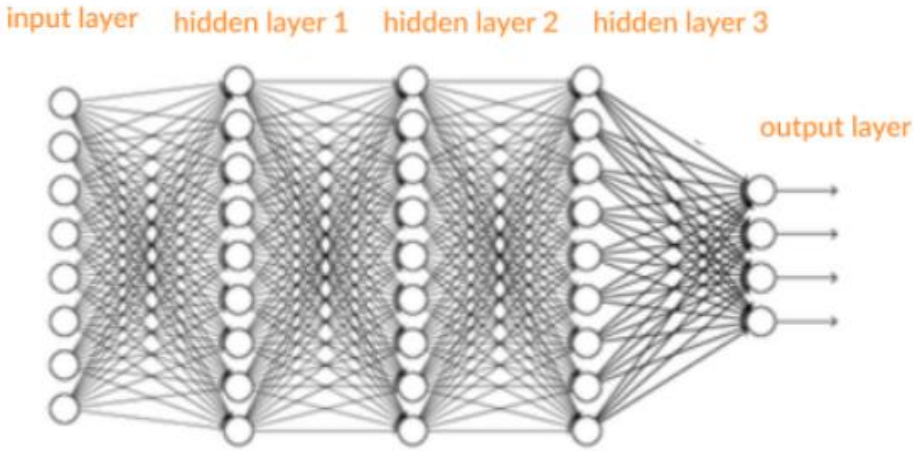
$$\cdots\cdots$$

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

ANN and Random Forest

- An artificial neural network is the computing system designed to simulate the way the human brain performs and processes the information. ANNs have self-learning capabilities that enable them to produce accurate results as more data becomes available.
- Random forest is a supervised learning algorithm. Random forest builds multiple decision trees, and they merge all together to get more accurate and stable results.





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.
- 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.

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, I am 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]

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 Label Encoder and then using one label encoder we can split the column according to the categories.
- 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.

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

Methodology

- Statistical analyses were performed using SciPy in windows. We used the Anaconda navigator for our code. Statistical analyses were carried out for all variables, using the Chi-Square test.
- 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.

Parameter Settings

- Multinomial 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.
- ANN-

    1. Keras- Sequential, Dense, KerasClassifier, np_utils

    2. 10 input nodes and testing with hidden layers, 6 output node

    3. Loss function as ‘sparse_categorical_crossentropy’, optimizer as adams, softmax activation function.

- Random Forest-

    1. sklearn library, using ensemble imported RandomForestClassifier

    2. n_estimator is 10 and random state = 10

Pearson Chi-Square Test

- 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.

r.No	Attributes	values	HI=14	HI=12	Hi=10	HI=16	HI=8	HI=6	Pearson Chi-square test	p-value
	Exp	>16 <16	512 21	161 8	178 4	86 3	61 5	6 0	4.4028	0.4925
	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
	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
	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. 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
	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

Confusion Matrix And Accuracy

- 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.

Accuracy-

1. Multinomial Linear regression – 92%
2. ANN – 94%
3. Random Forest – 75.6%

	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

Confusion Matrix for Random Forest

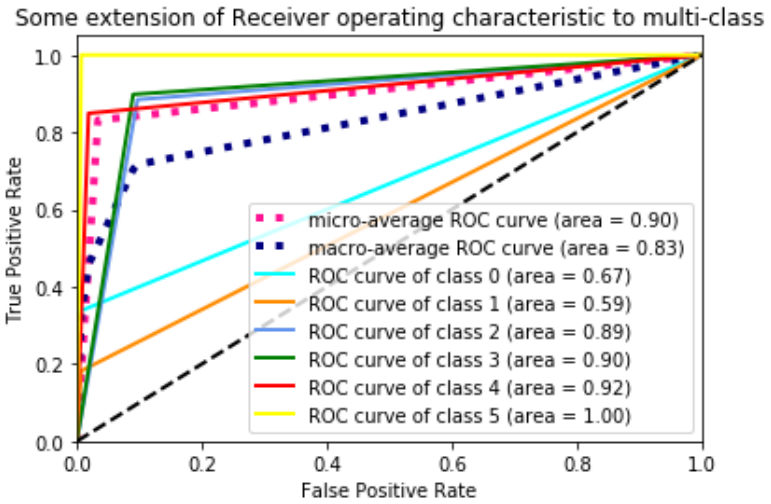
	6	8	10	12	14	16
6	0	2	0	0	0	0
8	0	3	5	0	0	0
10	1	1	23	0	0	0
12	0	0	0	57	0	0
14	0	0	0	0	47	1
16	0	0	0	0	2	18

Confusion Matrix for Multinomial Logistic Regression

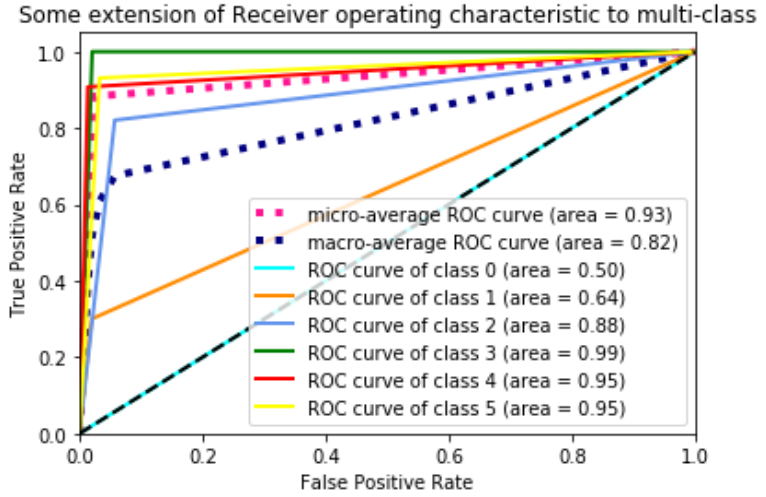
	6	8	10	12	14	16
6	0	1	1	0	0	0
8	0	2	6	0	0	0
10	0	2	19	3	1	0
12	0	0	5	43	9	0
14	0	0	0	4	42	2
16	0	0	0	0	4	16

Confusion Matrix for ANN

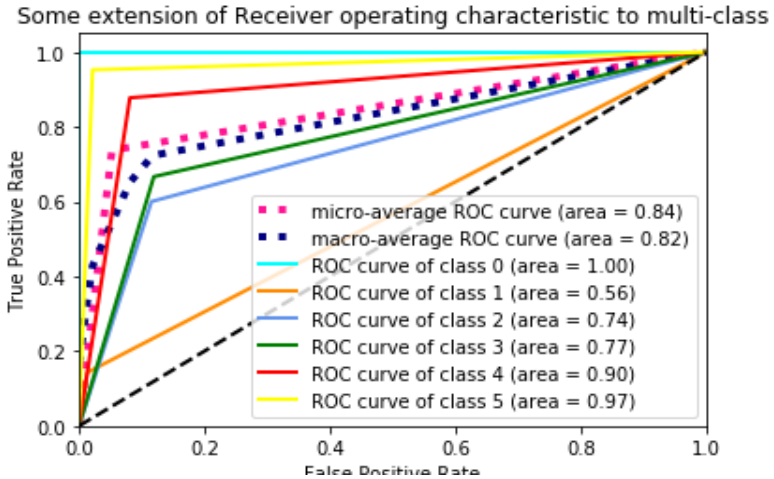
Receiver Operating Characteristics Curve



Artificial Neural Network



Multinomial linear regression



Random Forest

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.

Confusion Matrix and Ranking of Risk factors

- 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%.
- 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%.

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

Ranking for Multinomial logistic Regression

Conclusion and Future Scope

- 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.
- 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.
- Using machine learning model we can predict the health risk of a person and can also communicate what are the risk factor involved in a particular disease.

Thank You