Using machine learning in chat box for heart disease prediction

Lingyun Yang
University of Nottingham
Psxly4@nottingham.ac.uk

ABSTRACT

Heart disease is one of the deadly diseases with high incidence. But the prediction or diagnosis of heart disease requires complex procedures in hospitals. Therefore, this article attempts to collect some data using simple conversations and use machine learning models to predict the likelihood of a user's heart disease, so as to try to help possible high-risk groups to prevent it as soon as possible. The core idea of this project is using a chat box to have conversion with the user and collect some data from the conversion, then use bult-in machine learning models to make prediction. This paper contains two main contents, the design of the chat box and the design of the machine learning model.

1. INTRODUCTION

According to the CDC(The Centers for Disease Control and Prevention is the national public health agency of the United States) heart disease is one of the leading causes of death for people of most races in the US (African Americans, American Indians and Alaska Natives, and white people). However, the data from NHS England shows for patients waiting to start treatment at the end of November 2021, the median waiting time was 11.5 weeks. Such long wait times can miss the best prevention and treatment period for early-stage patients. Therefore, this research tries to use a simple method to allow users to know whether they are a high-risk group of heart disease and help users to prevent it as soon as possible.

2. RELATED WORK

2.1. K_NN Algorithm

Why K-NN? The k-nearest neighbors algorithm (k-NN) is a non-parametric supervised learning method first developed by Evelyn Fix and Joseph Hodges in 1951[KNN-first], The KNN algorithm is a type of lazy learning, where the computation for the generation of the predictions is deferred until classification. Although this method increases the costs of computation compared to other algorithms, KNN is still the better choice for applications where predictions are not requested frequently but where accuracy is important. In this project, the prediction is related to the user's health, so the accuracy rate is important, that’s why K-NN is chosen in this research.

2.2. SVC

Why SVM(SVC)? Support Vector Machine (SVM) was first heard in 1992, introduced by Boser, Guyon, and Vapnik in COLT-92 [svm-1]. Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. In another word, Support Vector Machine (SVM) is a classification and regression prediction tool that uses machine learning theory to maximize predictive accuracy while automatically avoiding over-fit to the data. This is exactly the goal that this research wants to achieve.

2.3. Chat Box

The chax box that used in this research is adapt from lab of module COMP4105 in University of Nottingham. Today, we will work on a simple chatbot agent that has a store of knowledge, uses that knowledge to decide what questions to ask, and adds to that store of knowledge by interpreting the user's responses.

3. EXPERIMENT

3.1. Tokenization using NLTK

According to the data from CDC. About half of all Americans (47%) have at least 1 of 3 key risk factors for heart disease: high blood pressure, high cholesterol, and smoking. Other key indicator include diabetic status, obesity (high BMI), not getting enough physical activity or drinking too much alcohol. There are 14 factors selected as the basis for the design question. These factors are also 14 features in the machine learning models. Table -1 shows some typical factor and there corresponded questions and answers.

Factors	typical questions	Answer rescticeted
BMI	could you please tell me your BMI?	float <= 50
MentalHealth	how many days during the past 30 days was your mental health not good? (0-30 days)	float <=30
DiffWalking	Do you have serious difficulty walking or climbing stairs?	Binary [yes,no]
Sex	are you male or female?	Binary [male,female]
AgeCategory	could you please tell me your age category, like 20-29 , 40-49	categorical
Diabetic	have you ever had diabetic?	Binary [yes,no]
SleepTime	how many hour do you sleep in a 24-hour period?	float <=24

Table – 1 Question design and answer scope

After the user enters an answer use natural language toolkit NLTK to tokenize the answer and tagging each word. After this match the target word from the tagged word. If there is an target word in the answer, then convert it into the correct format to be used as input to the model to make predictions. If user's answer out of the restricted answer, ask their user input answer again. Please refer to Table -2, the Pseudo code of chate box:

```
# initial question list and create unknown question list
Create QAdic = questions and answer dictionary
unknown list = list QAdic
```

```
# if there is unknown questions in list, start ask questions
if unknown list is not empty:
    ask question loop:
        question = random pick from unknown list
        #get the input from user, tokenize and tag answer
        answer = get input from user
        tokenize answer and tag them
        # Match the target key word from tagged answer
        targetAns = matched [ tagged word list]
        if the tagged word list includes "bye":
            break the loop and say bye to user
        else:
            if(matched):
                remove question from unknown list
                transfer matched key words to right format
                continue loop
            else(user input out of answer scope):
                ask user input the answer again
                back to the step answer = get input from user
                ...

         If(unknown list is not empty):
            continue the loop
        else:
            end the loop and say bye to user
```

Table -2 pseudo code of Chat Box

3.2. Machine learning Models

In this research, I build two models based on two different machine learning methods: the K_NN algorithm and the SVM algorithm. I also compare these two models based on the same dataset. This research uses a machine learning tool called scikit-learn to build and optimize the model. Scikit-learn (formerly scikits. learn and known as sklearn) is a Python library for machine learning. Support vector machines, random forests, gradient boosting, k-means and DBSCAN, are included, as well as various classification, regression and clustering algorithms. The NumPy and SciPy Python libraries are integrated with it.

3.2.1 Machine learning Models

The dataset used for this research is part of Personal Key Indicators of Heart Disease from Kaggle.com [Kaggle]. The original datasets from Kaggle.com contains 319794 instances. In order to reduce the computational load, I just randomly picked 1000 instances as a new datasets for this research. All data was split into training set and test set, the proportion of test size is 16%.

The Original dataset comes from the CDC and is a major part of the Behavioural Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to gather data on the health status of U.S. residents. As the CDC describes: "Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world.". The most recent dataset (as of February 15, 2022) includes data from 2020

3.2.2 Experiment

1) K_NN model

KNN manipulates the training data and classifies the new test data based on distance metrics. It finds the k-nearest neighbours to the test data, and then classification is performed by the majority of class labels. K value indicates the count of the nearest neighbours, and there are no pre-defined statistical methods to find the most favourable value of K. So first, randomly pick a K value and start the computing. After this I derive a plot between accuracy and K values in a defined range from 1 to 14(because the number of features is 14). Then choose the K value as having a maximum accuracy.

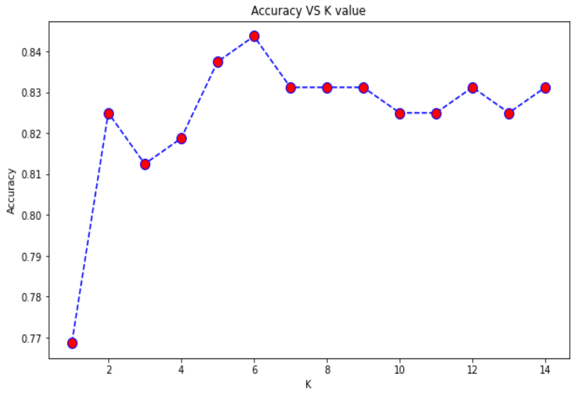


Figure – 1 Visualize plot between accuracy and K value

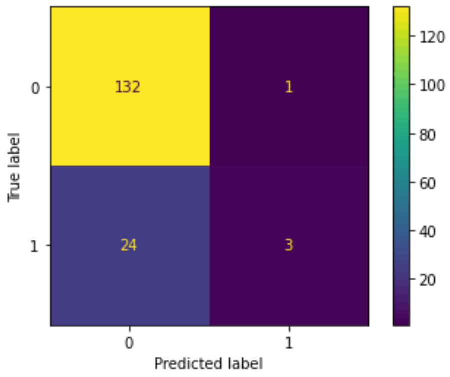


Figure – 2 Visualize plot confusion matrix , K = 6

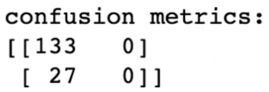
The result is shown in Figure – 1. From the plot, the maximum accuracy got in this experiment is 0.84 at K = 6. Therefore, set the K = 6 and calculate the confusion metrics. The result of confusion metrics is shown below. And visualization of confusion metrics is shown in Figure -2.

```
confusion metrics:
[[132   1]
 [ 24   3]]
```

2) SVC Model

Another model built in this research is based on machine learning method SVC(support vector classifier).

Support Vectors Classifier tries to find the best hyperplane to separate the different classes by maximizing the distance between sample points and the hyperplane. There are 3 main parameters in the Support Vector Classifier: Kernel, gamma and C. kernel kernel parameters selects the type of hyperplane used to separate the data. Using 'linear' will use a linear hyperplane (a line in the case of 2D data). 'rbf' and 'poly' uses a non-linear hyper-plane. Gamma is a parameter for non-linear hyperplanes. The higher the gamma value it tries to exactly fit the training data set. C is the penalty parameter of the error term. It controls the trade-off between smooth decision boundary and classifying the training points correctly. Set the param-eter as default setting, and computing an result. After that, use GridSearchCV method to optimize those parameters. Parameter range of C is[0.1,1,10,1000]. After calculation, the best parameters are { 'C': 0.1 } with a score of 0.7875. So set the C equals 0.1 get the confusion m-etrics shown below, and the visualization of confusion matrix shown in Figure -3.



```
confusion metrics:
[[133   0]
 [ 27   0]]
```

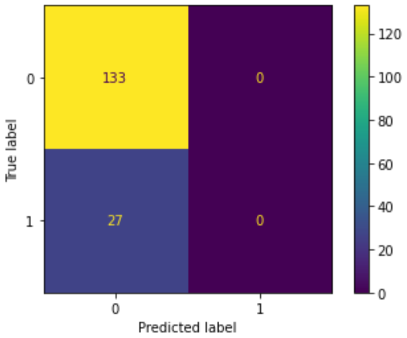


Figure - 3 Visualize plot confusion matrix

4. MODEL EVALUATION

In this research, two models were compared using Accuracy, Recall, Precision and AUC. The summary of Accuracy, Recall Precision and AUC is shown in Table - , and the visualized plot is shown in Figure -

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. The calculation formula of Accuracy is:

$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$. In this research, the accuracy of SVM model is 0.7875, the accuracy of K_NN model is 0.8438.

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The calculation formula is: $Precision = \frac{TP}{TP+FP}$. This formula directly shows what precision talks about. It about how precise/accurate the model is out of those predicted positive, how many of them are actual positive. Precision is a good measure to determine, when the costs of False Positive is high. The precision of SVM model is 0.18, the precision of K_NN model is 0.75.

Recall actually calculates how many of the Actual Positives that model capture through labeling it as Positive (True Positive). It is the ratio of correctly predicted positive observations to the all observations in actual class. The Recall of SVM model is 0.074, the recall of K_NN model is 0.111.

	SVM	KNN
Accuracy	0.7875	0.8438
Precision	0.1818	0.7500
Recall	0.0741	0.1111
AUC	0.4684	0.6331

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Postive Rate and False Positive Rate. Calculation formula are: $TPR = \frac{TP}{TP+FN}$, $FPR = \frac{FP}{FP+TN}$. An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The following figure shows a typical ROC curve.

AUC stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC. The AUC calculation results of these two models are 0.498 for SVC model and 0.633 for K_NN model From the point of view of AUC, the model built with K_NN algorithm have a better performance than the model built with SVC.

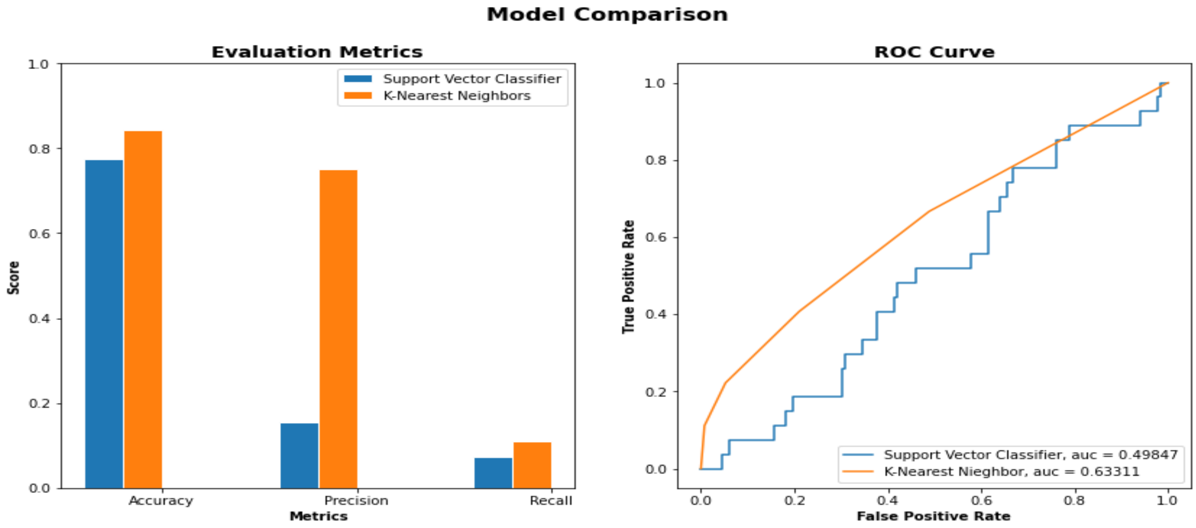


Figure – 3 Model comparison

5. FUTURE WORK AND CONCLUSION

In this research, a chat box and two machine learning model were applied in order to built a simple agent to predict the probability of someone have heart disease, so that users can be reminded to change their lifestyles as soon as possible for prevention, or get treatment. The model built on the K_NN algorithm performs better in both Accuracy, Precision, Recall and AUC than the model built on the SVM algorithm.

In the future, there are a few things that can be improved and developed if more time is available. First, the chat box now just have 14 questions built in with almost no interactive between users and the agent. So a more interactive chatbox might be better at engaging users to chat, increasing the likelihood of collecting more data. In addition, during the construction of the machine learning model, more training data sets and more features can also increase the accuracy of prediction.

REFERENCES