Using machine learning in chat box for heart disease prediction

COMP4105
Lingyun Yang -20341447

05/25/2022

Overview

- Background Introduction
- Project Object
- Experiment
  - Chat box building
  - Model building and parameter optimization
- Model Evaluation
- Feature work & Conclusion

Background Information

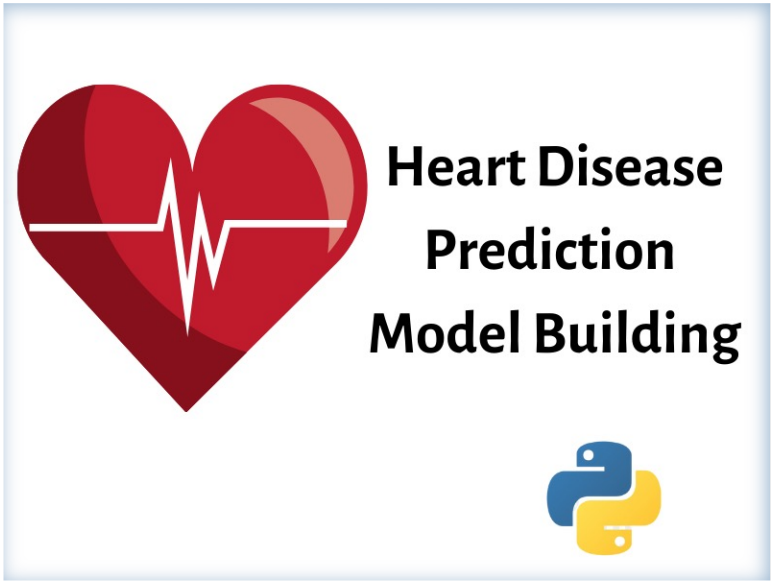
- According the CDC and NHS England reports, heart disease is one of the deadly disease with high incidence[1].
- The median number of waiting time in NHS England was 11.5 weeks[2].
- Such long wait times may miss the best prevention and treatment period for early-stage patients.

[3]

Project Object

Heart disease prediction model

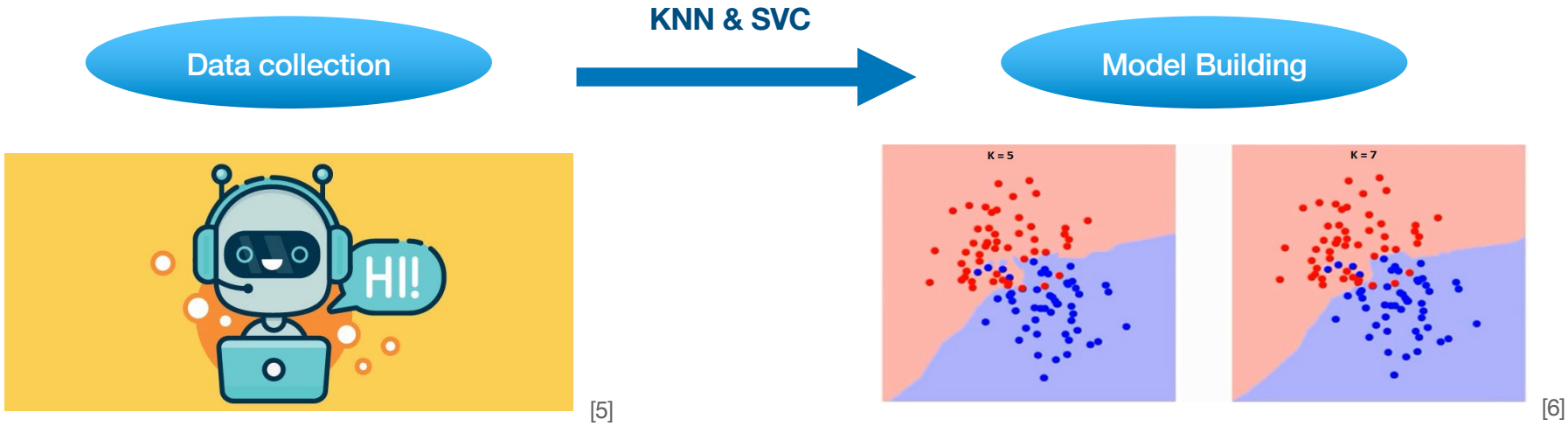
- Help possible high-risk groups to prevent heart disease by building a heart disease prediction model



[4]

Experiment

Key Approach

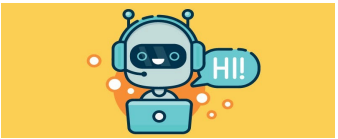


[5] [6]

- collect data using simple conversations and apply machine learning methods KNN and SVC to build the model.

Experiment

Chat Box design



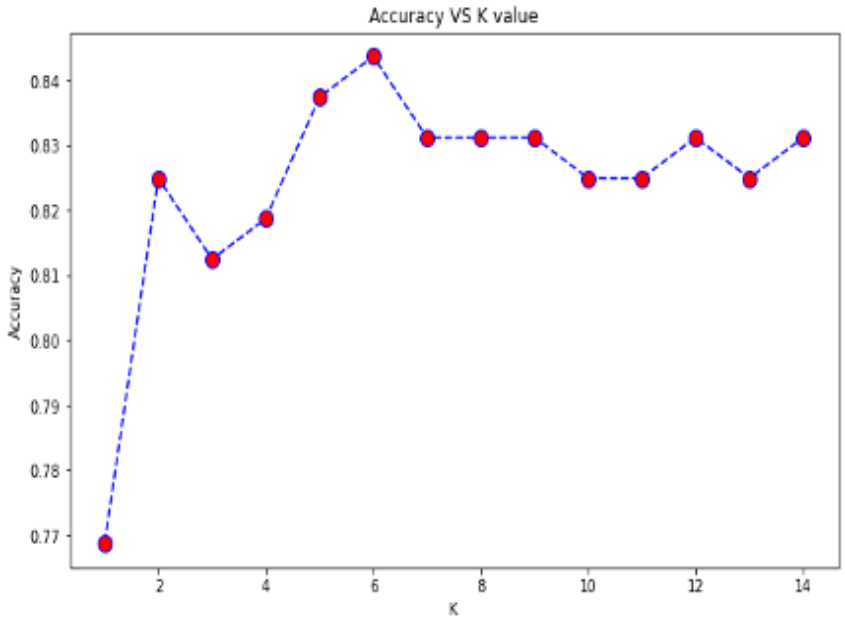
- Design and group questions based on the answer type: binary, numerical, categorical etc.
- NLTK toolkit to tokenize the answer and tagging each word.
- If there is an target word in the answer, then convert it into the correct format to be used as input to the model to make predictions.
- Error handling: If user's answer out of the restricted answer, ask the user answer this question again;

```
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
```

Experiment

K_NN Model and optimization

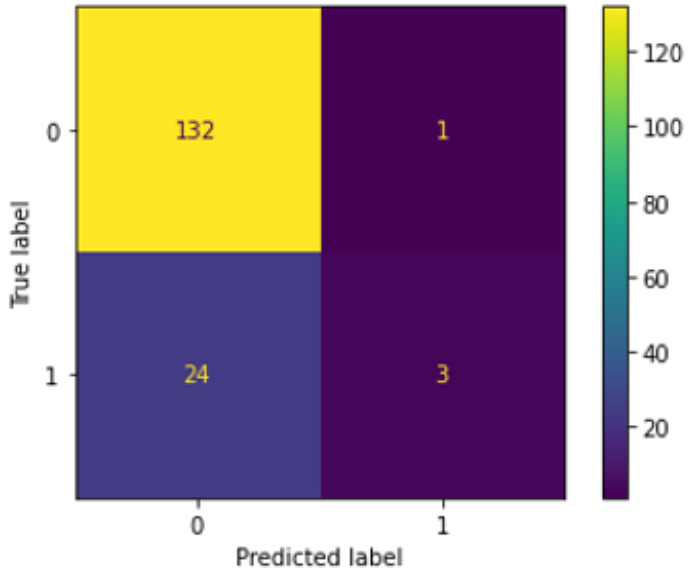
- KNN manipulates the training data and classifies the new test data using distance measures and then classifies the data based on the majority vote.
- No pre-defined statistical methods to find the best K.
- Randomly pick a K value and start the computing.
- Iterate over all possible K values.



Model Experiment

K_NN Model, Confusion matrix

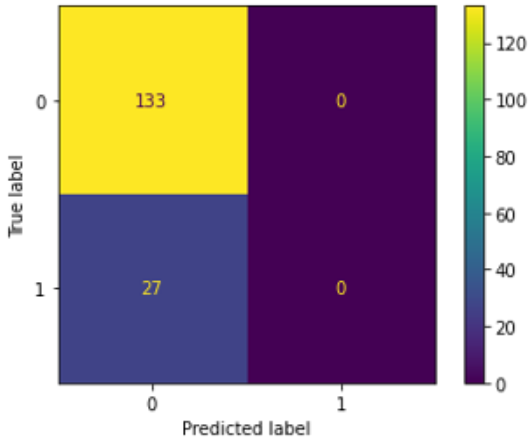
- Set the K = 6 and calculate the confusion metrics.



Model Experiment

SVC Model, Confusion matrix

- Support Vectors Classifier attempts to identify the optimal hyperplane for classifying data by optimising the distance between sample points and the hyperplane.
- C is the mistake term's penalty parameter. It regulates the trade-off between a smooth decision boundary and the proper classification of training points.
- Set Parameter range of C [0.1,1,10,1000]. The best C is 0.1 with an accuracy of 0.7875. And get the confusion matrix.



Model Evaluation

Evaluation Metrics

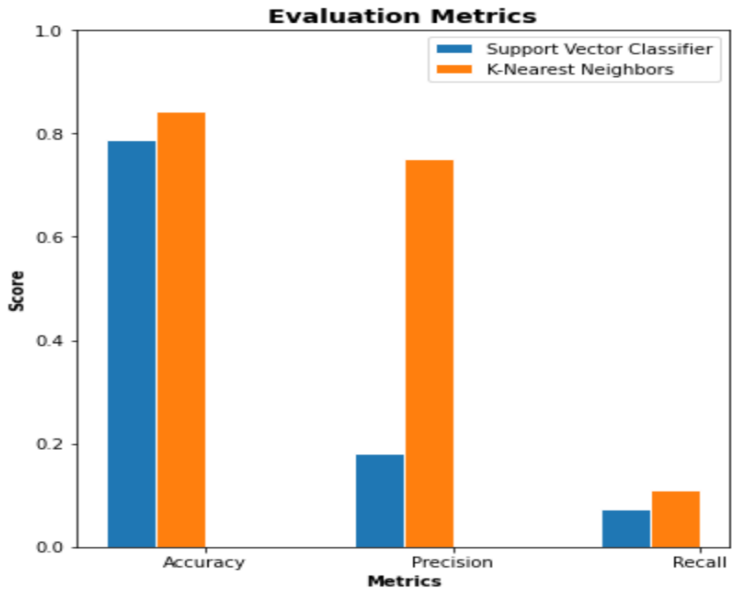
	SVM	KNN
Accuracy	0.7875	0.8438
Precision	0.1818	0.7500
Recall	0.0741	0.1111
AUC	0.4684	0.6331

- Accuracy is ratio of properly predicted observations to the total number of observations $Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$
- Precision is the ratio of correctly predicted positive observations to the total predicted positive observation

$$Precision = \frac{TP}{TP+FP}.$$

- Recall actually calculates how many of the Actual Positives that model capture through labeling it as Positive

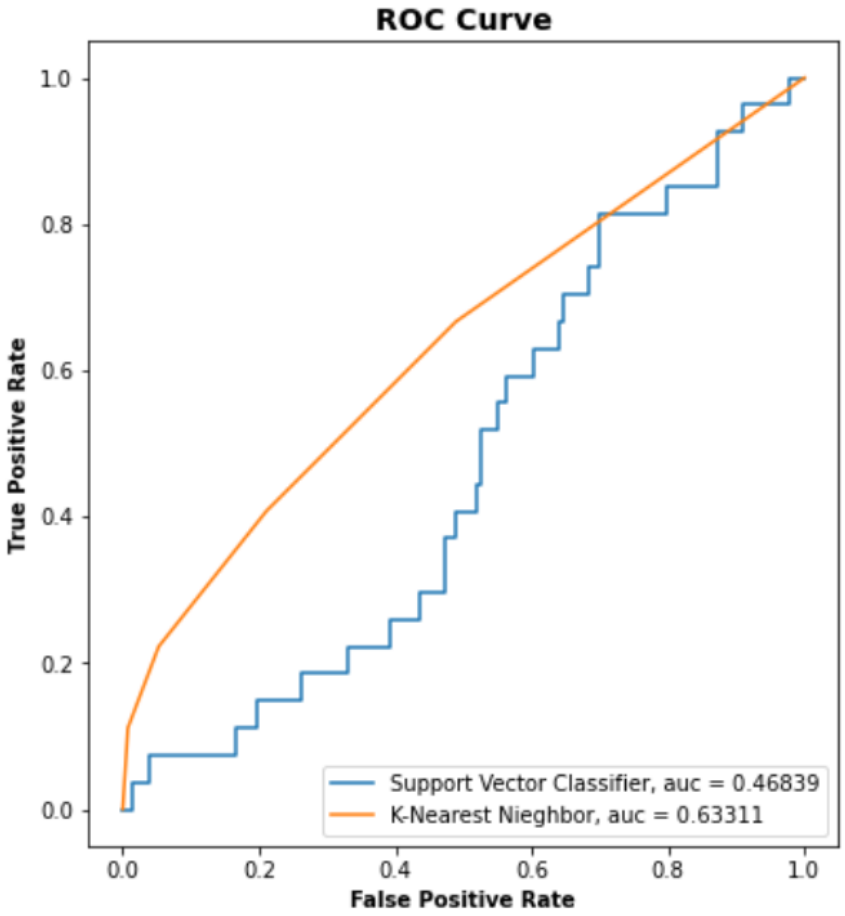
$$Recall = \frac{TP}{TP+FN}$$



Model Evaluation

ROC and AUC

- ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds
- AUC stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC.



Future work

Chat box

- More user-friendly
- More robust

More interactive between users and chat agent.

Reducing error sensitivity of chat box.

Prediction Model

- Minimising Data bias
- Improve accuracy rate

Select an almost equal number of positive and negative instances

More training samples

Reference

- [1] PYTLAK, K. 2022. *Personal Key Indicators of Heart Diseas.* kaggle.com/datasets/kamilpytlak/....
- [2] BAKER, C. 2022. NHS Key Statistics: England, February 2022.
- [3] Stephanie Stephensh.(March 2018). Worth the Wait? 15 Ways to Reduce Patient Wait Times. healthecareers.com/articles/healthcare-news...
- [4] PRIYANKA SHARMA.(2019). Heart Disease Prediction in Python. machinemantra.in....
- [5] Amnah khatun.(March 2018). Chatbot: the Next Big Thin. chatbotsmagazine.com/....
- [6] Tavish Srivastava. (March 2018). Introduction to k-Nearest Neighbors: A powerful Machine Learning Algorithm (with implementation in Python & R). analyticsvidhya.com/blog/....

Thanks!