Predicting diseases using the NHANES dataset

Name: Yuan Gao

ID: 300485853

Group 10: Rose de Anthony

Submission Date: Oct 26, 2021

Executive Summary

NHANES is the short name from The National Health and Nutrition Examination Survey and it is collected by the National Center for Health Statistics (NCHS), *which provides statistical information to guide actions and policies to improve the public health of the American people[1]*. Its source is on the website, https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx.[2]

In this project, I used the NHANES dataset to predict if a patient has Cardiovascular disease(CVD). This dataset can be split into five sections, which are Demographics, Dietary, Examination, Laboratory, and Questionnaire.

In each section, I applied the feature importance method from XGBoost to find the important features which can contribute to the risk of CVD. Then I extracted these features and merged the five sections by their unique IDs of the patients(each unique ID corresponds to one and only one patient). There were 20 important features and their definitions are given in '2.3 Attributes and meaning of variables in the Detailed Analysis'.

After merging, there are some missing values and especially one feature(DMDYRSUS — Length of time the participant has been in the US,) has too many missing values. I deleted this feature and also deleted any patients with any missing values. After that, I have 59214 patients and 19 features as the final dataset.

I split the dataset as training set and test set. The training dataset was used to learn the pattern by model. The testing dataset was used to evaluate the model, how good is the model, or if the model is robust when the model predicts the unseen data.

In the dataset, patients are not being diagnosed with CVD accounting for around 80% of the whole dataset, and being diagnosed with CVD accounting for around 20%. This is called imbalanced data. The model will more likely predict being diagnosed with CVD rather than not being diagnosed with CVD. Therefore, I used SMOTE method to fix this issue by adding 25615 patients being diagnosed with CVD to the training set.

I compared 14 models, such as Random Forest, Decision Tree, Lightgbm, to find the best performance model. Finally, the Random Forest model is a very robust model that has more accuracy than the other models to predict if a patient has CVD in this dataset. Therefore, I used the Random Forest model to predict CVD.

I did not do any tuning of the hyper-parameters, because it is not necessary in this case. The model has very good performance for predicting in the unseen dataset. The AUC score(range from 0 to 1, the higher, the better) is 0.98 and PR AUC(range from 0 to 1, the higher, the better) is 0.98. It means my model has around a 98% accuracy rate to predict if a patient has CVD, which is a very robust and good performance model.

There is no Privacy, Security issues. But there are some potential ethics issues about the dataset. Because the data are from US people if some people apply this data to the people who are not US people, which might have an ethical issue. Also, it has an ethical issue due to insufficient data about some small amount race.

1. Background

Cardiovascular diseases (CVDs) are the leading cause of death globally. An estimated 17.9 million people died from CVDs in 2019, representing 32% of all global deaths. Of these deaths, 85% were due to heart attack and stroke. It is important to detect cardiovascular disease as early as possible so that management with counselling and medicines can begin.[3] As the huge amount of data growing, the field of the data science provides a great convenience method to detect cardiovascular disease.

The aim of this project is to use NHANES data to predict if people have the cardiovascular disease.

2. Data Description

2.1 NHANES Data

NHANES is the short name from The National Health and Nutrition Examination Survey and it is collected by the National Center for Health Statistics (NCHS), *which provides statistical information to guide actions and policies to improve the public health of the American people[1]*. Its source is on the website, https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx.[2]

The data can be split into five sections, which are Demographics, Dietary, Examination, Laboratory, Questionnaire. The NHANES data was initiated in 1999 and collected information from 5000 people each year and is still ongoing. Because of the Covid-19, this program was suspended in field operations in March 2020, therefore, the 2019-2020 data was not completed.

The demographics section contains the demographics of the patients, gender, age, etc. The dietary section contains the results of questionnaires about it provides the samples' dietary patterns, weight, how much nutrition intake for instance. The laboratory and examination sections contain medical diagnostic information obtained by medical practitioners from the patients. The questionnaire section contains the results for many different questions about what medical conditions people have been diagnosed with.

NHANES data are used to answer a lot of questions about diseases in reality, such as diabetes, anemia, eye diseases, hypertension, and so on.

The data in NHANES includes numerical data and categorical data. I do not find any errors in the data.

When merging data, I used the primary key, SEQN.

2.2 Missing values (the white part is missing values)

There are 49693 rows and 33 columns in Demographics dataset, 15% missing values in total.

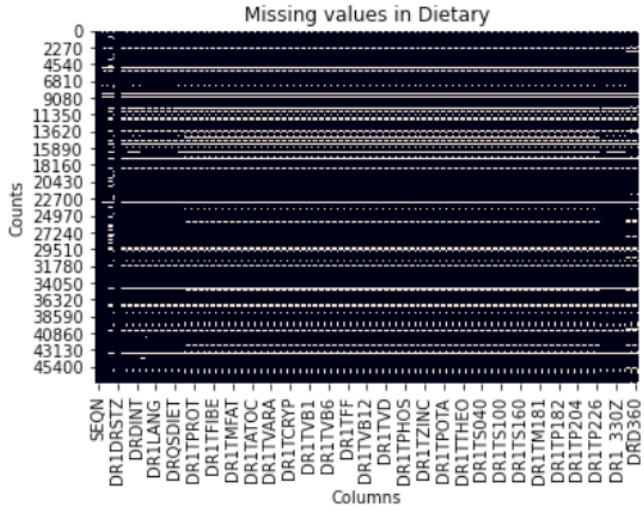


Figure 1

There are 47652 rows and 85 columns in Dietary dataset, 10.5% missing values in total.

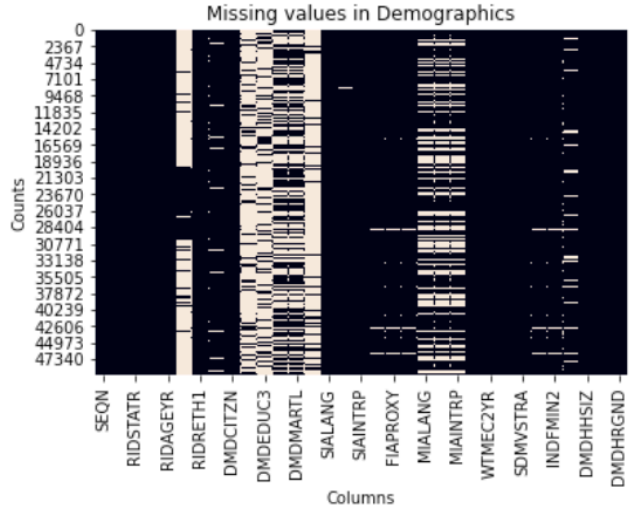
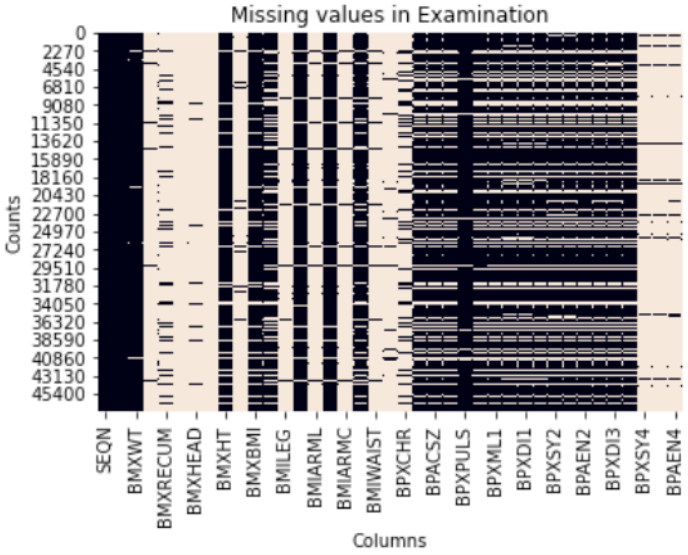
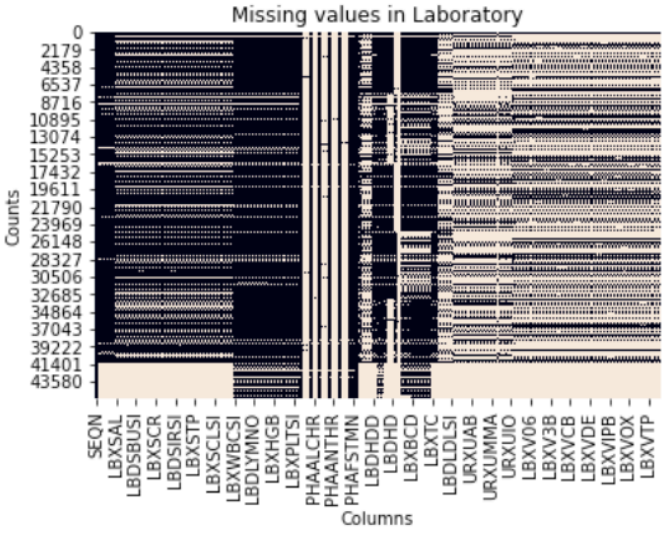


Figure 2



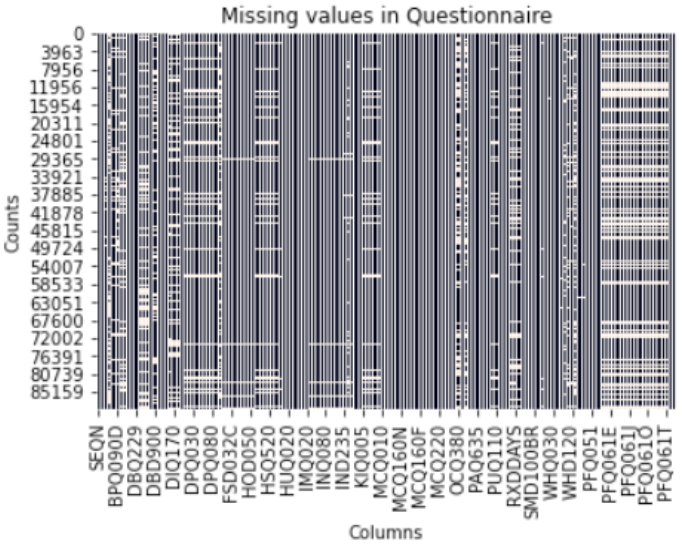
There are 47652 rows and 39 columns in Dietary dataset and 47% missing values in total.

Figure 3



There are 45744 rows and 172 columns in Laboratory dataset, 46% missing values in total.

Figure 4



There are 74989 rows and 153 columns, 9% missing values in total.

Figure 5

2.3 Attributes and meaning of variables in the Detailed Analysis

- DMDYRSUS(deleted after) - Length of time the participant has been in the US, numerical data
- RIAGENDER - Gender(1-male, 2-female), binary data
- RIDAGEYR - Age in years at screening, numerical data
- INDHHIN2 - Annual household income, numerical data
- DBQ095Z - Type of table salt used, categorical data
- DRQSPREP - Salt used in preparation, categorical data
- DRD340 - Shellfish eaten during past 30 days(1-Yes, 2-No). binary data
- DRD360 - Fish eaten during past 30 days(1-Yes, 2-No), binary data
- DRQSDIET - On special diet?(1-Yes, 2-No), binary data
- BPACSZ - cuff size(cm) — cuff width * cuff length , categorical data,
- BPXML1 - blood pressure level - maximum inflation levels(mm Hg), numerical data
- BMDSTATS - Body Measures Component Status code — analysts with a quick method of identifying survey participants with complete or partial body measurement data., categorical data
- BPXPLUS - Pluse regular or irregular, binary data
- LBXGH - Glycohemoglobin (%), numerical data
- LBDMONO - Monocyte number (1000 cells/uL), numerical data
- LBDBANO - Basophils number (1000 cells/uL), numerical data
- LBXHBS - Hepatitis B Surface Antibody, numerical data
- LBXHA - Hepatitis A Antibody, numerical data
- OCD150 - Type of work done last week, categorical data
- RXDDRGID - What generic drug were taken, categorical data

3. Ethics, Privacy and Security

There are no Privacy and Security issues.

However, there are some potential ethical issues. Firstly, because the data are from US people if some people apply this data to the people who are not US people, which might have an ethical issue.

Secondly, because there are different numbers in each race, therefore, it has an ethical issue due to insufficient data about other Hispanic, shown in Figure 6 below.

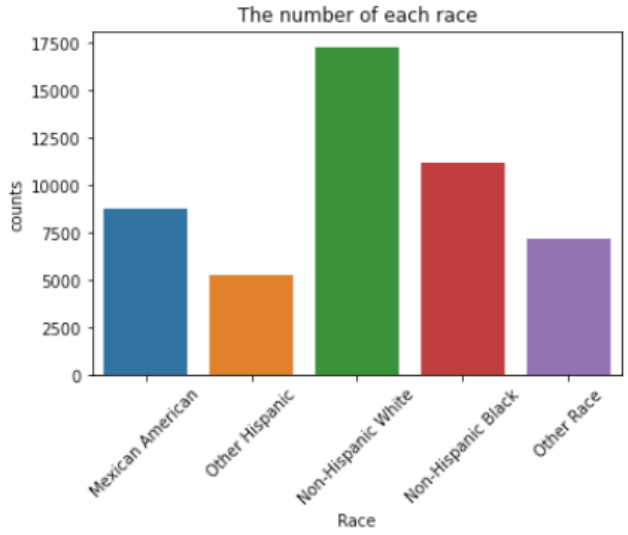


Figure 6

4. Exploratory Data Analysis

4.1 Feature selection from dataset

I used feature importance from XGBoost (version: 1.4.2) to limit the features. *Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value, the more important the feature.[4]*

I chose the features of feature importance score which is bigger than 0.05 and also the features from EDA. After selection, I kept 20 features(Table 1).

After the feature importance, I extracted these features from five sections and then used inner join to merge them by their unique keys ('SEQN'), each "SEQN" corresponds to one and only one patient.

4.2 Handle Missing Values after feature selection

Figure 7 presents the missing values after merging. The light part is the missing value.

There are two types of missing values. The first one is the patient did not answer the related question or did not do the related test. For example, if asked the patient if this

patient drank any alcohol before, this patient did not answer anything and left this question empty. This is the first type of missing value.

The second type of missing value is when they answer the questionnaire but fill out ‘refused’ or ‘do not know’. I treated these as missing values.

I deleted the DMDYRSUS column because it has so many missing values shown on the Figure 7. I also dropped any rows if that row has any missing values. After I dropped the missing values, there are 59214 rows and 20 columns in total. It turns out I deleted 12977 patients.

Features	The score of feature importance
DMDYRSUS	0.85
RIAGENDER	0.75
INDHHIN2	0.65
RIDAGEYR	0.62
LBDMONO	0.55
BPACSZ	0.24
DBQ095Z	0.162
BPXML1	0.14
BMDSTATS	0.1
LBXHBS	0.095
OCD150	0.095
DRQSPREP	0.09
DRD360	0.078
BPXPLUS	0.07
LBDBANO	0.07
RXDDRGID	0.07
LBXHA	0.065
DRD340	0.06
DRQSDIET	0.05
LBXGH	0.05

Table 1

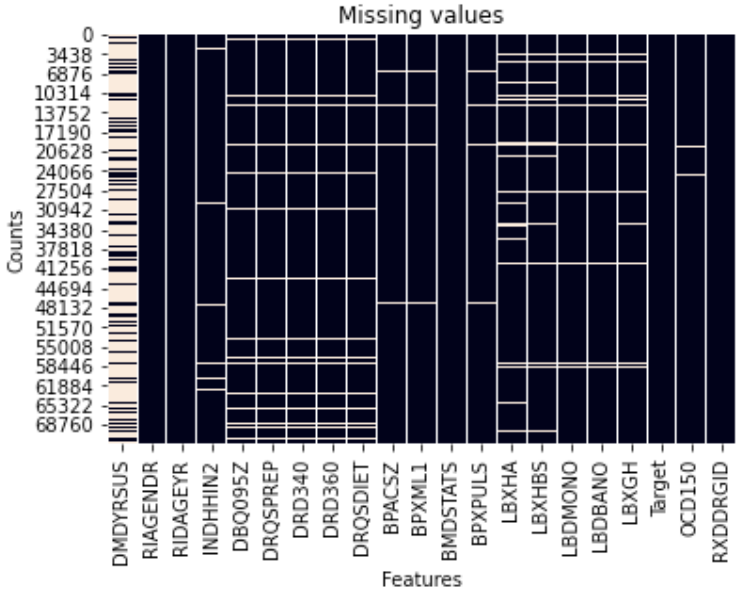


Figure 7

4.3 The proportion of if being diagnosed with CVD

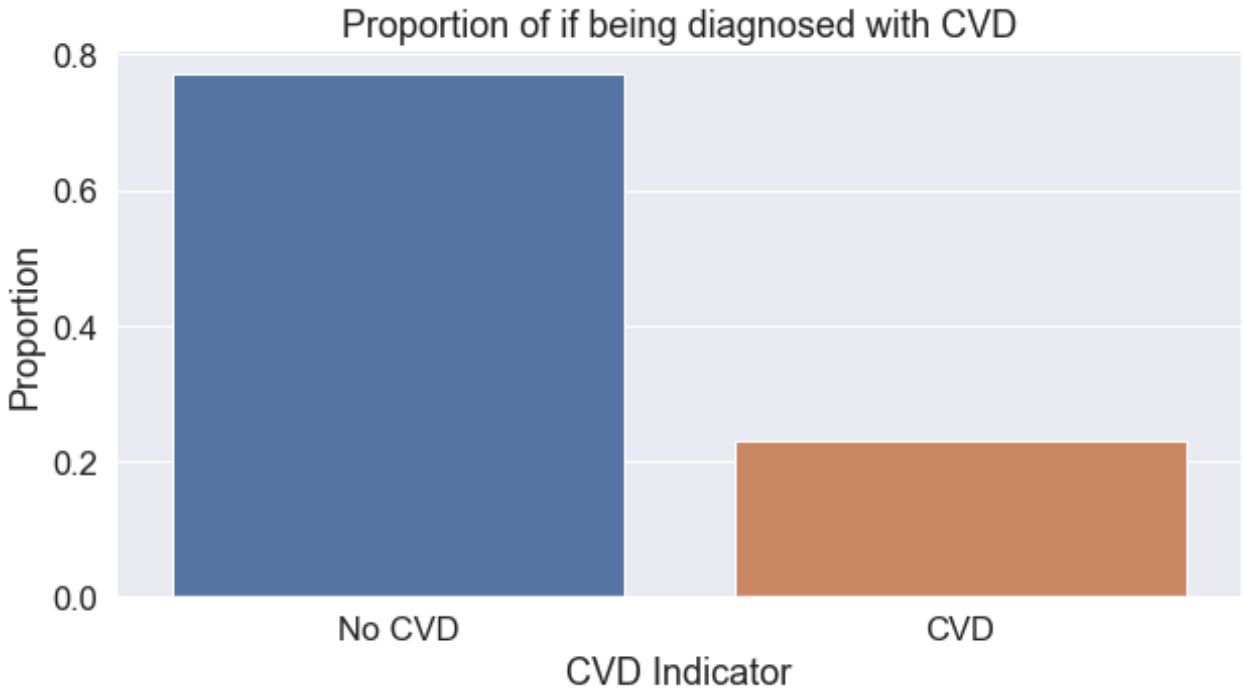


Figure 8

From figure 8, not being diagnosed with CVD is accounting for 77% of the whole dataset, and being diagnosed with CVD is accounting for 23% of the whole dataset.

Because there is an unequal distribution of Target in the dataset, the dataset is imbalanced. Because the model learns more from the majority class, therefore, the model is likely to classify the class as the majority class rather than the minority class.

4.4 Is there a correlation between gender and CVD?

From Figure 9, the proportion of male being diagnosed with CVD is more than that of the female being diagnosed with CVD. The proportion of males not being diagnosed with CVD is less than that of females not being diagnosed with CVD

I ran a chi-square test to check if there is a significant relationship between gender and CVD. The chi-square statistic = 678, p-value = $1.5^{*}\ 10^{-149}$. We have sufficient evidence to reject H0, which means there is a statistically significant relationship between gender and CVD.

4.5 Is there a correlation between age and CVD?

From figure 10, as the age increases, the proportion of people who have CVD increases. At the age of 78 to 80, the proportion of people who have CVD significantly increases. And these people account for half of people who do not have CVD in the same age range.

Then, I applied a chi-square test to check if there is a significant relationship between age and CVD. I cut off the age into 4 groups, 20-35, 35-50, 50-65, 65-80.

The chi-square statistic is 5857 and p-value $< 10^{-149}$. We have sufficient evidence to reject H0, which means there is a statistically significant relationship between age and CVD.

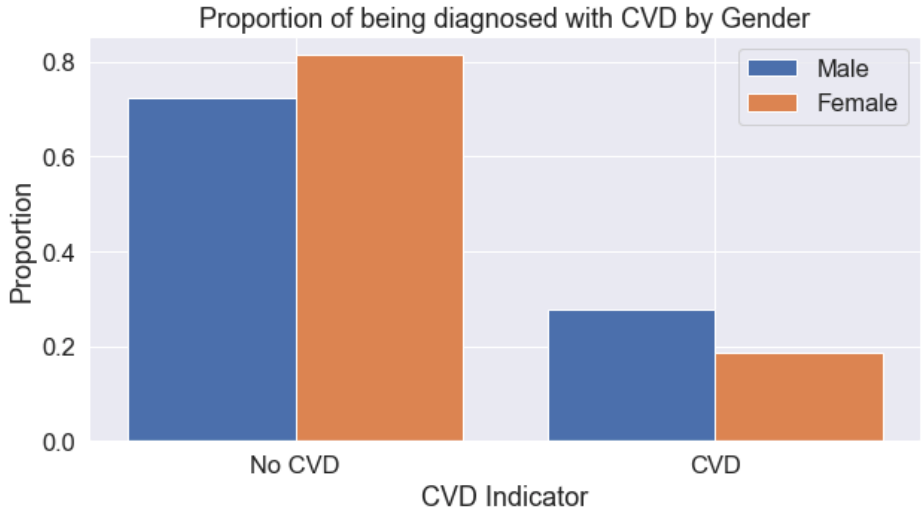


Figure 9

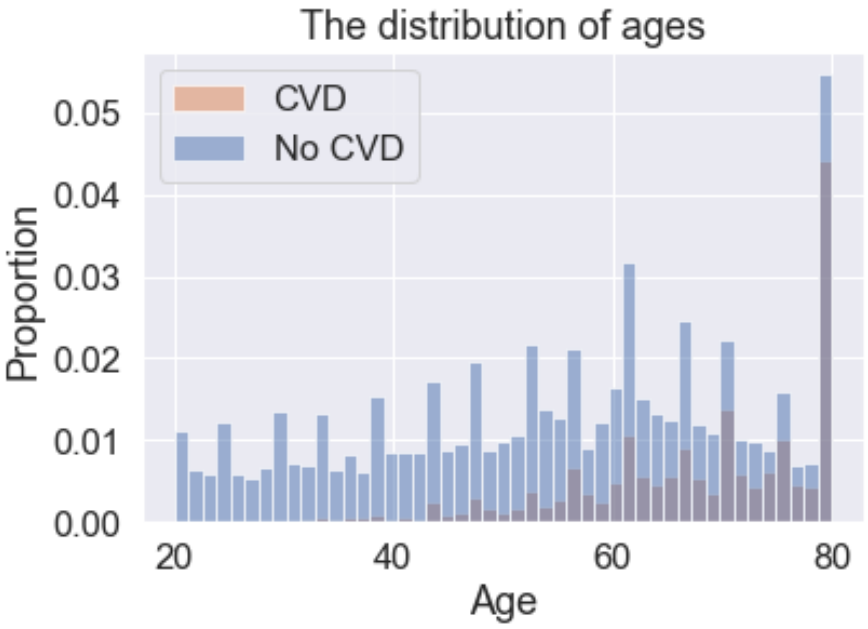


Figure 10

5. Detailed Analysis Results

5.1 Libraries

The library I used is PyCaret. The version is 2.3.3. *PyCaret is an open-source, low-code machine learning library in Python that automates machine learning workflows. PyCaret is essentially a Python wrapper around several machine learning libraries and frameworks such as Scikit-learn, XGBoost, LightGBM, CatBoost, spaCy, Optuna, Hyperopt, Ray, and many more. [5]*

In this project, I used compare_models from PyCaret, which can train the data from around 14 models and find the best performance model.

5.2 ML Modelling

The datasets were split as training and testing datasets using 80/20 train/test split.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	PR_AUC	TT (Sec)
rf	Random Forest Classifier	0.9556	0.9807	0.8341	0.9680	0.8960	0.8680	0.8720	0.9642	19.5220
et	Extra Trees Classifier	0.9346	0.9710	0.7719	0.9308	0.8439	0.8030	0.8086	0.9348	31.1300
knn	K Neighbors Classifier	0.8894	0.9618	0.9396	0.6902	0.7958	0.7224	0.7386	0.8800	16.5550
dt	Decision Tree Classifier	0.8995	0.8671	0.8071	0.7671	0.7864	0.7208	0.7213	0.6634	6.7830
lightgbm	Light Gradient Boosting Machine	0.8237	0.8549	0.5291	0.6397	0.5791	0.4689	0.4723	0.6514	5.9500

Figure 11

The training dataset was used to learn the pattern by the model. The testing dataset was used to evaluate the model, how good is the model, or if the model is robust when the model meets the unseen data.

Due to the imbalanced dataset, SMOTE was used in the training dataset to produce a balanced dataset. Before doing SMOTE, I have 47371 patients in the training set, which not being diagnosed with CVD accounted for 77%, and being diagnosed with CVD accounted for 23%. After doing SMOTE, I have 72986 patients in the training set, which not being diagnosed with CVD accounting for 50% and being diagnosed with CVD accounting for 50%. By SMOTE, I synthesized 25615 patients being diagnosed with CVD to the training set, increased half of the original training set.

The performance metric I used was AUC. *The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes. [6]*

From Figure 11, each row presents the performance of the model in each performance metrics. The table presents the top 5 AUC of 14 models. The best performance model is Random Forest. It also has the best performance in all metrics except Recall. Therefore, I used Random Forest as my final model.

5.3 Random forest

Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result.[7]. A decision tree consists of three components: decision nodes, leaf nodes, and a root node. A decision tree algorithm divides a training dataset into branches, which further segregate into other branches. This sequence continues until a leaf node is attained. The leaf node cannot be segregated further.[8]

5.4 ROC AUC plot and PR AUC plot

From the plots below, we can see the average AUC is 0.98(Figure 12) and Average precision is 0.97(Figure 13) for my model. The training model performs very well.

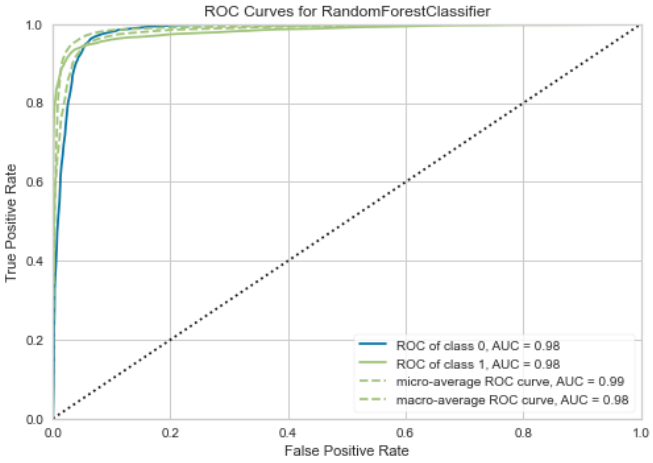


Figure 12

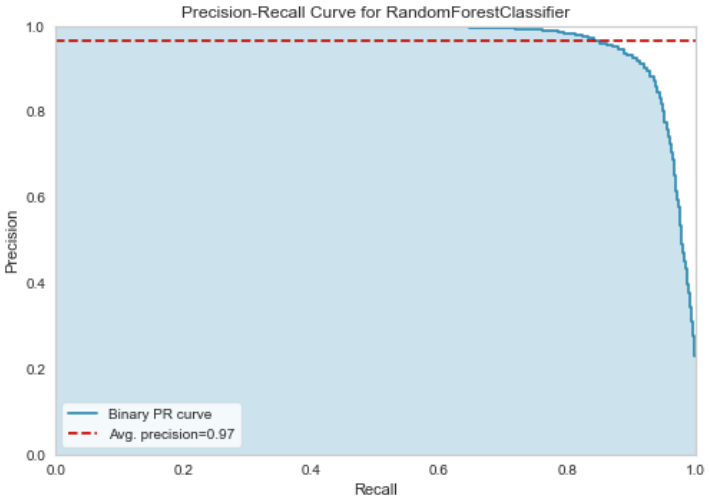


Figure 13

5.5 Parameter of Random forest

I used the default parameters. It is shown below (Figure 14). The reason I only used the default parameter is this model has already had very high scores in AUC and PR AUC. It also performs robust on unseen data (shown on "Predict on unseen data"). And the hyper-parameter tuning is slow and did not change much.

```
RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                       criterion='gini', max_depth=None, max_features='auto',
                       max_leaf_nodes=None, max_samples=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=100,
                       n_jobs=-1, oob_score=False, random_state=5636, verbose=0,
                       warm_start=False)
```

Figure 14

5.6 Uncertainty

Figure 15 presents a table with 10-fold cross validated performance metrics along with the trained model object.

For the trained model (Random Forest), the 95% confidence interval of AUC score is from 0.97 to 0.99. The 95% confidence interval of PR-AUC score is from 0.96 to 0.97.

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	PR_AUC
0	0.9566	0.9833	0.8434	0.9625	0.8990	0.8715	0.8746	0.9687
1	0.9608	0.9776	0.8513	0.9744	0.9087	0.8839	0.8872	0.9629
2	0.9557	0.9797	0.8303	0.9723	0.8957	0.8677	0.8721	0.9652
3	0.9484	0.9789	0.8158	0.9524	0.8788	0.8463	0.8504	0.9550
4	0.9551	0.9804	0.8303	0.9693	0.8944	0.8661	0.8703	0.9630
5	0.9593	0.9850	0.8487	0.9699	0.9053	0.8795	0.8827	0.9695
6	0.9587	0.9812	0.8436	0.9727	0.9036	0.8775	0.8811	0.9661
7	0.9527	0.9738	0.8160	0.9734	0.8878	0.8581	0.8634	0.9578
8	0.9551	0.9822	0.8318	0.9679	0.8947	0.8663	0.8704	0.9646
9	0.9541	0.9853	0.8303	0.9648	0.8925	0.8636	0.8675	0.9696
Mean	0.9556	0.9807	0.8341	0.9680	0.8960	0.8680	0.8720	0.9642
SD	0.0034	0.0033	0.0118	0.0063	0.0083	0.0104	0.0100	0.0046

Figure 15

5.7 Predict on unseen data

The AUC score on unseen data is 0.99 and PR AUC on unseen data is 0.98.

5.8 Bias

There is a potential bias in the result. I synthesized the 35% data in the training set, The more data synthesised, the lower the chances of the data being representative of a wider population, and the less useful the algorithm is likely to be for making predictions about new subjects that weren't in the original dataset. Although it has a good performance in unseen data split from the original dataset, there is a potential bias in the new subject.

6. Conclusions and Recommendations

I used 19 features to predict Cardiovascular disease by Random Forest model. The model has similar AUC and PR scores from in-sample and out-sample, which means the model is very robust.

The limitation of this project is the data is based on US people rather than all people in the world. Therefore, if we use this model to predict people who are not from the US, the prediction might have a bias or be less accurate.

In the light of feature importance and EDA, gender has an effect on having CVD. Male is more likely to be diagnosed with CVD than female.

As shown in my analysis, my model is quite reliable in the detection of CVD in patients expect the potential bias. We can apply this model to the website, a questionnaire to detect if having CVD. But some features from our data are from the lab dataset and exam dataset, which is very hard to be known by patients. Therefore, for further work, I might delete the hard-finding features and only use easy-finding features to predict the disease in a reasonable range of accuracy.

Reference

[1] https://en.wikipedia.org/wiki/National_Center_for_Health_Statistics

[2] https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx

[3] 11 June 2021, https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[4] May 12, 2018, Stacey Ronaghan, https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3

[5] https://github.com/pycaret/pycaret

[6] June 16, 2020, Aniruddha Bhandari, https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/

[7] July 22, 2021, Niklas Donges, https://builtin.com/data-science/random-forest-algorithm?__cf_chl_captcha_tk__=pmd_wFbzZlX9npXaJG_BTV8cDpUt__AoG7gsXMsBvF3YEo8-1635067802-0-gqNtZGzNA1CjcnBszQnl

[8] Dec 11,2020, Onesmus Mbaabu, https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/