# **COGS 118A- Project Checkpoint**

Project code is in another file named [Cogs\_118A\_Final\_Project\_Code.ipynb](https://github.com/COGS118A/Group045-Sp22/blob/main/Cogs_118A_Final_Project_Code.ipynb) under same folder

# **Names**

* Bohan Lei
* Zheran Li
* Ryan Choi
* Yiyao Liu
* Duye Liu

# **Abstract**

Our goal in this project is to build prediction models that predict whether or not someone has heart disease based on their health status and ethnographic data. We will clean the 2020 CDC survey data of 400k adults and choose the features from HeartDisease, BMI, Smoking, AlcoholDrinking, Stroke, PhysicalHealth, DiffWalking, Sex, AgeCategory, Race, Diabetic, PhysicalActivity, GenHealth, SleepTime, Asthma, KidneyDisease, and SkinCancer. We will train models including KNN, Decision tree, decision boundaries, and compare their performance primarily indicated by their classification accuracy. For each model, the optimal hyperparameter such as the k value in KNN and maximum depth of the decision tree was found using GridSearch.

# **Background**

One study that is similar to our project objective is using a machine learning model for predicting out-of-hospital cardiac arrests based on meteorological and chronological data. In this study, the researches used eXtreme Gradient Boosting algorithm to generate a model that predict daily Out-of-hospital cardiac arrest based on OHCA nationwide registry and high-resolution meteorological and chronological datasets from Japan[[1]](#footnote-0). Their result have shown that combining meteorological and chronological variables in the machine learning model have the best predictive accuracy in both the training and testing datasets[[2]](#footnote-1). Their research outcome have also indicated that factors like holiday, weekend, low ambient temperature, and large interday or intraday temperature difference are strongly associated out-of-hospital cardiac arrest[[3]](#footnote-2).

Similarly, a recent study done by the Cedars-Sinai Artificial Intelligence in Medicine division[[4]](#footnote-3) found that by combining CTA and F-NaF PET , two advanced imaging techniques together, the team was able to develop machine learning models that improve the prediction of heart attacks. The team took data from 293 patients from the ages of 56 - 74 years in the span of 53 month, and created three models to predict the possibility of future heart attacks[[5]](#footnote-4). By comparing with the other two models they developed based on baseline characteristics and the quantitative plaque analysis variables from CTA, they found that the model that combines CTA and F-NaF PET produce the most accurate result[[6]](#footnote-5).

# **Problem Statement**

According to the CDC and WHO, heart disease is the most common cause of death in the world, responsible for 16% of total deaths. There are some common key risk factors that could increase one’s chance of getting a heart disease including, but not limited to, high blood pressure, high cholesterol, physical inactivity, and smoking. Understanding what key risk factors lead to heart disease and avoiding them will make it much easier to prevent getting a heart disease. The original data given by kaggle contains 401,958 rows and 279 columns where each column corresponds to different factors that could lead to heart attack (questions on survey) and each row corresponds to values (answers for questions). Using this dataset, one could apply machine learning algorithms such as Decision Trees and KNN to detect patterns within the data, which will help us to understand which factors have the least/greatest impact on heart disease. Then, doctors could advise people who have a high chance of getting a heart attack (e.g. family history) what actions to take to lower their chance of getting a heart disease.

# **Data**

Please check our code for the data cleaning process.

In the data cleaning section we have changed the data into machine learning friendly form. We changed all boolean values into binary 0/1, and also changed the range of age into numerical values. After that we looked at the correlation between HeartDiease( our y column) and other features in order to reduce the dimension and complexity of our model.

In this project, we use the Personal Key Indicators of Heart Disease Dataset. The link to the dataset is below:

<https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease>

This dataset is based on the 2020 annual Center of Disease Control and Prevention (CDC) survey of around 400k adults related to their health status. The dataset has in total 18 variables and 320,000 observations. Each observation consists of the results of the 18 variables. The variables are: HeartDisease, BMI, Smoking, AlcoholDrinking, Stroke, PhysicalHealth, DiffWalking, Sex, AgeCategory, Race, Diabetic, PhysicalActivity, GenHealth, SleepTime, Asthma, KidneyDisease, and SkinCancer. Since the goal of our project is to find the factors that lead to heart disease, this data sets will best support the need of our project. We can use correlation map to see which factor (such as “Smoking”,”alcholDrinking”, or “stroke”) have the highest correlation to heartdiesase. Then we can used the data set to apply to machine learning algorithms to detect patterns within the data.

Since this dataset is clean enough, with no missing data, we don’t need to further do any data cleaning process. The most important variable is the HeartDisease variable, as that is the variable that we are going to predict by our model. We will use 1 to represent the result of “yes“， and 0 to “no”. This transformation will also be applied into other variables that contain only “yes” and “no”. This includes Smoking, AlcoholDrinking, Stroke, DiffWalking, Diabetic, PhysicalActivity, Asthma, KidneyDisease, and SkinCancer. Also, for sex, we denote 1 as “male” and 0 as “female”. For AgeCategory, since it will be hard for a model to predict a category, we will just use the starting age of the age group. For example, for category “75-79”, we will just replace it as 75. For race and general health, since there are too many categories, we will either use small numbers to represent each category, or use one-hot coding to transfer it into the numbers with 1 and 0.

# **Proposed Solution**

We plan to implement and compare the performance of different machine learning algorithms. Since this is a classification problem with binary data and we have many times more observations than variables, we make use of decision trees(random forest) algorithms, decision boundaries, and K Nearest Neighbor (KNN) on our cleaned data. An ideal model will have a higher chance of identifying people with high risk of heart disease from those who actually have higher risk. Then those individuals can receive a warning and seek medical help.

Since we have over ten columns to consider we might need to reduce the size of data. We will first check the relations between each variable, and try to delete the most unrelated variables. If it still cannot reduce many columns, we will use Principal Component Analysis (PCA) to reduce dimensions. Our solution should be tested by either cross validation or just apply it onto our test dataset, since we have a very large dataset. At last we will try to compare our results with the existing methods in papers or other individual’s posted solutions on Kaggle to further validate the result.

# **Evaluation Metrics**

We consider classification accuracy to be the most important performance measure of our model because it gives a straightforward measurement of how many correct predictions the model made out of all predictions. However, we also notice that heart disease is a rare situation, so our data is imbalanced. We will try to avoid the “accuracy paradox” and keep track of the model’s ability to identify people of high risk by also evaluating our model’s sensitivity, which is defined by the true positives out of all positives. Both accuracy and sensitivity are common evaluation metrics and we will get a sense of which model is better using these two scores.

# **Preliminary results**

Firstly our code is included in another .ipynb file as mentioned at the beginning of this document.

When we were fitting our data we used grid search to determine the best tree depth and generated a plot indicating the mean score across different tree depths. Then we trained our decision tree algorithm with the best parameter and got a score of 0.72, which we think could be worked on more. And also we tried KNN classifier with different k values and we are going to look into the result in the following weeks. For the details pelase check on our code

# **Ethics & Privacy**

The ethics problem is divided into 5 phases: data collection phase, data storage phase, data analysis phase, data modeling phase, and deployment phase.

For the data collection phase, since this dataset was collected by CDC and published for public use, we can assume the collecting agency was held to official standards. We can assume that it could represent the population in the US, and it does not have a strong collecting bias.

For the data storage phase, first of all, since this is a public data collection, we do not need to consider the data privacy problem. Also, since this is the data collected in 2020, we also do not need to be concerned about the problem of the updated data, since the update was already fixed.

For the data analysis phase, the data itself could lead to some kind of biases against a specific group of people. We will try our best to reduce such biases by doing visualizations and observations to supervise the learning process.

For the data modeling phase, the models we generate might also have prediction bias against specific groups of data. We will make use of the best unbiased model for our prediction to suppress such bias.

For the deployment phase, some of our models might be used for future prediction which generate prediction bias. We will address this problem at the end of our model build.

# **Team Expectations**

* Be on time for meetings
* Respond to messages on Discord
* Finish assigned tasks on time, seek help if needed
* Divide tasks evenly
* Ask on the group chat before any changes are made

# **Project Timeline Proposal**

We will basically meet every week on Thursday night, but the meeting time is flexible if anyone has timeline problems. We will basically first finish the project proposal, and start to wrangle and explore the dataset in our notebook. After getting feedback from the proposal, we will start to implement algorithms and models into the dataset, and compare the results of them. After that, we will start to write the final report. The specific timeline is shown below.

| **Meeting Date** | **Meeting Time** | **Completed Before Meeting** | **Discuss at Meeting** |
| --- | --- | --- | --- |
| 4/18 | 8 PM | Brainstorm topics/questions | Determine best form of communication; Discuss and decide on final project topic; discuss hypothesis; begin background research |
| 4/23 | 5 PM | Finish the proposal by part | Final check the project proposal |
| 4/28 | 8 PM | Explore the dataset and find the possibility to wrangle the data | Start planning the models will be used in the project, working on data cleaning individually |
| 5/5 | 8 PM | Finalizing the models will be used in the project and start working on them | Each member picks a part for the implementation. One person work on finaling the data cleaning, two person works on Decision trees, another two person work on KNN. |
| 5/12 | 8 PM | Finish implementing models and data cleaning | Compare the results of accuracy of each model, discuss the problems we met, and start to write the final project report |
| 5/19 | 8 PM | Finish fixing the possible error in the code of the model. | Discuss the error of the models, and how to improve the accuracy score. And use grid search to set hyperparameters. |
| 5/26 | 8PM | Finish and finalizing decision tree, KNN and decision boundary. | Start working on writing the report |
| 6/2 | 8PM | Finish writing the report | Final check with TA |
| 6/8 | Before 11:59 PM | NA | Turn in Final Project |

# **Footnotes**

1. Machine learning model for predicting out-of-hospital cardiac arrests using meteorological and chronological data

<https://heart.bmj.com/content/107/13/1084>

# Machine learning and advanced imaging improve prediction of heart attacks

<https://physicsworld.com/a/machine-learning-and-advanced-imaging-improve-prediction-of-heart-attacks/>

1. <https://heart.bmj.com/content/107/13/1084> [↑](#footnote-ref-0)
2. <https://heart.bmj.com/content/107/13/1084> [↑](#footnote-ref-1)
3. <https://heart.bmj.com/content/107/13/1084> [↑](#footnote-ref-2)
4. <https://physicsworld.com/a/machine-learning-and-advanced-imaging-improve-prediction-of-heart-attacks/> [↑](#footnote-ref-3)
5. <https://physicsworld.com/a/machine-learning-and-advanced-imaging-improve-prediction-of-heart-attacks/> [↑](#footnote-ref-4)
6. <https://physicsworld.com/a/machine-learning-and-advanced-imaging-improve-prediction-of-heart-attacks/> [↑](#footnote-ref-5)