**Project Proposal**

Team 17

Team members:

Leshan Zhao (lzz0064@auburn.edu), Maram Aldiabat (mza0200@auburn.edu), Noah Heckenlively (nah0039@auburn.edu); Leshan zhao will be coordinator and primary contact.

## 1. Introduction

- Background

  Nowadays, Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Identifying those at highest risk of CVDs and ensuring they receive appropriate treatment can prevent premature deaths. A large proportion of patients died because they receive treatment too late. So, if we can find out CVD patients earlier and ensure that they receive treatment in time, we can prevent many deaths.

- Problem to work on

  However, cardiac tests, such as cardiac CT or MRI, are usually expensive, which may not be affordable for some families. And even if the governments afford their costs, it could be quite a big amount of financial expenditure.

## 2. Solution

- Build a MLP model: how

To solve this problem, we plan to build a MLP model in **python**, which that will take the physical indexes and potential behavioural risk factors as input, and then predict whether a people is at high risk of CVDs and reveal behavioural factors that may increase the risk of heart diseases.

- Model input and output

  To be more specific, the input of our MLP model will be the physical indexes and potential behavioural risk factors such as whether a peoples smokes, and the output will be 0 or 1, indicating whether a people is at high risk of having cardiovascular diseases.

- Potential users and Significance:

  This classifier can be used as a risks screening tool for governments and hospitals, who can use the model to identify those citizens who are at highest risk of CVDs and suggest them to do the cardiac tests, so that both governments and patients could save a lot of money.

## 3. Methods

- Apply MLP to solve this problem: how does it helpful

  As we can see, we are facing a classification task, so our MLP model will be used to classify whether a people is at high risk of having cardiovascular diseases and find out those factors that may increase the risk of heart diseases. However, we can't ensure whether the features all have linear correlations with the risk. So, with MLP, we can overcome the limitation of linear models so that we may find out the more universal types of functional relations.

- Potential challenges

  When building the model, we may face challenges including: Noise in data and Data issues (nans, outliers, …) especially in Age, Weight, and Height information; size of dataset may be too small to cause overfiting; poor correlation of features with target which lead to low accuracy; generalization and overfitting, maybe the nature of dataset cause to some problem in the accuracy.

- Dataset to use:

  The Kaggle dataset is used in this project to see cardiovascular disease risk screening. This dataset contains 70k instances and 11 features that help in the prediction process like age, height, weight, Glucose, smoking, and whether the patient is an alcohol taker.

- How will we create/build our dataset?

  Since we're using a Kaggle dataset "cardiovascular disease", we won't need to build one.

- Hyperparameters in the model:

  Firstly, there is a need to define the specific hyperparameters to tune our model, like fine tune in cross-fold validation which helps to avoid overfitting by making the testing part 30% and 60% for training and 10% for validation part. Secondly, define the learning rate and the number of epochs of training. Finally, define the error rate to stop training our model when it reaches a specific level of error rate.

**4. Usefulness of classifier**

If we can achieve 80% f1 score, then this model could be used by various stakeholders to inform important decisions. Hospitals and governments could use this information in order to determine where a limited supply of tests may be distributed. Understanding that said tests cost money, insurance companies would be able to use the model in order to help them figure out what their costs on these tests will be so that they may determine how to price their insurance plans.

Also, if we are able to determine features that provide a lot of information gain to the model, an individual may be able to use that to help reduce health risks. For example, if exercise happens to have a significant negative correlation with cardiovascular disease, then a person can reduce risk by exercising more. Similarly, insurance companies could offer pricing deductions for people that try to minimize their own risk of cardiovascular disease, because that would save the insurance company money on paying for tests.

**5. Timeline**

We have constructed a rough timeline to decide tentatively when we will finish what goal:

| Date | Goal |
|---|---|
| Nov 10 | Project Proposal Report Due |
| Nov 10 | Finish data preprocessing (categorical one hot, deal with NaNs and outliers, feature selection, train-test-validation split) |
| Nov 13 | Implement MLP forward direction (no need for hyperparameters yet) |
| Nov 16 | Implement backward propagation (hyperparameters implemented) |
| Nov 18 | Evaluation/Validation (CFV) |
| Nov 29 | Final bug fixes and turn in project |