**Project Proposal**

Team 17

Leshan Zhao ([lzz0064@auburn.edu](mailto:lzz0064@auburn.edu)), Maram Aldiabat ([mza0200@auburn.edu](mailto:mza0200@auburn.edu)), Noah Heckenlively ([nah0039@auburn.edu](mailto:nah0039@auburn.edu))

1. Introduction

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.

…

1. Methods
2. Hyperparameters :

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.

1. Usefulness

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.

1. Dataset

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.

1. Timeline

|  |  |
| --- | --- |
| Date | Goal |
| Nov 10 | Project Proposal Report Due |
| Nov 10 | 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 |