735 Project Proposal

Group 4 Members: Renee Ge, HyungGyu Min, Lingbo Zhou, Songchen Zhuo

Dataset Link: https://www.kaggle.com/datasets/simaanjali/diabetes-classification-dataset

Data Description:
The data set contains clinical data obtained from a cohort of 5,132 patients. Variables encompass critical demographic and physiological including age (Age), gender (Gender), body mass index (BMI), total cholesterol level (Chol), triglycerides level (TG), high-density lipoprotein level (HDL), low-density lipoprotein level (LDL), creatinine level (Cr), blood urea nitrogen level (BUN), and an indicator of having diabetes (Diagnosis). For the purposes of our project, age will be the primary response variable for predictive modeling.
This data set does not contain any missing value. Gender is represented by 'F' indicating female and 'M' representing male; however, one individual was recorded as 'f'. We believe this was an input error, and will categorize this individual as female for consistency.

Background:
The concept of 'biological age' reflects the extent to which an individual is impact by aging-driven biological changes and can capture physiological deterioration better than chronological age (your age using your date of birth). Finding markers of biological age would allow for more efficient implementation of health-related interventions. As such, we would like to build a model attempting to predict age using our diabetes dataset.

Our Study Aims:
We plan to build a statistical and a machine learning model and compare them for predicting age from cardiovascular and kidney function data, while recording the code in an R package. The results will provide insights into the relationship between age and health indicator.
We will first divide the data into training and testing sets. For the statistical model, a multiple linear regression model will be built from scratch to predict age, while for the machine learning model, a random forest model will be used for age prediction. We will then assess the predictive performance of the linear regression and random forest models on the test set using loss functions like mean squared error and compare the results.