Predicting Cardiovascular disease and its risk factor using machine learning algorithms

1. Problem:

Existing models for predicting cardiovascular disease (CVD) risk often rely on traditional risk factors such as age, gender, hypertension, and smoking status. However, these models may overlook the complex interplay between socioeconomic factors and environmental influences on CVD risk. Our research aims to develop a novel machine-learning model incorporating socio economic indicators, environmental factors, and traditional risk factors to predict CVD risk more accurately. By doing so, we aim to identify previously unrecognized risk factors and improve the effectiveness of preventive interventions.

2. Data:

We will use the Cardiovascular disease detection data. It contains 5111 rows of individual patients, which span 12 columns of vital data including ID for each patient, BMI, average glucose level, stroke, smoking status, history of hypertension and heart disease; as well as their sociodemographic status like age, gender, working type, residence type, marriage status. Sociodemographic variables are not medical statistics but researchers should be attentive to how these social and environmental factors affect cardiovascular disease outcomes. We utilized the one hot encoding to represent categorical variables as numerical values in our models, specifically variables of gender, marriage type, residence type, smoking status. These are encoded by 0 and 1, smoking status is further divided into formally smoked, never smoked, current smoking status to investigate how different levels of smoking affect the CVD risk.

3. Approach

We will start by cleaning the data and handling missing values. After the data is pre-processed we will be using the following machine learning methods.

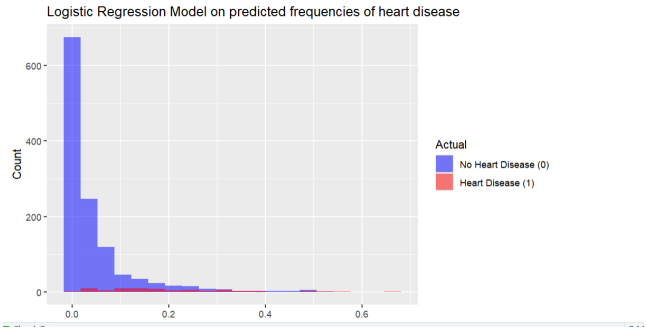


Figure 1
Logistic regression: Since a lot of variables are categorical, we decided to utilize logistic regression to predict the probability of the outcome. We have run two logistic regression models, discovering how some of the key indicators affect predicting and model performance, such as that age, glucose levels, gender, and smoking status. The model fits well, with significant coefficients showing variable importance in predicting heart disease risk.

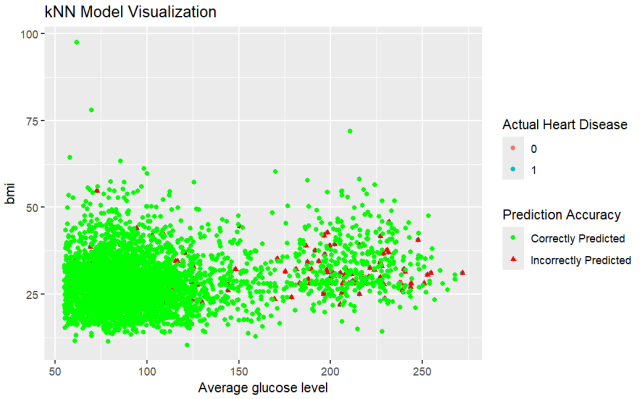
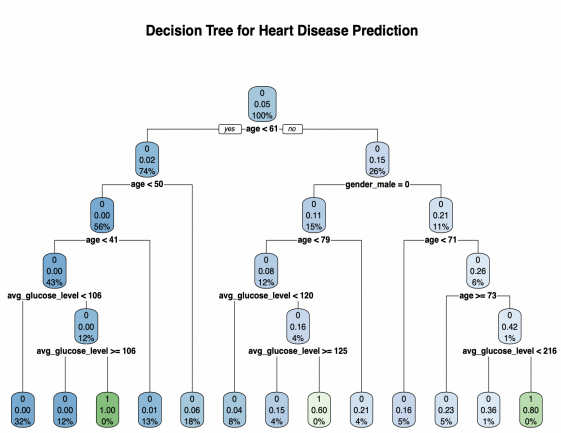


Figure 2
K-Nearest Neighbors (KNN): KNN method train a number of target variables within the dataset and predict the probability of heart disease by considering k nearest neighbor in the dataset. To find hyperparameters, we utilized the cross validation to find the optimal value of k that has the best performance (lowest error) on the unseen data (k = 9).This particular model that is being shown has represented both actual heart disease status and



prediction accuracy by the variables “average glucose level” and “bmi”.

Figure 3
Decision tree: The decision tree has components of nodes and leaves. Each leaf represents the probability of presence/absence of heart disease and the nodes are input parameters. This decision tree model primarily utilizes age and average glucose levels as its main criteria, highlighting age and glucose levels as critical factors in assessing heart disease risk.. Then younger patients are further classified by narrower age and glucose levels, whereas elderly patients are analyzed based on gender, further age categories & glucose levels.

Random forest: Random forest is an ensemble method that combines multiple decision trees to improve prediction accuracy. We can use the default number of trees (500) and maximum features (sqrt(n_features)).

Ensemble method: We utilized multiple models including logistic regression, kNN, random forest, decision tree to combine prediction together. The predictions of various models are stored in a matrix. For each row of the matrix, we calculated the mean of the predictions across all models. If the mean prediction is greater than 0.5, they are classified as positive for heart disease (1), otherwise, negative (0)
Our ensemble model exhibited limited effectiveness, achieving an accuracy of only around 5.4%. Challenges such as potential level mismatch between predictions and test data need to be addressed to improve the reliability of the ensemble approach. Further refinement and tuning are necessary to optimize the ensemble model for our task of heart disease prediction.

4. Evaluation:

We will evaluate our approach using a combination of traditional performance metrics (e.g. accuracy, precision, recall and F1-score) and advanced techniques such as Receiver Operating Characteristic (ROC-AUC) curve analysis. A comparison of different models presented in the table1, showcasing their performance metrics side by side.
Logistic regression shows high overall accuracy and sensitivity, with slightly lower AUC than kNN, which has perfect recall but zero specificity, indicating it fails to predict negative cases accurately. The Decision Tree and Random Forest models display balanced accuracy and precision but get zero specificity, similar to kNN.

Model	Accuracy	Sensitivity	Specificity	AUC	Precision	Recall	F1-Score
Logistic	0.9475	0.9992	0.0435	0.8896	0.9482	0.9992	0.9730
kNN	0.9460	1.0000	0.0000		0.9460	1.0000	0.9722
Decision Tree	0.9421	0.9959	0.0000	0.8473	0.9458	0.9959	0.9702

Random Forest	0.9460	0.9992	0.0145	0.8462	0.9467	0.9992	0.9722

Table 1 :Evaluation of test data

5. **What was completed, what was changed and why?**

   - Completed:
     We've effectively created and assessed machine learning algorithms for forecasting the risk of cardiovascular disease by integrating both conventional risk factors and socio-economic/environmental markers.

   - Changed:
     Originally, logistic regression and KNN were decided for modeling, but as the analysis progressed, decision trees and random forests were introduced, enhancing the model's effectiveness. This adaptation likely stemmed from the necessity to explore diverse modeling approaches to better grasp the intricacies of CVD risk prediction.

   - Why?:
     Incorporating decision trees and random forests facilitated a broader examination of the data and enhanced predictive accuracy. These models provide valuable insights into the significance of features and their interactions, thereby enriching our comprehension of the fundamental drivers behind CVD risk.

6. **Next steps: Based on what we've learned from the project, here are some suggestions for a student taking up a similar project.**

   - Feature Engineering: Explore additional feature engineering techniques to create new variables or transform existing ones. This could involve interaction terms, polynomial features, or domain-specific transformations that capture unique aspects of CVD risk.

   - Data Augmentation: Consider augmenting the dataset with additional socio-economic or environmental data sources to enrich the feature space. This could include data on air quality, access to healthcare, or socio-economic indices at the community level, providing deeper insights into the contextual factors influencing CVD risk.

   - Collaborative Ethical Framework: Collaborate with domain experts, such as cardiologists or public health researchers, to gain deeper insights into the clinical relevance of the predictive features and model outputs while ensuring ethical considerations. This collaborative approach not only refines the modeling approach but also ensures that the models are fair, transparent, and sensitive to potential biases, especially regarding the interpretation and utilization of socio-economic variables. By integrating expertise from diverse domains and prioritizing ethical guidelines, the research can uphold patient privacy, equity in healthcare outcomes, and practical applicability of the findings.

Group Members : Varun Putta, Zhiyi Ying, Hang Lei