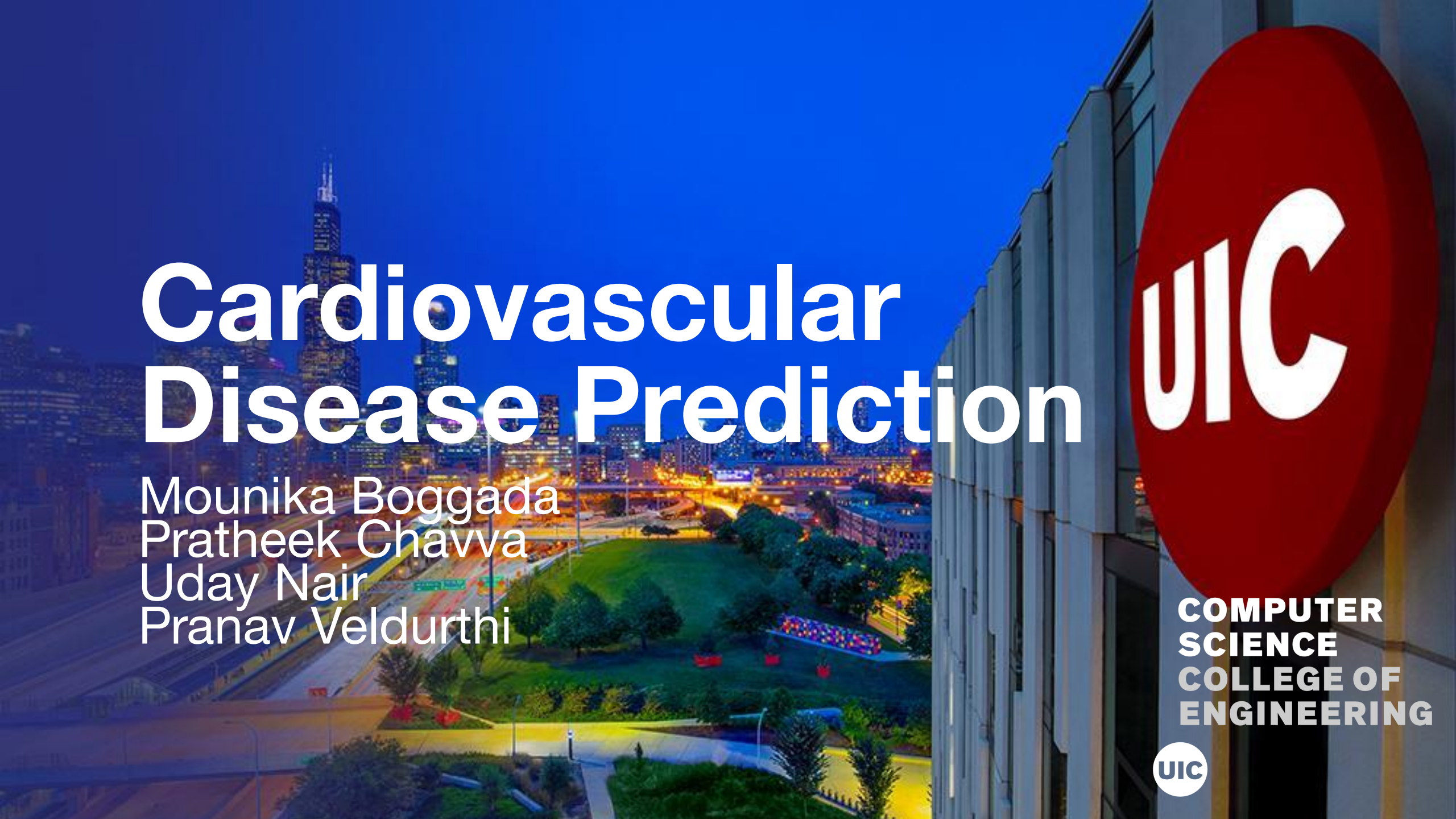


Cardiovascular Disease Prediction

Mounika Boggada
Pratheek Chavva
Uday Nair
Pranav Veldurthi

UIC

COMPUTER SCIENCE
COLLEGE OF ENGINEERING

UIC

Introduction

Global Health Impact

- Statistics and Facts: Cardiovascular diseases (CVDs) are the number one cause of death globally, taking an estimated 17.9 million lives each year according to the World Health Organization.
- Early detection and prevention are crucial for reducing mortality rates

Challenges in Detection and Treatment

- Limited access to healthcare facilities and resource constraints hinder early detection and effective treatment of CVDs

Role of Predictive Analytics

- Predictive analytics models play a vital role in identifying individuals at high risk of developing CVDs.
- Personalized Treatment Plans: Showcase the potential for machine learning models to tailor treatment plans based on individual risk profiles, improving patient outcomes.

Problem Statement

Objective:

To develop a predictive model for forecasting cardiovascular diseases (CVDs) based on various predictors.

Approach:

- Developing a data-driven approach for predicting cardiovascular disease.
- Utilizing Exploratory Data Analysis (EDA) to uncover patterns and insights.
- Applying machine learning techniques such as Random Forest and Gradient Boosting for predictive modeling.

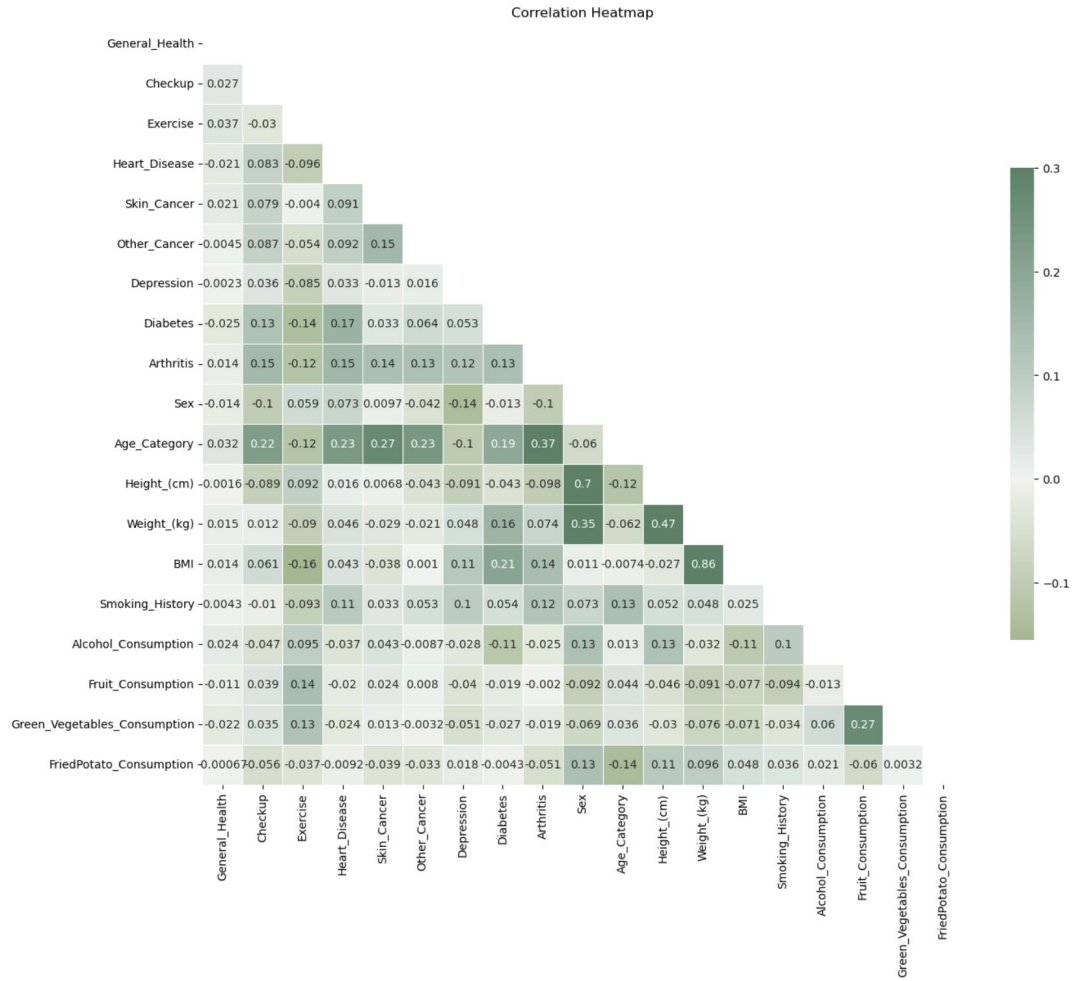
Expected outcome:

Improved accuracy in identifying high-risk individuals, enabling early detection and personalized interventions for better health outcomes and optimized resource allocation.

Dataset

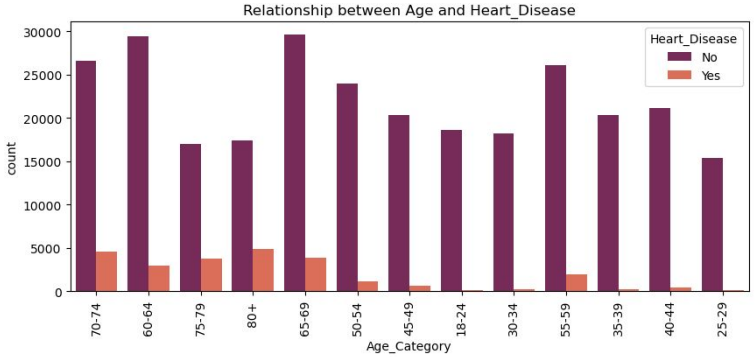
- BRFSS, the top US health survey, gathers state data via phone on residents' health behaviors, conditions, and preventive service use.
- The dataset boasts diverse features such as age, sex, overall health, checkups, exercise, smoking history, and the occurrence of diseases like heart disease, skin cancer, other cancers, diabetes, and arthritis.

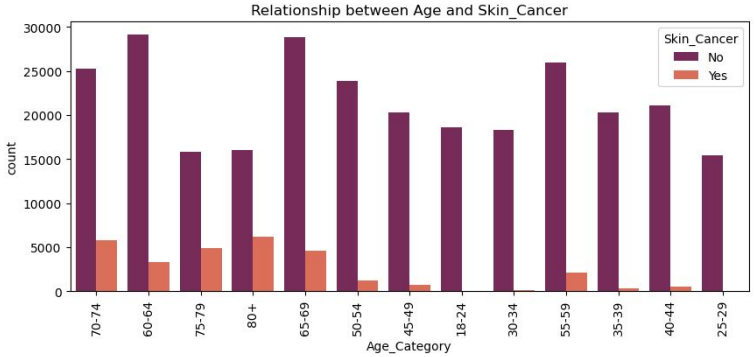
Exploratory Data Analysis



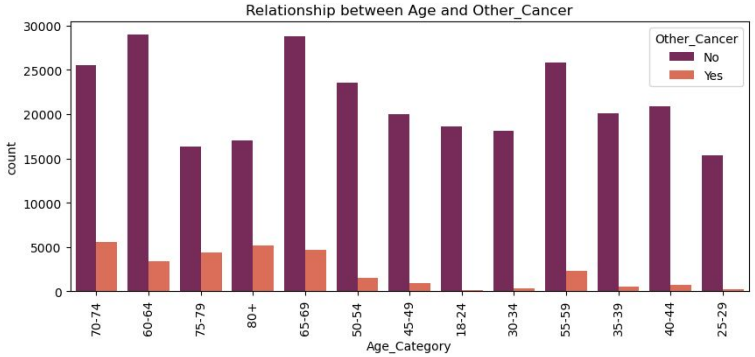
Correlation Matrix

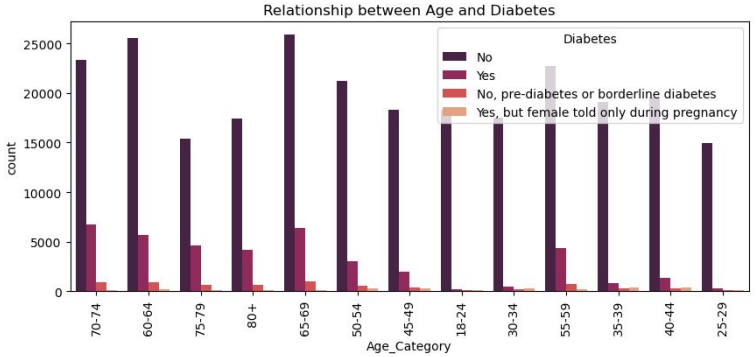
UIC COMPUTER SCIENCE

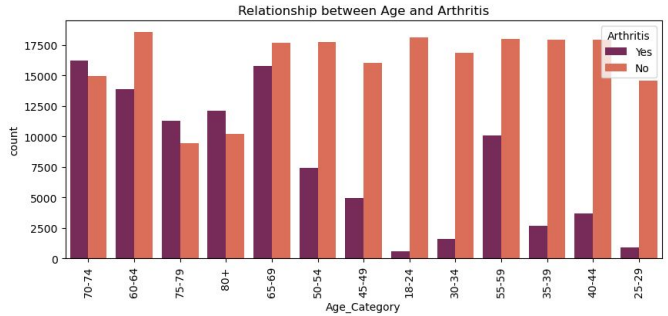


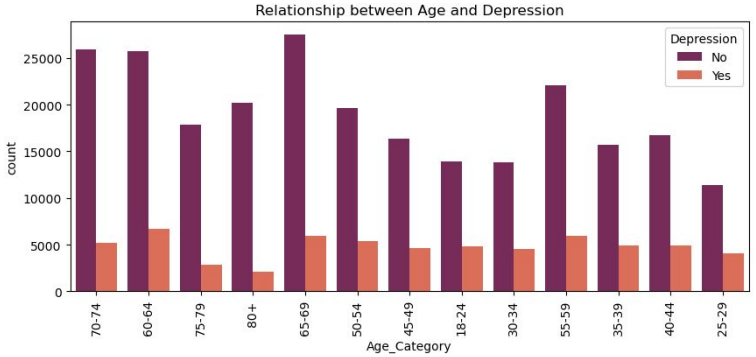


Relation between Age and Diseases

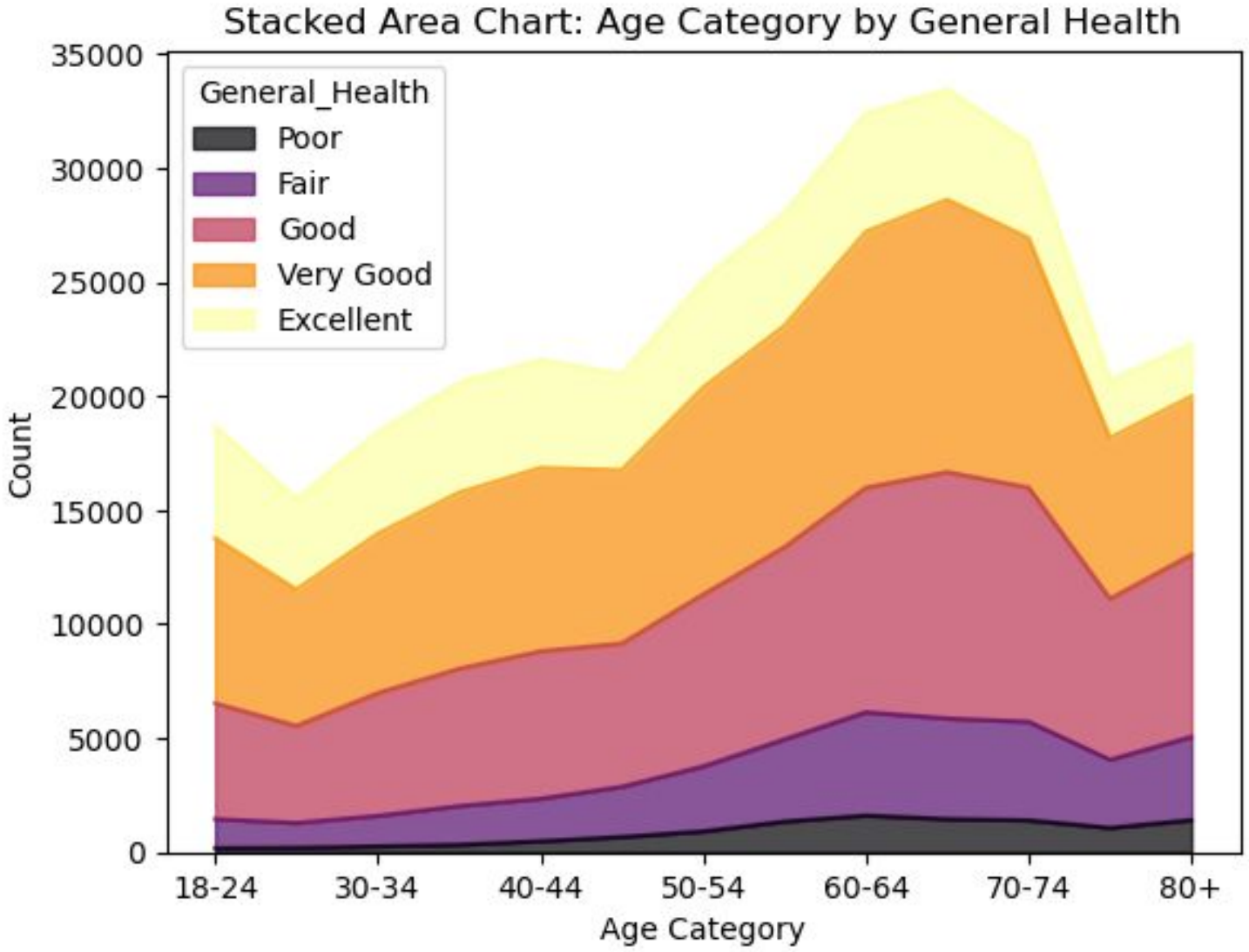








UIC COMPUTER SCIENCE


Stacked Area Chart: Age Category by General Health
General_Health
Poor
Fair
Good
Very Good
Excellent
Count
35000
30000
25000
20000
15000
10000
5000
0
18-24
30-34
40-44
50-54
60-64
70-74
80+
Age Category

Data Preprocessing and Feature Engineering

- Removal of null values, duplicates and outliers.
- Categorical encoding for columns:
  - BMI Categorization
  - Checkup Frequency Mapping
  - One-hot encoding for ‘Sex” column
  - Binary encoding for disease columns
  - Ordinal encoding for ‘General Health’, ‘Age Category and ‘BMI Category’

Models Used

K-Nearest Neighbors (KNN)

- Classifying based on the majority vote of the nearest neighbors
- KNN is a non-parametric model, makes no assumptions about the underlying data distribution
- It doesn't build a model during the training phase; instead, it stores all the training examples in memory.
- During the prediction phase, KNN searches for the k nearest neighbors of the new data point and determines its class label or output value based on those neighbors.

Models Used

Random Forest Classifier

- ensemble learning method that combines multiple decision trees to create predictive model
- built-in feature importance measure to assess the relative importance of each feature in predicting heart disease
- effectively handles class imbalances by resampling techniques or by adjusting class weights during training
- Randomized Search CV (Cross-Validation) to efficiently search for the best combination of hyperparameters

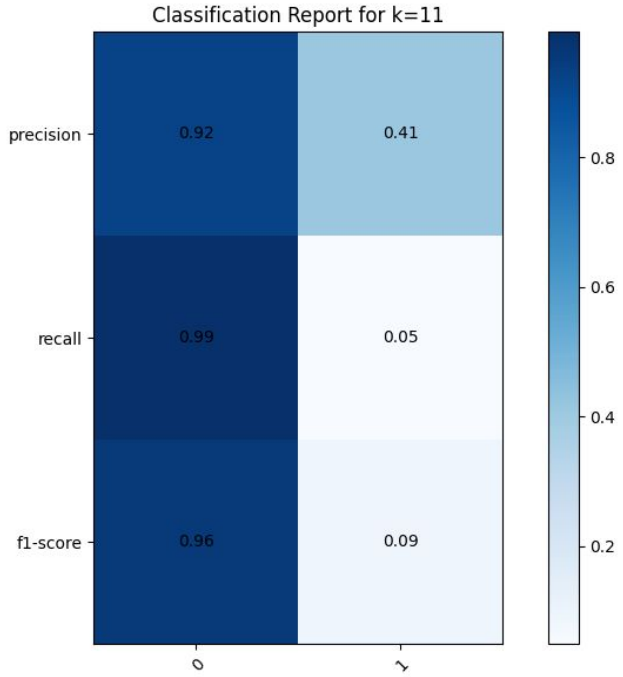
Models Used

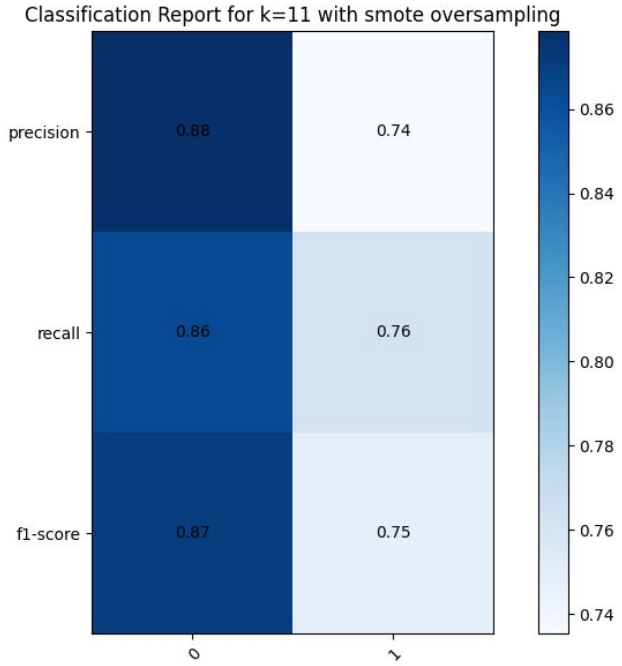
Gradient Boosting

- Ensemble learning technique combining multiple weak learners
- Sequentially adds new models to correct errors made by previous models and captures complex relationships in the data
- GridSearchCV for hyperparameter optimization to fine-tune the learning rate, max depth of trees, and number of estimators.
- evaluates all combinations of hyperparameters specified in the grid, enabling thorough exploration of the hyperparameter space

Results

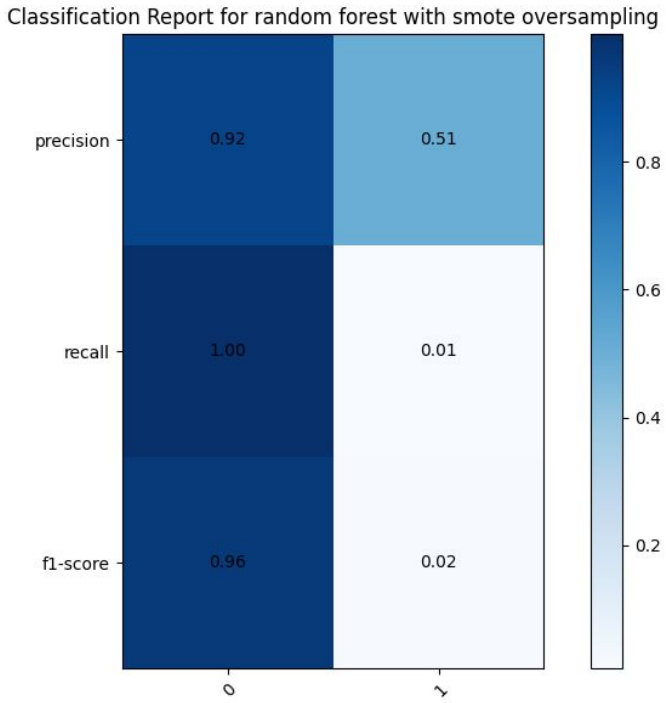
K-Nearest Neighbors (KNN)

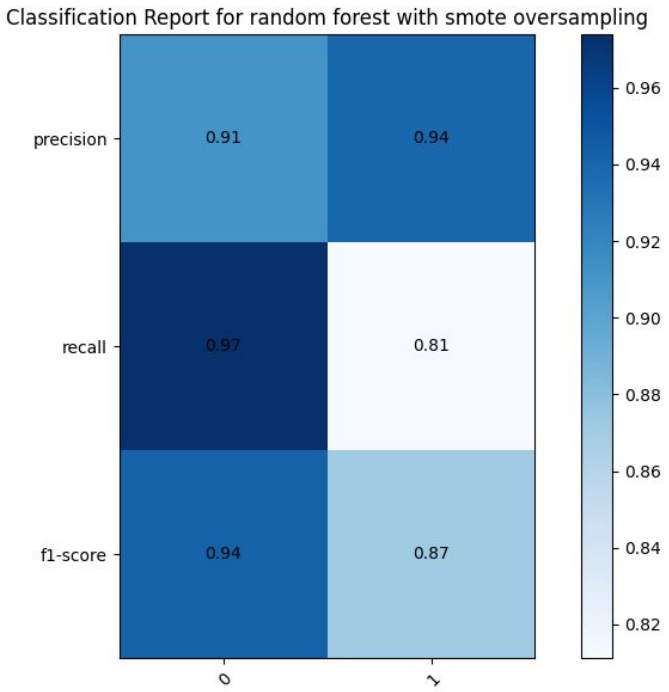




Results

Random Forest Classifier

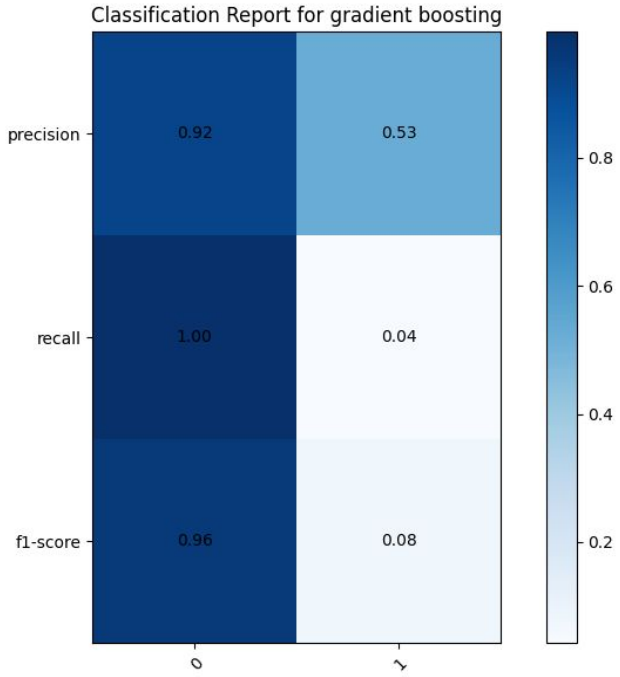


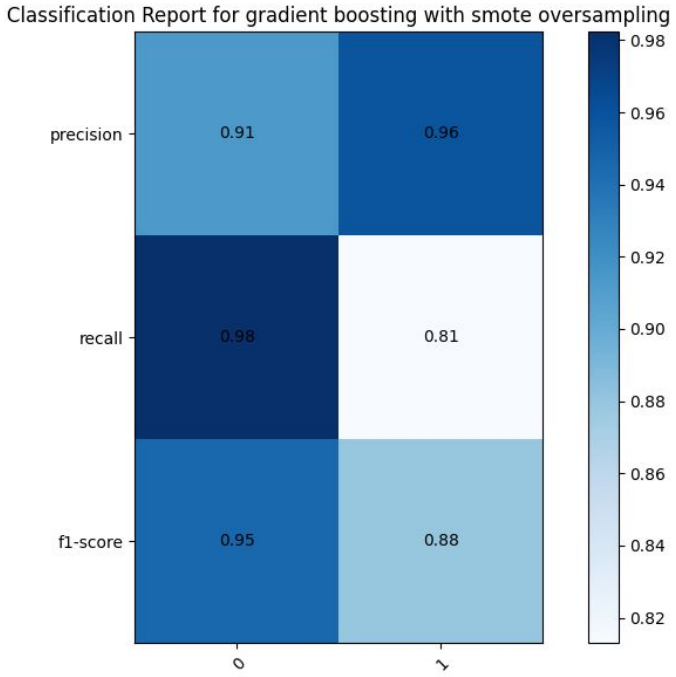


UIC COMPUTER SCIENCE

Results

Gradient Boosting





Conclusion

Model	Training time	Inference time	Memory
Gradient Boosting	3348.8 ms	6.5 ms	0.27 MB
KNN	0.1 ms	11.7 ms	39.5 MB
Random Forest	13465 ms	8 ms	5.6 MB

- Gradient Boosting is the best model as it achieves high accuracy comparable to Random Forest while having significantly lower training time, and memory usage.

Thank You

Questions?