CAPSTONE PROJECT

CARDIOVASCULAR RISK PREDICTION

Presented By:

1. Name of Student : ANAND KUMAR DALWAIE
2. College Name : ANNAMACHARYA INSTITUTE OF TECHNOLOGY & SCIENCES TIRUPATHI
3. Department : ARTIFICIAL INTELLIGENCE & DATA SCIENCE

edunet foundation

OUTLINE

- Problem Statement
- Proposed System/Solution
- System Development Approach
- Algorithm & Deployment
- Result
- Conclusion
- Future Scope
- References

PROBLEM STATEMENT

EXAMPLE :

- The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts.
- The classification goal is to predict whether the parent has a 10-year risk of future coronary heart disease (CHD).
- The dataset provides the parents' information. It includes over 4,000 records and 15 attributes. Variable Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors.

PROPOSED SOLUTION

The solution involves developing a machine learning model to predict the 10-year risk of coronary heart disease (CHD) based on patient data, which includes demographic, behavioral, and medical factors.

1.Data Preprocessing:

- **Data Cleaning:** Removed null values and duplicates.
- **Balancing the Dataset:** Used SMOTE and under-sampling to address class imbalance, ensuring the model can effectively predict CHD.

2.Feature Engineering:

- Created new features like Hypertension, Diabetes Severity, and Smoking Factor to enhance the model's predictive power.

3.Model Development:

- **Algorithm:** Used **XGBoost** for its efficiency in binary classification.
- **Training:** Applied GridSearchCV for hyperparameter tuning to optimize model performance.
- **Evaluation:** Focused on Recall to minimize false negatives, crucial for early diagnosis of CHD.

4.Deployment:

- **Scalability:** Deployed on a cloud platform, allowing healthcare providers to input data and receive CHD risk predictions through a user-friendly interface.

SYSTEM APPROACH

1. **Data Preprocessing:**
   - Data cleaning (handling missing values and duplicates)
   - Balancing the dataset using SMOTE and random under-sampling
2. **Feature Engineering:**
   - Creating new features like 'Hypertension,' 'Diabetes Severity,' and 'Smoking Factor'
3. **Modeling:**
   - XGBoost for prediction
   - Grid search and cross-validation for parameter tuning
4. **Technology Stack:**
   - Python for data processing and modeling
   - Scikit-learn for machine learning
   - Pandas and NumPy for data manipulation

ALGORITHM & DEPLOYMENT

Algorithm Selection:

- XGBoost is chosen for its high performance in binary classification tasks and its ability to handle large, structured datasets efficiently. It is particularly effective in managing imbalanced datasets, reducing overfitting through regularization, and providing fast predictions.

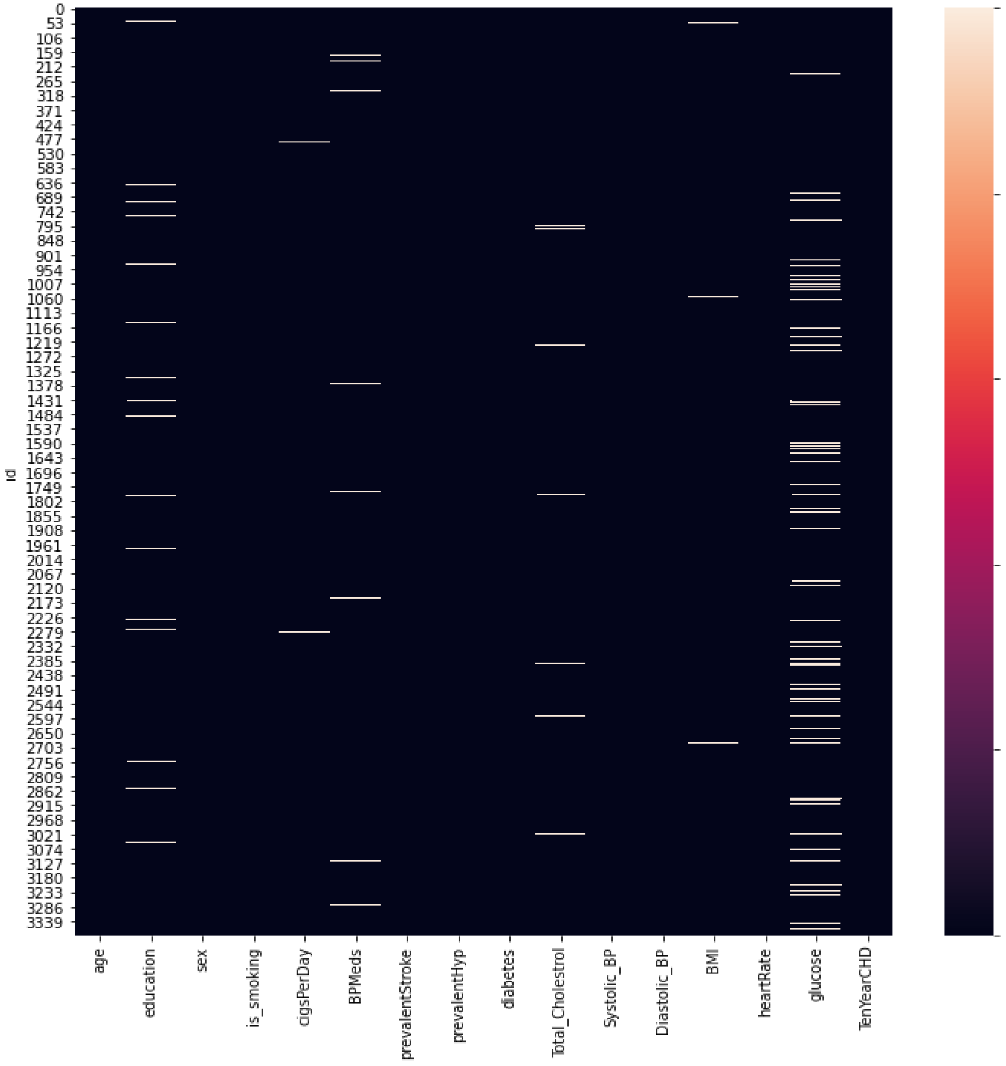
Training Process:

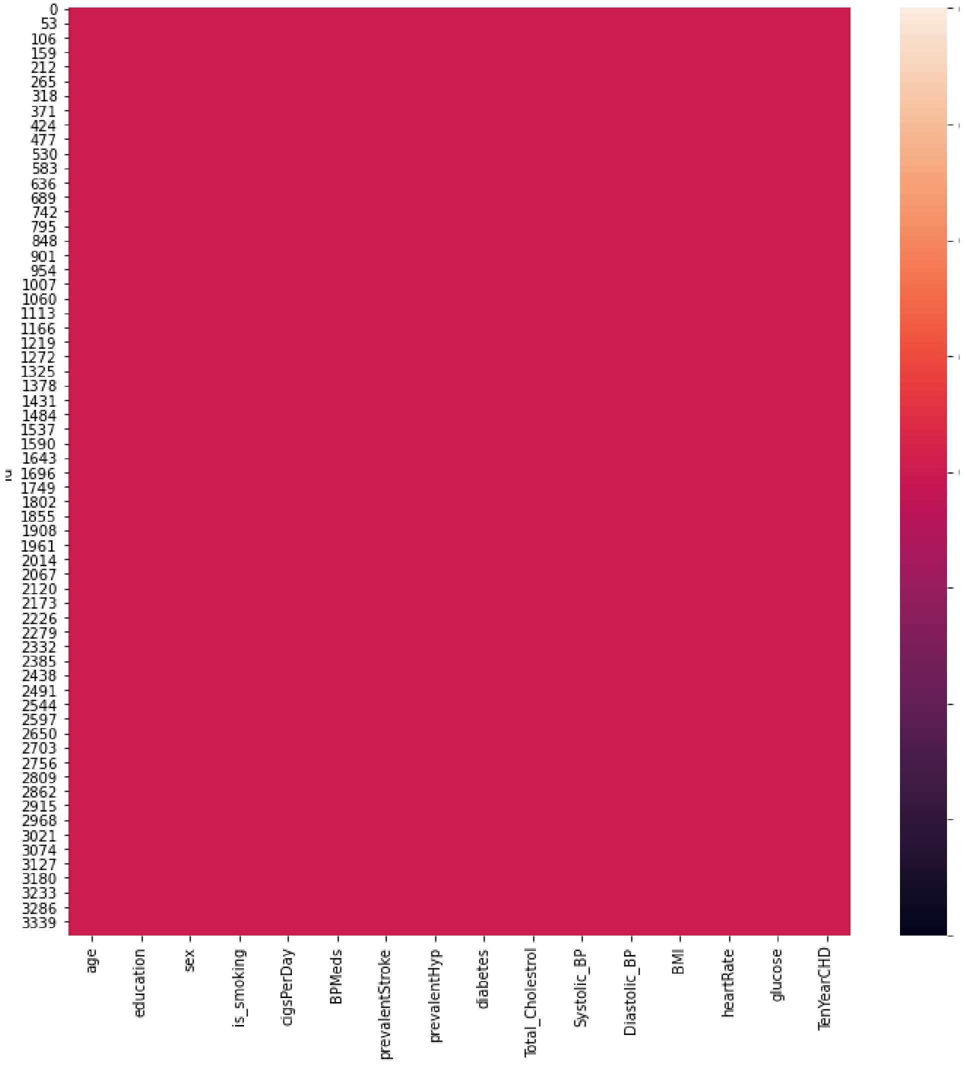
- The model is trained using a balanced dataset achieved through SMOTE and random under-sampling.
- Hyperparameter tuning is conducted using GridSearchCV, while cross-validation ensures the model generalizes well.
- Emphasis is placed on Recall to minimize false negatives, critical for accurate CHD risk prediction.

Deployment:

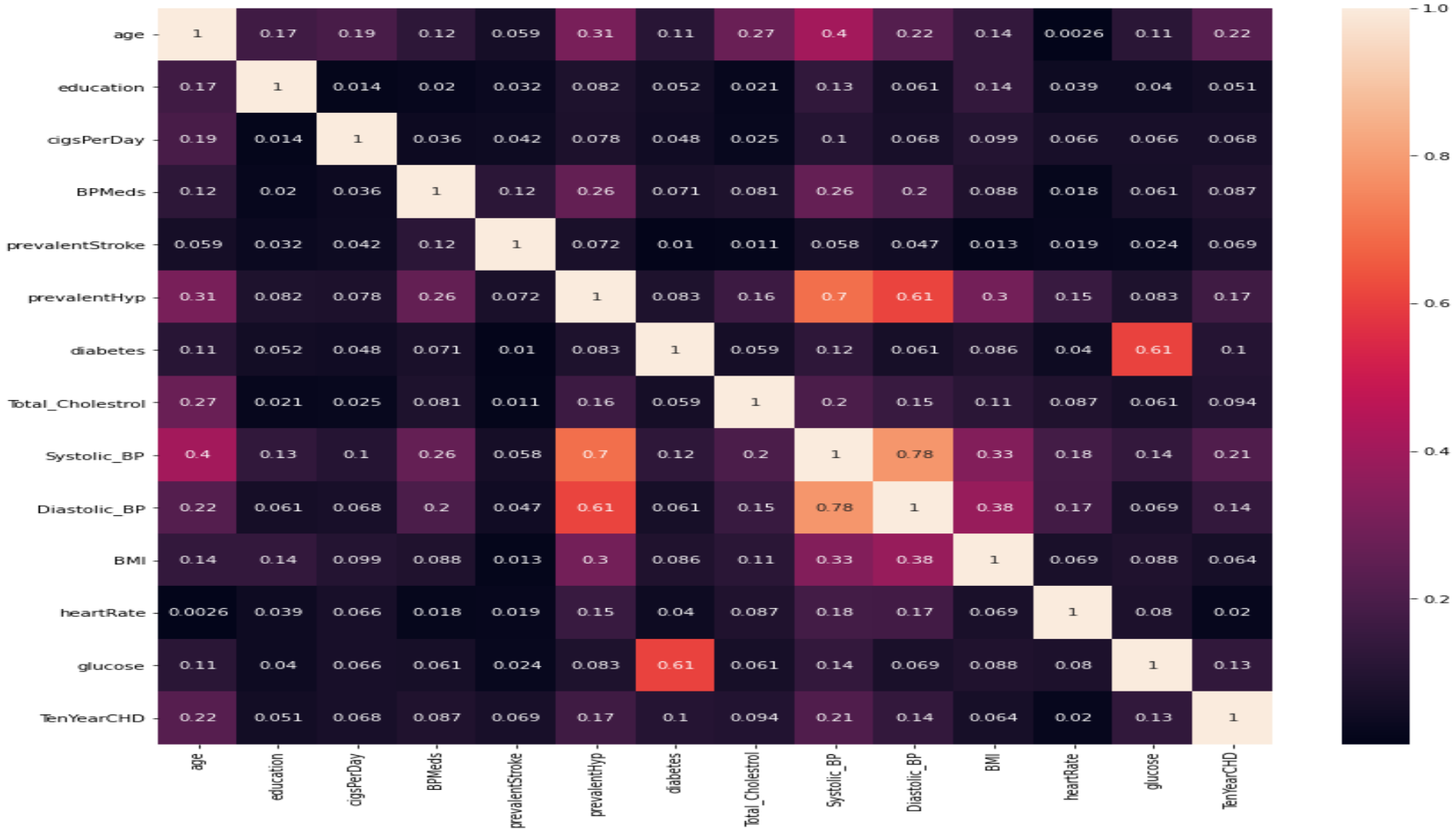
- The model is deployed on a cloud platform (e.g., AWS, Azure), ensuring scalability and real-time data processing.
- A user-friendly interface allows healthcare providers to input patient data and receive immediate CHD risk predictions.
- The system is designed for scalability and reliability, with continuous monitoring and updates to maintain accuracy in real-world applications.

MISSING VALUES & AFTER BILLING NAN VALUES



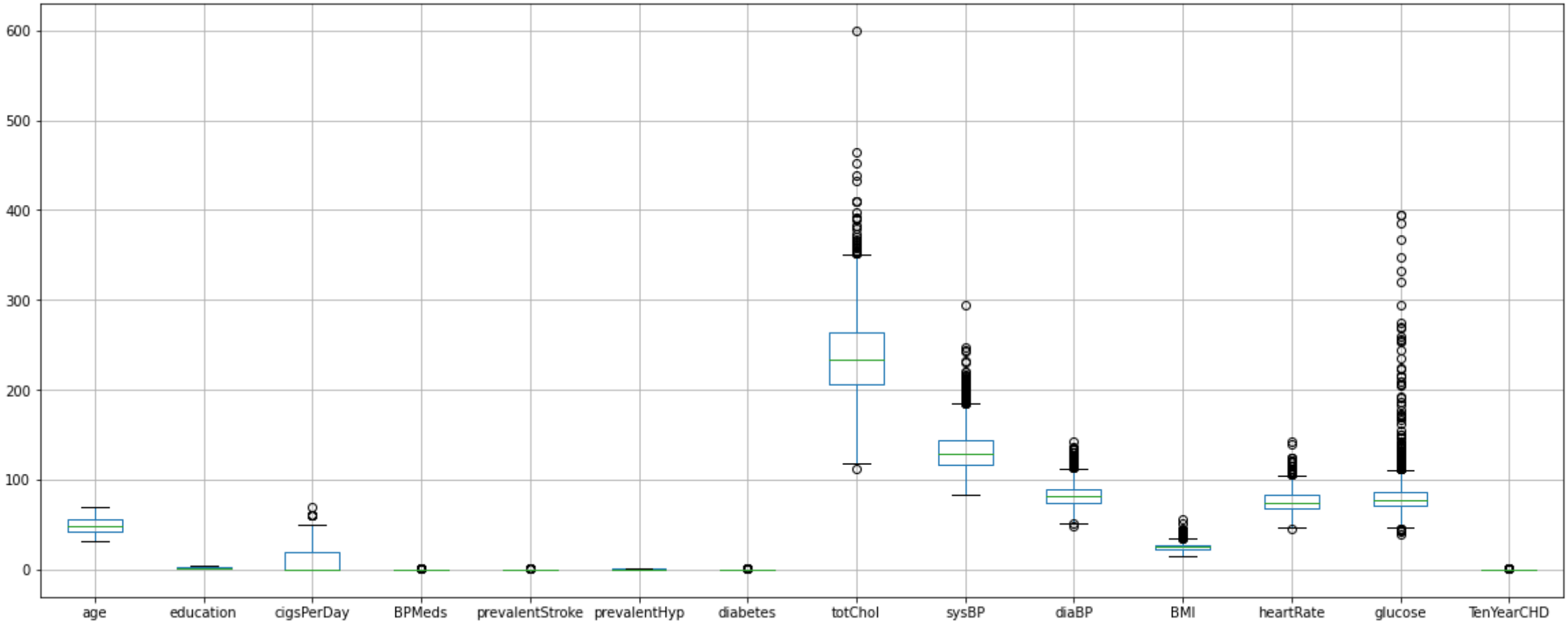


CORRELATION BETWEEN FEATURES

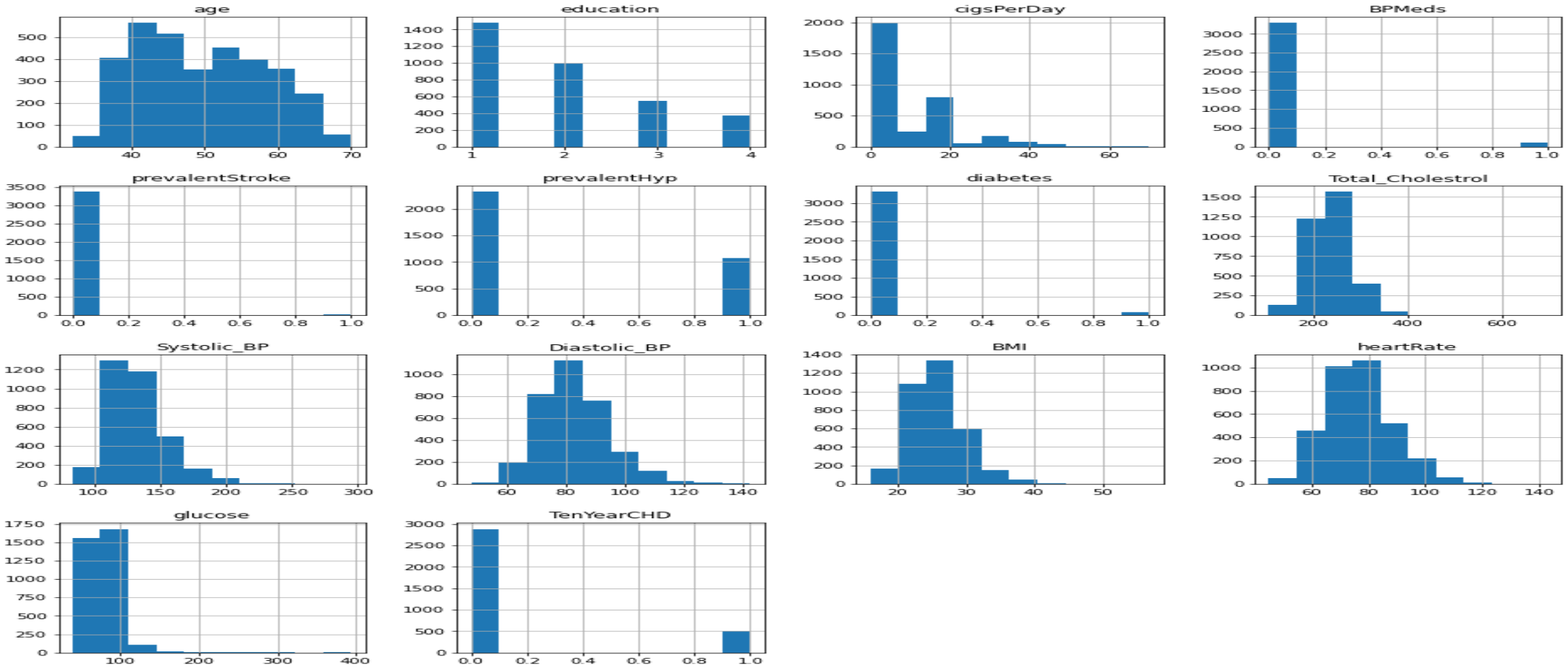


edunet foundation

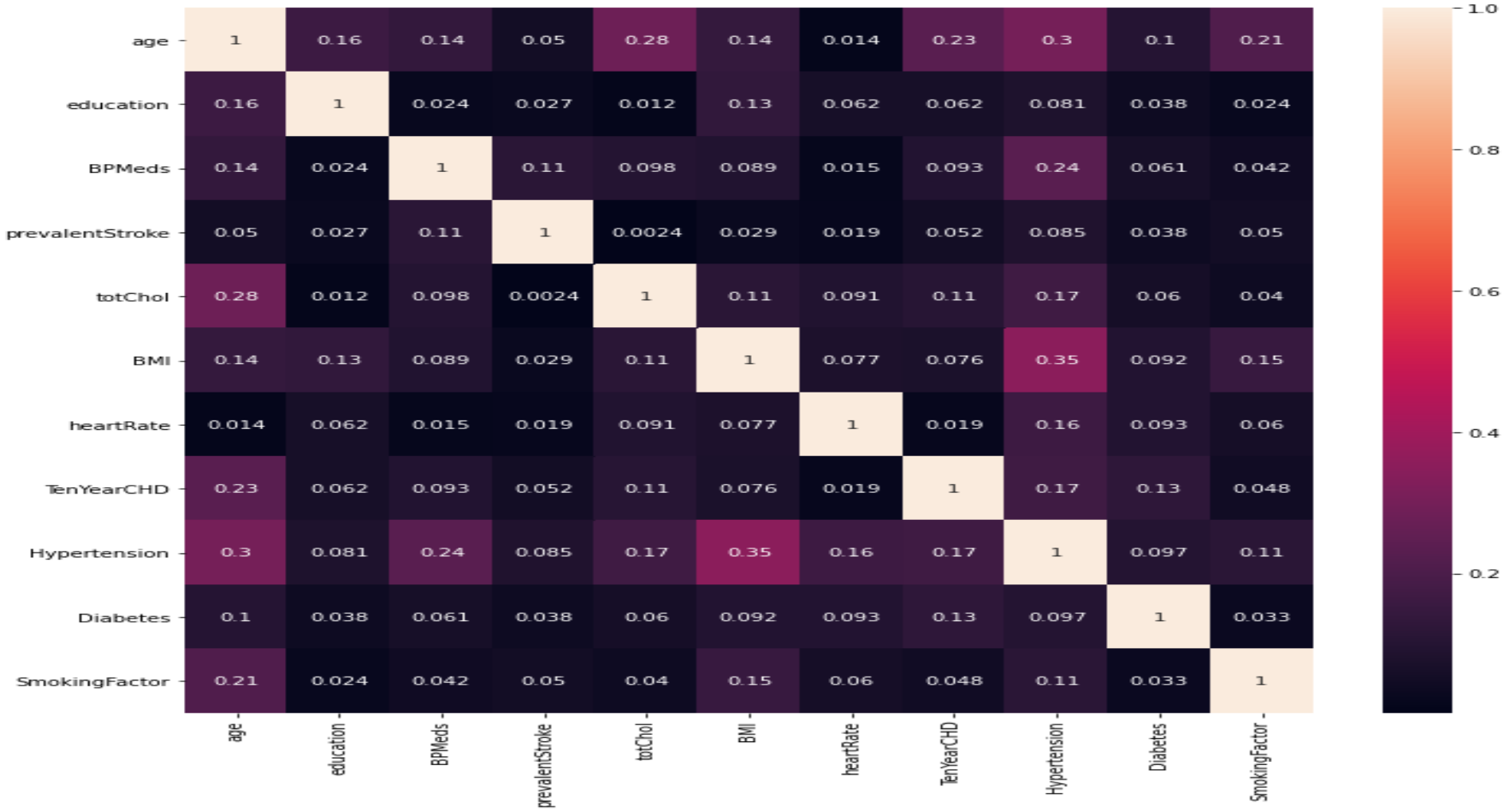
OUTLIER DETECTION



DISTRIBUTION OF DATA

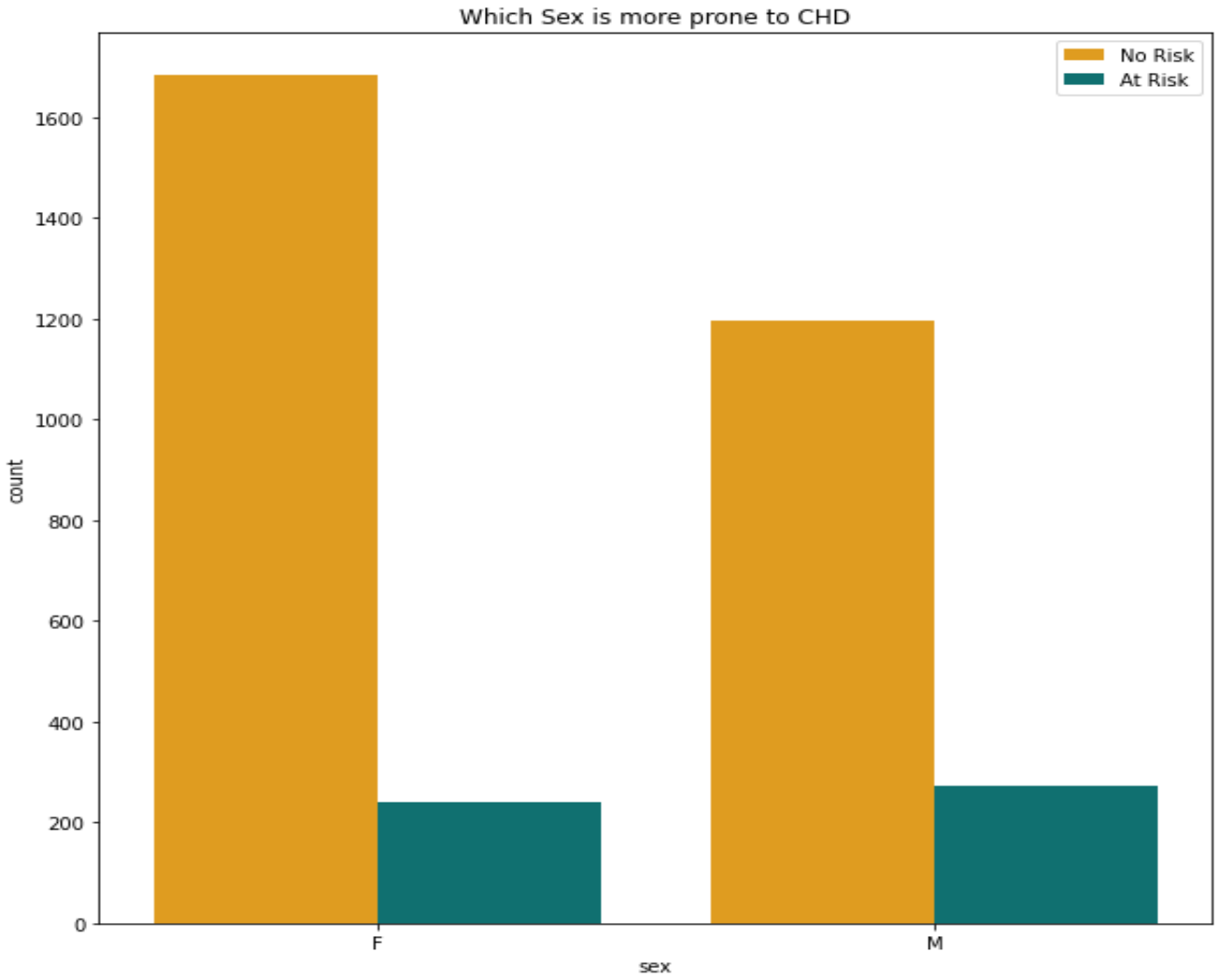


CORRELATION BETWEEN FEATURES AFTER FEATURE ENGINEERING

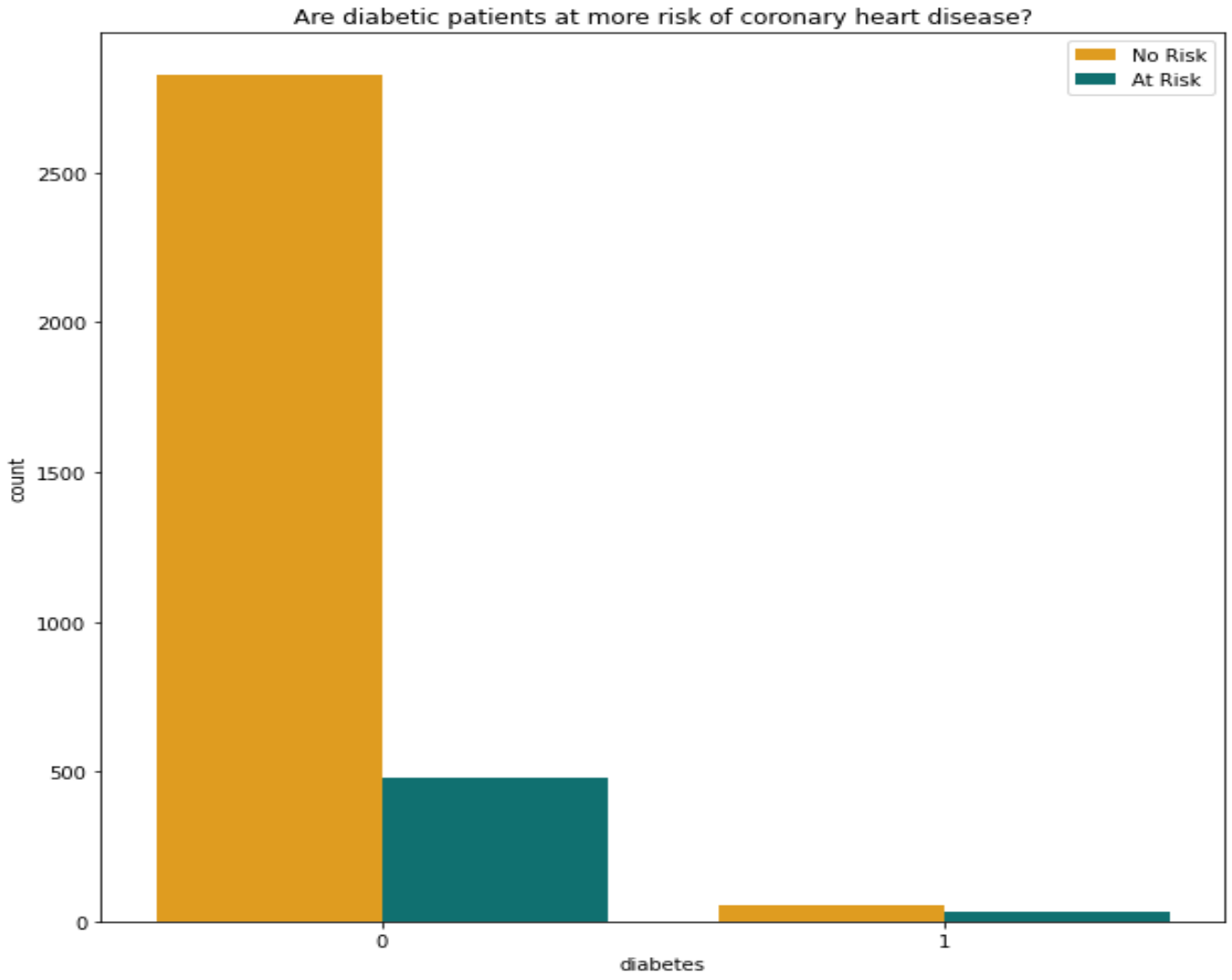


edunet foundation

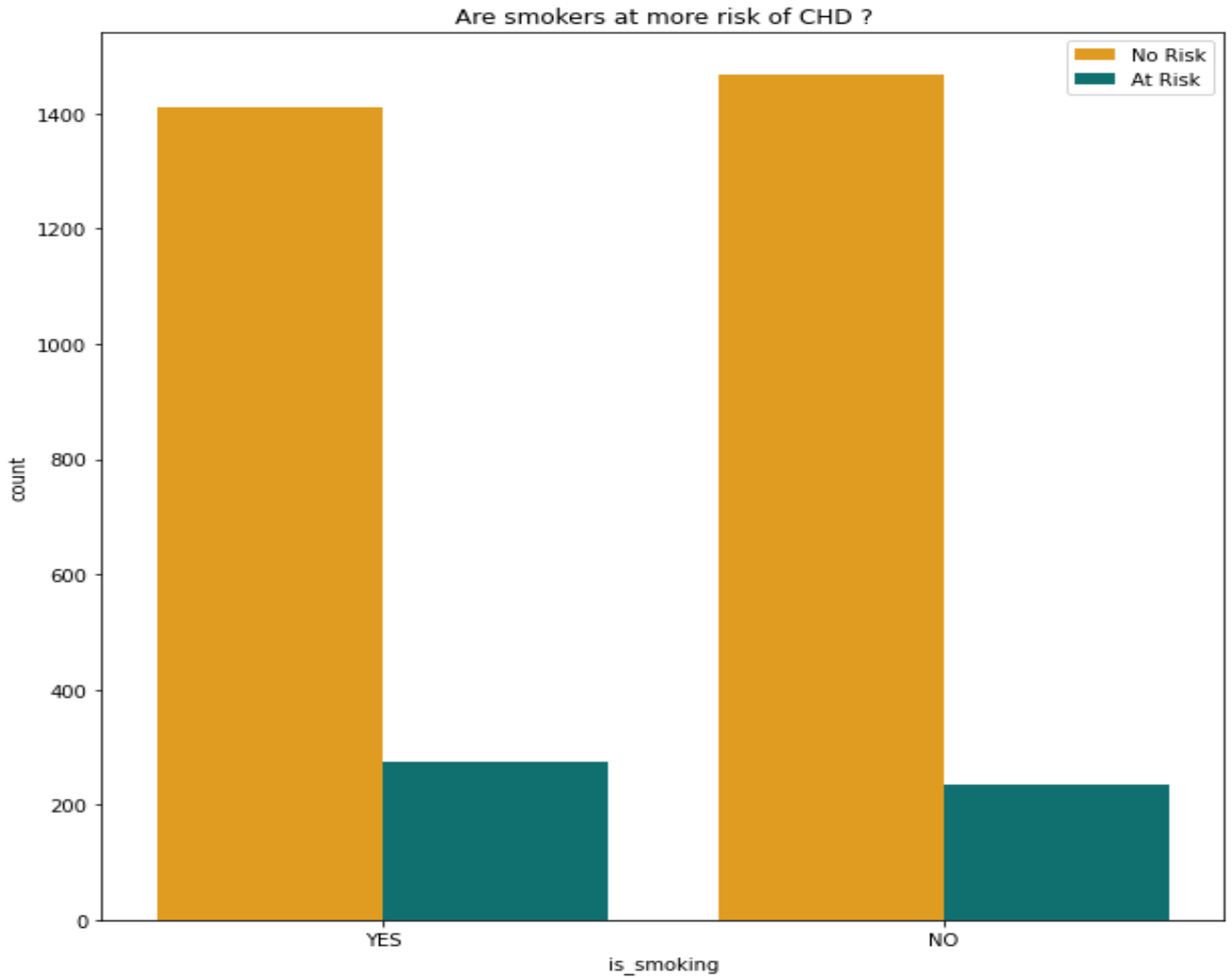
WHICH SEX IS MORE PRONE TO CHD ?



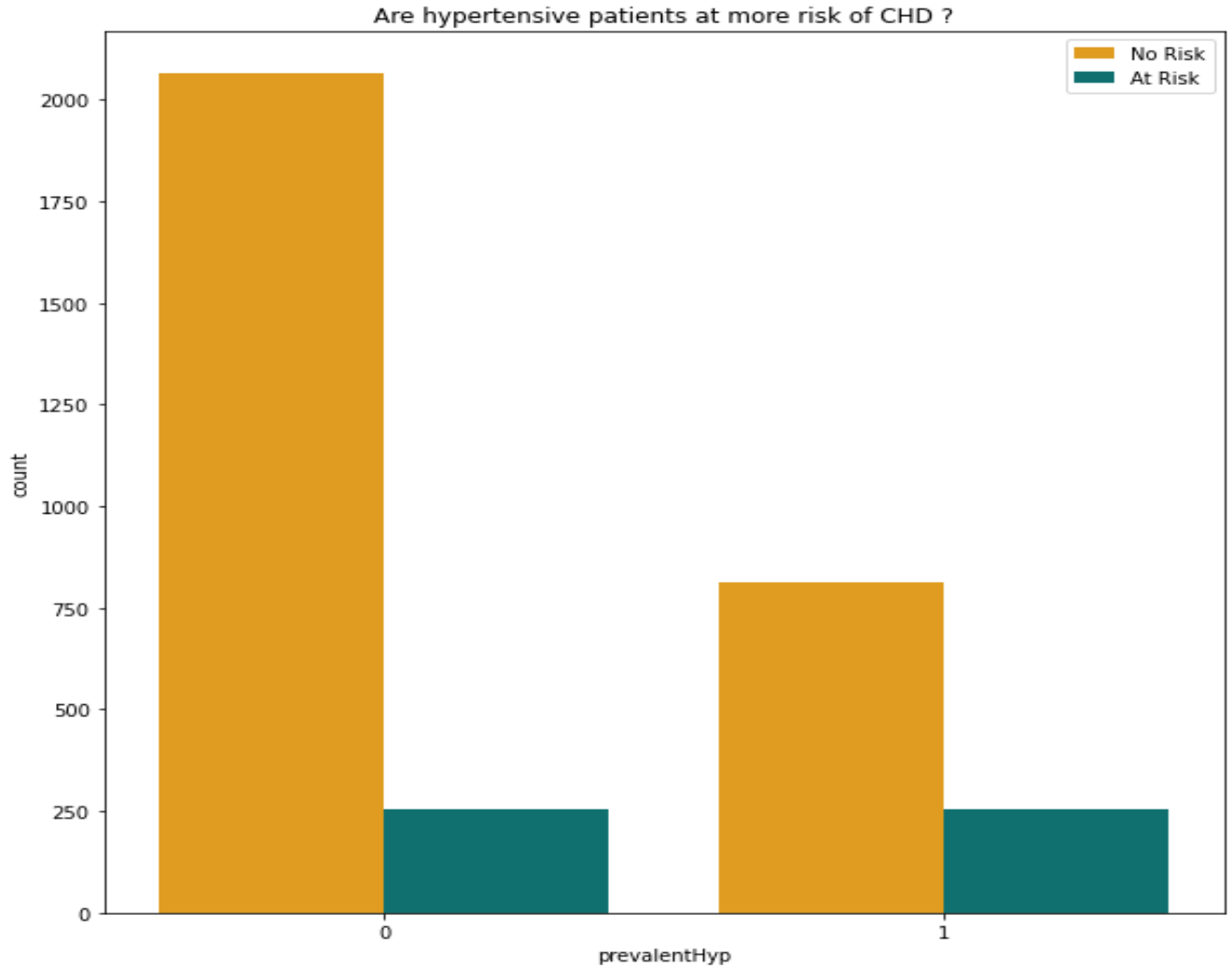
ARE DIABETIC PATIENTS AT MORE RISK OF CHD ?



ARE SMOKERS AT MORE RISK OF CHD ?

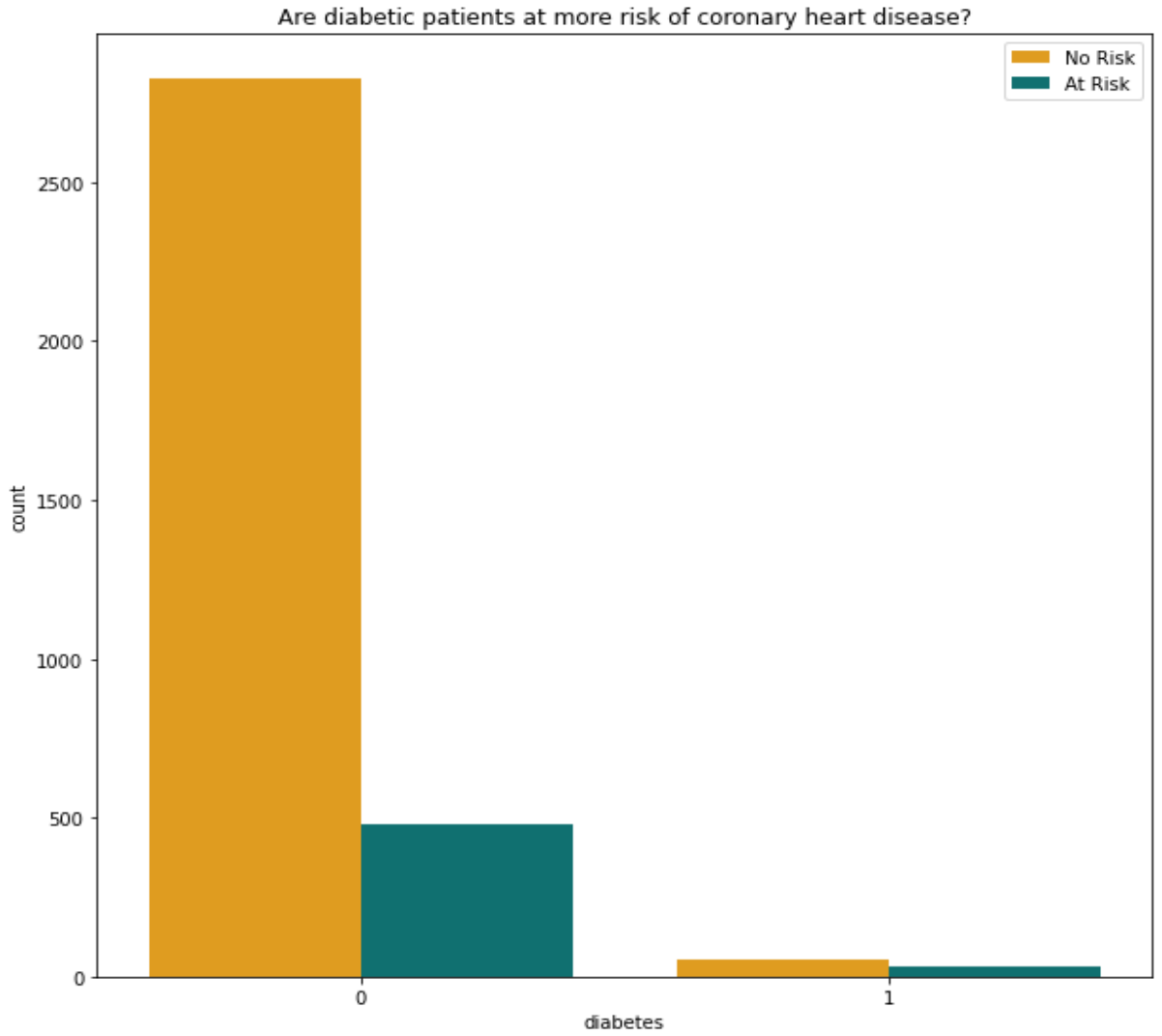


ARE HYPERTENSIVE PATIENTS AT MORE RISK OF CHD?

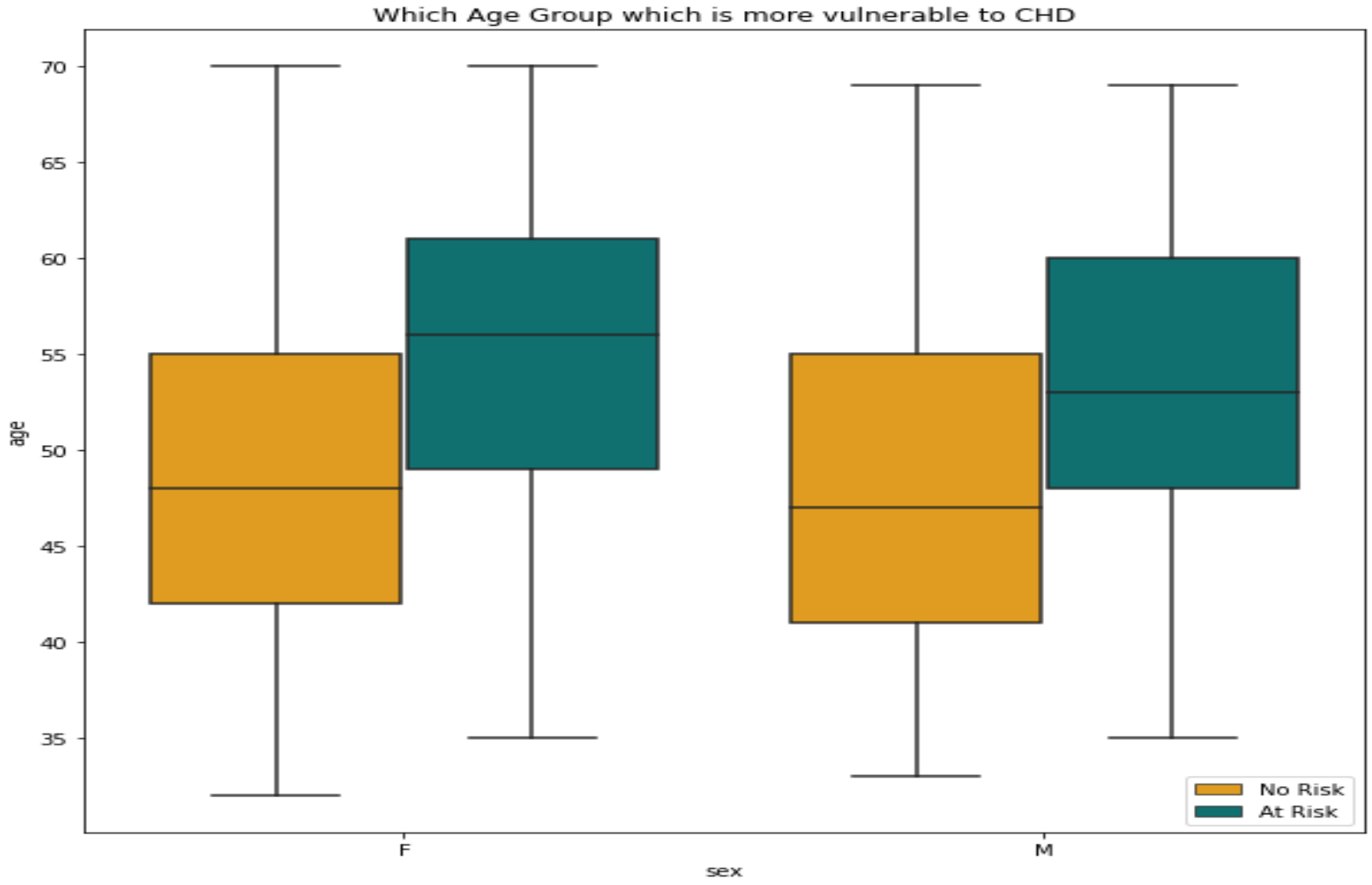


edunet foundation

ARE PATIENTS ON BP MEDICATION AT MORE RISK OF CHD ?

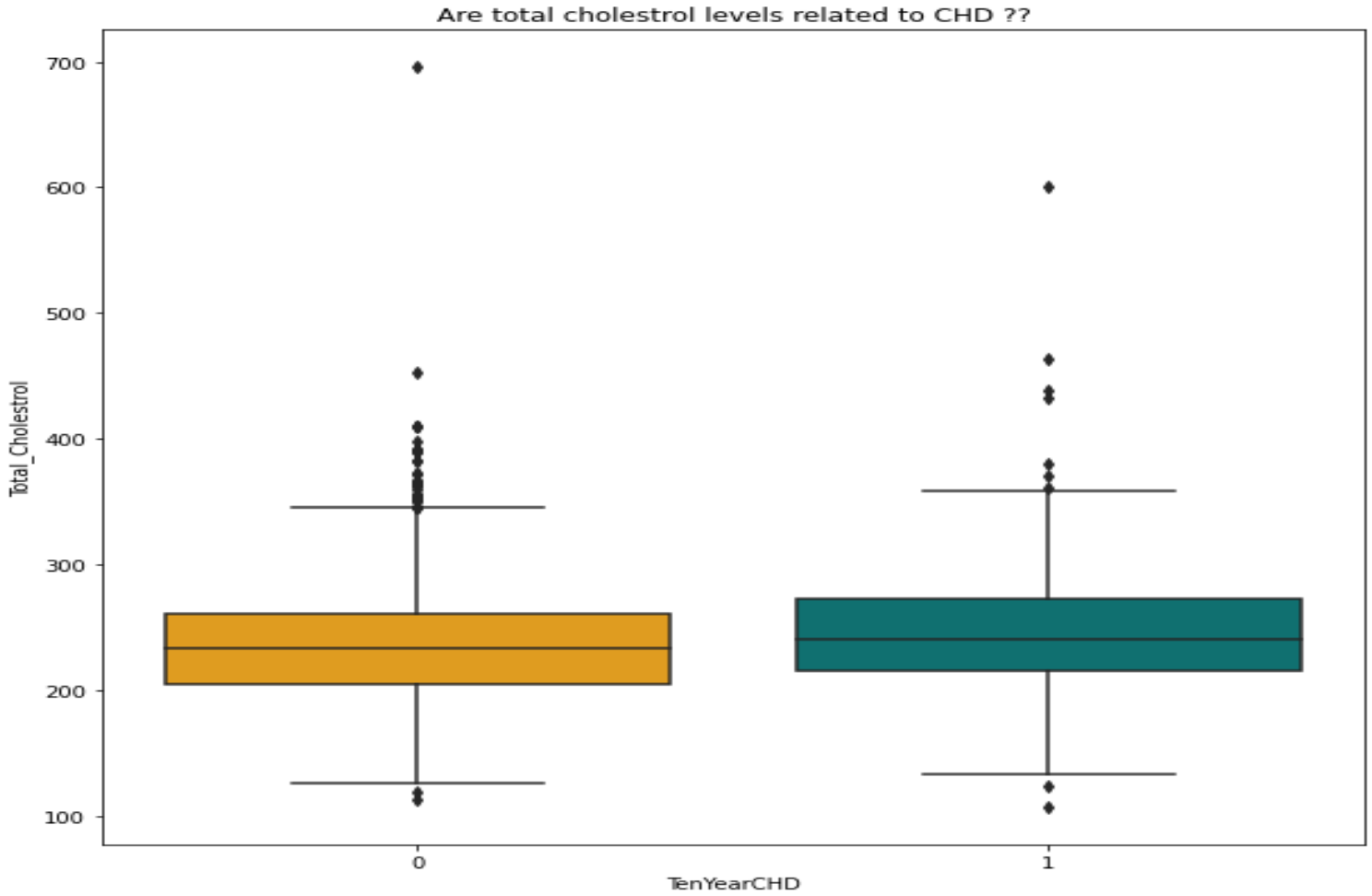


WHICH AGE GROUP IS MORE VULNERABLE TO CHD ?



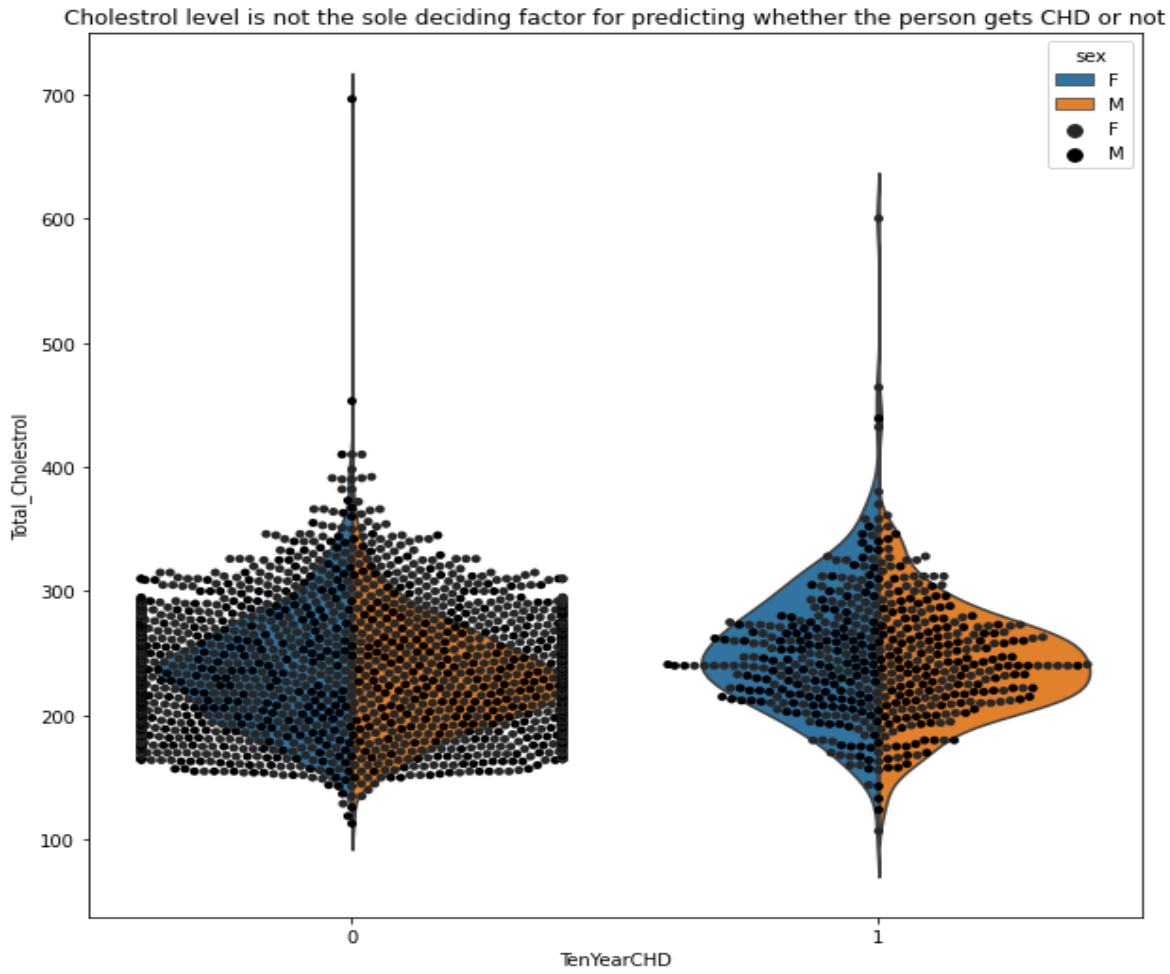
edunet foundation

ARE TOTAL CHOLESTEROL LEVELS RELATED TO CHD ?

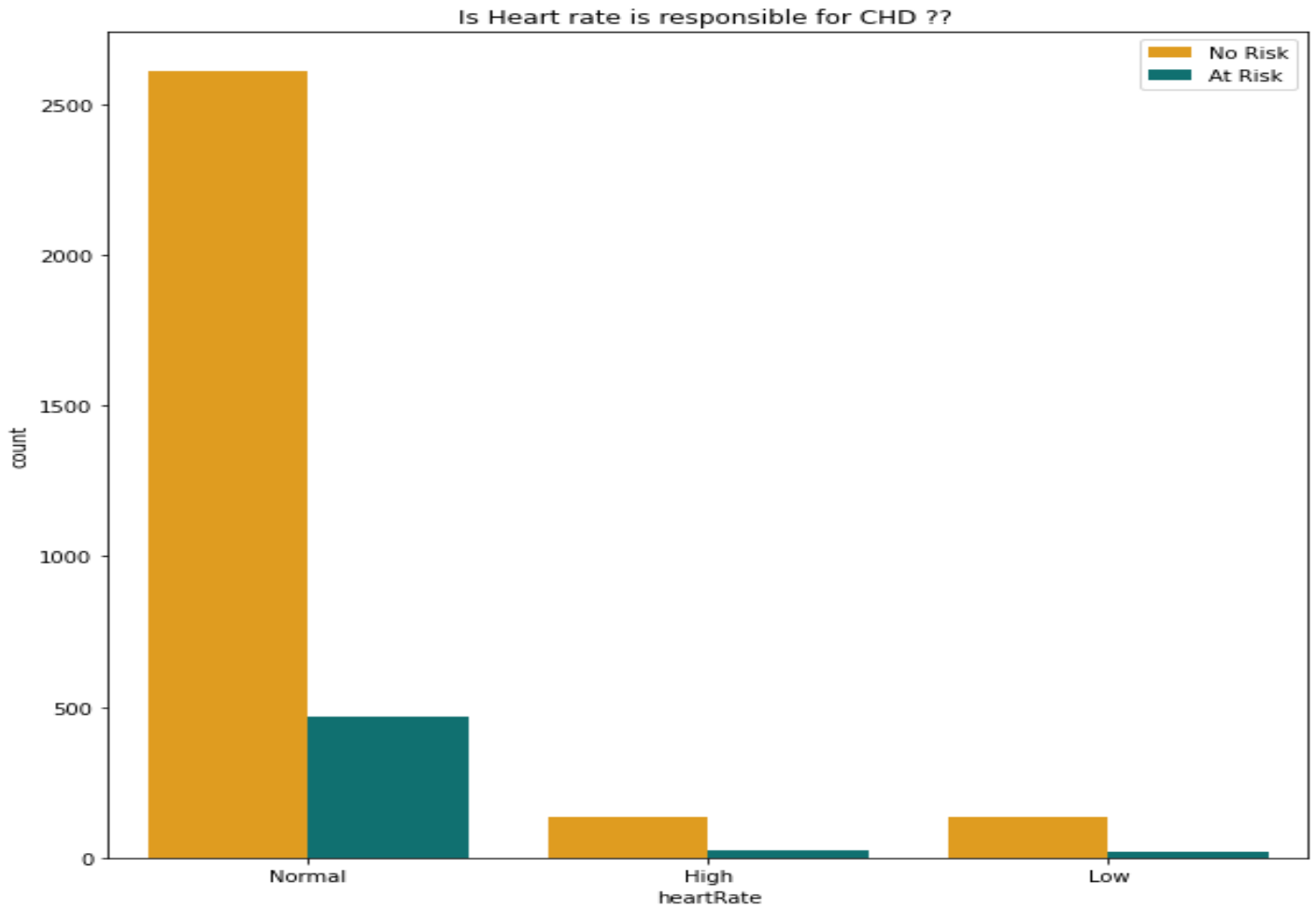


edunet foundation

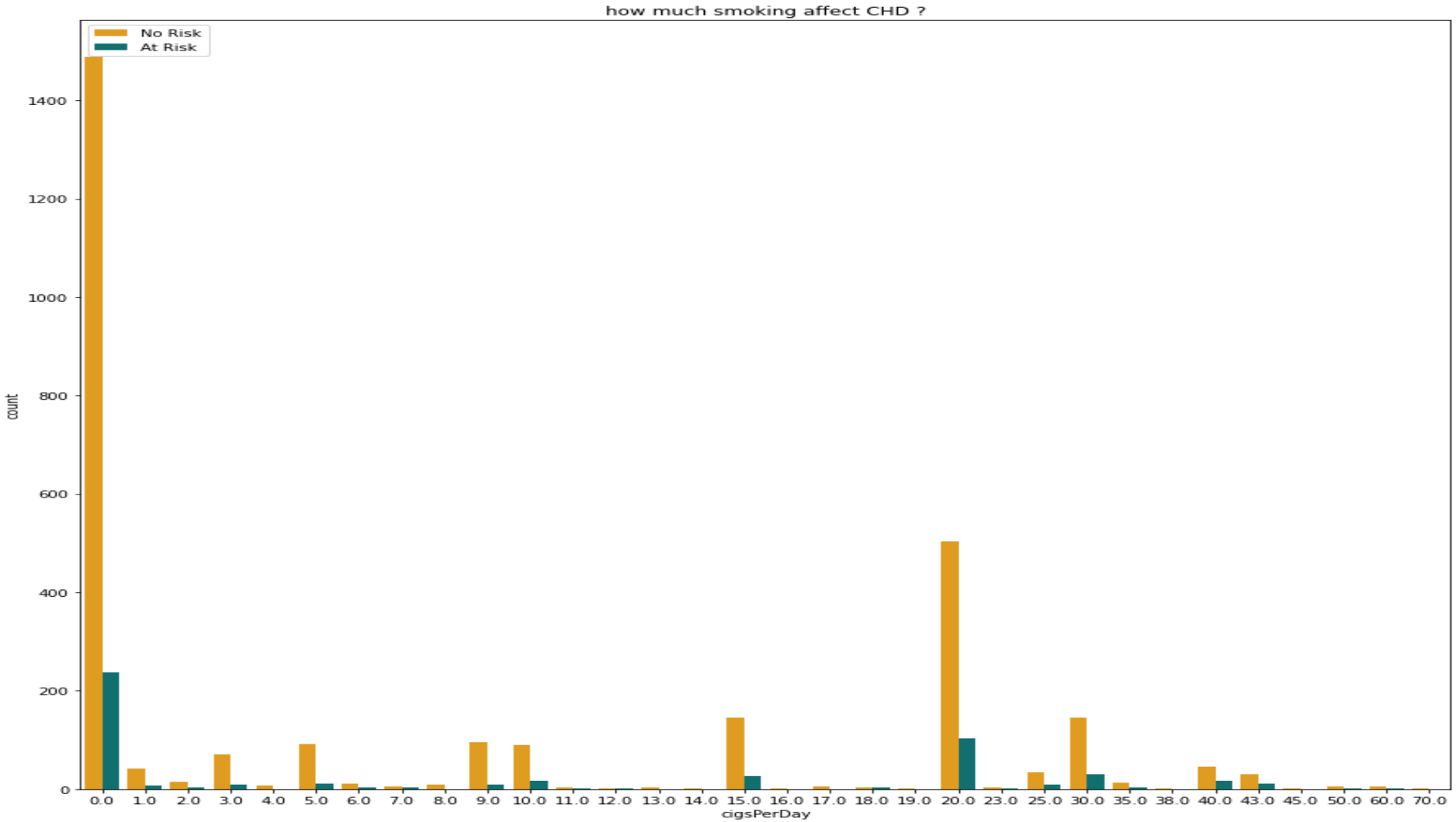
CHOLESTEROL LEVEL IS NOT THE SOLE DECIDING FACTOR FOR CHD



CAN HEART RATE POSSIBLY DEFINE THE RISK OF CHD ?

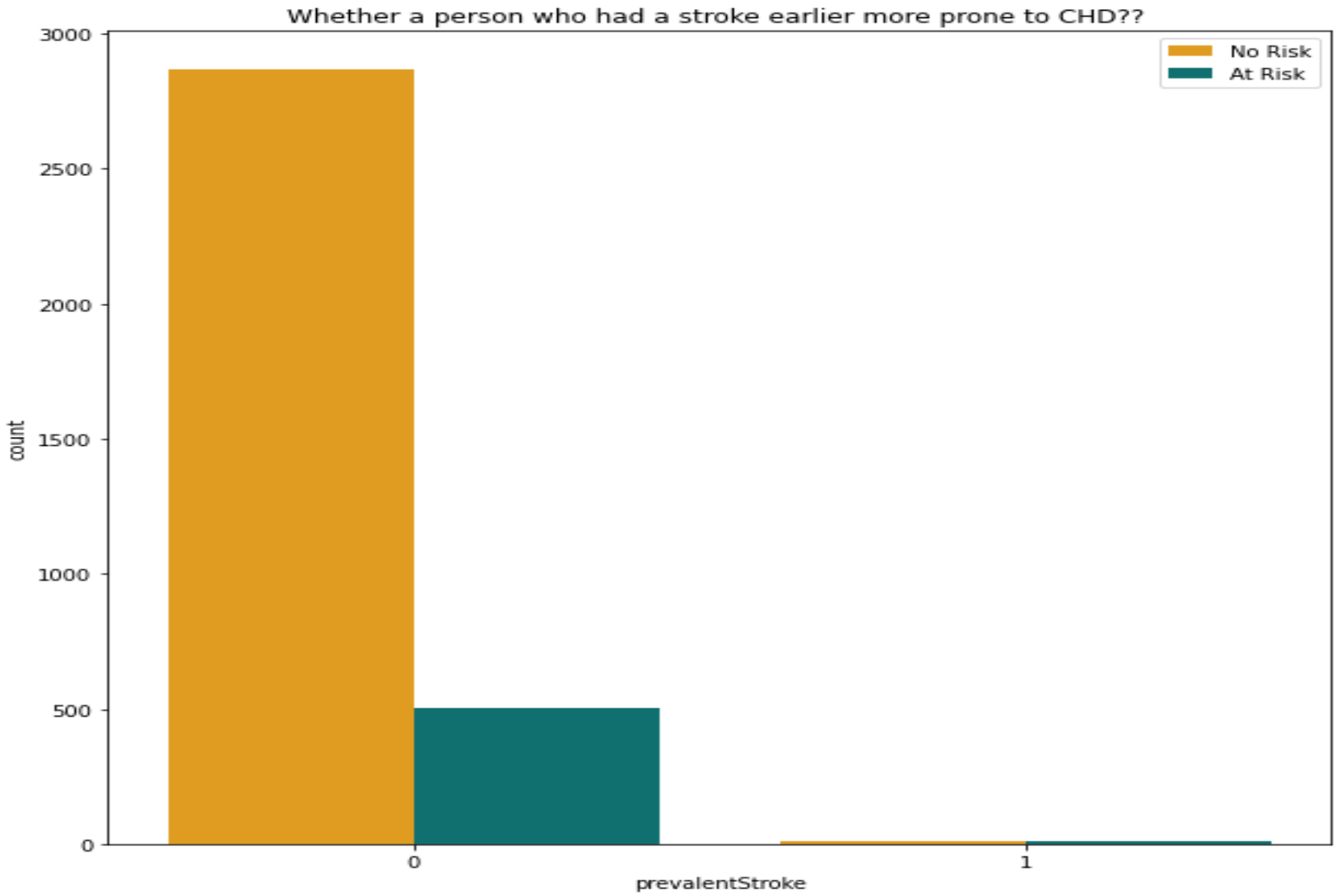


CAN SMOKING NUMBER OF CIGARETTES PER DAY LEAD TO CHD?

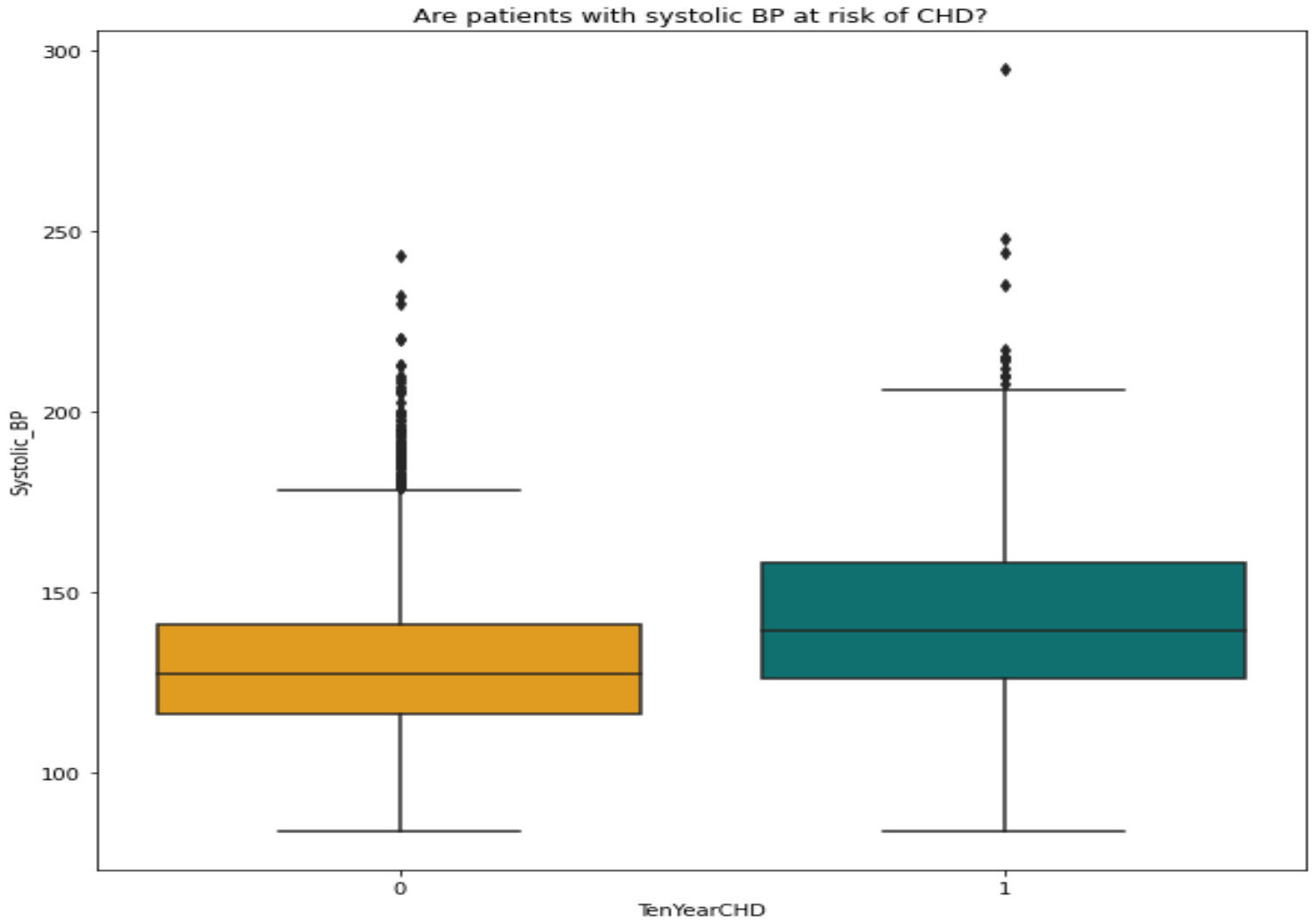


edunet foundation

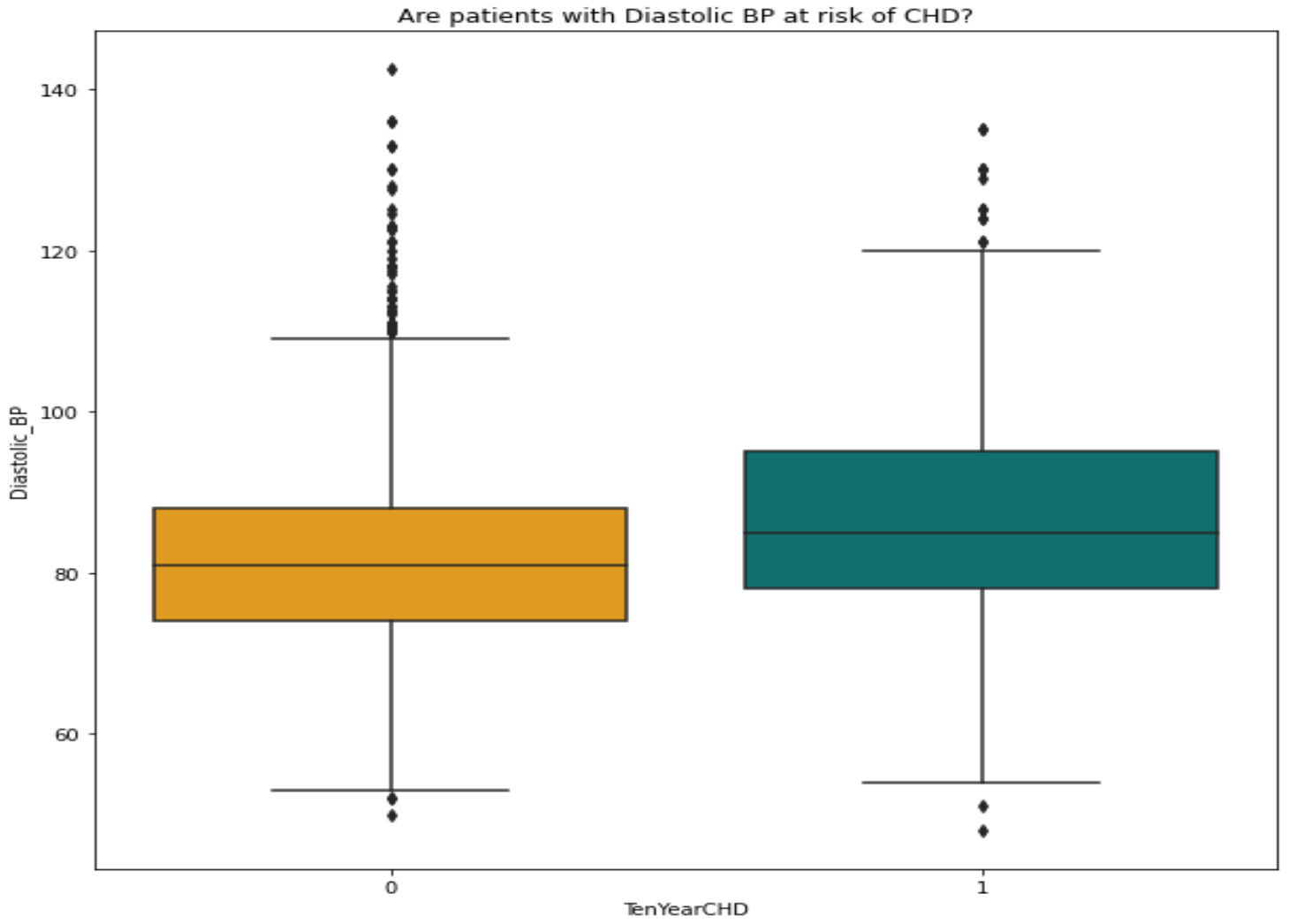
ONE WHO HAD A STROKE EARLIER MORE PRONE TO CHD ?



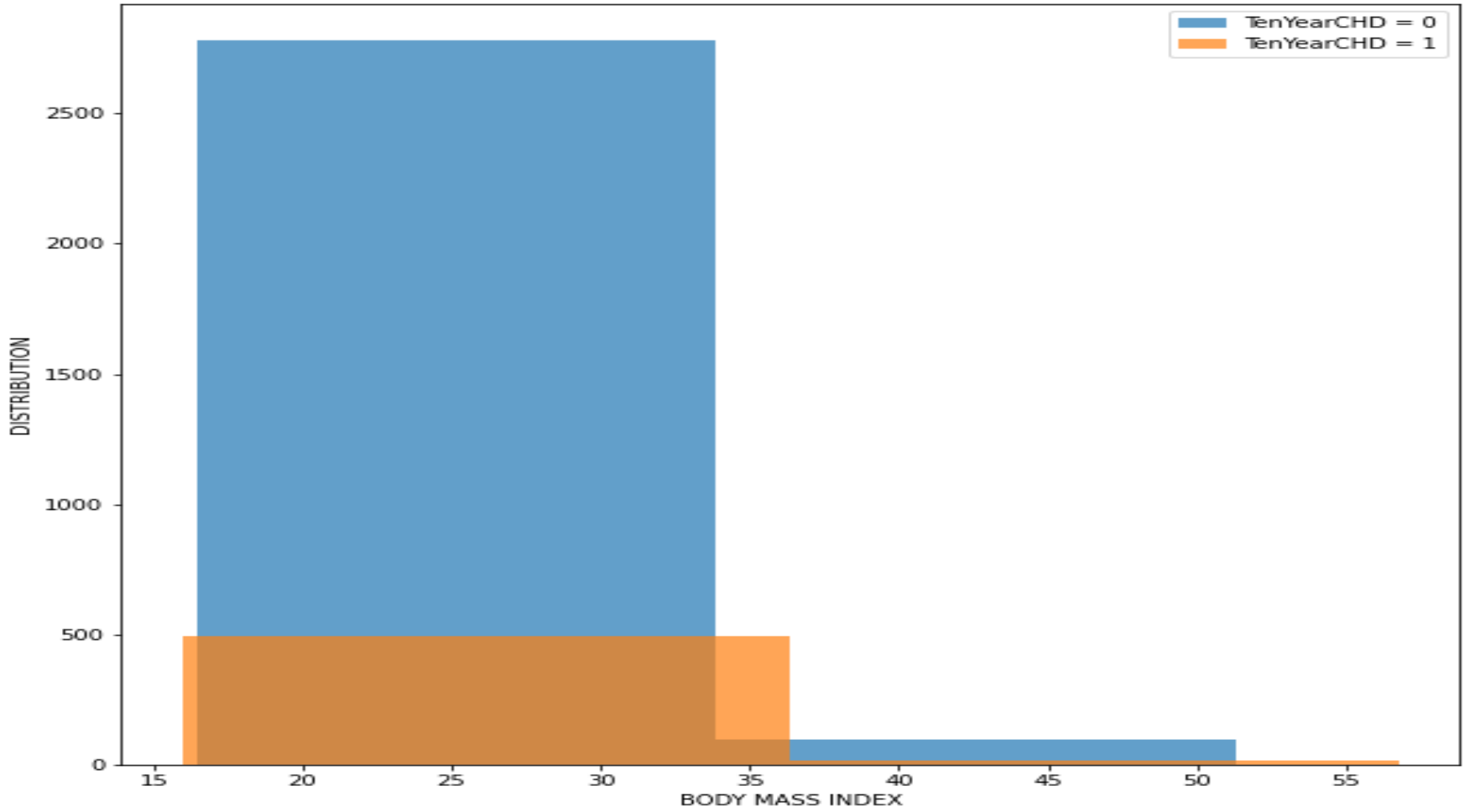
ARE PATIENTS WITH SYSTOLIC BP AT RISK OF CHD ?



ARE PATIENTS WITH DIASTOLIC BP AT RISK OF CHD ?

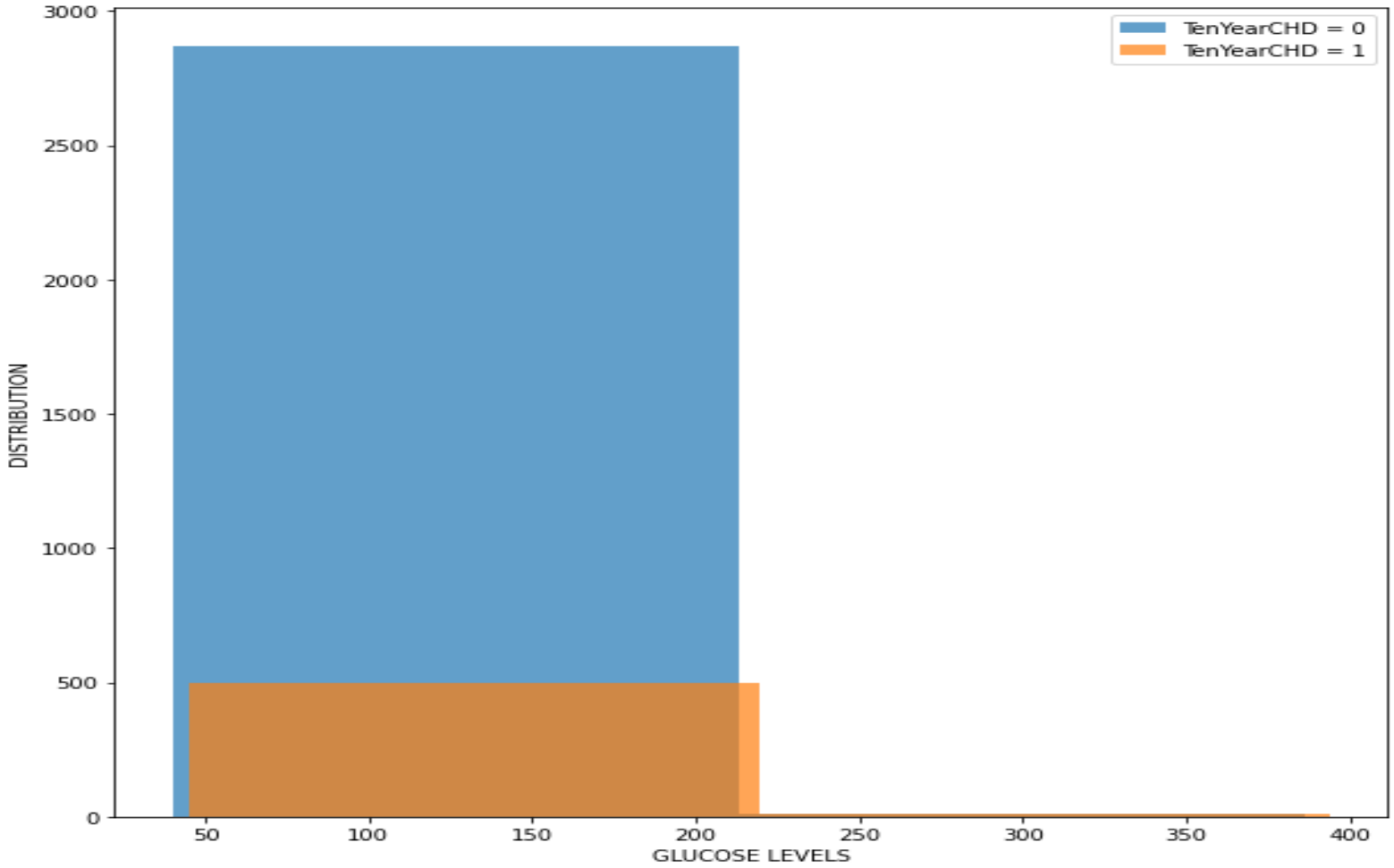


IS PATIENTS BMI IMPORTANT TO SHOW THE RISK OF CHD ?

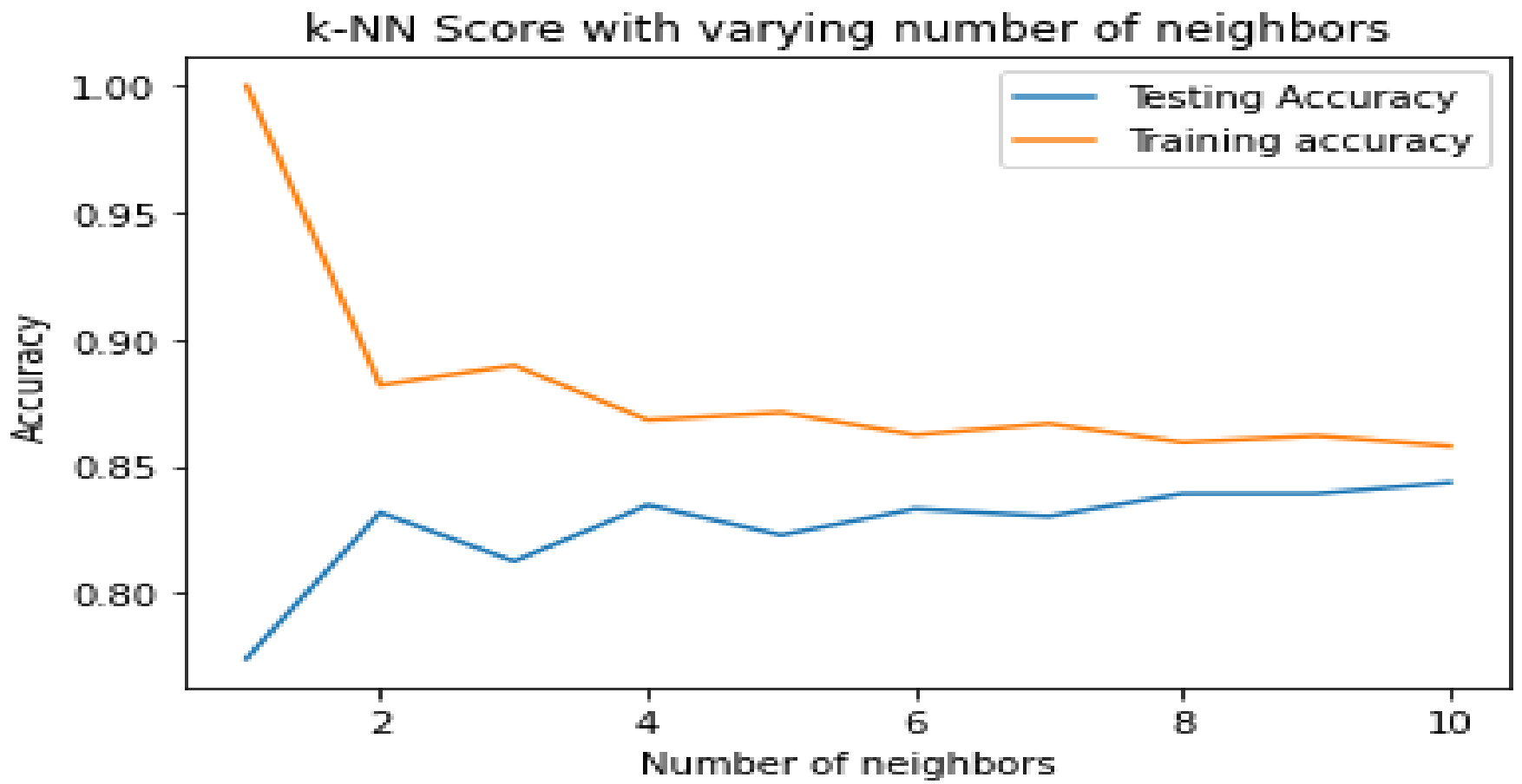


edunet foundation

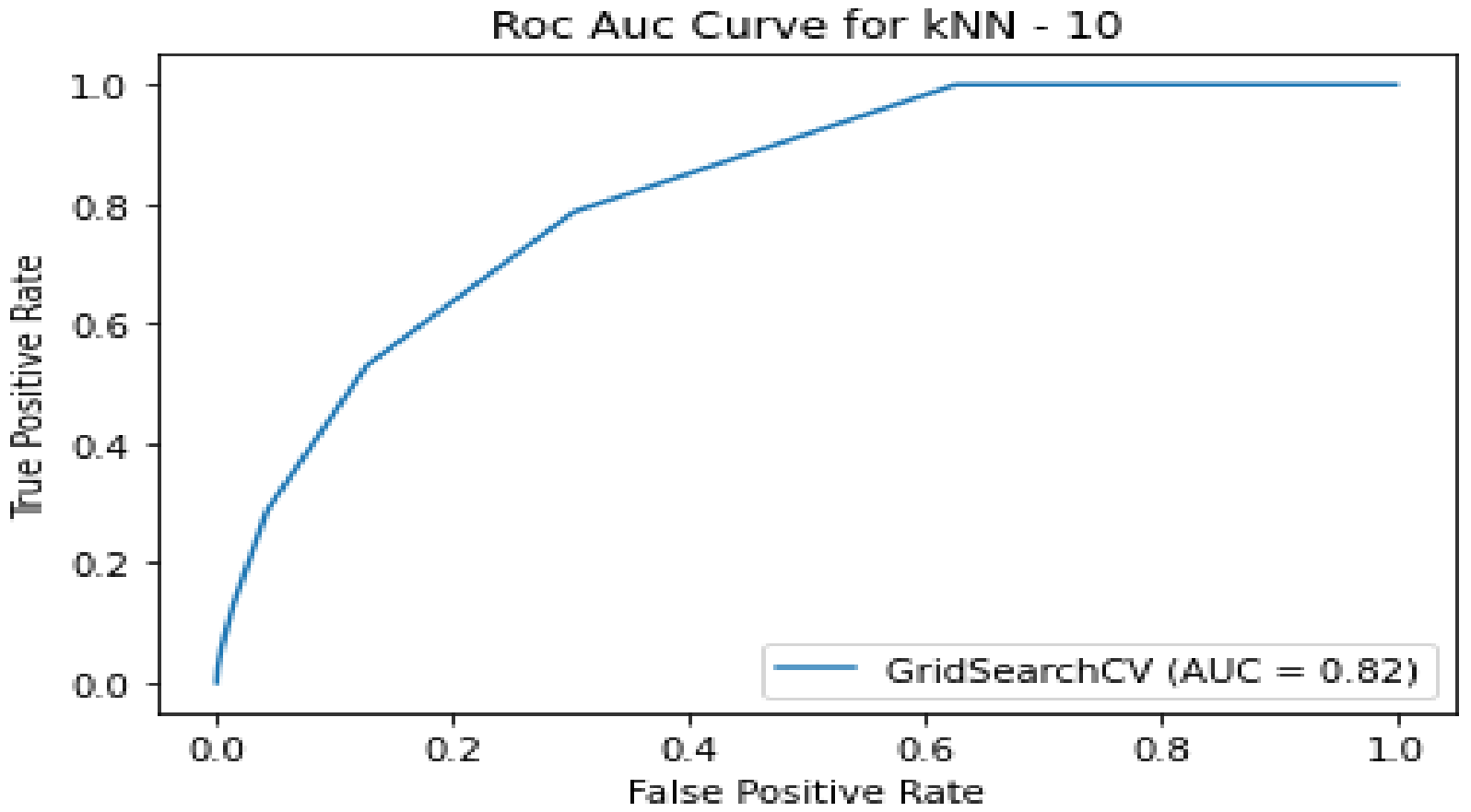
CAN PATIENTS' GLUCOSE LEVELS SHOW THE RISK OF CHD ?



K-NN SCORE WITH VARYING NUMBER OF NEIGHBORS



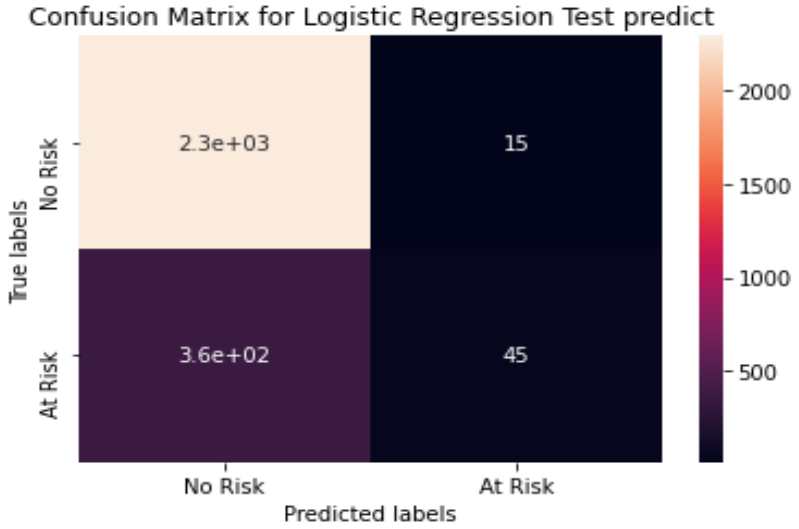
ROC AUC CURVE FOR KNN

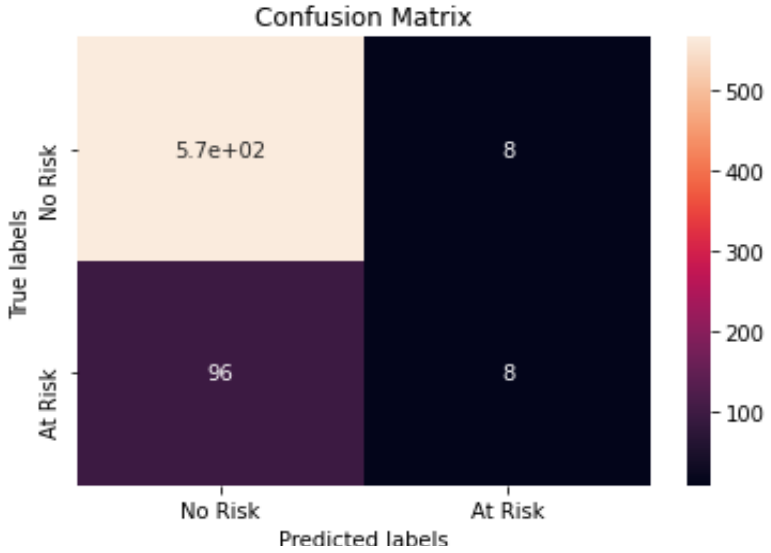


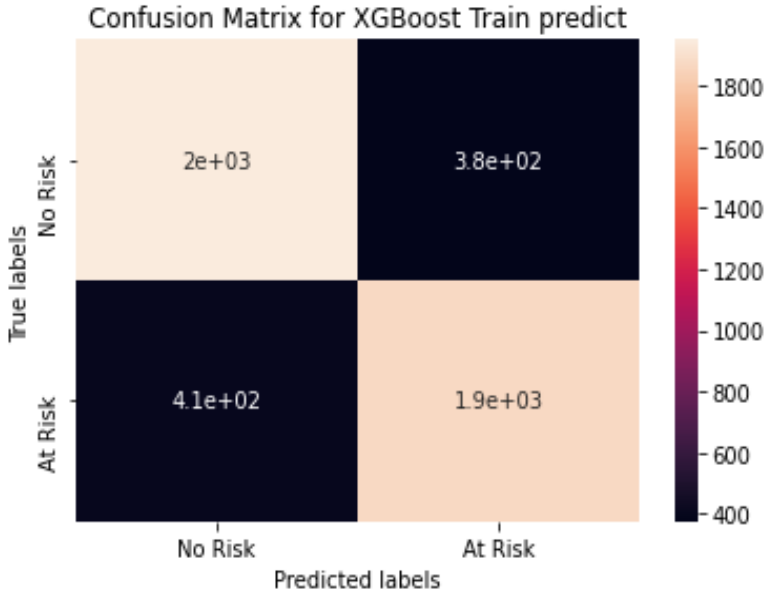
THE BEST FITTING MODEL: -

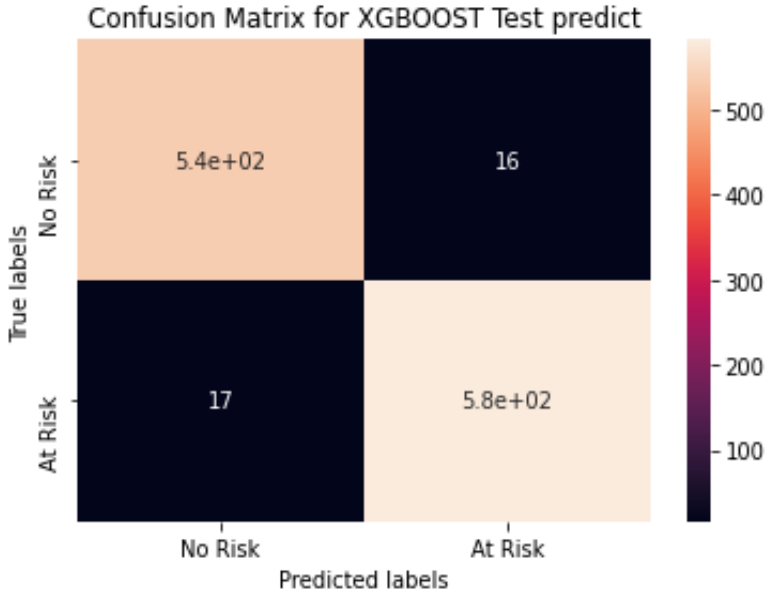
Sr.no	ML Model	Test Accuracy Score	Train Accuracy Score
1.	Naive Bayes Classifier	81	83
2.	KNN	84	86
3.	Logistic Regression	84	86
4.	Decision Tree	75	76
5.	Random Forest	89	99.8
6.	Gradient Boost	87	90
7.	XGBoost	97	83

CONFUSION MATRIX



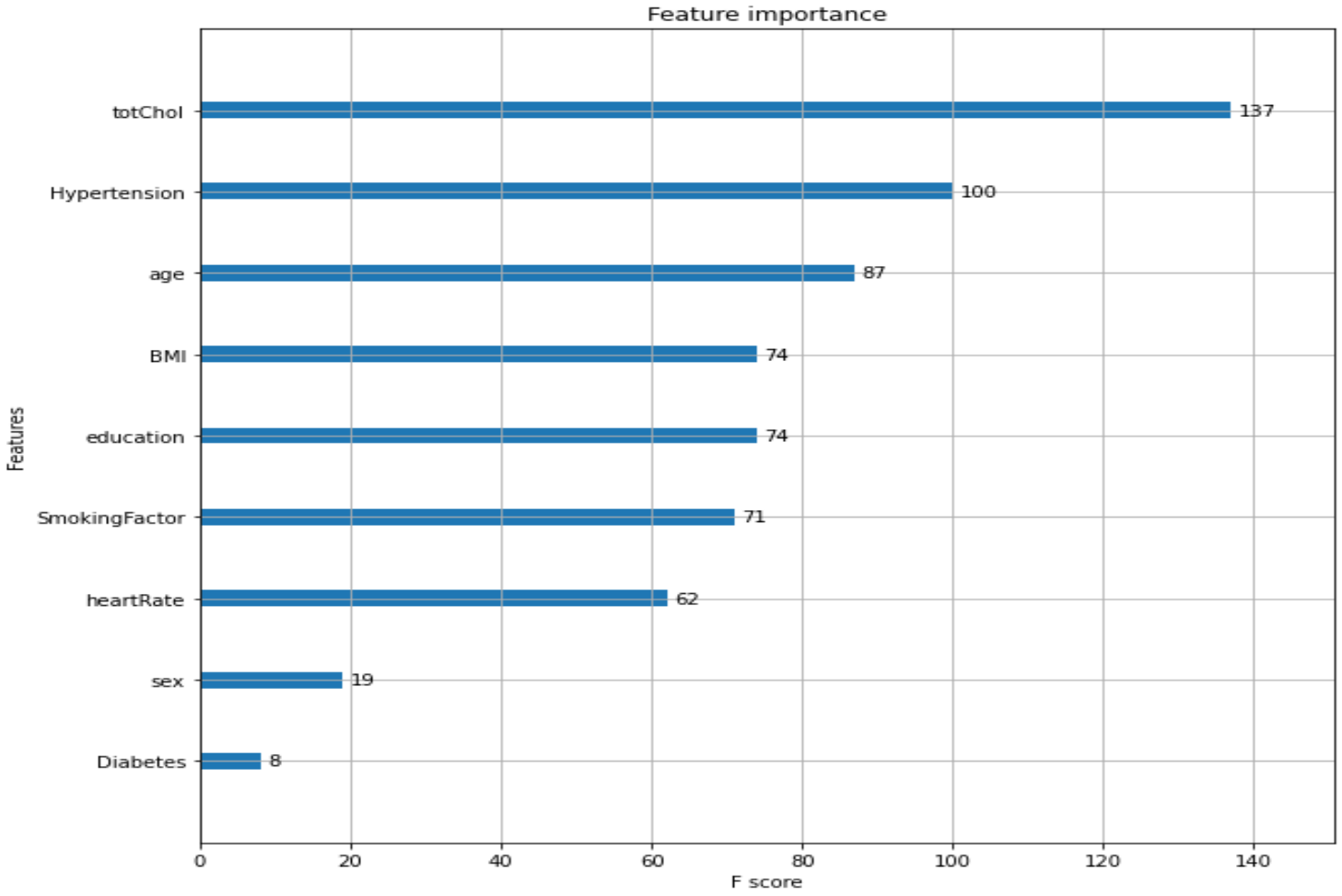






edunet foundation

THE FEATURE IMPORTANCE



RESULT

A cardiovascular disease detection model has been built using no of ML classification modelling techniques.

This project once deployed can possibly help predict the parents for cardiovascular disease based to their past medical history Blood pressure, Body mass index, Sugar levels etc.

The algorithms used in building the model are Logistic regression, Decision trees, KNN, Random Forest classifier, Naive bayes classifier, Gradient boost and XGboost.

The top three models with best accuracy are Gradient boost, Random Forest & XGboost with accuracy of 87%, 89%, and 97%, respectively.

CONCLUSION

- The project successfully developed a predictive model for CHD risk that can assist healthcare professionals in early diagnosis and treatment planning. The use of XGBoost, combined with careful data preprocessing and feature engineering, resulted in a model with high recall, crucial for minimizing the risk of undiagnosed CHD cases..

FUTURE SCOPE

- Incorporating additional risk factors such as genetic data
- Implementing real-time data processing for continuous risk assessment
- Expanding the model to predict other cardiovascular diseases

REFERENCES

KAGGLE DATASET:

https://www.kaggle.com/datasets/christofel04/cardiovascular-studydataset-predict-heart-disease

GITHUB :

https://github.com/AnandKumar56/Anand--Cardiovascular-Risk-Prediction

edunet foundation

COURSE CERTIFICATE 1

In recognition of the commitment to achieve professional excellence

Artificial Intelligence Fundamentals
IBM SkillsBuild

ANAND KUMAR DALWAIE

Has successfully satisfied the requirements for:

Artificial Intelligence Fundamentals

Issued on: 03 AUG 2024

Issued by IBM

Verify: https://www.credly.com/go/7N3S5KVU

IBM

COURSE CERTIFICATE 2

In recognition of the commitment to achieve professional excellence

Cloud Computing Fundamentals
IBM SkillsBuild

ANAND KUMAR DALWAIE

Has successfully satisfied the requirements for:

Cloud Computing Fundamentals

Issued on: 03 AUG 2024

Issued by IBM

Verify: https://www.credly.com/go/a1OUn57F

IBM

THANK YOU

edunet foundation