Exploratory and Predictive Analysis of Coronary Heart Disease & Diabetes

1st Arnav Jain
Computer Science, M.S. Student
Washington State University
Pullman, WA, USA
arnav.jain@wsu.edu
WSUID# 011809037

2nd Krupa Doranalu
Computer Science, M.S. Student
Washington State University
Pullman, WA, USA
krupa.doranalu@wsu.edu
WSUID# 118908406

***Abstract*—This project report presents a comprehensive exploration and predictive analysis of health outcomes, focusing specifically on Coronary Heart Disease (CHD) and Diabetes. This research leverages advanced analytical techniques and machine learning methodologies to gain insights into the factors influencing these critical health conditions.**

The exploratory phase involves an in-depth examination of diverse datasets encompassing demographic information, lifestyle factors, and general human health indices like BMI, age, etc. Descriptive statistics and data visualization techniques are employed to discern patterns, correlations, and potential risk factors associated with CHD and Diabetes. The subsequent predictive analysis utilizes machine learning algorithms to develop robust models capable of forecasting health outcomes based on the identified variables.

It contributes to the existing body of knowledge in health analytics by providing actionable insights for preventive healthcare measures. The findings are intended to aid healthcare professionals, policymakers, and researchers in understanding the dynamics of CHD and Diabetes, ultimately supporting the development of targeted interventions and personalized healthcare strategies.

***Index Terms*—Coronary Heart Disease (CHD), Diabetes, Framingham Heart Study, Exploratory Data Analysis (EDA), Predictive Modeling, Machine Learning, Logistic regression, Random Forest, Support Vector Machines (SVM), kNN, Decision Trees, Correlation Analysis, Visualization**

I. Introduction

This project stems from the critical need to delve into the intricate patterns and potential predictors associated with Cardiovascular Heart Disease (CHD) and Diabetes. As pervasive global health issues, these diseases not only pose significant threats to individual well-being but also strain healthcare systems worldwide.

The motivation behind this project lies in the desire to unravel hidden insights within the data, offering a deeper understanding of the factors influencing the occurrence of CHD and diabetes. By leveraging advanced data analytics techniques, our aim is to contribute to the ongoing scientific discourse on preventative measures and personalized healthcare strategies.

The main technical challenges involve navigating through the complexity of health datasets, addressing missing values, and developing robust predictive models that can handle the multifaceted nature of these diseases. Our solution approach integrates rigorous exploratory data analysis with sophisticated predictive modeling techniques, including machine learning and statistical approaches, to extract meaningful patterns and enhance the predictability of CHD and diabetes.

Our work awaits yielding both theoretical insights and empirical results, shedding light on potential risk factors and facilitating the development of targeted interventions for improved public health outcomes.

II. Problem Setup

The increasing prevalence of cardiovascular diseases, particularly coronary heart disease (CHD), and metabolic disorders like diabetes pose significant challenges to public health worldwide. Identifying the intricate relationships between various risk factors and health outcomes is crucial for devising effective preventive strategies and personalized interventions.

The primary problem addressed in this project is twofold: first, to conduct an in-depth exploratory analysis of the Framingham Heart Study dataset to uncover patterns and associations related to CHD, and second, to extend this analysis to a dedicated diabetes prediction dataset. The goal is to develop predictive models capable of identifying key risk factors and making accurate predictions regarding the likelihood of CHD and diabetes. Assumptions underlying this study include the assumption that the provided datasets are representative of the target population, that the selected features have significant predictive power, and that the relationships between variables remain consistent. Additionally, it is assumed that the models' generalizability extends beyond the dataset at hand. Ultimately, the project aims to contribute valuable insights to the understanding of cardiovascular health and diabetes, fostering the development of data-driven strategies for risk assessment and healthcare.

III. Solution Approach

In the initial phase of this project, a naive solution approach was adopted to tackle the analysis and modeling tasks on the CHD and Diabete dataset. For the Framingham dataset, we performed a basic exploratory data analysis (EDA) and modeling using common statistical measures and off-the-shelf machine learning algorithms. Similarly, for the diabetes

prediction dataset, we eployed a straightforward predictive analysis using conventional modeling techniques.

However, this rudimentary approach had its limitations. This solution lacked depth in terms of feature engineering, model optimization, and consideration of potential biases in the datasets. The models generated from this basic approach were not robust enough to handle the complexity and nuances present in the datasets, resulting in suboptimal performance and unreliable predictions. Even though the CHD data provided us with some really high accuracy scores they were too good to be true given the imbalance in data.

To refine the solution and overcome these drawbacks, we devised an enhanced approach. Feature engineering techniques, such as imputation for missing values and normalization of variables, were employed to preprocess the datasets. We introduced feature engineering through log transformations and standardization, ensuring a better representation of the underlying data distribution as well. TWe revisited the model selection process with a more systematic exploration of algorithms and careful tuning of hyperparameters to improve predictive performance. We used 5 different ML models namely. Logistic Regression, SVM, kNN, Decision Trees, and Random Forest. IN the Framingham Dataset in paritcular, we fine tuned the logistic regression model further using SMOTE to overcome the imbalance nature of the data.

The refined approach demonstrated significant improvements in model reliability and results having handled some of the variables and constraints of the data. The enhanced solution not only addressed the initial drawbacks but also provided a more comprehensive understanding. Despite this, our approach does have its own limitations. The models are still sensitive to variations in data quality, and potential biases in the datasets might affect the generalizability of the predictions. The while the predictive scores for the diabetes dataset are high, more data preprocessing and hypertuning of the model will give us more realistic and usable results.

IV. Exploratory Data Analysis (EDA)

The aim of our analysis here was to investigate a range of health-related factors and their interconnections to classify diabetes accurately. This comprehensive examination will not only provide insights into the patterns and trends in diabetes risk but will also create a solid base for further research. Specifically, research can be built on how these variables interact and influence of CHD and Diabetes occurrence and progression, crucial knowledge for improving patient care and outcomes in this increasingly critical area of healthcare. Below are some of our key results and more have inferences and visualizations have been made in our project code file on github, the link to which is provided in the References section -

A. Coronary Heart Disease

1) ***Dataset****:* The Framingham Heart Study, initiated in 1948, is one of the longest-running and most comprehensive car- diovascular studies globally. It is is a long term prospective study of the etiology of cardiovascular disease among a population of free living subjects in the community of Framingham, Massachusetts. This dataset includes data on over 4,240 participants and has 16 attributes (Some key features being Age, Sex, BP medication, Total Cholestrol, etc.). The longitudinal nature of the data enables the analysis of temporal relationships and disease progression.

2) ***Descriptive Statistics****:*

- **Age :** We can see that minimum age of subject found in given records is 32 while Max. being 70. So our values are ranging from 32 to 70.
- **cigsPerDay :** Subject smoking Cigarettes per day is as low as nill while we have 70 Cigs. per day making the Peak.
- **totChol :** Minimum Cholesterol level recorded in our dataset is 107 while Max. is 696.
- **sysBP :** Minimum Systolic Blood Pressure observed in Subject is 83 while Max. is 295.
- **diaBP :** Minimum Diastolic Blood Pressure observed in Subject is 48 while Max. is 142.
- **BMI :** Body Mass Index in our dataset ranges from 15.54 to 56.
- **heartRate :** Observed Heartrate in our case study is 44 to 143.
- **glucose :** Glucose sugar level range is 40 to 394.

3) ***Correlation Matrix****:* Correlation plot gives us valuable information regarding Relation within Attributes. It can either be Negative or Positive or Null. We need to always keep 1 feature from 2 strongly correlated ones but since we want to perform EDA we'll keep all and drop them before modelling.

- **currentSmoker and cigsPerDay** has strong correlation of 77 (Scaled for better Observations)
- **prevalentHyp vs sysBP / diaBP** are having positive correlation of 70 and 62.
- **glucose and diabetes** are postively correlated.
- **sysBP and diaBP** are also having positive correlation.

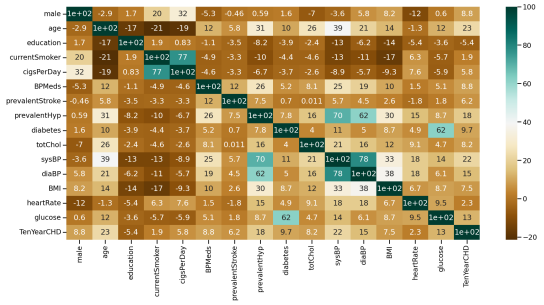

Fig. 1. Correlation Matrix of CHD

4) ***Distribution of key features****:* The plotted distributions reveal insightful patterns for key features. Glucose levels exhibit a right-skewed distribution, indicating a majority with normal levels but a notable minority with elevated readings. Total cholesterol follows a similar trend, with a right-skewed distribution emphasizing a substantial subgroup with higher

cholesterol. Systolic blood pressure and BMI distributions also skew to the right, suggesting elevated levels in a significant portion of the population. Conversely, diastolic blood pressure shows a close-to-normal distribution, indicating a prevalent normal range among subjects. Heart rate distribution appears approximately normal, with the majority falling within a healthy beats-per-minute range. Overall, these observations underscore the prevalence of elevated risk factors in a notable segment of the population, emphasizing potential cardiovascular health concerns.

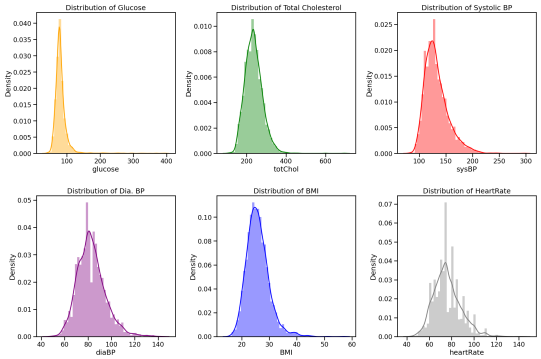

Fig. 2. Density Distributio

*5) **BP Analysis**:* The analysis of systolic blood pressure across age groups reveals a clear upward trend, indicating that median systolic blood pressure increases with age for both men and women. Notably, men exhibit higher systolic blood pressure than women on average within the same age brackets. For instance, the median systolic blood pressure for men aged 60-69 is 130 mmHg, compared to 120 mmHg for women in the same age group. In summary, the graph illustrates the age-related rise in systolic blood pressure and the gender-based distinctions, emphasizing that older adults generally have higher systolic blood pressure, and men tend to have higher average values than women across age groups.

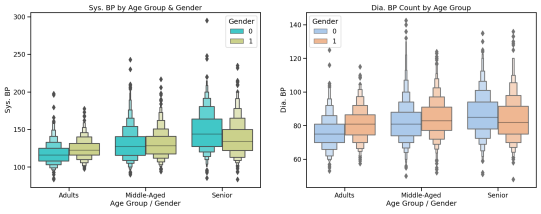

Fig. 3. (a) Sys. BP Vs. Age Group and Gender (b)Dia. BP Cunt Vs. Age Group

*6) **Cigarettes per day Analysis**:* The graph reveals distinct density patterns within different age groups:

- **Adults:** Median values show lower kernel density, with the 75% IQR following suit. In contrast, the 25% IQR exhibits higher density, indicating a specific distribution.
- **Middle-Aged:** Notably, the 25% IQR and median display higher kernel density, while the 75% IQR shows a lower density, suggesting a unique distribution pattern.
- **Seniors:** The senior age group demonstrates a distinct pattern, with the median and 25% IQR closely aligned and exhibiting higher kernel density. Conversely, the 75% IQR displays lower density, providing insights into the density distribution within this group.

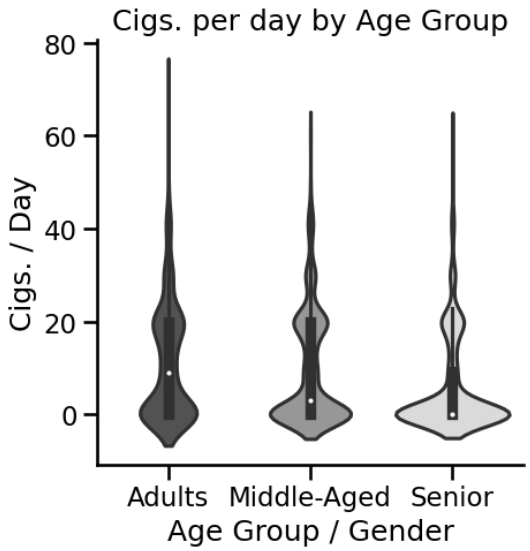

Fig. 4. Cigarettes per day Vs. Age Group

B. *Diabetes*

*1) **Dataset**:* This dataset is specifically curated for predicting the likelihood of diabetes occurrence, a prevalent metabolic disor- der with significant public health implications. It incorporates a range of features related to patients' health and lifestyle, enabling the development of predictive models for early di- abetes detection. It contains data of patients diagnosed with diabetes and non-diabetic individuals. The dataset consisting age, gender, body mass index (BMI), hypertension, heart dis- ease, smoking history, HbA1c level, and blood glucose level. The target variable is the presence or absence of diabetes

*2) **Descriptive Statistics**:* Notable statistics reveal a diverse age distribution with a mean of 41.80 years and a standard deviation of 22.46 years. The prevalence of hypertension and heart disease is relatively low, with means of 0.08 and 0.04, respectively. BMI exhibits a mean of 27.32 and a standard deviation of 6.77, with a wide range from 10.01 to 95.69. HbA1c levels and blood glucose levels display means of 5.53 and 138.22, respectively, showcasing variations in metabolic markers. The incidence of diabetes, with a mean of 0.09, indicates a relatively low prevalence in the dataset. These insights offer a comprehensive overview of the health characteristics, emphasizing the importance of considering factors such as age, BMI, and metabolic markers in the analysis of health-related datasets.

*3) **Correlation Matrix**:* The correlation matrix indicates strong positive correlations among age, BMI, HbA1c level, blood glucose, and diabetes status, suggesting higher values

in these variables are associated with an increased likelihood of diabetes. A weak positive correlation exists between gender and diabetes status, implying a slight male predisposition. These associations could stem from causal relationships, confounding factors like age, or possibly random chance, particularly in the weaker gender-diabetes correlation.

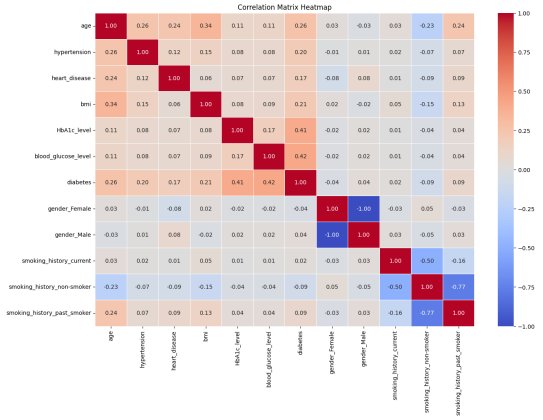

Fig. 5. Correlation Matrix for Diabetes dataset

*4) **Age Distribution**:* The 2 figures below represent the age distribution of samples under consideration that have (1) and do not have (0) diabetes.

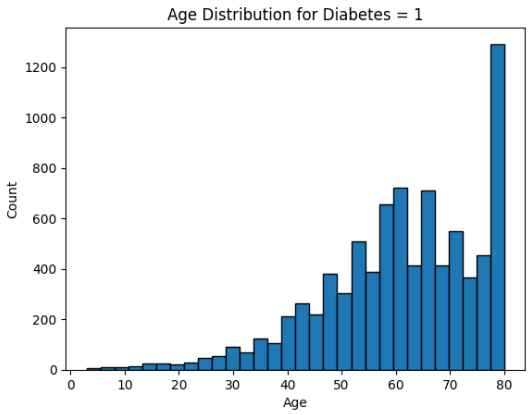

Fig. 6. Age Distribution when Diabetes = 1

*5) **Age Vs. BMI by Diabetes Classification**:* The scatter plot depicting age versus BMI versus diabetes reveals a positive correlation among these variables, indicating that older individuals are more likely to have both higher BMI and diabetes. The plot illustrates a stronger correlation between BMI and diabetes, emphasizing the significance of BMI as a more influential risk factor for diabetes compared to age. This shows that diabetes risk increases with higher BMI, underscoring the role of obesity in diabetes development.

*6) **BMI Vs. Diabetes Classification split by Gender**:* The analysis of BMI against diabetes classification split by gender shows that females are more likely to develop diabetes than males at a lower BMI. This is evident by the fact that the BMI

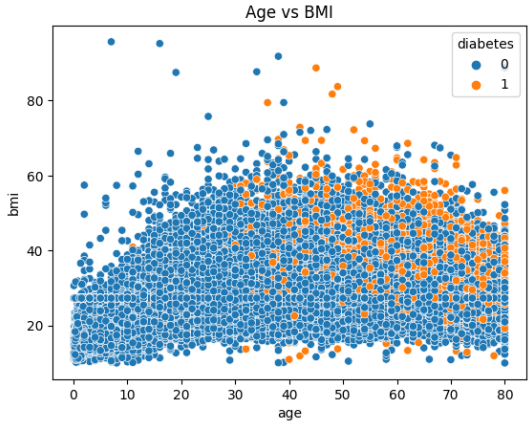

Fig. 7. Cigarettes per day Vs. Age Group

cutoff point for diabetes classification is lower for females than for males.

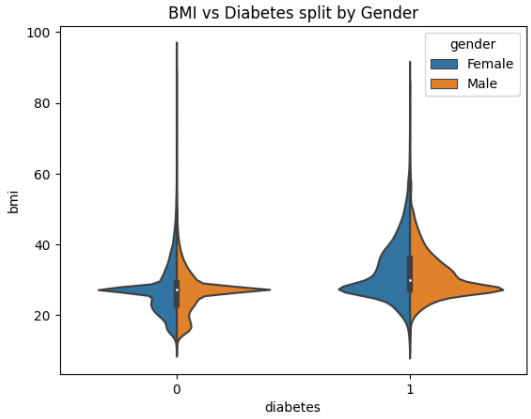

Fig. 8. Cigarettes per day Vs. Age Group

*7) **Interaction between Gender, BMI, and Diabetes**:* The boxplot shows the distribution of BMI by gender and diabetes status. The median BMI is higher for people with diabetes than for people without diabetes, for both males and females. However, the difference in median BMI between people with and without diabetes is greater for females than for males. This suggests that females are more likely to develop diabetes at a lower BMI than males.

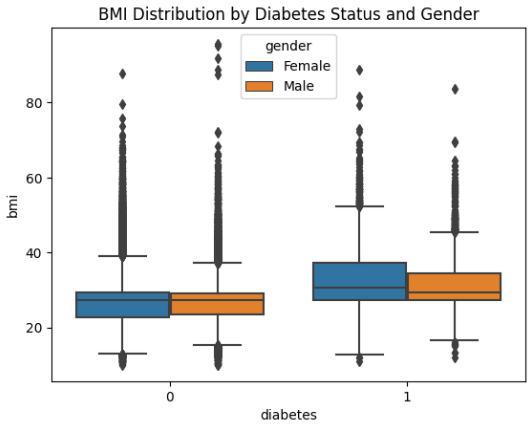

Fig. 9. Cigarettes per day Vs. Age Group

V. Predictive Analysis

We performed predictive analysis in a systematic approach to develop models that can predict the likelihood of CHD and Diabetes in a patient. Our approach involved the below step-by-step process -

1) **Data Collection and Exploration:** We collected the dataset from kaggle and performed the exploratory data analysis as discussed in the previous section to gather useful insights and understand the data in hand.
2) **Data Cleansing and Preprocessing:** Upon studying the data we realized that it had several missing data and noise that needed to be addressed which is why we preprocessed the data by handling the outlier, performing feature scaling, and encoding categorical variables to prepare the data for modeling.
3) **Feature Engineering and Selection:** We incorporated feature engineering to further refine the data and make it usable to build a classification model by employing imputation strategy for missing values, logarithmic transformation and standardization on continuous feature variables, and normalization of variables. We then identified and selected relevant features like Age, Sex, total cholesterol and more for the predictive model, focusing on variables with the highest impact on predicting coronary heart disease and diabetes.
4) **Data Splitting:** The datasets were divided into training and testing sets, with the training set used to train predictive models and the testing set for evaluation.
5) **Model Selection and Training:** After having a fully readily available data to work with, we adopted 5 different classification models to predict the health outcomes, namely, Logistic Regression, SVM, KNN, Decision Trees, and Random Forest. These models were then trained using the training dataset, adjusting hyperparameters and configurations for optimal results.
6) **Model Evaluation:** We finally rigorously evaluated all the models to check which gave us the most robust model for the given data by using metrics like accuracy, precision, recall, and area under the ROC curve, ensuring a thorough assessment of their effectiveness.
7) Fine-tuning and Optimization: For the Framingham Heart Study dataset, we found Logistic Regression to have the highest accuracy of about 86% so we built further on that model by adding class weight parameter to handle the imbalance nature of the dataset. The observations made from the metrics then led us to evaluating the model further by testing after over-sampling. So we performed oversampling using SMOTE, and re-evaluated the model.

Below are the results from our predictive analysis for the 2 health outcomes (Confusion Matrices with Recall, Precision, and F1 scores can be seen as outputs in our code in GitHub) -

A. *Coronary Heart Disease*

*1) **Confusion Matrix: Logistic Regression (Pre oversampling)**:* The accuracy score in an imbalanced dataset can be deceptive, as it may mask poor performance on the minority class. While achieving an accuracy of 0.8603 might seemed high (accuracy observed when training the model on the imbalanced data), the model's misclassification on the minority class became apparent. Cross-validation scores help assess over/underfitting, and the classification report emphasizes the importance of precision and recall, revealing a notable imbalance in recall scores between the majority and minority classes. This error can be highlighted as Type 1 errors in negative diagnoses and misclassifications in positive diagnoses.

To deal with this we added a class weight parameter. The addition of the Class Weight Parameter to the estimator significantly improved the recall score, addressing the issue of Type 2 errors in the confusion matrix. Then, for the positive class, the model correctly classified 108 instances with a reduction in misclassifications to 44. The decrease in Type 2 errors is crucial, as misclassifying actual cases of CHD as negative could pose serious consequences, making the model more suitable for deployment in production.

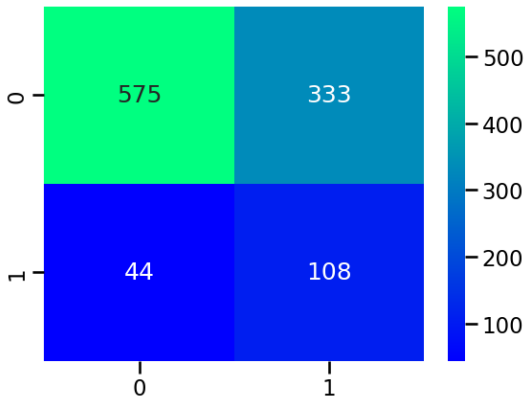

Fig. 10. Confusion Matrix: LR (Post adding class weight: balanced)

*2) **Confusion Matrix: Logistic Regression (Post over-sampling using SMOTE)**:* We then implemented SMOTE to create synthetic observations based on existing minority observations. Implementing SMOTE with a sampling strategy of "not majority" effectively balanced the class distribution, resulting in 2688 examples for each class. The confusion matrix shows improvement in Type 1 errors, with 581 correct classifications for the negative class and a reduction to 44 misclassifications for the positive class. However, Type 2 errors remain a concern, emphasizing the need for further refinement to ensure the model's reliability for production deployment.

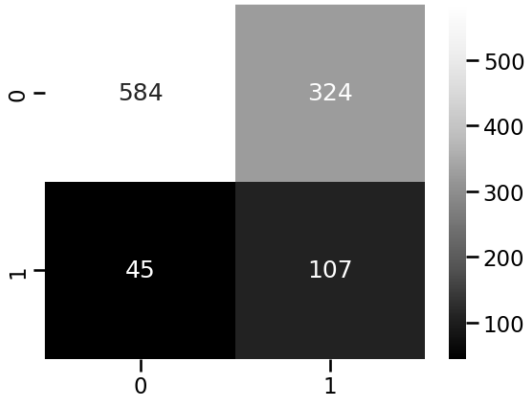

Fig. 11. Confusion Matrix: LR (Post oversampling using SMOTE)

*3) **ROC-AUC (Receiver Operating Characteristics - Area Under Curve)**:* ROC-AUC is a performance metric for classification, quantifying a model's ability to distinguish between classes. A higher AUC indicates better separation between true positives and false positives. The ROC curve, depicting true positive rate (TPR) against false positive rate (FPR), serves as a visual representation. In medical contexts, a superior AUC implies the model's effectiveness in distinguishing individuals with a condition from those without.

The ROC-AUC scores for two models indicate their performance, with the Logistic Regression model using "balanced" class weight achieving a decent score of 0.7376. Following closely is the Logistic Regression post Over-Sampling with a score of 0.7369. Higher ROC-AUC scores signify better model performance.

*4) **Accuracy Scores**:* The graph below in the figure shows the accuracy scores of different classification models. The x-axis shows the classification model, and the y-axis shows the accuracy score. The graph shows that the Logistic Regression model has the highest accuracy score of 86.04% and Decision tree model gives us the least accuracy score of 77.17%

*5) **Cross-Validation Scores**:* The graph in the figure below illustrates the comparision between the CV scores. The x-axis shows the model type, and the y-axis shows the cross-validation score. The cross-validation score is a measure of how well a model is able to generalize to unseen data.

The graph shows that the Logistic Regression model has the highest cross-validation score 84.84%, and Decision Tree has the least with a value of 76.10%. This suggests that the Logistic Regression model is the best model for generalizing to unseen data from the Framingham Heart Study dataset.

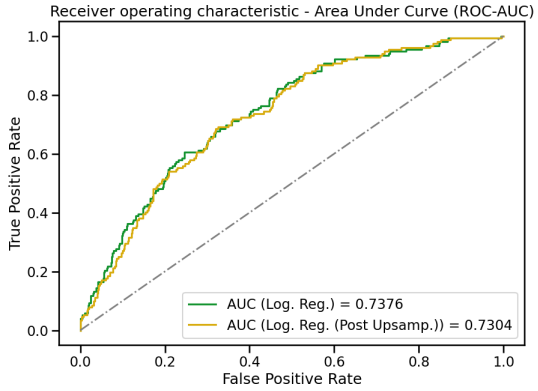

Fig. 12. ROC

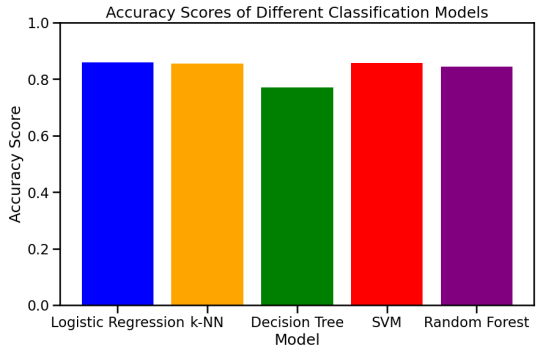

Fig. 13. Comparing Accuracy scores of all 5 models

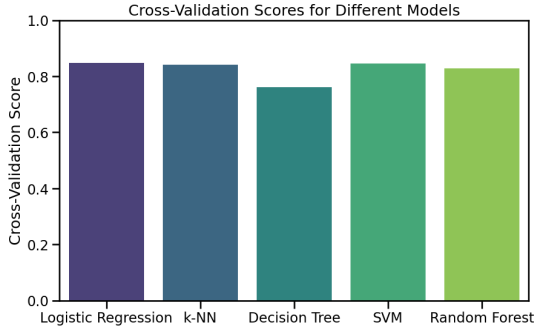

Fig. 14. Comparing CV scores of all 5 models

B. *Diabetes*

*1) **Accuracy Scores**:* The graph in the figure below illustrates the comparison between the accuracy scores. The x-axis shows the model type, and the y-axis shows the accuracy score. The accuracy score is a measure of how well a model is able to generalize to unseen data.

The graph shows that the Random Forest model has the highest accuracy score 96.5% This suggests that the Random

Forest model is the best model for generalizing to unseen data from the diabetes dataset.

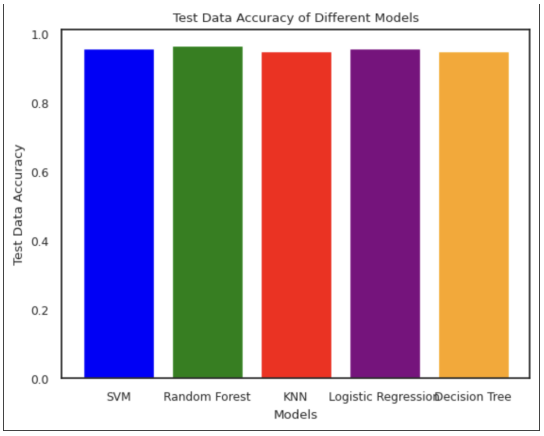

Fig. 15. Comparing Accuracy scores of all 5 models

*2) **F1 Scores**:* The graph in the figure shows the F1 scores of different classification models. The x-axis shows the classification model, and the y-axis shows the F1 score. The graph shows that the Random Forest model has the highest F1 score of 77.6%.

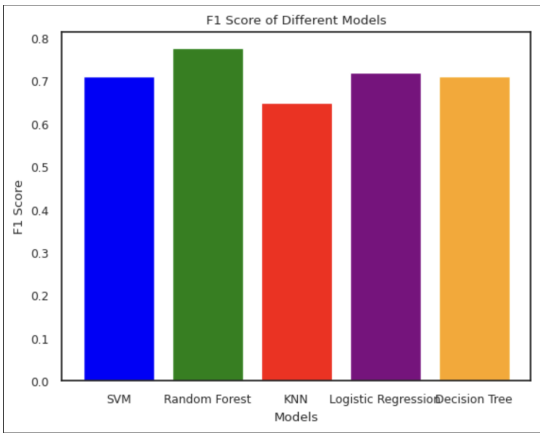

Fig. 16. Comparing F1 scores of all 5 models

VI. Future Work

While our work has made significant strides in leveraging machine learning and data science techniques to explore and predict health outcomes, there are several avenues for future research and refinement. One key area for potential enhancement involves the incorporation of more advanced machine learning models and ensemble techniques. Exploring deep learning architectures, such as neural networks, and ensemble methods like stacking, could provide a more nuanced understanding of complex interactions within the datasets. Additionally, the evaluation of emerging algorithms and methodologies could further enhance the predictive accuracy and robustness of the models.

Our work has primarily focused on static analyses, treating health-related variables as fixed over time. Future work could involve dynamic modeling, considering the temporal evolution of risk factors and their impact on health outcomes. Time-series analysis, recurrent neural networks (RNNs), and other dynamic modeling approaches could offer insights into the evolving nature of cardiovascular health and diabetes prediction. Incorporating more granular temporal information may also contribute to the identification of critical intervention points for preventive healthcare strategies.

Furthermore, the scalability and generalizability of the developed models should be a focal point for future research. Extending the analyses to larger and more diverse datasets, encompassing various demographic and geographic contexts, would enable the creation of models that are more universally applicable. This expansion would not only enrich the feature space but also open avenues for the exploration of dynamic changes in health parameters over time.

VII. Conclusion

In conclusion, we represented a comprehensive exploration of the Framingham Heart Study dataset and a dedicated diabetes prediction dataset. Through extensive exploratory data analysis (EDA) and the application of five distinct classification algorithms, we have endeavored to unravel patterns, relationships, and predictive insights within the domains of coronary heart disease (CHD) and diabetes. The integration of advanced data science methodologies has allowed us to contribute meaningful insights to the broader field of health informatics.

The findings from our exploratory analysis shed light on the intricate relationships between diverse health variables, providing a nuanced understanding of the risk factors associated with CHD and diabetes. The predictive models generated through the implementation of machine learning algorithms offer a glimpse into the future of healthcare, showcasing the potential for early risk assessment and intervention strategies.

The future of this field holds exciting possibilities, from the incorporation of advanced machine learning models to the integration of diverse data sources for a more holistic understanding of health.

References

[1] Kannel, W. B., Castelli, W. P., Gordon, T., & McNamara, P. M. (1979). Serum cholesterol, lipoproteins, and the risk of coronary heart disease. The Framingham Study. Annals of Internal Medicine, 90(1), 85-91.

[2] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, 15, 104-116.

[3] Al-Masni, M. A., Al-Absi, H. R., Kang, B. H., & Wei, D. (2018). Identifying type 2 diabetes risk factors using data mining techniques. Journal of Healthcare Engineering, 2018, 7 pages.

[4] Choi, E., et al. (2021). Deep learning for risk prediction of coronary heart disease using the Framingham Heart Study dataset. Journal of the American Heart Association, 10(9), e019916.

[5] Sun, J., et al. (2022). A hybrid machine learning model for diabetes prediction based on clinical and genetic data. Journal of Diabetes Research, 2022.

[6] Github Repository link - Click here!