National College of Ireland

Advancing Chronic Disease Analytics by Predicting Cardiovascular Disease Risk Based on Demographic and Health Factors in the US

MSc Research Project
Data Analytics

Tej Patel Yeliyuru Ramu
Student ID: x23216077

School of Computing
National College of Ireland

Supervisor: Cristina Hava Muntean

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Tej Patel Yeliyuru Ramu
Student ID:	x23216077
Programme:	Data Analytics **Year:** 2025
Module:	MSc Research Project
Supervisor:	Cristina Hava Muntean
Submission Due Date:	25/01/2025
Project Title:	Advancing Chronic Disease Analytics by Predicting Cardiovascular Disease Risk Based on Demographic and Health Factors in the US
Word Count:	8463 **Page Count** 19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Tej Patel Yeliyuru Ramu
Date:	25/01/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Advancing Chronic Disease Analytics by Predicting Cardiovascular Disease Risk Based on Demographic and Health Factors in the US

Tej Patel Yeliyuru Ramu
X23216077

Abstract

Cardiovascular Diseases (CVD) still remains as one of the leading causes of death in the United States. The current predictive models often give different results about how age, gender, race, and location impact the risk of CVD. Studying and research on these factors can support our healthcare providers in targeting disease prevention on high-risk populations in the US and help reduce the risk of CVD in the society. In this research we start by analysing how demographic and other health related factors are dependent on CVD risk, focusing on the demographic diversity in the US. Our main goal is to develop different analytical machine learning models and compare them with different parameters that finds out significant risks across these demographic groups, which could improve the understanding of CVD risk distribution. We use a large healthcare dataset from a US government website named US Chronic Disease Indicators and apply different data mining, machine learning models such as Logistic regression, Gradient boosting, Random Forest, SVM, KNN and analysis techniques to see investigate these factors are related to CVD. We measured and analysed the model's accuracy, sensitivity, and specificity. Logistic regression, Random Forest and Gradient boosting demonstrated commendable consistency with a score of 0.84 across all metrics, such as accuracy, precision, recall and F1 score showing a balance between model complexity and its performance. The findings of this study will give us a data driven basis for modifying CVD risk and which could increase the effectiveness of preventing healthcare programs in different groups across the United States.

1 Introduction

Globally, cardiovascular disease (CVD) is the second most fatal disease in the world after that of cancer with high mortality rates in most developed nations such as the US, with 32 percent deaths caused due to CVD Walther, (2023). A few familiar risky components like smoking, obesity, high blood pressure result in CVD, however the influence of age, gender, race, even geography still remain insufficiently covered. Some research done states that men have greater heart disease risk in younger ages, while racial inequalities along with discrimination among African Americans and Hispanics leads to hypertension conditions Harvard Health. (2019). Other than that, geographical differences due to nutrition habits, way of life and healthcare status also modify some outcomes. Although age, sex, and genetic predispositions are known CVD risk factors, studies examining the specific impacts of race

and regional variations within the US remain limited. Filling these gaps is fundamental not only for improving the performance of machine learning models in regard to predicting the risk of CVD, but also to provide possibilities to apply personalized and targeted relevant campaigns to the public's health so that the prevalence of the disease is reduced significantly

Cardiovascular disease accounts for high costs in treatment, hospitalization, or long-term care and it impacts an overweight and aging population that should benefit from prevention efforts. This can be addressed through early identification of high-risk individuals, allowing healthcare providers to take preventive strategies such as lifestyle alterations, medications, routine monitoring that can effectively lower the overall burden of CVD. Current diagnostic capabilities rely on traditional methods, like medical imaging and laboratory tests, that have downsides, they require skillsets that are not always present, and often analyze current symptoms, meaning that they may miss more subtle risk factors over time Gabriel S Tajeu et al. (2024). However, ML models present a great alternative as they explore large, complicated data sets and reveal what is concealed. Such models have the ability to link together structured health data like cholesterol levels, blood pressure, or medication with unstructured treatment populations such as age, sex, and place. This broad-based approach makes risk assessment better since some individuals at risk may otherwise be missed. As more data is being generated, predictive power of ML models is also getting enhanced leading to a more efficient, targeted strategy in CVD's preventive measures and reducing the stress on the healthcare system.

This research aims to demonstrate whether demographic indicators work in predicting CVD risk amongst the U.S. adults with the study scope including variables such as age, sex, race, and region i.e. East or West. Also different Machine learning models such as Logistic regression, random forest, gradient boosting, SVM and KNN are built and compared to predict the outcome by utilizing the CDI dataset (the United States Chronic Disease Indicators), this research utilizes the above mentioned machine learning algorithms, which mostly relies on binary classification to establish CVD risk using blood pressure and heart disease mortality rates as the key factors. Such demographic factors are very important as they have been shown to bear a lot of weight on health experiences with some populations being more at risk for CVD due to genetic factors or social, environmental or even health care influences. This paper uses qualitative and quantitative methods to achieve its objectives by analysing these variables, which enables the study to be able to determine high risk individuals and appropriate strategies to be used with those population. The goal of the investigation is to focus on the connection between data analysis and its application in public health, providing knowledge that could assist in lowering the rate of CVD and ensuring quality healthcare. The objective of this study is to prevent the advancement of chronic diseases by identifying the combined influence of health risk factors and demographics, which is less well-known in the literature at the moment

This study aims at answering two main research questions in relation to risk prediction of CVD. The first question is *To what extent do demographic factors, including age, sex, race, and regional location (East or West), statistically impact the likelihood of cardiovascular disease among adults in the United States? What measurable patterns and regional disparities are observed in CVD risk prediction based on these factors*. The second question is *Among machine learning models such as logistic regression, decision trees,*

random forests, support vector machines and K-NN which model demonstrates the highest accuracy, precision, recall, and F1 score in predicting cardiovascular disease using health and demographic data from U.S. adults? The objective of the research is to train and introduce several ML algorithms based on demographic and health data categories to be explored include logistic regression, random forests, Gradient boosting, support vector machines and KNN. Therefore, the research also seeks to evaluate the effectiveness of these models and determine the best algorithm to use in estimating CVD risk so the healthcare stakeholders can design specific interconnected mechanisms aimed at reducing CVD prevalence in the United States.

This report investigates how machine learning can be used in healthcare to improve the prediction of CVD. Section 2 examines the current studies on CVD risk, the methods of machine learning, and aspects relating to demographics which sets a base for this study. In section 3 is about methodology in which various forms of advanced algorithms in machine learning are used and creative procedures applied for creating the predictive models. Section four discusses the work method which involves data preparation, choosing of features and strategies to assess for creating a strong model. The fifth section shows us the outcomes of different models against each other in terms of their performance and ability in predicting CVD risk. The last section concludes how the models are evaluated and can be utilized practically within the healthcare domain and suggestions aiming advancement in future research to mitigate threat from CVD.

2 Related Work

Alanazi (2022) studies the application of machine learning approaches to classify and forecast chronic illnesses. The author discusses the fact that chronic diseases must be diagnosed in their early stages in order to avoid negative effects. In the proposed model feature extraction is performed using CNN and KNN for disease prediction. The types of the data used are both structured and unstructured, hence the paper shows that the higher accuracy is reached by using both data types to make predictions, than using Naive Bayes or logistic regression. The study focuses on CNN and KNN, by limiting exploration of other advanced algorithms that might give better results. Also, depending on the unstructured data create biases if the data is not uniformly collected or preprocessed.

As seen in the article titled “Chronic Kidney Disease Prediction Using Robust Approach in Machine Learning” (Anurag *et al*, 2023, p 01) have developed a machine learning model in which it predicts CKD based on laboratory data. The paper uses the features selection methods with classification methods to enhance the prediction accuracy and obtains a performance of 89 percent. The authors show us the connections between the correct CKD forecasts and the stoppage of the disease and its impact on the patient's health. They stated that with the possibility of continuously introducing new data into the ML models, it could be useful in helping make better decisions for treatment of CKD. But using such datasets limits the application of the models to different populations or situations and it does not discuss the challenges that may arise when deploying such models in real world clinical procedures. And, because it was not practically validated its findings are limited in actual practice.

In their work in Prediction of the chronic kidney disease with implementing Artificial Neural Network algorithms Madhavi Devi Botlagunta *et al.* (2023) shows how the ANN model predicts CKD given data such as age, blood pressure, or creatinine levels. This model was evaluated on 400 samples with 25 characteristics and produced an accuracy of 99% and 0.99 F1 score. The authors prove that applying the ANN models for early detection high risk individuals can be treated earlier and enhancing the treatment results. They also employed features selection and model selection in order to improve the accuracy of the prediction. However, the model rely on a very small dataset raises concerns about its ability on large populations, and the study does not show the practical challenges of applying ANN systems in real world.

(Chioma Susan Nwaimo *et al.* 2024, p 01) in their paper "Transforming Healthcare with Data Analytics, Predictive Models for Patient Outcomes" read about the importance of predictive modelling in improving healthcare. It also shows the importance of improving different patient related resources, including electronic health records and other genetic information. To increase the patient's outcome as well as staff productivity in order to enhance personalized care. The authors explain how predictive analytics can be used in clinical management including disease avoidance and identifying the right resources at the right time. They also answer the data collection difficulties and ethical concerns in relation to the use of these technologies. However, there is a specific issue about the majority of the factors covered in the report ignoring the problem of how variations in access to healthcare technologies and analytics tools could increase gaps in the areas that are still in question

The comparison of analysis of predictive healthcare models, PARAMO and machine learning for CKD prediction has two different approaches to healthcare analytics. According to Kenney Ng *et al.* (2013) the PARAMO platform uses a Map Reduce approach to analyze EHR data making 800 models in three hours only. Sultana Habiba *et al.* (2024) suggest a machine learning framework for CKD prediction with the accuracy of more than 98% using six important characteristics. Therefore, both methods increase predictive healthcare measures by optimizing speed and precision. Whereas for PARAMO, high performance platforms can be inaccessible to smaller institutions, and CKD model largely relies on specific datasets cannot be used in other cases. The Analyses presented in these works shows the need of developing scalable and flexible approaches to the analysis of health care systems and their performance.

(Divya Jain and Vijendra Singh, 2018) focus on feature selection and types of classification techniques important to chronic disease prediction. They focus on the process of feature extraction and techniques of increasing the classification rate, such as traditional, adaptive and parallel classification systems. They also argue that when combining two or more approaches to conflict, the results are far better. In another study, (Dr Macha Sarada and Dr Ralla Suresh, 2024) evaluated the application of machine learning where CNN stands for Feature Extraction, and KNN as Disease Classification. This is due to which their models are based on this specific type of data which works better than other models, including Navie Bayes and decision trees. As a result, both papers show how advancements in feature selection and machine learning algorithms enhance clinical decision making and early diagnosis. It also has certain drawbacks, such as the need for a large and high-quality dataset

and also handling multiple data of different format in accurate predictions is also another problem.

In the study of "Intelligent Digital Platform for Diabetes Management Using Big Data Technology" Xiangyong Kong *et al.* (2022) introduced a unique system that uses big data methods with Hadoop to address diabetes. It advances the integration of data, privacy and medical research. It also gives solutions for diagnosis and treatment through predictive modelling. In a similar way, Ying Zheng *et al.* (2024) "Risk Prediction Models of Depression in Older Adults with Chronic Diseases" focuses more on machine learning based risk prediction models for depression in older people. By adding a few characteristics into the selection of risk factors, the authors were able to find and classify the comparison between the male and female groups. The two papers discuss how big data and machine learning could enhance the health care system but there is space for improvement. The diabetes study is not so clear on how it might vary across different diseases, while the depression study has issues of imbalance datasets and predictability. The future work on this study could address these gaps to improve clinical uses.

M. Vijay Kumar *et al.* (2024) in their work "Chronic Disease Prediction Using Machine Learning" discuss the ensemble learning algorithms like random forest, gradient boosting and XGBoost which are used to predict chronic diseases including the diabetes, heart disease, and liver disorders. In the same way, Gregorius Airlangga (2024) in his study of "Comparative Analysis of Machine Learning Models for Intrusion Detection in IoT Networks" focuses on the machine learning models such as gradient boosting, random forest, logistic regression, and multi-layer perceptron to identify the attacks on IoT's. These two studies show us how ensemble learning can be improved in performance and accuracy for cybersecurity and healthcare. However, there are other issues must be dealt with, including overfitting, specialization, biases, and the lack of validation with real world data. Other important issues are statical views like, how well the solution extends across different problem situations, and how solutions are measured. Future study should answer these problems to improve the model's stability and usage.

"Machine Learning Algorithms to enhance the Value of Predictive Analytics in Chronic Disease Management-Potential of Pharmacogenomics" written by (Gopesh Kumar Bharti and Deshmukh Vaishnavi Jaikumar, 2024) has the ability of integrating with a pharmacogenomic approach to deal with the chronic disease prevention. They have used the Deep Convolutional Neural Networks (DCNN) to predict results based on genetic data. They also show that by using Chronic Disease Kaggle dataset, the DCNN model has the better accuracy than traditional ML methods. The authors point out the proposed method based on the use of an external DCNN to other similar ways of solving the problem and use them for the development of individual treatment protocols. Improved decision-making requires integrating genotypes and EHR. However, the study has limits such as the need for more data and the problems faced in environments with less resources.

2.1 Summary of literature review

The existing literature discusses different approaches with reference to forecasting chronic diseases like CNN, KNN, ANN, ensemble learning techniques and many more. Some

research findings show us that combining feature extraction and classification techniques has enhanced the prediction accuracy, with CNN and KNN models with good results. My research covers a key gap by comparing various machine learning models to find the best method, by using a larger number of current demographics and health indicators to estimate the CVD risks. The findings show that there are possibilities of developing more accurate and effective patient risk analyses to support clinical decision making. Also, data set inaccuracies, sample selection bias, and the issues that arise when applying predictive models in medical practice is addressed.

3 Research Methodology

Is it possible to precisely predict the risk of CVD using machine learning techniques? How do factors such as age, gender, ethnicity, and geographic location increase CVD risk among adults in the United States? Also, which machine learning models can provide the most reliable predictions, and how do they differentiate in the terms of accuracy, sensitivity, and specificity? An organized approach is required to answer these issues, and using the CRISP-DM framework is helpful.

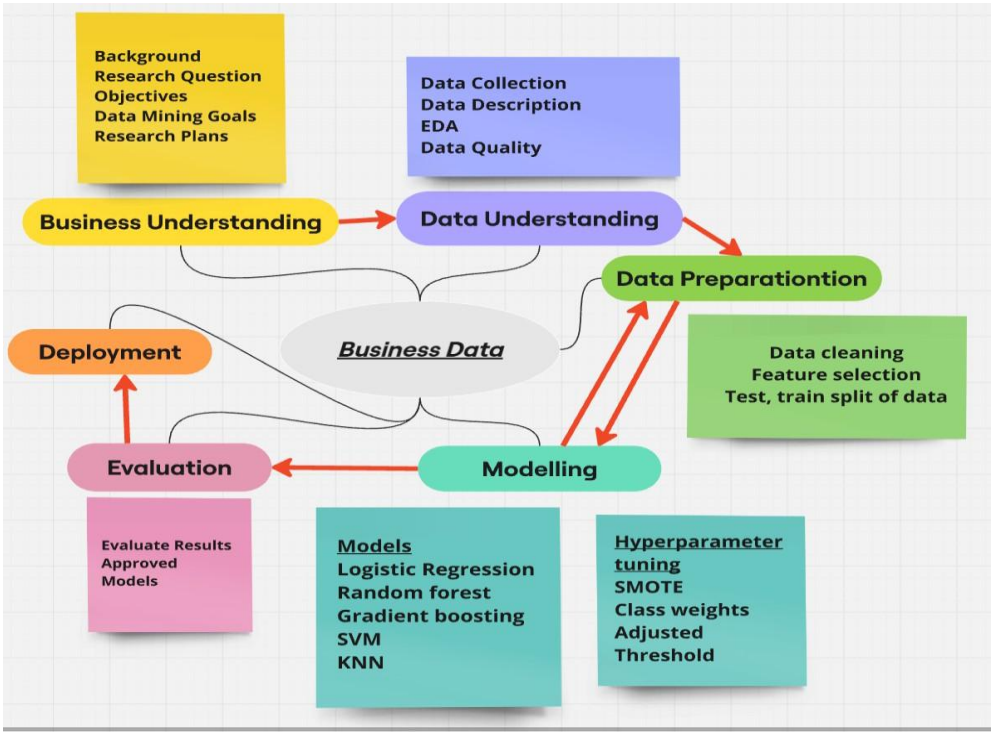

Figure 1: Diagram explaining the adoption of CRISP - DM Methodology with its steps involved

As shown in Figure 1, I have chosen CRISP-DM for this project as it provides a structured way to resolve data research issues. The process begins with Business Understanding in which I identify the purpose of forecasting CVD risk. Then, in Data Understanding I investigate the demographic and health factors that affect CVD. Data Preparation ensures that the data is clean and ready to be used for machine learning. During Modelling I have chosen

5 ML models and compared them to see which one is most suitable and finally in Evaluation and Deployment the best models are chosen and considered for its practical use in healthcare.

3.1 Business Understanding

It is important to understand the CVD risk since it is social issue and involving clinical characteristics such as age, sex, race, and region. The demographic influences are relevant to this study when considering the CDI dataset. The goal of applying machine learning technologies is mainly to identify and highlight the riskiest elements in order to develop additional preventive measures. When performing the study, these guidelines are followed during data preparation and modelling, as well as the evaluation of the predictive models.

3.2 Data Understanding

3.2.1 Data Collection and Data Description

The dataset being used in this study is the "US Chronic Disease Indicators" dataset which is taken from the data.gov website Services, (2024). It was published by the Centres for Disease Control and Prevention and the data has information from the year 2015 to 2022. It contains around 310,481 entries with 34 features. These features include many health indicators like blood pressure, cholesterol levels, and heart disease mortality rates. The Demographic data such as age, sex, race, and region are also included. This diversity of data helps analyze the link between demographics and CVD risk. The dataset was available in CSV format making it easy to use with analysis tools. It provides us a wide range of populations across the U.S. This data is collected using standardized health indicators and reporting systems.

The CDI dataset assures consistency and quality of data collected across various regions. The data covers for a long time enabling us to make appropriate conclusions with regards to its trends. This data is more useful for studying how different demographic factors influence the risk of CVD. The sample size of 310,481 records makes it possible to perform an in-depth analysis of the results. This is helpful for developing different machine learning models and it will also be useful in evaluating the CVD risk in various groups using these models.

3.2.2 Exploratory Data Analysis

EDA helps us in understanding the dataset structure, revealing the patterns and detecting if there are any variances. Initially I started by checking the missing values and found that columns such as Response, StratificationCategory2 and Stratification3 had no data, whereas DataValue and LowConfidenceLimit contained some gaps. This shows us the inconsistencies in our dataset. Visualizing the dataset revealed that categories such as Diabetes and Mental Health had appeared more often than others. Analysing the locations found irregular distribution across areas with some states providing more data.

Figure 2 shows the histograms used to determine how the numerical variables were distributed. One of the distributions DataValue and HighConfidenceLimit is skewed and has a number of outliers which can be seen in the above figures. According to the generated

heatmap, which shows us only relatively small linear dependence between the numerical variables, the majority of the variables were statistically independent. The data distribution through time looked to be evenly distributed over the years and time intervals between them showing temporal coverage. But differences in the demographic factors such as race or age developed especially between the stratified groups. Frem these insights provided we had a deeper look into the structure of the data and where the data need to be improved.

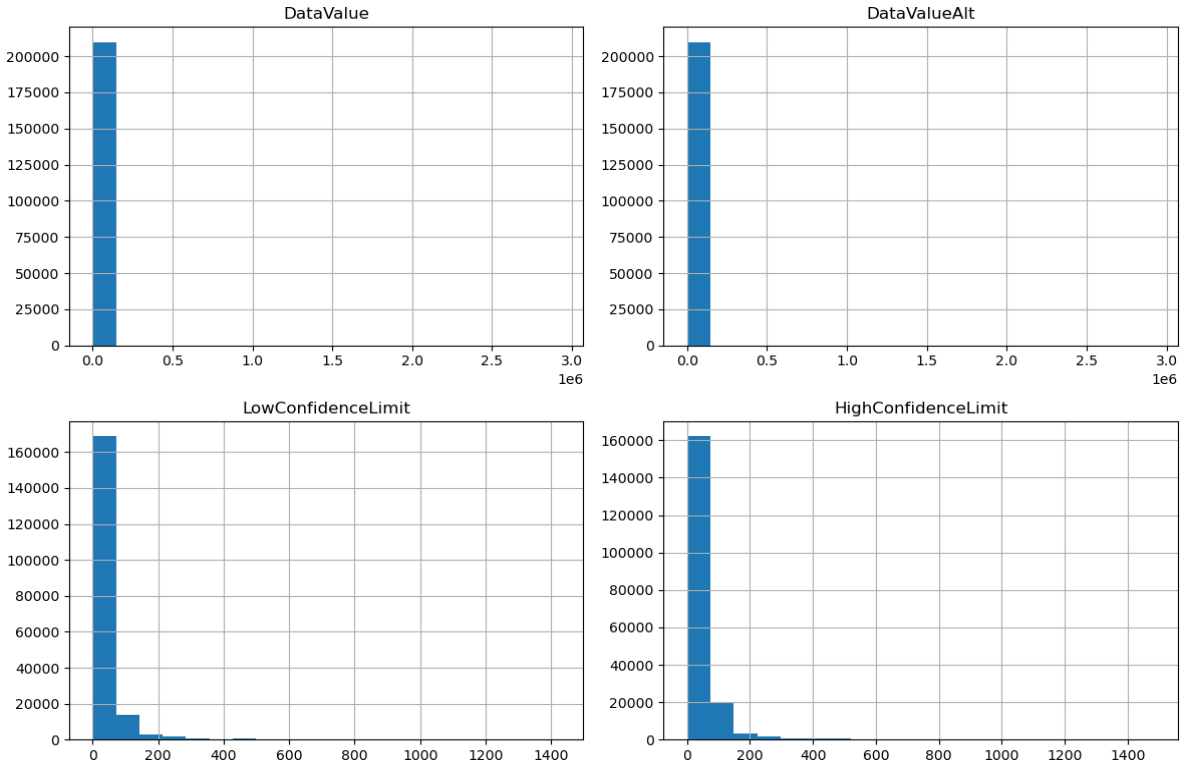

Figure 2: Distribution of Numerical Variables

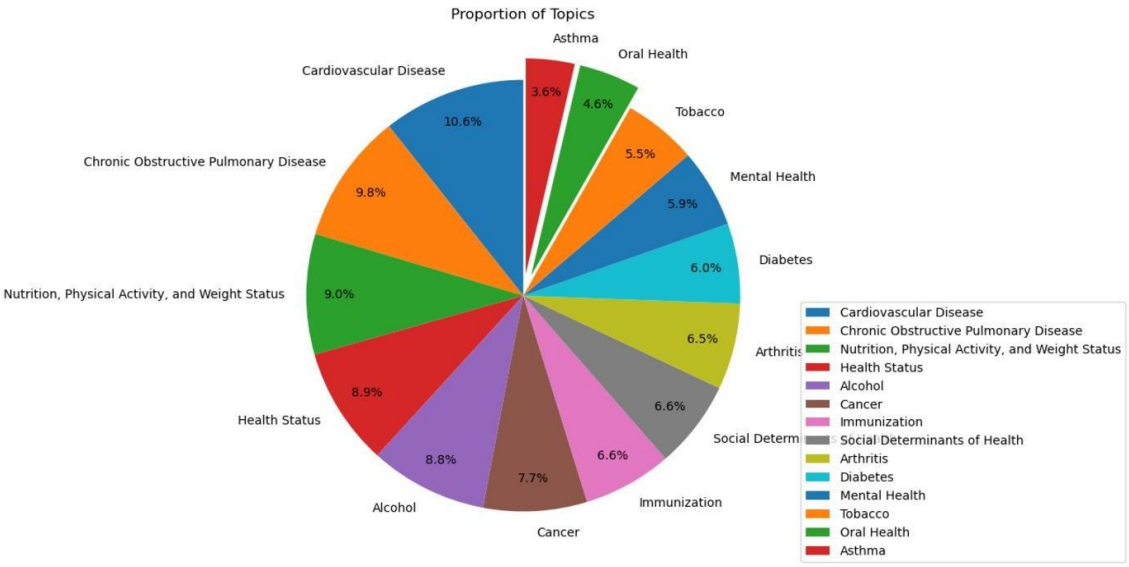

Figure 3: Proportion of Different topics in the dataset

In figure 3, the pie chart shows the distribution of topics in the dataset, with cardiovascular disease being the largest category at 10.6%, followed by chronic obstructive pulmonary disease & nutrition weight status at 9% each. A very smaller proportions are seen for topics such as Asthma 3.6% and Oral Health 4.6%. The chart shows us the dataset focus on different public health concerns while showing untouched areas. It provides insight into the analysis on major topics like cardiovascular diseases.

3.3 Data Preparation

Data cleaning is an important phase for our study in the data analysis process. It makes sure that the dataset is free of errors and consistent. The dataset had missing values, duplicates and inaccuracies that could have affected the results. The dataset had missing values in many columns. To fill in the gaps caused by the missing values, I considered the median of each column. In the Age column where a large amount of the data was missing, the median of the Ages 35 was used to fill in the gaps. This method was chosen because the median was less affected by extreme values, which makes it a accurate way to fill the missing data in the skewed distributions.

After dealing with missing values the next was to remove duplicate records. The dataset had some rows that were identical in all the columns. Duplicates were directly deleted to make sure the datasets quality was good and confirmed that each entry was unique. After the data cleaning process, the total number of rows in the dataset decreased from 310,481 to 310,460.

Another crucial step in data preprocessing is the Feature selection, which helps in improvement of the model by eliminating the noises and increasing the training time features. In the CDI dataset there were a huge number of topics related to different health problems, as out study focus mainly on the CVD risk, only those columns were filtered considered. The demographic information such as Age, Sex, Race and Region were included because they provided the important insights into risk factors. Other demographic variables which were non-informative were excluded. The feature selection and feature elimination method were also based on the correlation between each of the features and the target variable which included only those characteristics that correlate with CVD risk. This will reduce the model building process, improves efficiency and might increase the predicting accuracy of machine learning models.

After performing data cleaning and feature selection from the dataset, the total of 310,460 rows of data were split into the data for training and testing so that the model should be able to generalize on any unseen data. Initially the data was split into 70 and 30 ration, and later modified. The data from 248,368 rows was used to train the developed models and 62,092 rows were used to test the model's assessment. StandardScaler was done to both the training and testing datasets before model building, as SVM and KNN depends on the magnitude of its features. This is to make sure that the model has first been trained on say 70 percent of the data and the remaining 30 percent used to evaluate the model's ability to make the predictions of cardiovascular disease risk and its performance when tested on data it has not previously seen.

The data was pre-processed as needed for machine learning models other than handling missing values and dividing data into train and test datasets. Categorical variables such as Sex, Race and Region were encoded. As machine learning models need numerical input variables these categorical features were converted to numerical by using the one hot encoding method. The variable Sex was converted into two binary values 1 for Male and 0 for Female. And the qualitative variables Race and Region were also first converted into numerical variables. Another process was normalization of data. Age had higher variability while others had low variability. Since all the features should be given the same importance, the data was normalized to bring all the values to a small scale often in the range of 0 and 1.

3.4 Modelling

In this study I have used Logistic regression to predict the patients risk of CVD based on several health and demographic variables from the dataset. As logistic regression is a popular classification algorithm, it works good for binary outcomes, in which the objective is to predict one of the two classes. The way Logistic regression works is by creating a connection between the probability of an event and the independent variables. It's mainly used because of its ability to predict the health risks and determines the chances of having CVD or not in a patient

Random Forest is used for this research because of its properties such as an ensemble method for decision trees. It also aims at reducing the error rate of the model by averaging the results across multiple trees. This is mainly useful when dealing with many input features, here the Random Forest operates through development of many decision trees using randomly selected samples of data and then making average predictions. This helps us in developing a stable model for estimating the risk of CVD by overcoming aspects of nonlinearity in the features Shelf (2024).

Gradient boosting is another powerful Machine learning algorithm which is considered for this research. It was chosen because of its ability to handle both the regression and classification tasks effectively. It works by adding new trees to correct the errors that occurred in the previous ones, which makes it a perfect model for predicting the outcomes with complex patterns. In this research, Gradient boosting will detect the correlations between many health and demographic parameters and refine the model for many iterations to improve accuracy in predicting the CVD.

Support Vector Machine is considered for this research as it is suitable for high dimensional spaces as well as in the classification of large data. SVM has been told to have high accuracy of classification in machine learning problems. Much of the previous work is done with simple models which include Logistic Regression or even Random Forest. This is because SVM can be time consuming for large data sets and working with a complex health identifier needs the right parameters. However, the use of SVM model for this research is because of its capability of handling nonlinear relationships while at the same time being able to handle noisy data.

KNN is used in this study because of its simple use and can deal better with classification problems. It has been applied only in a few healthcare scenarios, its performance in identifying risks of CVD has not been explored. While several models have

previously been used in different research works, KNN has not been widely applied to CVD prediction because most other studies consider more complicated models such as Random Forest or Gradient boosting. Another drawback of using this model is that when working with big data, the computation of distance between the test sample and the entire sample is very time consuming. This was the reason why KNN is used for this research and to find its capability if it fits well in the low data pattern which is very important in predicting the risk of CVD

4 Implementation

This section presents the implementation of the different machine learning models considered for this research project. It describes how the machine learning models have been implemented for its performance evaluation to predict the risk of CVD. It trains and evaluates a few models on several metrics like accuracy, precision, recall and F1-score, after the division of the dataset into a training set and a test set. The metrics will help us guide the decision in model selection. The data was split into different ratios such as 70-30, 80-20 and 75-25, and several preprocessing techniques were used to optimize these models, including scaling and SMOTE.

Before starting with the implementation phase let's look at the tools and libraries used throughout the research. The dataset was initially stored in a CSV file and was first reviewed and organized using Microsoft Excel. Later, it was imported into Jupyter Notebook which was the main platform for writing and running Python code for this research

In this study the feature selection and the design of features are more important for CVD risk assessment. I have performed correlation analysis to check for the presence of a relationship between features and the target variable, also to remove unimportant predictors, if there are any. In the domain, specific filtering methods were done by consideration to age, sex, race and area codes since they are known to influence cardio health. Using variable importance measures from the Random Forest and Gradient Boosting models, important features were identified which have high performance power. There is also a work of diversity, seeing how features perform for various CVD subgroups as a check that any input to models contributed to meaningful prediction and as a check of equity. In the case of feature transformation, constructing nominal features, sex, race, and region was encoded with the one-hot encoder, while age and health indicators for the patients were scaled uniformly with other feature sets for uniformity with different machine learning algorithms. This process has not only removed useless variables from the dataset, but also set them up for optimal training and prediction of the model which fits our research objectives perfectly.

In this whole research, a list of libraries was used to manage different parts of the analysis. Initially, Pandas was used for the data manipulation helping to read the CSV file and deal with the missing values in our dataset. NumPy was used for the numerical calculations during the data preprocessing. To visualize the distributions in our dataset I used Matplotlib and Seaborn in which seaborn gave some more advanced visualizations to see the relationships in the dataset. Scikit - learn was the main for the machine learning tasks making model training and testing for algorithms like Logistic regression, Random Forest, Gradient Boosting, SVM, and KNN. Also, it was used in the evaluation matrices such as accuracy,

precision, recall, F1-score, and ROC-AUC. It even helped with hyperparameter tuning with GridsearchCV and RandomizedSearchCV and helped in splitting data to train and test sets. To deal with the class imbalance, Imbalance _ learn was applied to use Synthetic Minority Over Sampling Technique (SMOTE). All these libraries ensured the data handling, model training and evaluation was carried out efficiently for this research

5 Evaluation

5.1 Model testing setup and evaluation

5.1.1 Performance Analysis with 70-30 Data Split

The first step in modelling was to split the dataset in a ratio of 70-30, where 70% would be used for training and the remaining 30% for testing. This split is normally done in machine learning so that the model generalizes well to the unseen data. Class 0 represents non-CVD cases, and class 1 represents CVD cases. The results after models were evaluated had a lot of concerns. From the table 1 of Comparison of Model Performance Metrics, Logistic Regression had an accuracy of 84.06%, but looking into it in detail, it is seen that the model has not been able to identify non-CVD cases in class 0 since its precision, recall, and F1-score for Class 0 are 0. This also shows us that the model has a bias toward predicting Class 1 CVD. Random Forest achieved an accuracy of 43.02% which was significantly lower while having a recall of 0.98 for Class 0 but with a very low precision of 0.22. These results showed a high rate of false positives, which means there was a tendency to predict Class 0. The performance of the SVM model was very poor with an accuracy of 15.94%, and this was because of its poor prediction capability with class 1. KNN had achieved an accuracy of 81.78% with a low recall for Class 0 showing the poor identification of non-CVD cases by the model. These results indicate that the initial split had caused class imbalance, which also affected the performance of these models in terms of correctly predicting non-CVD cases.

Table 1: Comparison of Model Performance Metrics (70-30 Split)

Algorithm (70-30 split)	**Accuracy**	**Precision**		**Recall**		**F1-Score**	
		Class					
		0	**1**	**0**	**1**	**0**	**1**
Logistic Regression	0.8406	0	0.84	0	1	0	0.91
Random Forest	0.4302	0.22	0.99	0.98	0.33	0.35	0.49
Gradient Boosting	0.8393	0.38	0.84	0.01	1	0.02	0.91
SVM	0.1594	0.16	0	1	0	0.28	0
KNN	0.8178	0.31	0.85	0.12	0.95	0.17	0.9

5.1.2 Model Behavior with Increased Training Data (80-20 Split)

Next, the data was split into 80 and 20 in which 80% of the data was for training and 20% for testing. This gets more data fed into training and allows these models to learn more about the data. As seen in Table 2 the accuracy of Logistic Regression was 84.55% and recall in class 1 goes up. Precision in class 0 was really low at 0.69, the recall in class 0 drops drastically to

0.03. It means that even with the increased size in the training set the model didn't work well with the class imbalance. Random forest gave similar results as Logistic Regression, it had an average accuracy of 84.04%. Recall for Class 1 stayed high at 1.0 while precision for Class 0 was very low again which says that the model completely failed at predicting non-CVD cases. Gradient Boosting was as strong as the Random Forest with an accuracy of 84.04% performing equally in both precision and recall. The SVM model slightly increased its accuracy to 84.08% with still issues in imbalance, especially in predicting Class 0. KNN had unchanged accuracy of 81.21% performing equally in precision and recall. Here the performance of most models improved in comparison with the 70-30 split, but the issue of imbalance was still the same.

Table 2: **C**omparison of Model Performance Metrics (80-20 Split)

Algorithm (80-20 split)	**Accuracy**	Precision		Recall		F1-Score	
		Class 0	**Class 1**	**Class 0**	**Class 1**	**Class 0**	**Class 1**
Logistic Regression	0.8455	0.69	0.85	0.03	1	0.06	0.92
Random Forest	0.8404	0.27	0.84	0.01	1	0.02	0.91
Gradient Boosting	0.8404	0.27	0.84	0.01	1	0.02	0.91
SVM	0.8408	0.39	0.85	0.02	0.99	0.04	0.91
KNN	0.8121	0.23	0.85	0.08	0.95	0.12	0.89

5.1.3 Balancing Classes with SMOTE Implementation (80-20 Split)

After applying the SMOTE on the training data for the 80-20 split, to handle the class imbalance and generate samples from the minority Class 0, the resulting models give better performance in learning from the minority class. The accuracy for Logistic Regression dropped significantly after the SMOTE was used, all the way down to 54.2%. But the recall for Class 0 increased to 0.5, the precision for Class 1 was worse, and the model seemed unable to provide a better balance between the two classes. This drop shows that there was overfit to samples and not good enough hyperparameter tuning in this area. Random Forest, Gradient Boosting, and SVM showed similar results after SMOTE. With an improved recall for Class 0, the overall performance did not see much improvement. These results indicate to us that SMOTE did not effectively solve the problem of imbalance without further optimization of the models. A clear numeric representation of the above data is shown in the below Table 3. KNN had an improved recall for Class 1 after SMOTE and precision for Class 0 remained low, saying that the class imbalance issue was still the same for a few models even after balancing the data.

Table 3: **C**omparison of Model Performance Metrics with SMOTE technique

Algorithm (80-20 split)	**Accuracy**	Precision		Recall		F1-Score	
		Class 0	**Class 1**	**Class 0**	**Class 1**	**Class 0**	**Class 1**
Logistic Regression	0.542	0.17	0.5	0.26	0.86	0.55	0.67
Random Forest	0.434	0.21	0.96	0.35	0.98	0.34	0.5
Gradient Boosting	0.43	0.21	0.98	0.35	0.99	0.33	0.49

SVM	0.434	0.21	0.96	0.35	0.98	0.34	0.5
KNN	0.812	0.23	0.08	0.12	0.85	0.95	0.89

5.1.4 Hyperparameter Tuning and Threshold Adjustment (75-25 Split)

In this final step, Logistic Regression, Random Forest, Gradient Boosting and SVM were tuned for their hyperparameters using GridSearchCV while KNN was tuned using RandomizedSearchCV. In this step the fine tuning of the exact hyperparameters is done to get a better performance from the developed models. Also, here the data split of 75-25 is applied to get the performances for different training and testing proportions. After tuning hyperparameters Logistic Regression, Random Forest, and Gradient Boosting was with a stable output of 84.5% accuracy for each of them. Recall for Class 1 was better, and F1-scores were higher, showing a better balance between precision and recall. Setting the decision threshold to 0.3 has optimized the precision and recall for these models, mainly in the case of Logistic Regression and Gradient Boosting. On adjusting the class weights SVM improved the recall for Class 1 and still performed badly on the Class 0 precision. KNN after applying the SMOTE technique and adjusting the threshold, it had a good recall on Class 1 and precision for Class 0 is still low. Results show us that the tuning hyperparameters and the changing thresholds have improved the performance of our models differently. The best overall balance came from both Logistic Regression and Random Forest with Gradient Boosting with a higher F1-score.

The model evaluation has shown us that model selection, class balancing techniques such as SMOTE, and hyperparameter tuning were much more important in terms of better performance. The choice of splitting the data as 75-25 with tuning of the hyperparameters and changing the thresholds resulted in the best overall performance of the models to predict the CVD risk. These changes were mainly helpful for the Logistic regression and Gradient Boosting models. Also, these models had higher F1 scores with a better balance between the precision and recall. However, it was not enough to make the KNN and SVM work well due to its class imbalance and the nature of the algorithms.

5.2 Results and Evaluation

In the evaluation section we go into detail study of the final model outputs and their ability in

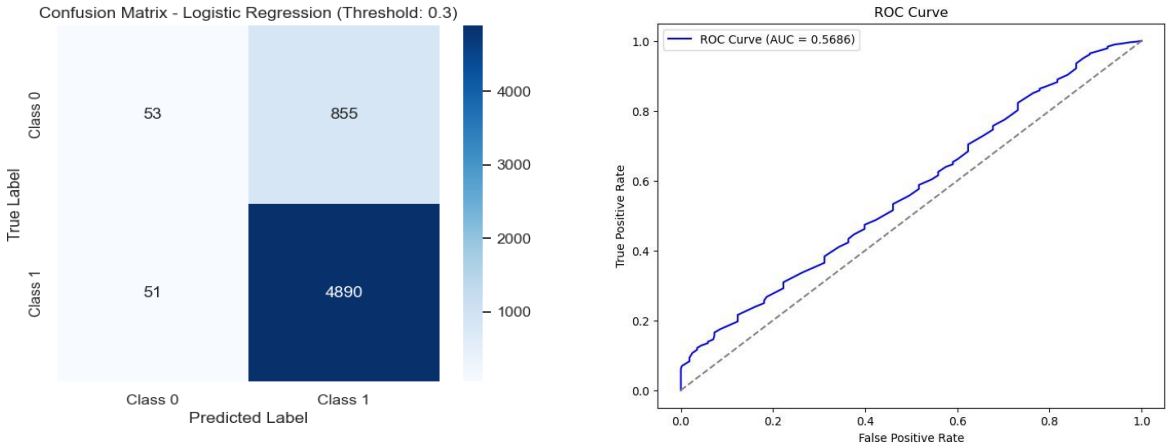

Figure 4: Confusion matrix and ROC curve of Logistic Regression Model

prediction of CVD occurrences in the US. We discuss different metrics of evaluation including accuracy, precision, recall, F1-score, ROC-AUC and computational efficiency. Comparing all these models will be done on a uniform threshold of 0.3 in Logistic Regression, Random Forest, and Gradient Boosting models.

The other two models, such as Support Vector Machine and K-Nearest Neighbors, each have some modification in their configurations. The performance metrics of the models are summarized in the below table.

Table 4: **C**omparison of Model Performance Metrics with Hyperparameter Tuning

Algorithms (75-25 Split)	Optimization Techniques	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Logistic Regression	**Threshold Adjustment (0.3)**	0.8451	0.8512	0.9897	0.9152	0.5686
Random Forest		0.834	0.8538	0.9694	0.908	0.6563
Gradient Boosting		0.8451	0.8512	0.9897	0.9152	0.6669
SVM	**Class Weighting**	0.5451	0.858	0.5529	0.6725	0.5207
KNN	**SMOTE**	0.6257	0.881	0.644	0.7441	0.6647

It is seen from the above table 4, both the Logistic Regression and Gradient Boosting have gained a high accuracy and outstanding F1 score, with both the models having an accuracy of 84.51% and the F1 score standing at 0.9152. They have a very high recall of 0.9897, which is very good in identifying the positive examples of CVD. While the recall for both models was high, the precision was kind of comparatively lower which measured at about 0.85. These will be addressed by the medical experts who indicate that some false positives might be there among the predictions. The Random Forests algorithm was performing ok with an accuracy of 83.40%, but it had a lower recall when its compared with the Logistic Regression and Gradient Boosting methods.

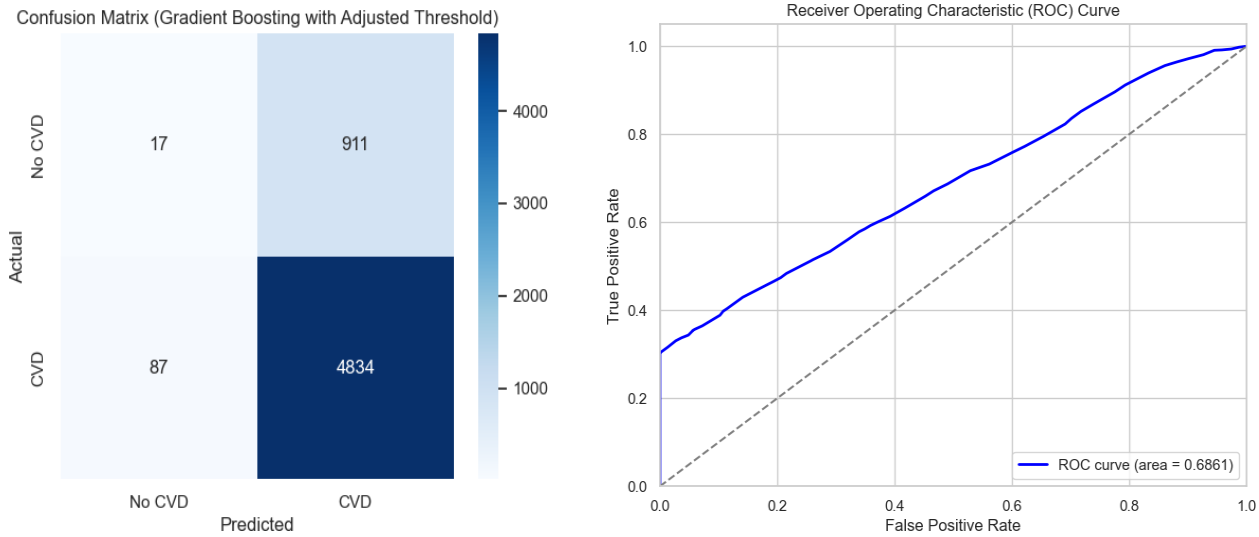

Figure 5: Confusion matrix and ROC curve of Gradient Boosting Model

On the other hand, the performance of the Support Vector Machine was relatively poor, with an accuracy of about 54.51% with a low recall of 0.5529. Even if the class weights are considered during training, low performance to the class imbalance problem is possible. In the same way, the K-Nearest Neighbors also resulted in a poor accuracy of 62.57% and a recall of 0.6440, even though when class imbalance was addressed with SMOTE.

The computational efficiency of each model measured by training time is summarized in the table below.

Table 5: Computational efficiency of the models

Model	Time in Seconds
Logistic Regression	0.0242
Random Forest	0.4893
Gradient Boosting	0.7668
SVM	4.2716
KNN	0.0156

KNN was the fastest from the point of view of execution speed time. It took 0.017 seconds for training, while Logistic Regression took 0.028 seconds. The SVM model became the slowest at 4.17 seconds, which already showed us its inefficiency with the background of other predictors. Random Forest and Gradient Boosting also required a minimum training time of 0.54 and 0.82 seconds, respectively, but these models gave better performance with respect to the quality of prediction and for this training might have been longer.

The main objective of this research was to evaluate the efficacy of machine learning models in predicting the CVD risk within the United States, as measured by accuracy, precision, recall and F1 score. According to the evaluation metrics, Logistic Regression and Gradient Boosting were identified as the most proficient models for forecasting the occurrence of cardiovascular disease. Both models showed high levels of accuracy 84.51% and recall of 0.9897, reflecting their ability to accurately identify cases of cardiovascular disease. Both models are expressing the strong balance between Precision and Recall with the F1 score 0.9152, hence it is considered as the most appropriate models answering the research question.

Although Random Forest had generally good performance, its recall for Class 1 was low when compared with Logistic Regression and Gradient Boosting. Hence, it would be considered as acceptable in view of the prediction of the CVD cases. SVM performed very poorly, with its accuracy of 54.51% and recall of 0.5529, with not doing well for the class imbalanced dataset. And, even with the application of SMOTE, the performance was still poor using the KNN algorithm, with low accuracy and recall for Class 1, hence not as reliable as the other models and is not an acceptable model to predict the cases of CVD.

With respect to training efficiency, KNN took the least time followed by the logistic regression, but still KNN was not that great in terms of the quality of the predictions. On the other hand, Random Forest and Gradient Boosting took more time for training but showed the most accurate predictions and turned out to be more successful in forecasting the risk of

cardiovascular disease. It simply words, though the KNN and Logistic Regression were efficient with respect to training time, the other two models were far more capable in the generation of accurate predictions. Therefore, considering the performances of accuracy, recall, precision, and F1-score The Logistic Regression and Gradient Boosting models work out as the best to predict the risk of cardiovascular disease. Further optimization of these machine learning models or searching for other methodologies that involve ensemble techniques can be focused in studies with the aim of improving prediction accuracy.

5.3 Discussion

In this study, the performance of the machine learning models was quite well, with the optimized threshold of 0.3 from Logistic Regression, Random Forest, and Gradient Boosting. These models gave some very promising accuracies of 84 and 84%, precision of 0.85 and recall of 0.98 metrics useful in predicting the case of cardiovascular disease in the United States. However, there is quite a particular area for improvement, the addition of more variable types and better class imbalance management could provide us the results in higher performances. Although those models showed high accuracy, more improvements and investigations for other methodologies such as feature engineering or ensemble methods could also produce better results. Even if there is room for improvement, the current study has solved a gap in the field of cardiovascular disease prediction by considering the important demographic factors of age, sex, and geographic location into factors that were often overlooked in previous research.

To overcome these limitations, we could change a few steps in the methodology by making some alterations. Considering the populations of all states and including other factors such as income levels, education levels, nationality, genetic data and behavioral patterns will enhance the reliability and efficiency of those genetic AI models. Other techniques for this problem can include SMOTE for under sampling of a dataset or employing the cost sensitive learning algorithms. Speaking about the features of engineering we can get interaction term and incorporate between various sets of more complex factors like the health condition index by the neighborhood, or health care service availability index. The models can be introduced to SCEL, Deep Learning Models and Ensemble Models which might enhance its performance as a result of the patterns that the models identify and also due to model heterogeneity in terms of family they come from. It is also possible to expand the study beyond the accuracy measure range by including both MCC and balanced accuracy to enlarge the list of the reliable measures will enlarge the range of reliable measures. At the same time using SHAP or LIME will increase the number of methods for interpreting the results hence offering a broad solution to the flaws pointed out in the present study

The study shows that demographic factors such as age, sex, race and region are important in predicting the risk of cardiovascular disease, like what past research has not found. However, this study adds something new by using advanced machine learning models to combine these factors in a better way. Previous studies used simpler methods and looked at a few factors, whereas this study used more data and improved methodologies. This resulted in more accurate predictions showing us how improved models can assist the study of chronic diseases. This study does have some limits. As the dataset is purely based on US citizens, It focuses only on people in the US, so the results may not work for other countries as their

healthcare, culture and environment is completely different. Also, this study gives useful ideas about how demographic factors affect cardiovascular disease in the US. It also provides a starting point for future studies to look at these patterns in other countries.

6 Conclusion and Future Work

In Conclusion, this study has successfully met its goal by analyzing a dataset which had the demographic and health factors and by testing and comparing several machine learning models. The Logistic Regression, Random Forest, and Gradient Boosting models performed best, with accuracies of 84%, 83% and 84%. These models show us the importance for predicting who is at risk of cardiovascular disease, with well-balanced precision and recall scores showing their reliability. The study also addressed a major gap by including parameters such as age, gender and geographic location in the analysis, which were left out in similar studies. Thus, the findings show that machine learning can provide us useful insights to improve public health strategies.

The study showed positive results, but there's still room for improvement. Further studies could be considered by adding a few more factors such as the social and economic conditions or genetic data to improve the model's accuracy. Research methods like combination of different models or using advanced algorithms like deep learning, also enhance the results. As the study was based only on US data, it would be a good idea if it was tested with different data from other countries, to check if similar trends occur worldwide. Looking into how these models can be used practically in healthcare such as early diagnosis and treatment, it will make this research more applicable for preventing the risk of cardiovascular diseases.

References

Airlangga, G., 2024. *Comparative Analysis of Machine Learning Models for Intrusion Detection in Internet of Things Networks Using the RT-IoT2022 Dataset,* Indonesia:

MALCOM: Indonesian Journal of Machine Learning and Computer Science.

Alanazi, R., 2022. Identification and Prediction of Chronic Diseases Using Machine. *Hindawi Journal of Healthcare Engineering*, 2022(2826127), p. 9.

Anurag, 2023. *Chronic Kidney Disease Prediction Using Robust Approach in Machine Learning,* Punjab: International Conference on Innovative Sustainable Computational Technologies (CISCT).

Chioma Susan Nwaimo, 2024. *Transforming healthcare with data analytics: Predictive models for patient outcomes,* Illinois: GSC Biological and Pharmaceutical Sciences.

Gabriel S Tajeu, 2024. *Cost of Cardiovascular Disease Event and Cardiovascular Disease Treatment-Related Complication Hospitalizations in the United States,* US: National Library of Medicine.

Harvard Health, 2019. *Premature heart disease.* [Online] Available at: https://www.health.harvard.edu/heart-health/premature-heart-disease [Accessed 1 December 2024].

Jaikumar, 2024. *Significance of Machine Learning Algorithms to Improve Predictive Analytics in Chronic Disease Management through Pharmacogenomics,* Raipur: SEEJPH.

Kenney Ng, 2013. *PARAMO: A Parallel predictive Modeling platform for healthcare analytic research using electronic health records,* Nashville, TN: Journal of Biomedical Informatics.

M. Vijay Kumar, 2024. *Chronic Disease Prediction Using Machine Learning,* Springer Nature Link.

Madhavi Devi Botlagunta, 2023. *Prediction of Chronic Kidney Disease with Artificial Neural Network,* Chicago: International Conference on Recent Trends in Advance Computing (ICRTAC).

Services, 2024. *U.S. Chronic Disease Indicators.* [Online] Available at: https://catalog.data.gov/dataset/u-s-chronic-disease-indicators [Accessed 19 August 2024].

Shelf, 2024. *Random Forests in Machine Learning for Advanced Decision-Making.* [Online] Available at: https://shelf.io/blog/random-forests-in-machine-learning/ [Accessed 6 December 2024].

Singh, 2018. *Feature selection and classification systems for chronic disease prediction: A review,* Gurugram: Egyptian Informatics Journal.

Sultana Habiba, 2024. *Early Prediction of Chronic Kidney Disease Using Machine Learning Algorithms with Feature Selection Techniques,* Dubai: Research Gate.

Suresh, 2024. *Machine Learning-Based Identification and Forecasting of Chronic Illnesses,* Telangana: International Journal of All Research Education and Scientific Methods (IJARESM),.

Walther, 2023. *New Study Reveals Latest Data on Global Burden of Cardiovascular Disease.* [Online] Available at: https://www.acc.org/About-ACC/Press-Releases/2023/12/11/18/48/New-Study-Reveals-Latest-Data-on-Global-Burden-of-Cardiovascular-Disease [Accessed 25 October 2024].

Xiangyong Kong, 2022. *Disease-specific data processing: An intelligent digital platform for diabetes based on model prediction and data analysis utilizing big data Technology,* China: Frontiers in public health.

Ying Zheng, 2024. *Risk prediction models of depression in older adults with chronic diseases,* Bengbu: Journal of Affective Disorders.