Risk Factors for Cardiovascular Heart Disease

Arjun A U
Msccs(DA)
Dept.Computer Science
Rajagiri College of Social Sciences

Abstract : Analyzing the effect of lifestyle and environmental factors on the risk of cardiovascular disease. The aim of this project is to classify whether a person has heart diseases or not. Predicting the risks of different age groups based on their demographic characteristics such as gender, height, weight and smoking status. Detecting patterns between levels of physical activity, blood pressure and cholesterol levels with likelihood of developing cardiovascular disease among individuals .The project aims to build multiple classification models using Orange tool and select the model with the best accuracy out of it.

INTRODUCTION

Cardiovascular disease (CVD) is a group of conditions that affect the heart and blood vessels. The most common type of CVD is coronary heart disease (CHD), which is caused by a buildup of plaque in the coronary arteries that supply the heart with oxygen and nutrients. CVD is a leading cause of death worldwide, and its risk factors can be both modifiable and non-modifiable. Modifiable risk factors are those that can be changed or managed through lifestyle changes or medical treatment, while non-modifiable risk factors cannot be changed. Understanding the risk factors for CVD is important for prevention, early detection, and management of the disease. In this context, in the following discussion, we will explore the major risk factors for CVD. Risk factors for CVD include:

- Age: As people age, the risk of CVD increases.
- Gender: Men are at a higher risk of CVD than premenopausal women, but after menopause, the risk for women increases.
- Smoking: Smoking is a major risk factor for CVD, and it damages the blood vessels and can cause plaque buildup.
- High blood pressure: High blood pressure puts a strain on the heart and blood vessels and can lead to CVD.
- High cholesterol: High levels of LDL cholesterol (the "bad" cholesterol) can cause plaque buildup in the arteries.
- Physical inactivity: Not getting enough exercise can increase the risk of CVD.

LITERATURE REVIEW

1. In the Framingham study, risk factors for coronary heart disease in the elderly were investigated and reviewed.
2. The amount of oxygen consumed by people with ischemic heart disease is significantly influenced by heart rate..
3. Assessing teenagers' perceptions of the risk factors for cardiovascular disease and evaluating their possible impact on reported behaviours, such as exercise, smoking, and food, as well as their body mass index (BMI).
4. study dietary practises and assess them in light of cardiovascular risk status in Turkish teenagers between the ages of 12 and 19..
5. The main cause of death in the globe is cardiovascular disease. The abnormal lipid and cholesterol levels may be responsible for up to half of these fatalities.
6. This connection is particularly strong in patients with hypertension or diabetes, and the clustering of these risk factors appears to be the key determinant of cardiovascular morbidity associated with increased heart rate in these situations.
7. This cross-sectional study examined the connections between seven risk factors for coronary heart disease (CHD) in the general population and the psychosocial work environment, as measured by job demand-control (JDC) and effort-reward imbalance (ERI)..
8. It is generally accepted that hereditary factors have a role in cardiovascular heart disease (CHD). Several research have looked into how inflammatory markers like tumour necrosis factor (TNF-) and C-reactive protein (CRP) contribute to the development of cardiovascular disorders.
9. In Ireland, cardiovascular diseases account for roughly 46% of all fatalities. Changes in lifestyle may be helpful to prevent certain illnesses. Almost 195,000 deaths are predicted to occur by the year 2020. It is crucial to gauge how well-informed young people are about CVD in order to establish teaching programmes about it.
10. The public's concerns about their health have increased due to obesity, which has become the most significant cardiovascular risk factor, both in industrialised and developing nations.
11. Cardiovascular heart disease affects the majority of people worldwide. Early cardiac illness diagnosis helps to reduce mortality rates.
12. n this study, the quality of life (QOL) of heart illness patients with and without depression and related comorbidities is evaluated and compared.
13. The goal of this study was to gauge Saudi community knowledge about cardiovascular diseases (CVDs) and risk factors in Riyadh..
14. This study aims to evaluate adult population knowledge of cardiovascular disease and risk factors in southern Saudi Arabia. 1,049 individuals

completed a web-based cross-sectional survey in August 2021..

15. The study had two goals: first, to determine how well-informed the general public was about cardiovascular diseases (CVDs), and second, to offer written information on CVD risk factors, key symptoms, and preventive.
16. The goal of this review is to examine how nut consumption affects coronary heart disease risk factors and other cardiovascular risk variables..

IMPLEMENTATION

Tools Used – Orange(3.34.0) is an open-source data mining and visualization toolkit. It is used for explorative rapid qualitative data analysis and interactive data visualization

a) Data description

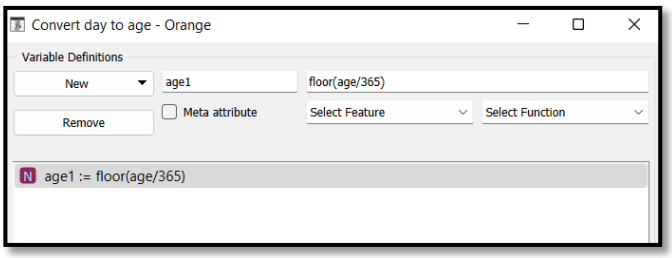
The dataset used in this project was obtained from Kaggle website. The details about the attributes given here are used for classifying whether the potential relations between risk factors and cardiovascular disease that can ultimately lead to improved understanding of this serious health issue and design better preventive measures This dataset consists of 70000 instances and 14 features. There are 9 numeric variables and 4 are categorical variables. The attributes in the datasets are:

- Age
- Gender
- Height
- Weight
- Ap_hi
- Ap_lo
- Cholesterol
- Index
- Id
- Gluc
- Smoke
- Alco
- Active
- Cardio

NO	Attribute	Description	Type
1	age	Age of the individual.	numeric
2	Gender	Gender of the individual	categorical
3	height	Height of the individual in centimetres.	numeric
4	weight	Weight of the individual in kilograms	numeric
5	ap_hi	Systolic blood pressure reading.	numeric
6	ap_lo	Diastolic blood pressure reading	numeric
7	Cholesterol	Cholesterol level of the individual	numeric
8	Gluc	Glucose level of the individual	numeric
9	Smoke	Smoking status of the individual	categorical
10	alco	Alcohol consumption status of yhe individual	categorical
11	active	Physical activity level of the individual	categorical
12	cardio	Presence or absence of cardiovascular disease	categorical

Pre-process

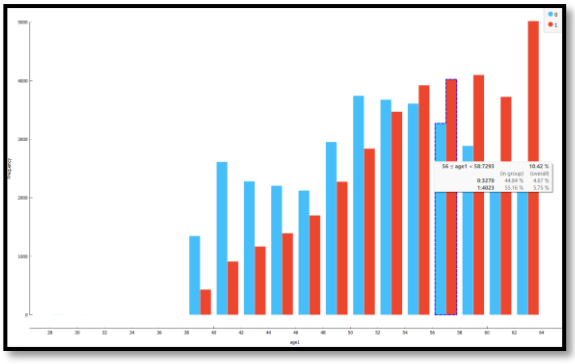
Here, in the dataset, as a part of pre-processing, In this dataset day is in the form of days. It can be converted to age using Feature Constructor



b) Data Exploration

In order to analyze any data and to bring out the important features, we need to explore the data.

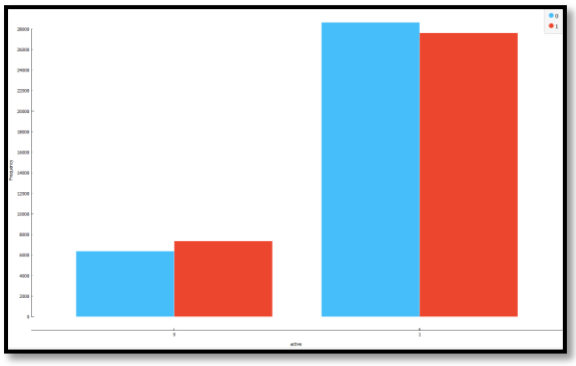
Fig : Age v/s Count



From the above plot, we can say than 63-64 years of age are more prone to cardiovascular diseases.

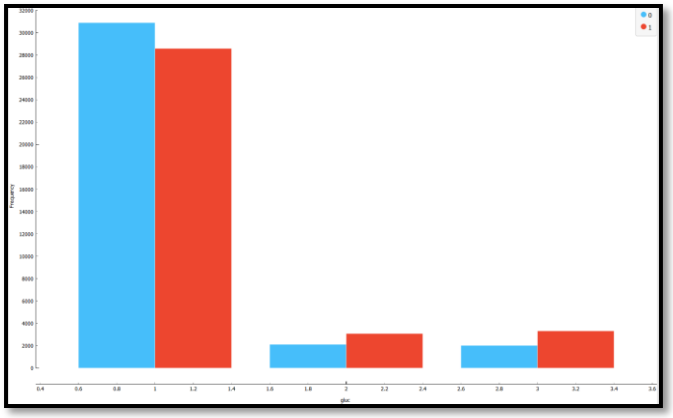
Ratio of diseased to healthy people increases with age.

Fig : Active v/s Counts



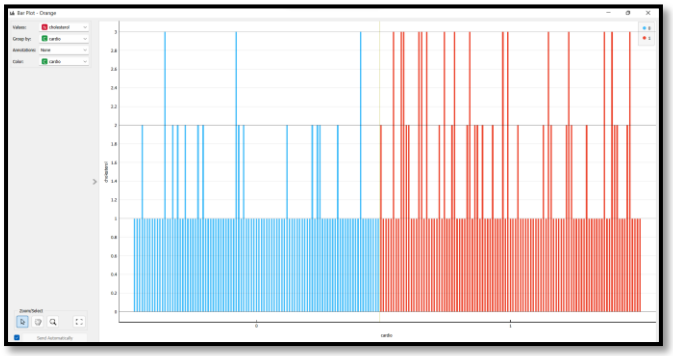
From the above plot, we can say that People are more healthy when they follow an active lifestyle.

Fig : Glucose level v/s Count



From the above plot, we can say that people with higher level of glucose are more prone to cardiovascular diseases.

Fig : cardio v/s cholesterol (Bar plot)

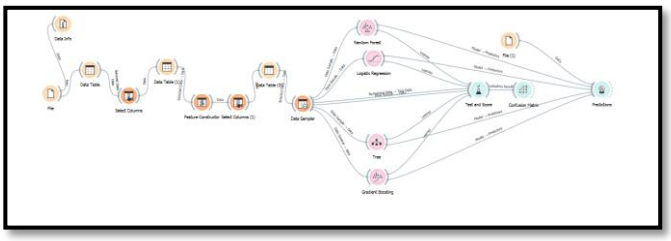


From the above, we can say that people with high cholesterol level(level 3) are more prone to cardiovascular diseases.

c) Classification Techniques

This project consists of a few classification techniques, such as Random Forest, Logistic Regression, Tree and Gradient boosting. There is no missing value in the dataset. The target variable chosen here is 'Cardiovascular, which has two values; 0 and 1. The dataset is split into 80:20 ratio for train and test. The following is the orange connection diagram:

Fig : Model Building Risk Factors for Cardiovascular Heart using Orange



i. Random Forest (RF)

It consists of a number of trees that work as a whole. Each tree in the random forest gives a prediction in which the one which has high votes will be the model prediction.
The random forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. The greatest number of trees in the forest leads to higher accuracy and prevents the problem of over fitting.

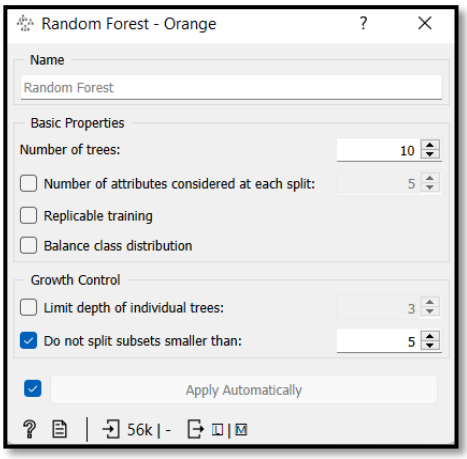


Fig : Random Forest Model Parameters

The model evaluation value and the confusion matrix of the Random Forest Algorithm is shown below:

Model	AUC	CA	F1	Precision	Recall
Random Forest	0.765	0.705	0.705	0.705	0.705

Fig : Random Forest Evaluation Parameter

ii. Logistic regression

In logistic regression, the dependent variable is categorical and takes only two values, typically 0 or 1, representing the presence or absence of an event or outcome. The independent variables can be either continuous or categorical, and their values are used to estimate the probability of the dependent variable.

The logistic regression model uses a logistic function or sigmoid curve to transform the linear combination of the predictor variables into a probability value between 0 and 1. The logistic function maps any real-valued input

into a range of 0 to 1, which makes it suitable for modeling probabilities.

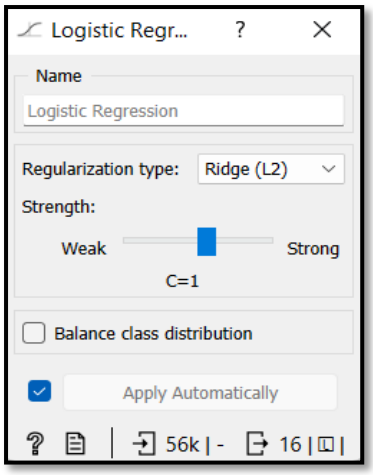


Fig : Logistic regression Model Parameters

The model evaluation value and the confusion matrix of the logistic regression is shown below:

Model	AUC	CA	F1	Precision	Recall
Logistic Regression	0.785	0.719	0.719	0.722	0.719

Fig : logistic regression Evaluation Parameter

iii. Tree

A algorithm called tree divides data into nodes based on class purity (information gain for categorical and MSE for numeric target variable). It comes before Random Forest. Both categorical and numerical datasets can be handled by Tree in Orange, which was created in-house. Moreover, it can be applied to problems requiring classification and regression.

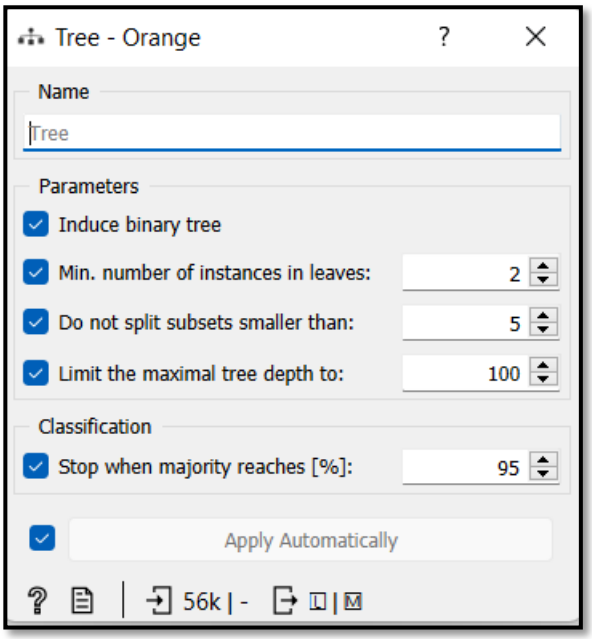


Fig : Tree Model Parameters

The model evaluation value and the confusion matrix of the Tree Algorithm is shown below

Model	AUC	CA	F1	Precision	Recall
Tree	0.668	0.647	0.646	0.649	0.647

Fig : Tree Evaluation Parameter

iv. Gradient Boosting (GB)

It is a popular boosting technique in which every predictor corrects its predecessor's error. The base of the gradient boosting is the CART (Classification and Regression Tress). Each tree predicts a label and the final prediction is through the formula:

```
y(pred) = y1 + (eta *  r1) + (eta * r2) + ..... + (eta * rN)
```

It can predict continuous target variable as well as categorical target variable. In categorical target variable the cost function is Log loss.

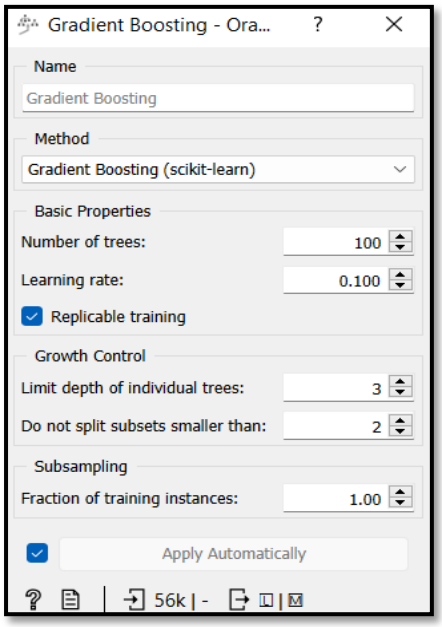


Fig : Gradient Boosting Model Parameters

The model evaluation value and the confusion matrix of the Random Forest Algorithm is shown below:

Model	AUC	CA	F1	Precision	Recall
Gradient Boosting	0.801	0.735	0.735	0.737	0.735

Fig : Gradient Boosting Evaluation Parameter

v. K- Nearest Neighbors Algorithm (KNN)

It is a supervised machine learning algorithm that provides a simple solution to both regression and classification. When a new sample comes, the algorithm begins by calculating the distance between the new sample and the existing data points and it is assigned to the group common to the k nearest selected neighbours.

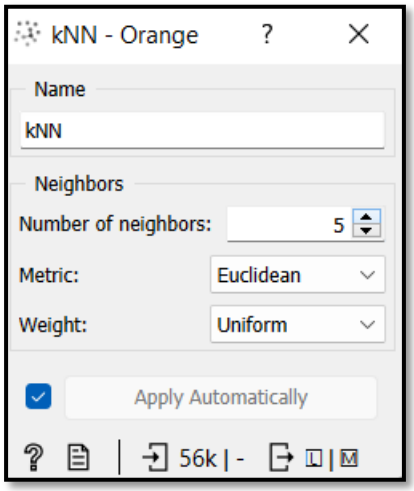


Fig : KNN Model Parameters

The model evaluation value and the confusion matrix of the KNN is shown below:

Evaluation results for target (None, show average over classes)

Model	AUC	CA	F1	Precision	Recall
kNN	0.739	0.690	0.690	0.691	0.690

Fig : KNN Model Evaluation Parameter

RESULT AND DISCUSSION

The dataset 'Risk Factors for Cardiovascular Heart Disease' consists of 53000 rows and 12 columns. The selected target variable is 'Cardio'. The data is split into 70:30 ratio for train and test.

Classification model applied in the datasets are Random Forest, Logistic regression, Tree, Gradient Boosting. The performance value obtained for each model is given below

Evaluation results for target (None, show average over classes)

Model	AUC	CA	F1	Precision	Recall
Tree	0.668	0.647	0.646	0.649	0.647
Random Forest	0.764	0.702	0.702	0.703	0.702
Logistic Regression	0.785	0.719	0.719	0.722	0.719
Gradient Boosting	0.801	0.735	0.735	0.737	0.735

Fig : Evaluation Result

Among the applied classification models, **Gradient Boosting** has highest accuracy which is **73.5%**. The confusion matrix of GB is given below

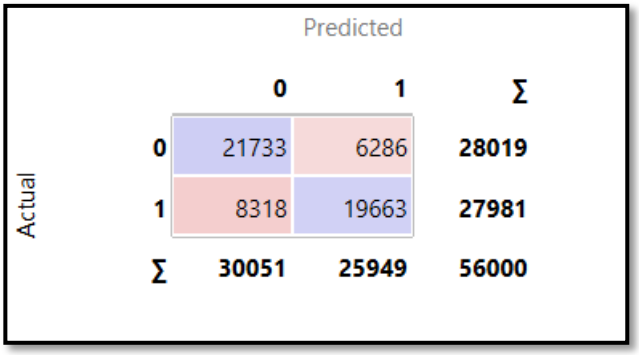


Fig: Confusion matrix of Gradient Boosting

An external test data was applied to test the classification using the GB model and the classification is as follows

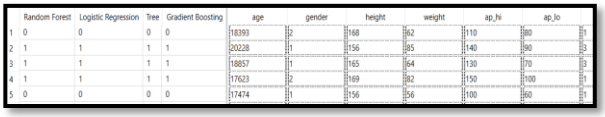


Fig : Classification outcome for the given data

Gradient Booster gives correct values of test data set

CONCLUSION

The main goal of this paper on Risk Factors for Cardiovascular Heart Disease. there are several risk factors for cardiovascular heart disease (CVD). Smoking, high cholesterol, high blood pressure, and age are all significant risk factors for cardiovascular heart disease. Smoking damages the lining of the blood vessels increases blood pressure, and reduces blood flow, while high cholesterol contributes to the build-up of fatty deposits in the arteries. High blood pressure strains the heart and blood vessels, and age increases the risk of developing CVD due to the natural aging process.

However, these risk factors can be managed through lifestyle changes such as regular physical activity, healthy eating habits, and weight management, as well as medication when necessary. Quitting smoking, managing cholesterol levels and blood pressure, and adopting healthy habits can significantly reduce the risk of developing CVD.

REFERENCES

1) https://www.researchgate.net/publication/11924514_Risk_Factors_in_the_Elderly_A_View_From_Framingham
2) https://www.researchgate.net/publication/30950868_Increased_heart_rate_as_a_risk_factor_for_cardiovascular_disease
3) https://www.researchgate.net/publication/8229234_Adolescent_assessment_of_cardiovascular_heart_disease_risk_factor_attitudes_and_habits
4) https://www.researchgate.net/publication/8402735_Determination_of_dietary_habits_as_a_risk_factor_of_cardiovascular_heart_disease_in_Turkish_adolescents
5) https://www.researchgate.net/publication/7000996_Risk_factors_for_coronary_heart_disease_in_patients_with_schizophrenia
6) https://www.researchgate.net/publication/5846520_Heart_Rate_as_an_Independent_Risk_Factor_for_Cardiovascular_Disease
7) https://www.researchgate.net/publication/233973066_A_cross-sectional_study_of_the_relationship_between_job_demand-control_effort-reward_imbalance_and_cardiovascular_heart_disease_risk_factors
8) https://www.researchgate.net/publication/283640855_The_role_of_gene_variants_of_the_inflammatory_markers_CRP_and_TNF-a_in_cardiovascular_heart_disease_Systematic_review_and_meta-analysis
9) https://www.researchgate.net/publication/319206816_Assessing_the_Knowledge_of_Cardiovascular_Disease_Among_Young_People_in_South_Dublin
10) https://www.researchgate.net/publication/330099579_The_impact_of_obesity_on_cardiovascular_disease_risk_factor
11) https://www.researchgate.net/publication/349758103_Performance_evaluation_of_supervised_and_unsupervised_machine_learning_algorithms_by_predicting_cardiovascular_heart_disease
12) https://www.researchgate.net/publication/342260660_Quality_of_life_among_patients_with_cardiac_disease_the_impact_of_comorbid_depression
13) https://www.researchgate.net/publication/342524961_Awareness_of_cardiovascular_disease_associated_risk_factors_among_Saudis_in_Riyadh_City
14) https://www.researchgate.net/publication/356432243_Assessment_of_Cardiovascular_Diseases_Knowledge_and_Risk_Factors_Among_Adult_Population_in_the_South_Region_of_Saudi_Arabia
15) https://www.researchgate.net/publication/352440866_A_Community-Based_Cross-Sectional_Study_Assessing_the_Level_of_Awareness_and_Insight_Related_to_Cardiovascular_Diseases
16) https://www.researchgate.net/publication/365481273_Effects_of_Nut_Consumption_on_Cardiovascular_Risk_Factors_and_Coronary_Heart_Diseases