A STUDY TO EVALUATE CARDIOVASCULAR DISEASE IN WORKING-AGE ADULTS

1. Introduction :-

The majority of deaths worldwide are caused by cardiovascular disease (CVD), which is increasingly impacting working-age individuals. It is normal for people in their 30s, 40s, and 50s to have cardiovascular events like heart attacks or strokes, despite the fact that CVD is frequently linked with age. Together with higher healthcare expenditures and lower productivity, this has enormous ramifications for both people and society.

The results of this study will offer important new knowledge on working-age people' cardiovascular health, which may help in the development of focused therapies to lower CVD risk factors and prevent cardiovascular events. The findings of the study might also have a big impact on public health policies and initiatives that attempt to lessen the cost of CVD on society.

2. Source and Description of the Data :-

The information came from Kaggle, a website for data science and machine learning enthusiasts to find and exchange datasets, kernels, and contests. Data on cardiovascular disease risk factors and outcomes for a sample of patients are included in the dataset, which was generated by Svetlana Ulianova.The data description is given in *Table 1* (Appendix section).

We have a total of 70000 observations among them there are some outliers and leverage records which may tamper our results. You can observe them in *Table 2* . So, we have removed the outliers and leverage records from the original data.

3. Methods :-

3.1 Data Cleaning :-

In every data analysis process, data cleansing is an essential step. To achieve accurate and trustworthy results in this study, we ran a number of data cleaning processes on the dataset. To ensure consistency and to make the data easier to examine, we first converted the factors into quantitative and categorical variables. For better analysis, we also changed the individuals' age from days to years and divided them into early (between 30 and 45 years) and late (between 46 and 67 years) midlife groups.

We also used weight and height values to calculate the BMI variable using the conventional formula $bmi = weight(kgs)/((height(cms)/100)^2)$. Using ranges of 70 to 160, 40 to 120, and 16 to 40, respectively, for systolic blood pressure, diastolic blood pressure, and BMI, we eliminated outliers and excluded participants with extreme results in order to assure the correctness of the data.

Lastly, in order to guarantee the models' dependability and accuracy, we divided the dataset into 25% for training and 75% for testing. For more details, see *Table 2*. We were able to acquire accurate and trustworthy results thanks to these data cleaning techniques, and we were then able to make valid inferences regarding the correlation between CVD risk variables and outcomes in working-age individuals.

Used Observations	65332
Missing Values	NA
Duplicates	NA
Odd Observation	4668
Training Data	48999
Testing Data	16333

Table 2

3.2 Primary Analysis :-

We performed summary statistics in the primary analysis for both categorical and quantitative variables. The minimum, maximum, mean, standard deviation, and median were determined for quantitative values. We calculated the overall subject population and the proportion of those with each kind of cardiovascular disease for categorical variables. We can better comprehend the overall distribution and variability of the data with the use of these statistics. For better understanding, observe *Table 3* (Appendix section).

According to the bar graph showing the prevalence of cardiovascular disease by gender and age, we discovered that early midlife was associated with lower rates of CVD in both men and women than late midlife. The identification of age-specific CVD risk factors and the formulation of effective preventative measures depend on the information provided. Refer *Figure 1.*

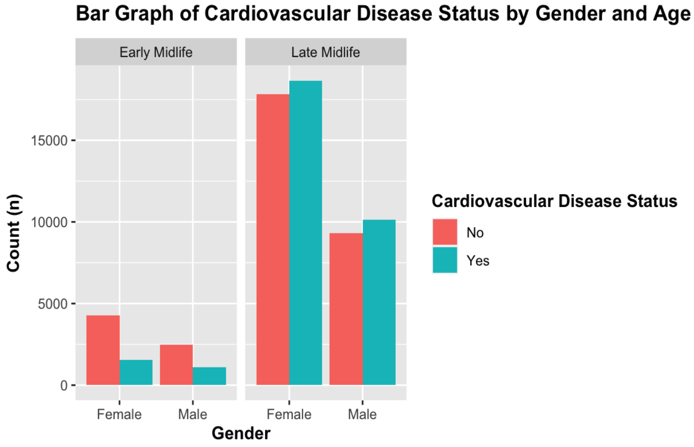

Figure 1

According to our study of the box plot of cardiovascular disease status by BMI and cholesterol, those with cholesterol levels that are much higher than normal and a BMI of over 25 have a greater chance of having CVD than people with lower cholesterol levels. cases. Refer *Figure 2.*

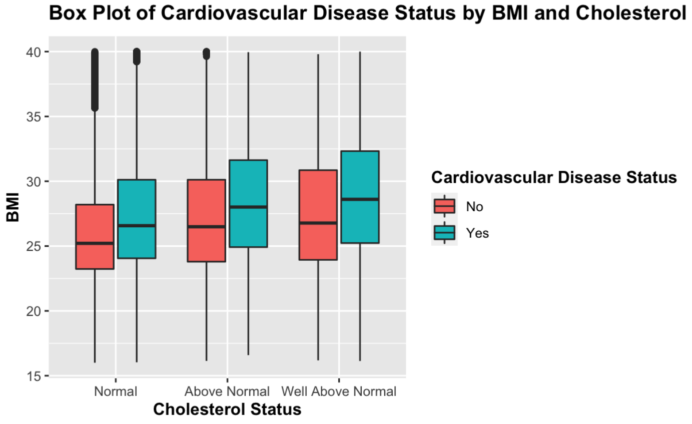

Figure 2

The association between blood pressure and CVD risk was examined using a scatter plot of cardiovascular disease states by blood pressure. The figure demonstrated that there were higher incidences of CVD in the systolic (125 to 160) and diastolic (70 to 100) blood pressure ranges. The correlation between blood pressure and CVD risk means that people at high risk for CVD should have their blood pressure closely monitored. Refer *Figure 3.*

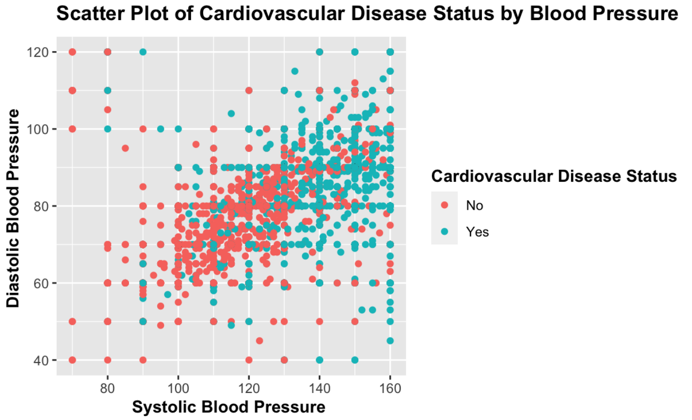

Figure 3

According to the bar graph showing the prevalence of cardiovascular disease according to blood sugar levels and physical activity levels, patients with abnormal glucose levels who did not exercise or exercised excessively had a greater chance of getting CVD. These results highlight the value of a good glucose balance and moderate physical exercise in lowering CVD risk. Refer *Figure 4*.

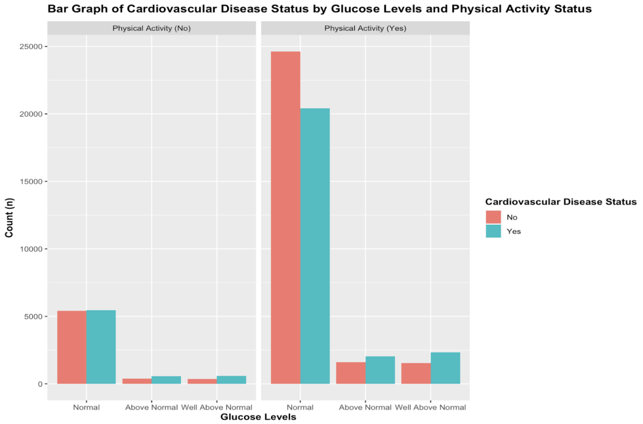

Figure 4

3.3 Variable Selection :-

According to the Boruto model, we have selected a few predictors. The Boruta model is a feature selection machine learning technique based on the random forest algorithm. By contrasting each feature against a collection of arbitrary shadow characteristics, it finds the most crucial variables in a complicated dataset. By using this strategy, overfitting is avoided and the most pertinent variables are included in the final model. The Boruta model has been successfully used in a number of different industries and has shown promise in raising the prediction model's accuracy. Refer *Table 4*.

According to AIC criteria, we have selected a few predictors. A statistical method for model selection is the Akaike Information Criterion (AIC). It gauges a model's quality by how well it can account for the number of parameters utilized and yet explain data variation. Because the AIC penalizes complicated models, it may be used to choose simpler models that provide a good explanation for the data. Better model fits are indicated by lower AIC values. In many different domains of statistical modeling, the AIC is commonly employed. Refer *Table 4*.

Variables selected using Boruto Model	age_cat, bmi, cholesterol, diastolic, smoke, Systolic
Variables selected using AIC Criteria	age_cat, systolic, diastolic, cholesterol, smoke, bmi, glucose, alco, active
Interaction between predictors	There is no collinearity and multicollinearity checked by correlation matrix and VIF.
Variables selected for final model	age_cat, systolic, diastolic, cholesterol, smoke, bmi, glucose, alco, active

Table 4

Correlation matrices and the variance inflation factor (VIF) were employed to examine predictor-predictor interaction. We learned that there is no interaction, and our final model includes the following predictors: age cat, systolic, diastolic, cholesterol, smoking, bmi, glucose, alcohol, and activity. For more detail, Refer *Table 4 and Figure 5*.

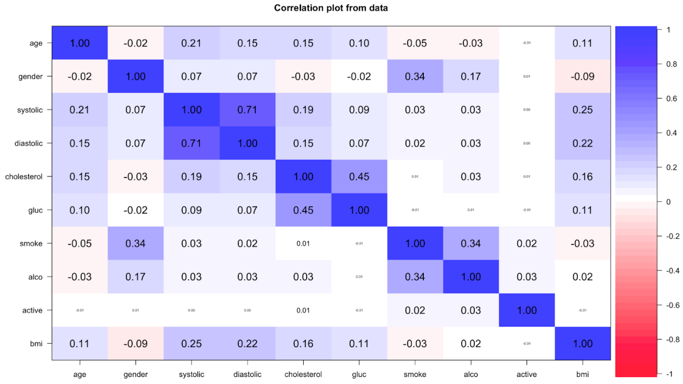

Figure 5

3.4 Model :-

We performed different types of machine learning models like random forest, C50, XGBoost, GLM, and Naive Bayes. Multiple logistic regression gives us the better model, which we can see by the ROC curve graph. Refer *Figure 6*.

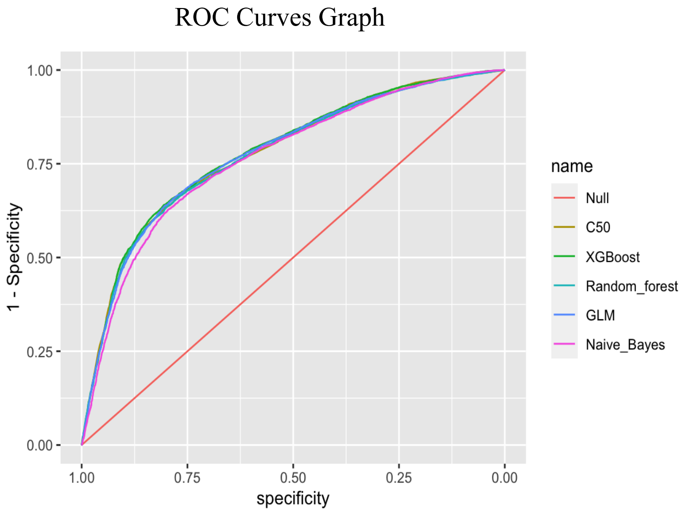

Figure 6

A statistical method known as multiple logistic regression is used to examine correlations between a binary dependent variable and a number of independent factors. It is a development of simple logistic regression, which uses a single predictor variable to describe the likelihood of a binary result. The evaluation of the effects of several independent factors on the likelihood of a binary result is made possible by multiple logistic regression. It is frequently utilized in disciplines like medicine, epidemiology, and psychology to pinpoint illness risk factors, forecast results, and comprehend the connections between variables. For modeling intricate interactions and making result predictions across a range of fields, multiple logistic regression may be a potent tool.

The general logit equation of multiple logistic regression is

$$log(\frac{\hat{p}(x_1,x_2)}{1-\hat{p}(x_1,x_2)}) = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + ... + \beta_k * x_k$$

, where β_0 is the intercept term, β_1 to β_k are the coefficients of the predictor variables x_1 to x_k, and logit(P) is the log odds of the result (i.e., the natural logarithm of the ratio of the probability of the event to the probability of its complement). While accounting for the other factors in the model, this equation may be used to calculate the impact of each predictor variable on the likelihood of the result.

4. Results :-

The logit equation of our model is

$$log(\frac{\hat{p}(x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9)}{1-\hat{p}(x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9)}) =- 10.22 + 0.03 * bmi + 0.68 * age\_catLate\ Midlife - 0.23 * activeYes + 0.06 * systolic + 0.01 * diastolic - 0.27 * alcoYes +.39 * cholesterolAboveNormal - 0.17 * smokeYes + 1.204897 * cholesterolWell\ Above\ Normal + 0.08 * glucAbove\ Normal - 0.34 * glucWell\ AboveNormal$$

Key metrics for assessing binary classification models like logistic regression include accuracy, sensitivity, specificity, AUC, and ROC. With accuracy assessing correct predictions, sensitivity measuring true positives, specificity measuring true negatives, and AUC measuring the area under the ROC curve, they assess the model's capacity to forecast a binary result. These metrics are displayed in *Table 4 and Figure 7.*

Accuracy	72.24%
Sensitivity	0.64
Specificity	0.79
AUC	0.7773

Table 4

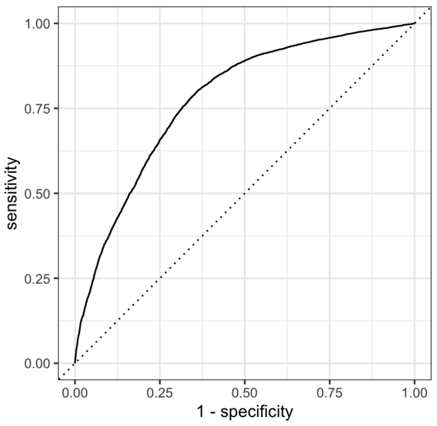

Figure 7

5. Discussion :-

5.1 Limitation :-

This study has certain restrictions, such as an unknown data source and the inability to examine by location. When each factor/variable is taken into account, the data is likewise unbalanced, and analysis is constrained by data from a single visit.

The research did not take into account additional confounding variables and factors like genetic information and stress. Also, for various age groups, the analysis's accuracy may differ.

5.2 Future Works :-

Sub-group analysis can be conducted. Improvements can be made to increase the accuracy by performing other models like DL, NN, etc.

Instead of excluding the odd observations, we can impute the data using various imputation methods (e.g., considering the worst observation or average values).

6. Conclusion :-

After analyzing the data of the study, it can be concluded that cardiovascular disease is a significant health concern for working-age adults. The study highlights the importance of identifying the risk factors associated with CVD and taking measures to prevent or manage the disease. The multiple logistic regression analysis performed in the study had an accuracy of 72% and an AUC of 0.79, indicating that the model has moderate predictive power.

According to the study's findings, key CVD risk factors include age, gender, BMI, cholesterol, blood pressure, glucose levels, and physical activity level. The unidentified data source and the unequal distribution of the variables are two shortcomings that the study has noted. Notwithstanding these drawbacks, the study contributes important knowledge on the incidence and risk factors of CVD in working-age individuals. Further research may be done to verify the results and raise the predictive model accuracy.

7. References :-

1.**https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset**

2.**https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)**

8. GitHub R - Code :-

1.**https://github.com/tvivekanandareddy/Project-Stat-631**

9. Appendix :-

Variable	Cardiovascular Disease (N = 31423)	No Cardiovascular Disease (N = 33909)	Total (N = 65332)
Age Category, n (%)			
Early Midlife	2643 (8.42)	6764 (19.94)	9407 (14.39)
Late Midlife	28780 (91.58)	27145 (80.06)	55925 (85.61)
Gender, n (%)			
Female	20203 (64.29)	22096 (65.16)	42299 (64.74)
Male	11220 (35.71)	11813 (34.84)	23033 (35.25)
BMI			
n	31423	33909	65332
(Min, Max)	(16.00, 40.00)	(16.00, 40.00)	(16.00, 40.00)
Mean (Std)	27.82 (4.52)	26.17 (4.19)	26.96 (4.42)
Median	27.19	25.39	26.13
Systolic Blood Pressure			
n	31423	33909	65332
(Min, Max)	(70.00,160.00)	(70.00, 160.00)	(70.00, 160.00)
Mean (Std)	131.7 (14.7)	119.1 (11.6)	125.2 (14.60)
Median	130.0	120.0	120.0
Diastolic Blood Pressure			
n	31423	33909	65332
(Min, Max)	(40.00, 120.00)	(40.00, 120.00)	(40.00, 120.00)
Mean (Std)	83.83 (8.87)	77.96 (7.94)	80.79 (8.89)
Median	80.00	80.00	80.00
Cholesterol, n (%)			
Normal	21031 (66.92)	28550 (84.19)	49581 (75.89)
Above Normal	4980 (15.84)	3614 (10.65)	8594 (13.15)
Well Above Normal	5412 (17.22)	1745 (5.146)	7157 (10.95)
Glucose, n (%)			
Normal	25876 (82.34)	30026 (88.54)	55902 (85.56)
Above Normal	2613 (8.31)	1984 (5.85)	4597 (7.03)
Well Above Normal	2934 (9.33)	1899 (5.60)	4833 (7.39)
Smoke, n (%)			
Yes	2608 (8.3)	3164 (9.33)	5772 (8.83)
No	28815 (91.70)	30745 (90.67)	59560 (91.16)
Alcohol Intake, n (%)			
Yes	1571 (5)	1875 (5.52)	3446 (5.27)
No	29852 (95.00)	32034 (94.47)	61886 (94.72)
Physical Activity, n (%)			
Yes	24796 (78.92)	27754 (81.85)	52550 (80.43)
No	6627 (21.08)	6155 (18.15)	12782 (19.56)

Table 3.

Variables	**Type**	**Description**
Id	Quantitative	Unique number of the patient.
Age	Quantitative	Age of the patient in days.
Height	Quantitative	Height of the patient in cms.
Weight	Quantitative	Weight of the patient in Kgs.
Gender	Categorical	Gender of the patient (1 - women, 2 - men).
Systolic blood pressure	Quantitative	Measures the blood pressure in your arteries when your heart beats.
Diastolic blood pressure	Quantitative	Measures the blood pressure in your arteries when your heart rests between beats.
Cholesterol	Categorical	Cholesterol level in the patient (1 - normal, 2 - above normal, 3 - well above normal).
Glucose	Categorical	Glucose level in the patient (1 - normal, 2 - above normal, 3 - well above normal).
Smoking	Categorical	Smoking status of the patient (0 - No, 1 - Yes).
Alcohol intake	Categorical	Alcohol status of the patient (0 - No, 1 - Yes).
Physical activity	Categorical	Physical activity status of the patient (0 - No, 1 - Yes).
Cardiovascular disease (Target Variable)	Categorical	cardiovascular disease status of the patient (0 - No, 1 - Yes).
Age_cat (Grouping Variable)	Categorical	Early Midlife (30 - 45 years), Late Midlife (46 - 65 years)
BMI (Derived Variable)	Quantitative	Derived from weight and height.

Table 1.