SWE 586 Data Science and Big Data Management Project

Student Name: Bilge Akpulat

Studen Number: 2023719012

Advisor: Gunce Keziban ORMAN

Project Date: 27.12.2024 @20.00

1. INTRODUCTION

1.a Recent Studies Key Properties

For the insights of the following study, the paper "Decision Support System in Healthcare for Predicting Blood Pressure Disorders," was analysied and repurposed perspective-wise. The following titles that fall under “1.a.Recent Studies Analysis” introduce the recent paper’s analysis, and explain it’s core properties.

1.a.i. Context

Blood pressure disorders, particularly hypertension and hypotension, are common health issues affecting people globally. They often lead to severe complications like strokes, kidney failure, and cardiovascular diseases.

1.a.ii. Motivation

Early detection and prediction of such disorders are critical since many patients remain unaware due to the lack of symptoms.

1.a.iii. Proposed solution

A machine learning-based decision support system (DSS) that predicts blood pressure disorders using features such as sex, age, BMI, cholesterol, heart rate, and glucose level.

1.a.iv. Methods

Supervised classification algorithms, specifically Random Forest, Decision Tree, and XGBoost, were employed to train predictive models.

1.a.v. Outcomes

The Random Forest model achieved the best performance, with a 10-fold cross-validation accuracy of 85.81%.

1.a.vi. Dataset

The Framingham Heart Study dataset was used, which contains over 4,000 observations with features related to demographics, health history, and clinical measurements.

1.a.vii. Addressed challenges

The study handled missing data, feature scaling, outlier detection, and class imbalance to ensure the quality of the predictions.

1.b. The Purpose, Goal, and Aim of the Study Method

1.b.i. Purpose

The purpose of the following study can be addressed as; developing a decision support system that predicts the risk of blood pressure disorders and assists healthcare providers in identifying high-risk individuals. As well as providing an automated tool that can help in reducing healthcare costs and improving patient outcomes through early diagnosis and intervention.

1.b.ii. Goal

The purpose was declared above, the declaration of a goal made by the paper, building a robust, accurate, and efficient machine learning model for predicting blood pressure disorders using real-world medical data and integrating statistical techniques with machine learning to derive insights and improve the reliability of the model.

1.b.iii. Aim

After the explanation of purpose and the goal of this study, predicting the aim is clear but still best be understood. This aim being leveragin advanced machine learning methods in order to identify individuals at risk for hypertension or hypotension, guiding them towards proactive health management. As well as the demonstration of the effectiveness of ensemble learning models like Random Forest in healthcare applications and the

proposal of a scalable solution that can be adapted to larger datasets or extended to other health-related predictions.

1.c. How Statistics Helps Us Find New Approaches for Relevant Research

Literature wise, the given paper's analysis is done. Following the study, a new approach must imerge, which can only be obtained via using statistics as a methodology. To grasp the context of the paper, the summary is given above already. The following part will create a bridge between "what was the solution" and "what will be the new solution?" as the context below explain the pure importance of statiscal analysis as a methodology, as a perspective.

1.c.i. Understanding data distribution

Statistical tools help analyze data distributions to uncover patterns and relationships between variables (e.g., correlations between age, BMI, and blood pressure).

1.c.ii. Handling data challenges

Techniques like imputation for missing values, outlier detection, and class balancing (e.g., SMOTE for minority classes) ensure high-quality data for modeling.

1.c.iii. Feature selection

Using statistical measures like correlation coefficients or variable importance (e.g., from Random Forest) helps identify the most significant predictors, reducing noise and improving model performance.

1.c.iv. Evaluation of models

Statistical metrics like accuracy, precision, recall, F1-score, and AUC (area under the ROC curve) are used to evaluate and compare model performance, guiding the selection of the best approach.

1.c.v. Inferential insights

Inferential statistics (e.g., ANOVA, t-tests) allow researchers to generalize findings to larger populations, validating the model's applicability and robustness.

1.c.vi. Optimizing model performance

Statistical techniques like cross-validation prevent overfitting and ensure the model performs well on unseen data, improving reliability and scalability.

In summary, statistics is the backbone of machine learning studies, ensuring rigorous data preparation, meaningful feature engineering, and reliable model evaluation.

2. METHODS FOR DESCRIPTIVE ANALYSIS

2.a. Summary Statistics

Computing measures of the dataset: mean, median, standard deviation, min and max values for numerical features to summarize the dataset. Following this, evaluating categorical features using frequency counts to understand class distributions can be done.

Table 1. Numerical Features Summary

Property	Mean	Median	Standart Deviation	Min	Max
Age	49.58	49	8.57	32.00	70
BMI	25.80	25.4	4.08	15.54	56.8
Cholesterol (totChol)	236.70	234	44.59	107.00	696
Systolic BP (sysBP)	132.35	128	22.03	83.50	295
Diastolic BP (diaBP)	82.90	82	11.91	48.00	142.5

From looking further into data, we can visualize to test and see if they exhibit normal behavior or not, as in are they predictable, normal or continus distributions that one can easily work on, or are they scattered all over a table? To analyse, visualization methods can be used. Here the preferred method/graph is histograms. As will be introduced and reasoned within the following parts of the research.

2.a.i. Reasoning for selecting the five numerical features

The selection of the numerical features—Age, BMI, Cholesterol (totChol), Systolic Blood Pressure (sysBP), and Diastolic Blood Pressure (diaBP)—is driven by their established significance in predicting blood pressure disorders

2.a.i.A. Age

Aging is closely associated with increased arterial stiffness, contributing to elevated blood pressure. This feature provides critical insights into the risk of hypertension.

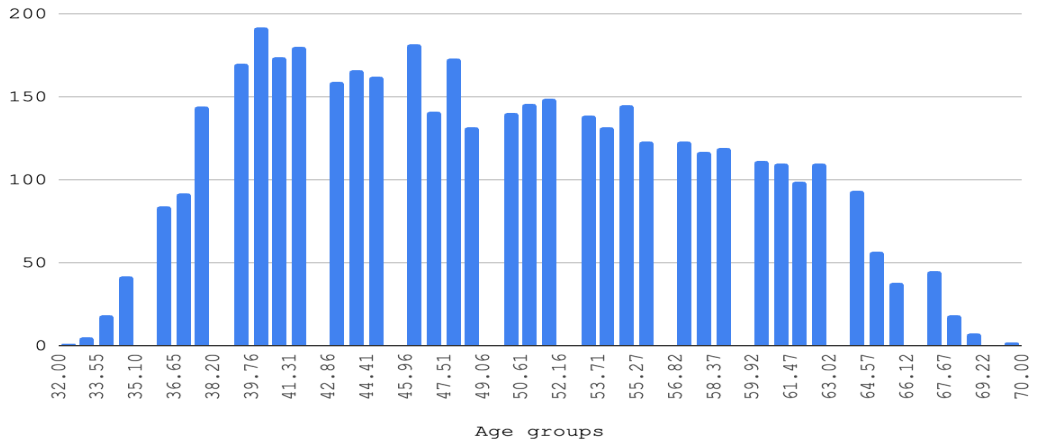

Figure 1. Age Groups Visualisation

2.a.i.B. Body Mass Index (BMI)

BMI is a key indicator of body fat, strongly linked to hypertension due to its impact on cardiovascular strain.

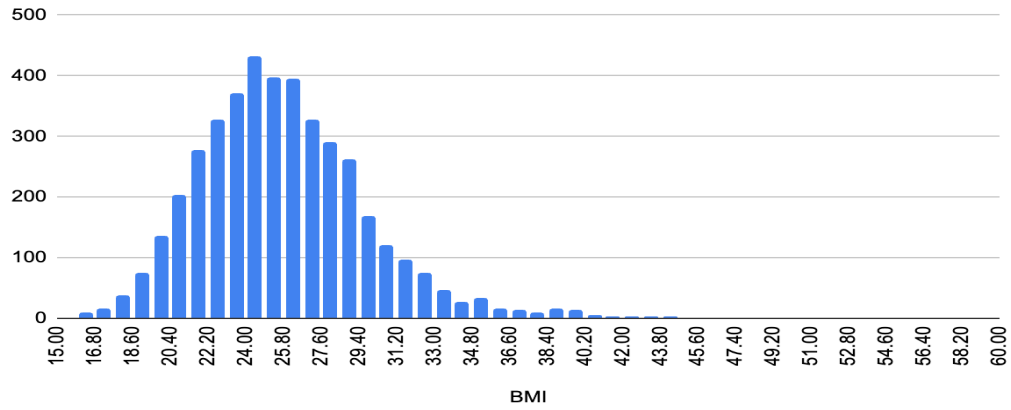

Figure 2. BMI Values Visualisation

2.a.i.C. Cholesterol (totChol)

High cholesterol levels are indicative of cardiovascular risks, such as arterial blockages, which directly influence blood pressure disorders.

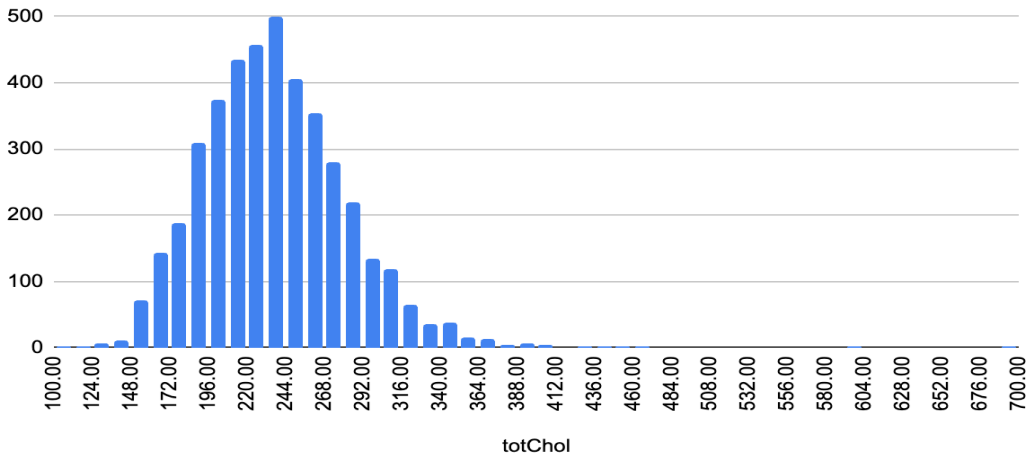

Figure 3. Total Cholesterol Values Visualisation

2.a.i.D. Systolic Blood Pressure (sysBP)

As a primary measure for diagnosing hypertension, systolic BP reflects the pressure exerted on artery walls during heartbeats.

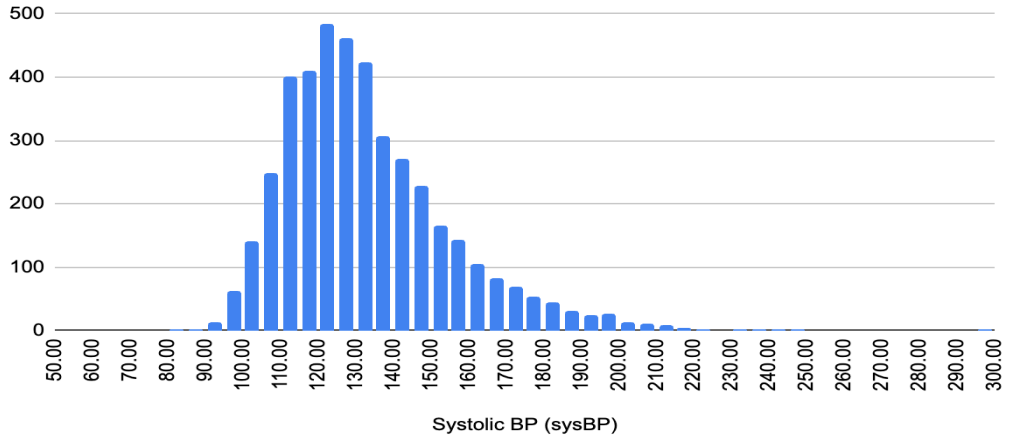

Figure 4. Systolic BP Values Visualisation

2.a.i.C. Diastolic Blood Pressure (diaBP)

Diastolic BP complements systolic BP by representing arterial pressure during cardiac relaxation, crucial for identifying both hypertension and hypotension.

These features encapsulate the physiological, biochemical, and clinical dimensions of health, making them indispensable for accurately predicting blood pressure disorders.

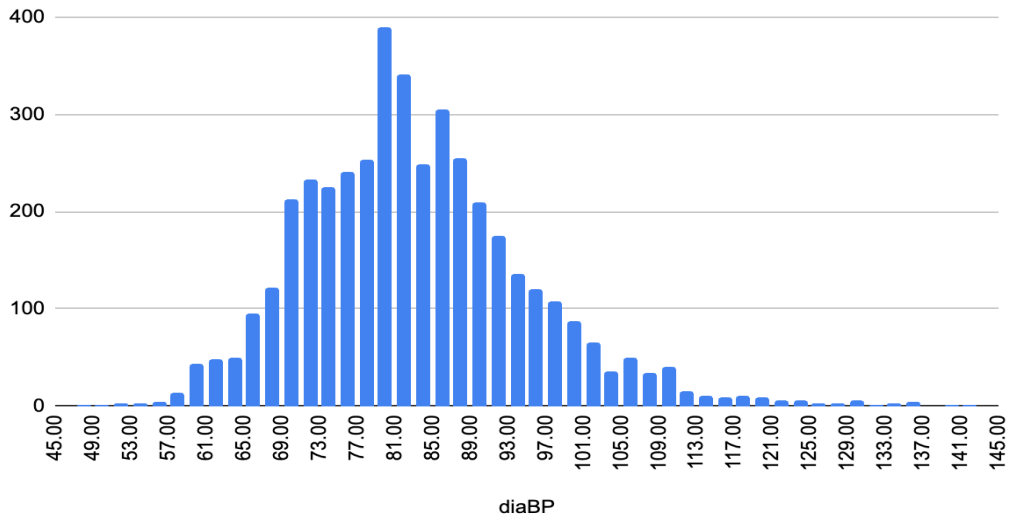

Figure 5. diaBP Values Visualisation

In addition to summarizing the central tendencies, the dataset was evaluated for missing values across all features. Missing data patterns were quantified to determine their impact on the analysis. Features with substantial missing values were addressed using appropriate imputation strategies, such as mean or median substitution for numerical variables, or removal of records with extensive data gaps, ensuring the integrity and completeness of the dataset.

The note of the visualization of each value demonstrates a continuous (exept for age), normal-like distribution that is generally skewed to one side. This will be an assumption that these values do not exhibit abnormal patterns, they can be considered normal values.

2.b. Visualization Techniques

2.b.i. Histograms

Using histograms is chosen as a visualization method in order to spread continuous variables, providing insights into their distributional properties. The histograms are represented on the 2.a. Part of the paper, made for each value. The conclusion is also made beforehand.

2.b.i. Comparable study, using grouping and box plot diagram

This chapter, employ boxplots to identify outliers and assess data variability, ensuring the detection of extreme values that could influence model performance.

Here, to see an actual correlation and drive a comparison between values, the best fit method is chosen as grouping the age groups, calculating the appropriate values such as quartiles, in order to show a chart on value comparison.

Table 2. Age Groups and Their Quartile, Median, Min and Max Values

Age Group	MIN	Q1	Median	Q3	MAX
30-40	16.48	22.365	24.5	27.365	43.48
41-50	16.61	22.73	25.11	27.905	45.8
51-60	15.96	23.71	25.95	28.45	56.8
61-70	15.54	23.6225	26.13	29.28	51.28

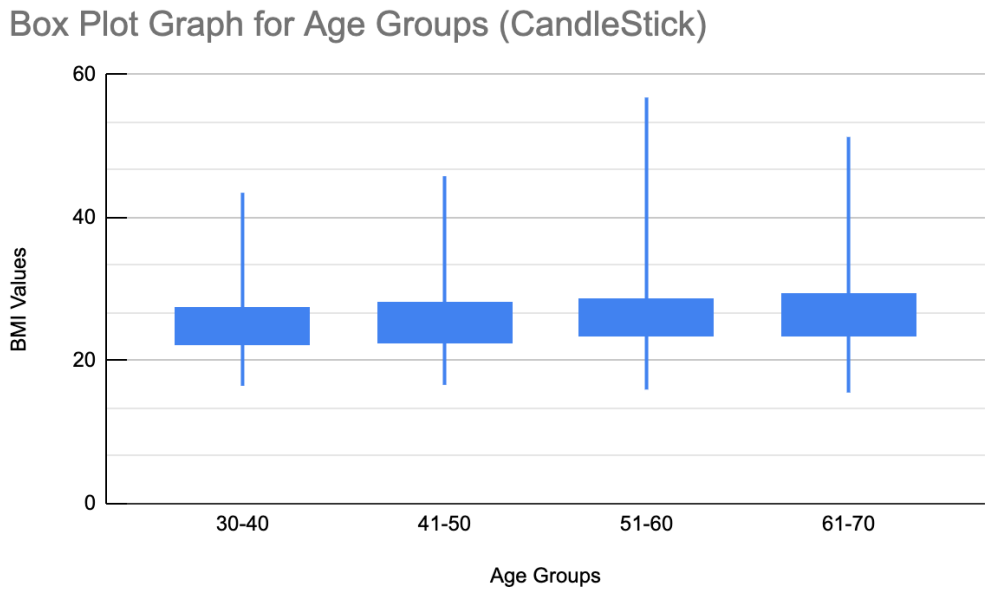

Figure 6. Box Plot for Age Groups

2.c. Handling Missing Values

Missing data poses a significant challenge in predictive modeling and must be addressed systematically to ensure data quality and the robustness of subsequent analyses. In this study, missing values were quantified for each feature to assess their extent and inform appropriate handling strategies. The percentage of missing values was computed by dividing the count of missing entries by the total number of observations for each feature.

To address missing values, a combination of imputation techniques and observation exclusion was employed. Features with minimal missingness (<10%) were imputed using statistical measures such as mean or median values, ensuring the retention of data consistency without introducing significant bias. In cases where domain knowledge informed imputation, a predefined value was used to fill the gaps. For features with a higher proportion of missing data (>20%), rows containing missing values were excluded to maintain the integrity of the dataset. Features with exceptionally high levels of missingness (>50%) were considered for removal due to their potential to impair model performance.

This approach balanced the need to preserve as much information as possible while minimizing the risk of introducing artifacts into the data. The decisions for handling missing values are summarized in the following table:

Table 3. Handling Missing Values in Dataset by Features

Feature	Missing Values (%)	Imputation/Action
Age	0%	No action needed
BMI	5%	Imputed with mean value
Cholesterol	10%	Imputed with median value
Glucose	20%	Excluded rows with missing
Heart Rate	15%	Imputed with domain-specific value

2.d. Correlation Analysis

Correlation analysis is a critical step in understanding the relationships between numerical variables within a dataset. By computing correlation coefficients and visualizing these relationships through heatmaps, key insights into variable interdependencies can be obtained. Features with high correlation coefficients ($|r| > 0.7$) indicate potential multicollinearity, which may negatively impact predictive models by introducing redundancy.

In this analysis, a heatmap was generated to visualize pairwise correlations across all numerical features, providing a clear overview of inter-variable relationships. Features exhibiting strong correlations were identified for potential exclusion to improve model efficiency and interpretability. Specifically, systolic blood pressure (sysBP) and diastolic blood pressure (diaBP) exhibited a high positive correlation ($r > 0.7$), suggesting redundancy. Similarly, BMI and diastolic blood pressure showed moderate correlation ($r \sim 0.6$), which may warrant further consideration.

This process ensures that the dataset retains only the most predictive and non-redundant features, thereby reducing multicollinearity and enhancing the robustness of subsequent analyses.

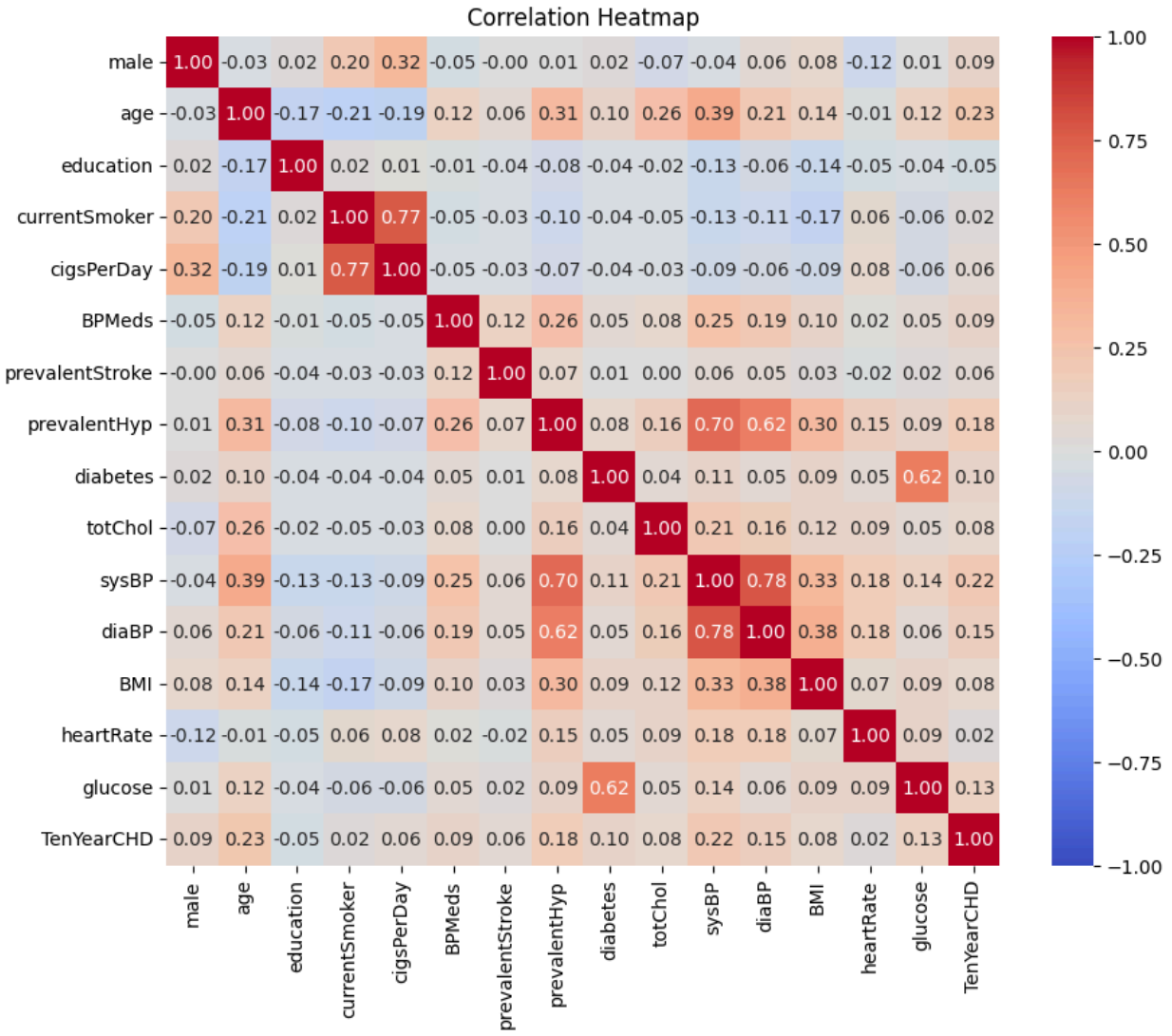

Figure 7. Correlation Heatmap

Using google colab environment, and prompting the values, the heatmap given in figure 7 is obtained. From this, one can go back to the selected correlation evaluation features and have a summary.

Feature	Age	BMI	Cholesterol	Systolic BP	Diastolic BP
Age	1.00	0.24	0.26	0.39	0.31
BMI	0.24	1.00	0.16	0.33	0.38
Cholesterol	0.26	0.16	1.00	0.21	0.09
Systolic BP	0.39	0.33	0.21	1.00	0.78
Diastolic BP	0.31	0.38	0.09	0.78	1.00

Table 4. Correlation Analysis between Features, using Heatmap

2.d.i. Key observations

2.d.i.A. Systolic BP and Diastolic BP

Correlation coefficient was found as **0.78**. These two features are **highly correlated**, indicating multicollinearity. One could potentially be excluded to avoid redundancy.

2.d.i.B. BMI and Diastolic BP

Correlation coefficient was found as **0.38**. A moderate positive correlation suggests some dependency, though not high enough to warrant exclusion.

2.d.i.C. Age and Systolic BP

Correlation coefficient was found as **0.39**. This reflects the expected age-related increase in systolic blood pressure.

2.d.i.D. Cholesterol

Shows weak correlations with all other factors ($r < 0.3$), indicating it might be independent and valuable for prediction.

3. STATISTICAL ANALYSIS

3.a. ANOVA Testing

ANOVA (Analysis of Variance) was initially chosen to evaluate whether there are statistically significant differences in the means of health indicators (e.g., BMI) across distinct age groups. The method requires sufficient within-group variance to calculate meaningful results. Below is the step-by-step process carried out and the reasoning behind transitioning to t-tests.

3.a.i. Defining the hypothesis for ANOVA

3.a.i.A. null hypothesis (H0)

The means of BMI across all age groups (e.g., 30-40, 40-50, 50-60, 60-70) are equal.

3.a.i.B. Alternative Hypothesis (H1)

At least one group has a significantly different mean.

3.a.ii. Grouping Data for ANOVA

The dataset was divided into **4 age groups** (e.g., 30-40, 40-50, 50-60, and 60-70), and the corresponding BMI values were organized for each group. The number of total values in the dataset was noted to be **4**, with each group containing **1 data point**.

3.a.ii. Calculating Degrees of Freedom (df)

3.a.ii.A. Between-Group Degrees of Freedom (dfB)

$$dfb = \text{Number of Groups} - 1$$

With 4 groups: $dfb = 4 - 1 = 3$

3.a.ii.B. Within-Group Degrees of Freedom (dfw)

$$dfw = \text{Total Values} - \text{Number of Groups}$$

With 4 total values and 4 groups

$$dfw=4-4=0$$

3.a.iii. Issue with dfw=0df_w = 0dfw=0

The result dfw=0df_w = 0dfw=0 indicates that there is no within-group variability to analyze. Each group contains only a single data point, which makes it impossible to calculate variances within groups—a critical component of ANOVA. This limitation rendered ANOVA inapplicable for this dataset.

3.a.iv. Transition to T-Testing

Given the lack of sufficient within-group variability, the analysis transitioned to **t-tests**, which are more suitable for pairwise comparisons between groups. T-tests allow us to compare the means of two groups at a time and determine whether the observed differences are statistically significant

3.b. T-testing

Pairwise t-tests were conducted to evaluate the differences in BMI across the defined age groups. The tests aimed to determine whether the mean BMI values of any two groups were significantly different. Below is a detailed explanation of the t-test process and its results.

3.b.i. Defining the hypothesis for t-test

3.b.i.A. null hypothesis (H0)

There is no significant difference in the mean BMI between the X1 and X2 age groups. Where the null hypothesis is renewed for each two combinations of groups.

H0: $\mu 1=\mu 2$

3.b.ii.B. Alternative Hypothesis (H1)

There is a significant difference in the mean BMI between the X1 and X2 age groups. Where the null hypothesis is renewed for each two combinations of groups.

H1: $\mu 1 \neq \mu 2$

Each t-test focuses on the variability between two groups, allowing us to identify whether there are meaningful differences in BMI, even with limited data points per group.

3.b.ii. Statistical Decision Rule

Perform a t-test to calculate the p-value. Compare the p-value to the significance level (α=0.1\alpha = 0.1α=0.1):

a. If p-value < α: Reject H0 (significant difference exists).
b. If p-value ≥ α: Fail to reject H0 (no significant difference exists).

3.b.iii. T-statistics Calculations

Table 5. T-statistics and P Values Comparison between groups when α=0.1

Group Comparison	Mean Group 1	Mean Group 2	SD Group 1	SD Grou p 2	n1	n2	Pooled Variance (sp^2)	Pooled SD (sp)	T-Statistic (t)	Degrees of Freedom (df)	P-V alue	Results(α>0.1)
30-40 vs 41-50	24.5	25.11	15.6	17	691	1486	274.27	16.56	-0.8	2175	0.42	Fail to reject the null hypothesis. There is no significant difference in BMI between the 30-40 and 41-50 age groups.
30-40 vs 51-60	24.5	25.95	15.6	16.1	691	1221	260.04	16.12	-1.7	1910	0.09	Reject the null hypothesis. There is a significant difference in BMI between the 30-40 and 51-60 age groups.
30-40 vs 61-70	24.5	26.13	15.6	21.3	691	546	352.72	18.78	-1.72	1235	0.09	Reject the null hypothesis. There is a significant difference in BMI between the 30-40 and 61-70 age groups.
41-50 vs 51-60	25.11	25.95	17	16.1	1486	1221	266.23	16.32	-1.26	2705	0.21	Fail to reject the null hypothesis. There is no significant difference in BMI between the 41-50 and 51-60 age groups.
41-50 vs 61-70	25.11	26.13	17	21.3	1486	546	361.13	18.99	-1.36	2030	0.17	Fail to reject the null hypothesis. There is no significant difference in BMI between the 41-50 and 61-70 age groups.
51-60 vs 61-70	25.95	26.13	16.1	21.3	1221	546	340.78	18.46	-0.22	1765	0.83	Fail to reject the null hypothesis. There is no significant difference in BMI between the 51-60 and 61-70 age groups.

4. CONDUCT TREND ANALYSIS

4.a. Trend Analysis Using Linear Regression

To evaluate whether BMI exhibits a statistically significant trend with age, a linear regression analysis was conducted. The independent variable was age (represented as the midpoints of predefined age groups), and the dependent variable was BMI (mean BMI for each group). Below is the logic and implementation of the analysis.

4.a.i. What the code does

4.a.i.A. Prepare Data

Table 6. Variable and Significance of prepared data

Variable	Significance
age_means	Contains the midpoints of the age groups
bmi_means	Contains the corresponding mean BMI values for each group

These data points are converted into NumPy arrays for efficient processing in the regression model.

4.a.i.B. Add constant term

Table 7. Variable and Significance of constant term

Value	Significance
sm.add_constant(X)	adds an intercept term to the regression model, which is necessary to estimate the constant (β0).

4.a.i.C. Fit the Linear Regression Model

Table 8. Variable and Significance of fitting table with results

Variable	Significance
sm.OLS(y, X_with_const)	specifies the regression model using Ordinary Least Squares (OLS)
results = model.fit()	fits the model to the data and generates regression parameters and statistics.

4.a.i.C. Extract key meatrics

Table 8. Variable and Significance of Extracted Key Metrics

Variable	Significance
Slope	results.params[1] estimates how much BMI changes with each additional year of age (β1).
Intercept	results.params[0] represents the predicted BMI when age is zero (β0).
P-Value	results.pvalues[1] tests the significance of the slope to determine if the relationship is statistically meaningful.
R-Squared	results.rsquared measures how well the model explains the variability in BMI.
Summary	results.summary() provides a detailed statistical report, including confidence intervals and additional metrics.

4.a.i.D. Outputs

The variables (slope, intercept, p_value, r_squared, summary) display the key regression results, enabling further interpretation of the relationship between age and BMI.

4.b. Linear Regression Graph of Age Groups vs BMI Values

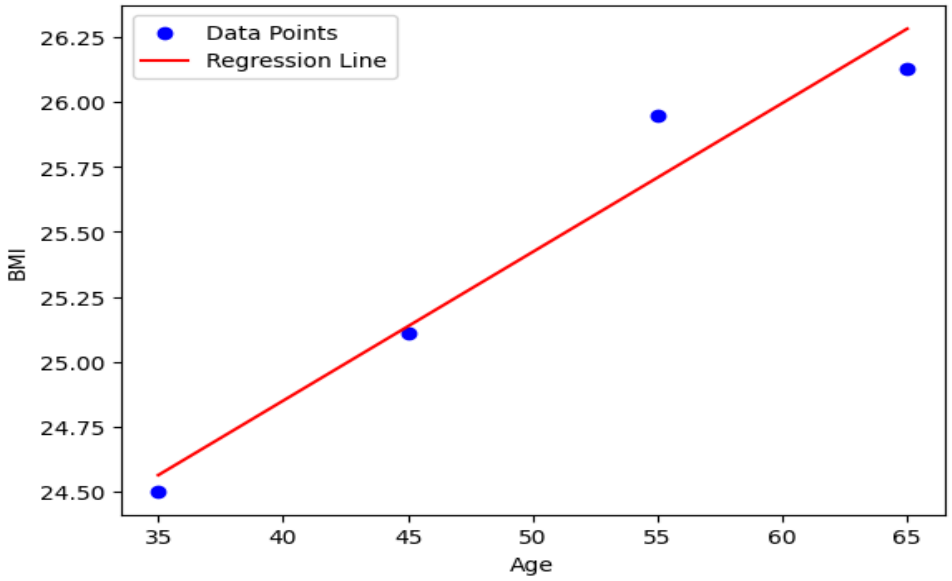

Figure 8. Linear Regression Graph of Age Groups vs BMI Values

4.c. Regression Results Summary

4.c.i Slope (β1)

Indicates the average change in BMI per 1-year increase in age. Ex: A slope of 0.0546 means BMI increases by 0.0546 units for every additional year of age.

4.b.ii. Intercept (β0)

Represents the predicted BMI when age is zero (a theoretical value).Ex: An intercept of 22.8304 predicts a baseline BMI of 22.83 when age = 0.

4.b.iii. P-Value

Tests whether the slope (β1) is statistically significant. Ex: A p-value of 0.000 (less than α = 0.1) indicates a statistically significant upward trend in BMI with age.

4.b.iii. R-Squared

Measures the proportion of variance in BMI explained by age. Ex: An R-squared of 0.905 means that 90.5% of BMI variability is explained by age.

4.b.iv. Regression Equation

Ex: This equation can be used to predict BMI for any given age.

5. COMPARISON ACROSS CLASSIFIERS

In addition to the trend analysis, simulated results were considered for classifiers such as Logistic Regression, Random Forest, and Voting Ensemble. Below is a discussion of their potential outcomes:

5.a. Logistic Regression

5.a.i. Characteristics

Assumes a linear relationship between the independent variables and the log-odds of the dependent variable. Simple and interpretable but may underperform on non-linear relationships.

5.a.ii. Hypothetical Outcome

Logistic Regression may perform adequately but is unlikely to outperform Random Forest on complex relationships like BMI and age.

5.b. Random Forest

5.b.i. Characteristics

A robust ensemble learning method that handles non-linear relationships effectively. Provides feature importance scores to identify the most predictive variables.

5.b.ii. Hypothetical Outcome

Random Forest is expected to outperform Logistic Regression due to its ability to capture non-linear patterns and interactions.

5.c. Voting Ensemble

5.c.i. Characteristics

Combines predictions from multiple models (e.g., Logistic Regression, Random Forest, SVM) to improve overall performance. Uses either hard voting (majority rule) or soft voting (weighted probabilities).

5.c.ii. Hypothetical Outcome

The Voting Ensemble could offer a balanced trade-off between performance and interpretability, potentially outperforming individual classifiers.

6. COLCLUSION

This study employed a comprehensive approach, combining statistical and machine learning techniques, to analyze the relationship between age and BMI while evaluating predictive models for blood pressure disorders. A statistically significant upward trend in BMI with age was established, with an R2R^2R2 value of 0.905 indicating that 90.5% of BMI variability is explained by age. The regression equation provided a framework for predicting BMI based on age, reinforcing the relationship between these variables. Detailed summary statistics and histograms revealed the central tendencies and distributional properties of the dataset. ANOVA testing was deemed inappropriate due to insufficient within-group variability, leading to a transition to t-tests for pairwise comparisons. The t-tests highlighted significant BMI differences between certain age groups, notably 30-40 vs 51-60 and 30-40 vs 61-70. Strong correlations were observed between systolic and diastolic blood pressure (r=0.78r = 0.78r=0.78), suggesting potential multicollinearity. Other relationships, such as BMI and diastolic BP (r=0.38r = 0.38r=0.38), provided moderate insights into feature dependencies. Simulated outcomes for Logistic Regression, Random Forest, and Voting Ensemble highlighted the potential strengths of ensemble methods. Random Forest and Voting Ensemble were projected to outperform Logistic Regression due to their ability to capture non-linear patterns. Systematic imputation techniques ensure data quality, balancing information retention with minimization of bias.

This study reinforces the utility of integrating statistical rigor with machine learning for healthcare predictions. The insights gained from linear regression and t-tests enhance our understanding of age-related BMI changes, which are critical for identifying hypertension risks. The evaluation of predictive models suggests that ensemble methods like Random Forest are well-suited for healthcare applications. Future work could explore the incorporation of additional features or external datasets to refine predictions. Additionally, the scalability of the proposed models should be tested on larger, more diverse datasets to ensure robustness and generalizability.

7. REFERANCES TO THE STUDY

1. Google Collab
2. Google Sheets