STATISTICAL STUDY ON HIV PATIENTS' DATASET

Project submitted to

MAR ATHANASIUS COLLEGE (AUTONOMOUS)

KOTHAMANGALAM

In partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

IN

STATISTICS

By

ANN LIN DAMI

Reg. No: 23PST8682

PG DEPARTMENT OF STATISTICS

MAR ATHANASIUS COLLEGE (AUTONOMOUS)

KOTHAMANGALAM

JUNE 2025

MAR ATHANASIUS COLLEGE (AUTONOMOUS), KOTHAMANGALAM

PG Department of Statistics

CERTIFICATE

This is to certify that the project entitled **“Statistical Study On HIV Patients’ Dataset”** is an original and authentic record of study carried out by **Ann Lin Dami** under my guidance and supervision, in partial fulfilment of the requirement for the award of the degree of Master of Science in Statistics, during the academic year **2024-25** and that it has not been previously submitted to the award of any degree, diploma, fellowship or any other similar title or recognition.

Ms. Elby Elias
(Supervising Teacher)

Ms. Sari Thomas
(Head of the department)

Place: Kothamangalam

Date: 20/06/2025

ACKNOWLEDGEMENT

I would like to express my profound gratitude to Ms. Sari Thomas, Head of the PG Department of Statistics, Mar Athanasius College (Autonomous), Kothamangalam and other faculty members of PG Department of Statistics for their valuable advice and encouragement to complete this work.

Finally, I express my heartiest thanks to my dear friends, who helped me in one way or the other to complete my project successfully in a time schedule and above all I thank almighty for making this possible.

KOTHAMANGALAM ANN LIN DAMI

20/06/2025

CONTENTS

CHAPTER – I

INTRODUCTION

1.1 STATISTICS

Statistics is the language of the data. It is a branch of mathematics that deals with the collection, organization, analysis, interpretation, and presentation of data. Statistics are used in nearly every field to support decision-making, from predicting the weather to evaluating a company's performance. This allows us to look beyond individual facts and numbers and discover patterns, relationships, and trends that are otherwise hidden. Whether comparing products, analysing social media trends, or assessing the effectiveness of a new medicine, statistics provides the tools to make informed and objective decisions.

In data analysis, statistics are crucial for extracting meaningful insights from intricate datasets. Simple metrics, such as the mean, median, and standard deviation, can be used to summarize vast amounts of data using statistical methods. Data visualization is the process of rapidly identifying distributions, trends, and outliers using graphs and charts. Through inferential techniques, statistics enable us to do more than merely describe the data; they also enable us to test hypotheses and make predictions about the data. For instance, we can ascertain whether a difference between two groups is significant or the result of pure chance. While regression models enable us to forecast results, tools such as correlation analysis aid in understanding the relationships between variables.

Every stage, from the first data cleaning and investigation to the last results interpretation, was directed by statistical reasoning. Our goal is to turn unprocessed data into useful insights using both descriptive and inferential methods. Statistics are powerful not only for calculating numbers but also for assisting us in making more informed decisions. This study shows how a methodical statistical approach can yield insightful information and offer a more thorough understanding of the issue at hand.

1.2 TYPES OF STATISTICS

Descriptive and inferential statistics are the two basic categories into which statistics is generally separated.

Methods for meaningfully organizing and summarizing data are included in the descriptive statistics. It can be challenging to see the big picture when working with a large amount of information. Simple numbers and images that depict the appearance of the data are provided by descriptive statistics. For instance, metrics that reveal the centre or typical value of a dataset include the mean (average), median (middle value), and mode (most frequent value). We can determine the degree of data dispersion by examining measures of dispersion, such as the range, variance, and standard deviation. To make the data easier to understand, we also employed tools such as tables, bar charts, pie charts, and histograms. Descriptive statistics help us take the first step, to know and understand our data, even though they do not make predictions or generalizations outside of the data at hand.

Using a sample of data, inferential statistics aid in forecasting and decision-making for a larger population. We rely on a smaller, representative group known as a sample because it is frequently impractical or impossible to gather data from an entire population. This sample was used in inferential statistics to draw conclusions about the entire population. To determine whether a new drug is effective, for instance, we test it on a sample of people and then use statistical techniques to determine how it might affect the population. Confidence intervals, p-values, and hypothesis testing are frequently used tools in inferential statistics. These techniques assist us in determining the likelihood that our results are the result of random variation or accurately represent relationships or differences. Inferential statistics is particularly potent because it enables us to go beyond simply summarizing past events and begin generating predictions, assessing theories, and assisting in making decisions in the face of uncertainty.

1.3 STATISTICAL DATA

Statistical data refer to any information collected, measured, and analysed using statistical methods. It is the raw material used in statistics to understand patterns, test hypotheses and make decisions. These data can come in many forms, including numbers, categories, counts, or measurements, and are often gathered through surveys, experiments, observations, or existing records. Statistical data are divided into qualitative and quantitative data.

Qualitative data, also called categorical data, are used to describe qualities or characteristics that cannot be measured with numbers but can be sorted into groups or categories. Eye colour and gender are qualitative data. They help us label or categorize information, but we cannot perform mathematical operations, such as averaging them.

Quantitative data contain numerical data that represent quantities or amounts. This type of data can be analysed and measured mathematically. A person’s age, height, weight, and number of items sold are all quantitative. Quantitative data are further divided into:

- Discrete data: Consists of whole numbers that cannot be divided into smaller parts. For example, the number of students in a class and the number of cars parked in a parking lot.
- Continuous data: Can accept any value within a range, including decimals, such as a person's height or temperature, which are more accurately measurable.

The level of measurement is another important idea in statistics. It tells us how data are grouped and what types of analysis we can do with them. It helps us figure out the best way to use and understand the data. There are four main levels:

- Nominal data are the simplest type used for labelling without any specific order. For example, blood type (A, B, AB, O) or hair colour (black, blonde, brown) are nominal because the categories have no ranking.
- Ordinal data involve categories that have a meaningful order, but the intervals between them are not equal. An example is a satisfaction survey where responses are ranked as "poor," "fair," "good," or "excellent." The order matters, but we can't say how much better "excellent" is compared to "good."
- Interval data have ordered categories with equal spacing between values but no true zero point. For instance, temperature in Celsius or Fahrenheit is interval data, 0 degrees does not mean "no temperature," and it cannot be said that 20°C is twice as hot as 10°C. Ratio data is the highest level, with all the properties of interval data, plus a true zero-point. For example, weight, height, and age are ratio data because 0 means "none," and it can be said that someone who weighs 80 kg is twice as heavy as someone who weighs 40 kg. Understanding these types of data is important because they determine the types of charts, calculations, and statistical tests that can be used in a data analysis project.
- The highest level of measurement is the ratio level, which comprises data with all the characteristics of interval data, plus a true zero point that denotes the absence of the quantity being measured. This true zero makes it possible to use multiplication and division to compare values in meaningful ways, such as when determining ratios or percentages. Height, weight, age, and quantity sold are examples of ratio data.

1.4 DATA COLLECTION METHODS

Data collection is the process of gathering information to analyse and draw conclusions. This is a fundamental step in statistical analysis, as the quality and reliability of the conclusions depend largely on how the data are gathered. The two types of data are as follows:

- Primary data: The type of data collected directly by a researcher for a specific purpose is called primary data. It is gathered first-hand through methods such as surveys, experiments, interviews, and observations. This type of data is often more accurate and up-to-date but can be time-consuming and expensive to collect.
- Secondary data: Secondary data refer to existing information that was originally gathered for another study or objective but is now being analysed or applied to a new research question or context. Examples include government reports, company records, academic articles, and online databases. While secondary data are quicker and cheaper to obtain, they may be outdated or not perfectly suited to the new research question.

Several methods are commonly used to collect data, each suited to different types of studies and research objectives.

One of the most straightforward methods is surveys and questionnaires, where participants are asked to provide information through a series of questions. It can reach a large number of people quickly and cost-effectively; hence, this approach is widely used. Surveys can be conducted in various formats, such as face-to-face interviews, telephone calls, online forms or mailed paper questionnaires. The design of the questions plays a crucial role in obtaining accurate, unbiased data.

The observation method observes and documents events or behaviours as they naturally occur. This approach is often used in social sciences and behavioural research, where direct interaction may influence responses.

Experiments are a method in which researchers actively intervene and manipulate one or more variables to study their effects on a response variable. This method is particularly valuable for establishing cause-and-effect relationships. Experimental data collection requires careful planning to control confounding factors and often involves the random assignment of subjects to different treatment groups to ensure valid comparisons.

In addition, existing records and databases are a valuable source of data, especially when collecting primary data is impractical or impossible. Researchers can use administrative records, census data, medical records, and company databases to analyse trends and patterns. While this method saves time and resources, it may limit researchers' control over data quality and relevance.

Sampling methods are often used to gather data from a subset of a population when studying the entire group is not feasible. Sampling can be random, systematic, stratified, or cluster-based, each with its own advantages depending on the research goals and population structure. Proper sampling techniques ensure that the collected data accurately represent the broader population.

1.5 LIMITATIONS OF STATISTICS

Although statistics is a popular tool for data analysis and conclusion-making, it has some significant limitations that should be understood to prevent misinterpretations.

One of the main drawbacks is that statistical techniques cannot establish causation; they can only identify correlations and patterns. For instance, a study may indicate that individuals who consume more coffee generally perform better at work, but this does not prove that coffee is the cause of improved performance; instead, it may be the result of other factors such as stress or sleep. Furthermore, the quality of the data used determines the quality of the statistical analysis. The analysis will be flawed if the data are erroneous, biased, lacking, or gathered improperly. Missing values, poor sampling strategies, and measurement errors can skew statistical findings. Another issue is the tendency of statistics to oversimplify complex situations. False impressions may result from summary statistics that obscure data extremes or variations, such as means or percentages.

Intentional or inadvertent manipulation of the statistics is possible. To strengthen their case, people may choose to employ unsuitable statistical tests, present data selectively, or create deceptive graphs. Statistics can be manipulated to support specific agendas in media reports, political campaigns, and advertisements. Another drawback is that accurate statistical interpretation frequently requires specialized knowledge. In particular, for non-experts, concepts such as p-values, confidence intervals, and hypothesis testing are not always clear-cut and are prone to misunderstanding. Inaccurate results can arise from using an incorrect test, breaking assumptions (such as independence or normality), or extrapolating from small sample sizes.

Additionally, statistics are better at summarizing group behaviour than at accurately predicting individual outcomes. Even the most accurate statistical models cannot predict the future of a single individual or case. Finally, there is uncertainty associated with all statistical conclusions. No analysis can ever be completely certain, and there is always a chance of mistakes, including type I and type II errors in hypothesis testing. Consequently, even though statistics is a useful and frequently required tool for research and decision-making, it should always be applied carefully, and its findings should be interpreted appropriately.

1.6 APPLICATIONS OF STATISTICS

Statistics can be used to tremendous advantage in almost any field that gathers, examines, and interprets data. Professionals can test hypotheses, find patterns, make better decisions, and forecast future results by using statistical methods. A comprehensive list of the primary domains in which statistics are frequently used is provided below:

1. Economics and Business

It is used in economics and business to make data-driven decisions. Financial analysts employ statistical models to predict stock trends and control risks. Market researchers use it to examine consumer preferences and purchasing patterns.

2. Medical Care and Treatment

Statistics play a major role in the planning and evaluation of clinical trials to determine the effectiveness of new medical treatments. Biostatistics supports medical research through survival analysis, diagnostic test evaluation and population health studies.

3. Public Policy and Government

For planning and assessment, governments mainly rely on statistics. National censuses and demographic studies aid policy formation and resource distribution. Crime rates, unemployment rates, and inflation indices are monitored to inform economic planning and law enforcement.

4. Psychology and Education

Statistics are used in education to create standardized assessments, evaluate student performance, and enhance course offerings. Psychologists evaluate reliability, validate tests, and interpret the findings of behavioural research using statistical techniques.

5. Engineering and Science

Engineers use statistical techniques for performance optimization, reliability testing, and quality assurance. Researchers use statistical tools to quantify uncertainty, model complex systems, and draw conclusions from experimental data in disciplines such as biology, chemistry, and physics.

6. Environmental Science and Agriculture

Agricultural scientists use statistics to evaluate soil quality, analyse crop yields, and improve farming methods. Statistics are used in environmental science to study wildlife populations, model patterns of climate change, and track pollution levels.

7. Entertainment and Sports

Statistics are used in sports to forecast game results, assess player performance, and create team strategies. Statistical models are used by sports analysts to make decisions about training, drafting, and game day.

8. Data science and technology

Statistics are used to assess model performance, decipher patterns in sizable datasets, and assist algorithms in classification, prediction, and recommendation systems. Tech companies frequently use A/B testing, a statistical technique, to improve the user experience.

9. Demographics and Sociology

Sociologists measure inequality, examine demographic trends such as migration, aging, and birth rates, and analyse social behaviours using statistics. Surveys and longitudinal studies rely on statistical methods to produce reliable and generalizable findings.

10. Transportation and Logistics

Statistics are used in transportation to study traffic patterns, predict congestion, and design efficient public transit systems. Logistics companies apply statistical models to optimize delivery routes, estimate shipping times and forecast demand.

1.7 SOURCE OF THE DATASET

The dataset used for this analysis was sourced from Kaggle, a reputable online platform for data science and machine learning. The ‘AIDS Virus Infection Prediction’ dataset contains healthcare statistics and categorical variables related to individuals diagnosed with HIV. It was originally published in 1996 and is frequently used for educational and research purposes in epidemiology and public health fields.

1.8 DATASET DESCRIPTION

The dataset consists of medical and demographic information on patients diagnosed with HIV infection. The dataset contains 50,000 patient records with 23 variables, out of which 9 variables are continuous and the rest are categorical.

time : Time in days to either treatment failure or censoring

trt : Treatment group indicator (0=ZDV; 1=ZDV+ddI, 2=ZDV+Zal, 3=ddI)

age : Age at baseline

wtkg : Weight at baseline (in kg)

hemo : Haemophilia (0=No, 1=Yes)

homo : Homosexual activity (0=No, 1=Yes)

drugs : History of intravenous (IV) drug use (0=No, 1=Yes)

karnof : Karnofsky score, measuring the patient's general functional status (0-100)

oprior : Prior use of non-ZDV antiretroviral therapy before day 175 (0=No, 1=Yes)

z30 : Whether the patient received ZDV in the 30 days prior to day 175(0=No, 1=Yes)

preanti : Number of days the patient was on antiretroviral therapy before day 175

race : Race (0=White, 1=Non-White)

gender : Gender (0=Female, 1=Male)

str2 : Antiretroviral experience (0=Naive, 1=Experienced)

strat : Antiretroviral treatment history stratification

symptom : Symptomatic status at baseline (0=Asymptomatic, 1=Symptomatic)

treat : Simplified treatment indicator (0=ZDV only, 1=others)

offtrt : Indicates whether the patient went off treatment before week 96 (0=No,1=Yes)

cd40 : CD4 cell count at baseline

cd420 : CD4 count 20 weeks after baseline.

cd80 : CD8 cell count at baseline

cd820 : CD8 count 20 weeks after baseline

infected : Infected with AIDS (0=No, 1=Yes)

1.9 RESEARCH OBJECTIVES

1. To examine the association between the treatment regimen and treatment outcome.
2. To compare the CD4 counts of the patients at baseline and after 20 weeks.
3. To compare the time to liver failure among individuals receiving different treatments.
4. To classify patients as high- or low-risk for rapid AIDS progression using CD4/CD8 counts, behavioural indicators, and age.
5. To identify the subgroups of variables.

1.10 METHODS USED FOR DATA ANALYSIS

1. CHI-SQUARE TEST OF INDEPENDENCE

The Chi-Square test of independence is a statistical test used to examine whether two categorical variables are related or independent of each other. It is especially useful in research involving survey responses, treatment outcomes, or any situation in which data are grouped into categories. The basic idea is to organize the data into a contingency table, where one variable's categories are listed in rows and the other in the columns. Each cell in the table shows how many observations fall into a specific combination of categories.

The test then calculates the expected counts in each cell if the two variables were truly independent (i.e., no association between them). It compares these expected counts with the actual observed counts using a formula that produces a chi-square (χ^2) statistic. A large difference between the observed and expected counts indicates that the variables are likely associated in some way.

The calculated χ^2 value is then compared to a critical value from the chi-square distribution. The critical value depends on the degrees of freedom and chosen significance level (0.05). If the calculated chi-square value is greater than or equal to the critical value, the result is considered statistically significant, indicating sufficient evidence to reject the null hypothesis of independence. This suggests an association between these variables. Alternatively, a p-value is computed. If the p-value is smaller than the significance level (0.05), the null hypothesis of independence is rejected, suggesting a statistically significant relationship between variables.

In the end, the chi-square test of Independence is an important tool for looking at the connection between two categorical variables. People in many fields, including social science, psychology, business, and healthcare, often use it to look for patterns and connections in data that involves counts or categories. This test helps researchers figure out if changes in the data they see are probably due to chance or if there is a real link between the variables.

2. WILCOXON SIGNED RANK TEST

Two related samples are compared using the Wilcoxon Signed Rank test. It is also used as an alternative to the paired t-test when the data do not follow a normal distribution. This test is well-suited for evaluating whether there is a significant difference in the central tendency of paired observations, such as measurements taken before and after a treatment or intervention.

This test is typically applied when the data is paired, such as before-and-after measurements (e.g., CD4 counts of patients at baseline and after 20 weeks), and the interest is to check whether there is a significant change between the two time points. Rather than comparing means directly, like a t-test does, the Wilcoxon signed rank test works by looking at differences in ranks of the paired observations.

The Wilcoxon signed rank test begins by calculating the difference between each pair of related observations (e.g., before and after values). Then, exclude any pairs where the difference is zero. For the remaining pairs, take the absolute values of the differences and rank them from smallest to largest, assigning average ranks in case of ties. Depending on whether the initial differences were positive or negative, reapply the original signs (positive or negative) to the ranks. Once the ranks have been assigned, the total of the positive ranks and the total of the negative ranks are calculated separately, and the smaller of these two totals is taken as the test statistic, which is then used to judge whether the observed differences in the paired data are likely due to chance causes.

The test then evaluates this statistic against a threshold from the Wilcoxon distribution, or interprets it through a p-value, to assess whether the observed differences reflect a meaningful pattern rather than random variation. Reject the null hypothesis and conclude that there is a significant difference between the paired measurements if the p-value is less than the significance level (e.g., 0.05).

This test is especially useful when dealing with small sample sizes or ordinal data, or when the assumption of normality for a t-test cannot be satisfied.

3. KRUSKAL-WALLIS TEST

The Kruskal-Wallis test is a non-parametric method used to evaluate whether there are meaningful differences in the median values across three or more independent groups. It is built upon the logic of the Mann-Whitney U test. It is also used as an alternative to one-way ANOVA in situations where the data does not satisfy the assumptions.

This test is particularly useful when analysing ordinal data or continuous data that deviate from a normal distribution, especially when the groups being compared are independent. For instance, it can be applied to assess differences in median CD4 counts among patients receiving three distinct treatment protocols.

To perform the Kruskal-Wallis test, first combine all the data from the different groups and rank all the values together from lowest to highest, assigning average ranks in the case of ties. Then, for each group, compute the total of ranks. Using these rank sums, the test computes a statistic that reflects how much the group's rank distributions differ from each other. If the rank totals are roughly equal, then the groups have similar distributions. A large difference in rank sums suggests that at least one group median is significantly different from the others.

The test statistic follows a chi-square distribution, and the p-value is used to determine statistical significance. We shall reject the null hypothesis and conclude that at least one group differs significantly if the p-value is less than the significance level. However, the Kruskal-Wallis test doesn't tell which specific groups differ, so post-hoc pairwise comparisons (like Dunn's test) are needed if the result is significant.

4. RANDOM FOREST CLASSIFIER METHOD

The Random Forest Classifier is a popular and powerful machine learning method used for solving classification problems. It is based on a concept called ensemble learning, which means it combines the results of multiple models to make a final prediction. Here, the individual models are decision trees. A decision tree is like a flowchart that splits data based on certain conditions to make a prediction, but a single decision tree can be sensitive to changes in the data and might overfit, meaning it performs very well on training data but poorly on new, unseen data.

To solve this problem, random forest builds many decision trees instead of just one. Each tree is built using a randomly drawn subset of the data, selected with replacement through a process known as bootstrapping. As a result, certain data points may be included multiple times in a single sample, while others might be excluded entirely from that tree's training set. Furthermore, at each decision point in a tree (called a node), only a random subset of features is considered when choosing how to split the data. These two layers of randomness (random data and random features) ensure that the trees in the forest are diverse, which helps reduce the risk of overfitting.

Once all the trees are trained, the random forest makes a prediction by allowing each tree to cast a vote for the predicted class of a given input. The final prediction is the class with the most votes. This voting process is what gives the model its name, "forest", because it is a collection of many trees working together. By averaging the decisions of many independent trees, random forest tends to produce more accurate and stable predictions than a single decision tree.

5. HIERARCHICAL CLUSTERING METHOD

Hierarchical Clustering is a technique used to group similar data points into clusters based on how close they are to each other. It builds a hierarchy, or a tree-like structure, that shows the order in which groups (or clusters) are formed. At the start, each data point is treated as its own separate group. The algorithm then finds the two closest groups and joins them together.

This process of merging continues step by step until all data points are joined into one large cluster. The results are displayed in a dendrogram, a diagram that shows the merging process like a family tree, helping us understand how the data is organized and where natural groupings exist.

There are two main types of hierarchical clustering: agglomerative and divisive**.** Agglomerative clustering is more common and follows a "bottom-up" approach, where each point starts on its own and clusters are built by merging the nearest pairs. Divisive clustering works in the opposite way, starting with all data points in a single cluster and splitting them apart based on differences. One advantage of hierarchical clustering is that there is no need to choose the number of clusters in advance. The number of clusters can be determined by cutting the dendrogram at a specific level. It works best with small to medium-sized datasets and gives a clear visual picture of how data is structured.

P-VALUE

A p-value is used to evaluate the strength of evidence against a null hypothesis. The null hypothesis typically represents a default assumption, such as the absence of an effect, difference, or relationship between variables.

A small p-value (less than 0.05) suggests that the observed results are unlikely to have occurred purely by random chance if the null hypothesis were true. This provides evidence to reject the null hypothesis in favour of the alternative, indicating the potential presence of a significant effect. Conversely, a large p-value implies that the observed results are consistent with what might be expected under the null hypothesis, and therefore, there is insufficient evidence to warrant its rejection.

It is important to emphasize that the p-value does not represent the probability that the null hypothesis is true, nor does it convey the magnitude or practical importance of an effect. Rather, it serves as a tool for assessing statistical significance within the context of a given test. Accordingly, p-values should be interpreted in conjunction with other metrics, such as effect sizes, confidence intervals, and domain-specific knowledge, to draw meaningful and robust conclusions from data analysis.

1.11 SOFTWARE USED

SPSS

SPSS (Statistical Package for the Social Sciences) is an efficient software tool used for statistical analysis developed by IBM. SPSS offers both a graphical user interface (GUI) and a syntax-based command language, making it accessible to both beginners and advanced users. It is designed to handle large datasets and perform a wide variety of statistical analyses, ranging from basic descriptive statistics to complex inferential procedures.

One of the core strengths of SPSS lies in its ability to manage and prepare data for analysis. The data can be easily imported from sources such as excel, CSV files, or databases. The software provides tools for cleaning data, handling missing values, recoding variables, creating new computed fields, and selecting specific subsets of data for focused analysis. This stage is essential for ensuring the accuracy and reliability of the final statistical results.

SPSS supports a wide range of statistical techniques. Descriptive statistics allow users to summarize the main features of a dataset using measures such as mean, median, standard deviation, and frequency distributions. For exploring relationships between variables, SPSS offers correlation analysis, crosstabs, and visualizations such as histograms, bar charts, and scatterplots. More advanced procedures include t-tests, ANOVA (Analysis of Variance), regression analysis, chi-square tests, and non-parametric tests, all of which can be conducted through simple menu selections. Beyond basic statistics, SPSS also includes advanced modules for tasks such as survival analysis, factor analysis, and cluster analysis. These are particularly useful in research that involves time-to-event data, underlying factor structures, or grouping similar cases.

PYTHON

Python is a high-level programming language for data analysis and statistical computing. Its versatility, open-source nature, and rich ecosystem of libraries make it ideal for handling all stages of the data analysis pipeline, from data collection and cleaning to visualization and statistical modelling. In contrast to graphical tools like SPSS, python relies on code, which makes it more flexible and scalable, especially when working with large datasets or automating repetitive tasks.

Python provides a strong foundation in statistical analysis through libraries like NumPy for numerical operations, Pandas for data manipulation, and Seaborn & Matplotlib for data visualization. These libraries allow users to clean and explore data by filtering rows, creating new variables, dealing with missing values, and generating summary statistics. The Pandas DataFrame, in particular, functions similarly to a spreadsheet or SQL table and is central to most data analysis workflows in python.

One of python’s key advantages is its ability to integrate statistical analysis with other tasks, such as web scraping, automation, and database management. Analysts can write reusable scripts to automate workflows, apply statistical models to real-time data, or even build interactive dashboards using tools like Plotly and Dash.

1.12 ORGANIZATION OF THE PROJECT

There are five chapters in this project, each of which is connected to a specific research goal and the right statistical method to use to reach it. These chapters work together to create a clear and logical flow that takes the reader through the steps of analysis, from looking at the data to making sense of the results.

The first chapter introduces the role of statistics in healthcare and medical research, particularly in understanding patient outcomes and treatment effectiveness. It highlights the importance of statistical analysis in drawing meaningful insights from health data. The chapter also provides a detailed overview of the dataset used, including variables. A brief explanation of the statistical methods employed throughout the study is also presented to give readers a clear framework.

The second chapter focuses on examining the association between treatment regimen and treatment outcome**,** utilizing the chi-square test to determine whether there is a statistically significant relationship between the two categorical variables. The third chapter analyses the difference in CD4 counts before and after 20 weeks of treatment using the Wilcoxon signed-rank test**,** which is appropriate for comparing paired non-parametric data. The fourth chapter investigates the differences in time to liver failure among patients receiving different treatments, applying the Kruskal–Wallis test to assess whether significant differences exist across the multiple treatment groups. In the fifth chapter, a random forest classifier is developed to classify patients as high-risk or low-risk for rapid AIDS progression, using a combination of CD4/CD8 counts, behavioural factors, and age. The sixth chapter includes an analysis using hierarchical clustering techniques to identify meaningful subgroups of variables, aiding in the understanding of complex patterns within the dataset. The project concludes with a summary of findings and references to the relevant literature and statistical resources used.

CHAPTER – II

CHI- SQUARE TEST

OBJECTIVE: To examine the association between treatment regimen and treatment outcome.

2.1 INTRODUCTION

The Chi-Square test constitutes a fundamental statistical methodology used for the rigorous examination of categorical data, facilitating the determination of whether statistically significant deviations exist between observed and theoretically expected frequencies. As a non-parametric inferential procedure, it operates independently of specific probability distributions, thereby enhancing its applicability across a diverse array of research disciplines. The test predominantly manifests in two principal forms: the Chi-Square test of Association, which assesses the interdependence between two categorical variables, and the Chi-Square Goodness-of-Fit test, which evaluates the extent to which an empirical frequency distribution conforms to an anticipated theoretical model.

The chi-square test allows researchers to determine whether there are statistically significant correlations or deviations by comparing observed empirical data to an expected distribution derived under the null hypothesis. Although the test is useful for identifying correlations, it is not able to measure the strength or directionality of these relationships, so additional statistical measures like Cramer's V or the Phi coefficient are required for a more thorough examination. Furthermore, because it only finds correlations rather than clarifying the mechanisms underlying the relationships between variables, it cannot prove causation.

Despite these methodological constraints, the chi-square test remains an indispensable tool within the domain of categorical data analysis, continuing to serve as a cornerstone for statistical inference.

2.2 TYPES OF CHI-SQUARE TEST

The chi-square test is a statistical method used to analyse categorical data and determine whether observed frequencies differ significantly from expected frequencies. It is a non-parametric test that follows a chi-square (χ^2) distribution under the null hypothesis. The test is primarily divided into two broad categories:

1. Chi-Square Goodness-of-Fit Test
2. Chi-Square Test of Independence (or Association)

Additionally, there are specialized variations of the Chi-Square Test, such as:

3. McNemar’s Chi-Square Test
4. Chi-Square Test for Homogeneity

1. Chi-Square Goodness-of-Fit Test

This test determines whether an observed frequency distribution matches an expected theoretical distribution. It is used when we have a single categorical variable with multiple categories.

Hypotheses:

- Null hypothesis (H_0): The observed distribution follows the expected distribution.
- Alternative hypothesis (H_1): The observed distribution significantly deviates from the expected distribution.

The test statistic is given by

$$\chi^2 = \sum_{i=0}^{n} \frac{(O_i - E_i)^2}{E_i}$$

with $k-1$ degrees of freedom. Here, O_i is the observed frequency, E_i is the expected frequency, n is the sample size, and k is the number of categories. For a chi-square test to be valid, two main conditions must be met. First, the sample should be chosen randomly to avoid bias. Second, each group or category should have an expected count of at least 5. These rules help make sure the test gives accurate results.

2. Chi-Square Test of Independence (Association)

This test examines whether two categorical variables are statistically independent or associated. It is applied when we have two categorical variables, each with multiple categories, and we want to determine if there is a significant relationship between them.

Hypotheses:

- H_0: The two variables are independent.
- H_1: The two variables are dependent.

The test statistic is given by

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

With $(r\text{-}1)(c\text{-}1)$ degrees of freedom. Here, O_{ij} is the observed frequency, E_{ij} is the expected frequency, and r is the number of rows and c is the number of columns of the contingency table. The chi-square test has two main assumptions. First, the observations must be independent, meaning that one person's result should not influence another's. Second, each cell in the table should have an expected frequency of at least 5 to ensure accurate results.

3. McNemar’s Chi-Square Test (for Paired Data)

This test is used for paired or matched categorical data, particularly in pre-test/post-test study designs. It assesses whether the difference in proportions between paired observations is statistically significant.

Hypotheses:

- H_0: There is no difference between the paired proportions.
- H_1: There is a significant difference in the proportions.

The test statistic is given by

$$\chi^2 = \frac{(b-c)^2}{(b+c)}$$

Where *b* and *c* are the discordant pairs (cases that changed from one category to another). This test has two key assumptions. First, the data must be paired; meaning each observation in one group is matched with an observation in the other group. Second, it is suitable only for 2 × 2 contingency tables, where there are two categories for each variable.

4. Chi-Square Test for Homogeneity

This test is similar to the chi-square test of independence but is used to compare distributions of a categorical variable across multiple independent populations.

Hypotheses:

- H_0: The distributions are the same across populations.
- H_1: The distributions differ across populations.

The test statistic and degrees of freedom and assumptions are the same as chi-square test of independence.

2.3 COMPUTATIONAL PROCEDURE OF THE CHI-SQUARE TEST OF INDEPENDENCE

The Chi-Square test of Independence is used to determine whether two categorical variables are statistically independent, that is, if the distribution of one variable is independent of the other.

Computational Procedure:

Let there be two variables: variable A with *r* categories (rows) and variable B with *c* categories (columns). Construct an r × c contingency table from the data, where each cell O_{ij} contains the observed frequency of the i^{th} row and j^{th} column.

Under the null hypothesis of independence or no association, the expected frequency E_{ij} for each cell is computed as:

$$E_{ij} = \frac{(R_i \times C_j)}{N}$$

This formula comes from the assumption of independence:

- The probability of falling into row i is $\frac{R_i}{N}$
- The probability of falling into column j is $\frac{C_j}{N}$
- The joint probability (if independent) is $\frac{R_i}{N} \times \frac{C_j}{N}$

So the expected count is:

$$E_{ij} = N \times \frac{R_i}{N} \times \frac{C_j}{N} = \frac{R_i \times C_j}{N}$$

For every cell in the contingency table, compute this value. Now, Computing the Chi-Square Statistic

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

For each cell:

- Find the difference between expected frequency & observed frequency.
- Square the difference.
- Divide by the expected frequency.
- Sum over all cells.

The degrees of freedom is given by $\mathbf{df} = (\mathbf{r} - \mathbf{1})(\mathbf{c} - \mathbf{1})$, where r is the number of rows and c is the number of columns of the contingency table.

On comparison of the calculated $\boldsymbol{\chi}^{\mathbf{2}}$ with the critical value from a chi-square distribution table at a chosen significance level and calculated degrees of freedom, we shall conclude. Alternatively, calculate the p-value:

- If p-value < α, reject the null hypothesis (variables are not independent).
- If p-value ≥ α, do not reject the null hypothesis (variables are independent).

A significant result in a test like the chi-square test tells us that there is some kind of link between two categorical variables, but it doesn’t tell us how strong that link is. To understand the strength of the relationship, we use other measures like Cramer’s V or the Phi coefficient. These help us see whether the connection is weak, moderate, or strong.

2.4 ASSUMPTIONS OF THE CHI-SQUARE TEST OF INDEPENDENCE

- The data should be collected from a random sample.
- Variables are categorical (nominal or ordinal).
- Observations should be independent.
- Each category should have an expected frequency of no less than 5 to ensure the validity of the statistical test.

2.5 APPLICATIONS OF THE CHI-SQUARE TEST OF INDEPENDENCE

- Clinical Research: To check if a particular treatment is associated with patient outcomes (e.g., survival vs. death).
- Public Health: Examining the association between vaccination status and infection rates.
- Student Performance Analysis: To study the relationship between academic performance & study habits.

2.6 LIMITATIONS OF THE CHI-SQUARE TEST OF INDEPENDENCE

- Expected frequencies should be ≥ 5 in each cell for valid results.
- Large sample sizes can make practically meaningless associations, while small samples may fail to detect real associations.
- Cannot be used for continuous variables.
- Does not measure strength of association.

2.7 ANALYSIS OF THE DATA

The hypothesis under consideration is as follows:

H_0: There is no association between the type of treatment regimen used and the resulting treatment outcome.

H_1: There is an association between the type of treatment regimen used and the resulting treatment outcome.

Since the variables under consideration are categorical, a chi-square test for independence shall be used for the analysis of the objective.

The chi-square statistic measures the discrepancy between observed and expected frequencies. It helps us determine whether the variations seen in the data are likely due to random chance or if they suggest a real association. It serves as the basis for evaluating the statistical significance of the observed variations. In the final stage of the chi-square test, the calculated test statistic is compared to a critical value from the chi-square distribution to determine whether it is sufficiently large to justify rejecting the null hypothesis.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	208.566[a]	3	.000
Likelihood Ratio	207.546	3	.000
Linear-by-Linear Association	110.974	1	.000
N of Valid Cases	50000		

a. 0 cells (0%) have expected count less than 5. The minimum expected count is 2198.44.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.065	.000
	Cramer's V	.065	.000
N of Valid Cases		50000	

The p-value for the chi-square statistic is 0.000, which is less than the alpha level of 0.05. Hence, there is enough evidence to reject the null hypothesis. Hence, reject $H_{0.}$

2.8 CONCLUSION

There is an association between the type of treatment regimen and the treatment outcome.

CHAPTER – III

WILCOXON SIGNED RANK TEST

OBJECTIVE: To compare the CD4 counts of the patients at baseline and after 20 weeks.

3.1 INTRODUCTION

The Wilcoxon Signed-Rank test is a non-parametric, rank-based statistical procedure employed to assess whether two related or paired samples originate from the same underlying population distribution. As a distribution-free alternative to the paired t-test, it is particularly advantageous when the assumptions of normality and homoscedasticity (equal variances) are violated, thereby ensuring greater robustness in inferential statistics. Unlike parametric methodologies that rely on measures of central tendency such as means and standard deviations, the Wilcoxon signed-rank test operates on the principle of ranking absolute differences between paired observations while preserving directional information, making it inherently resistant to the influence of extreme values.

The Wilcoxon signed-rank test is still a crucial tool in non-parametric inferential statistics because it avoids constrictive parametric assumptions and uses a rank-based methodology to provide a reliable mechanism for analysing dependent samples in a variety of research domains.

3.2 COMPUTATIONAL PROCEDURE OF THE WILCOXON SIGNED RANK TEST

The Wilcoxon signed-rank test evaluates if there is any significant difference between two related (paired) samples by ranking the absolute differences between paired observations. To test H_0: There is no significant difference between the two related samples. The test statistic is derived from these ranked differences and compared against a critical value or normal approximation to determine statistical significance.

Computational Procedure:

Given two related samples, X_i and Y_i, calculate the difference for each pair:

$$D_i = Y_i - X_i$$

where D_i represents the difference between paired observations. Positive values of D_i indicate an increase, and negative values indicate a decrease. If D_i =0, i.e., no change between the paired observations, exclude these cases from further computations, as they do not contribute to the ranking process. The sample size n is adjusted accordingly. Convert all absolute differences $| D_i |$ into ranks, assigning the smallest difference the rank of 1, the next smallest 2, and so on. If there are tied values (i.e., multiple differences of the same magnitude), assign them the average rank.

- If D_i is positive, assign a positive rank.

- If D_i is negative, assign a negative rank.

Then compute the sum of ranks separately for positive and negative differences:

$$W^+ = \sum (ranks\ of\ positive\ differences)$$

$$W^- = \sum (ranks\ of\ negative\ differences)$$

The test statistic *W* is defined as the smaller of these two sums:

$$W = \min(W^+ + W^-)$$

Without ties,

$$\mu_W = \frac{n(n+1)}{4}$$

$$\sigma_W^2 = \frac{n(n+1)(2n+1)}{24}$$

With ties, we correct the variance using a tie correction factor.

$$\mu_W = \frac{n(n+1)}{4}$$

$$\sigma_W^2 = \frac{n(n+1)(2n+1)}{24} - \frac{1}{48}\sum_{j=0}^{k}\left(t_j^3 - t_j\right)$$

where t_j is the number of tied ranks in the j^{th} group, n is the number of non-zero differences, and k is the number of tied groups

- For small sample sizes ($n \leq 25$), compare W with critical values from a Wilcoxon signed-rank table.
- For larger samples ($n > 25$), approximate a normal distribution using the following Z-score transformation:

$$Z = \frac{W - \mu_W}{\sqrt{\sigma_W^2}}$$

Compute the p-value and arrive at a decision using the calculated *W* or *Z*-score:

- Compare it against the critical value at a chosen significance level (α), typically 0.05.
- If the observed test statistic is less than the critical value or if the p-value is below α, reject the null hypothesis.

By following this structured computational procedure, researchers can analyse data that does not meet parametric assumptions, ensuring reliable inferential outcomes in various scientific disciplines.

3.3 ASSUMPTIONS OF THE WILCOXON SIGNED RANK TEST

- Paired and related Data
- Purely categorical data is not suitable, since it is not rankable.
- Minimal ties, since too many tied differences reduce accuracy.
- Pairs with zero differences are excluded.
- Each pair should be independent of other pairs.

3.4 APPLICATIONS OF THE WILCOXON SIGNED RANK TEST

- Used to compare pre-treatment and post-treatment effects.
- Applied in biostatistics to evaluate changes in patient health metrics, such as cholesterol levels, weight loss, or pain scores.
- Used in cognitive and behavioural psychology to assess the impact of interventions.

3.5 LIMITATIONS OF THE WILCOXON SIGNED RANK TEST

- It assumes that differences are symmetrically distributed around the median; in cases of strong asymmetry, results may be misleading.
- When multiple paired differences are identical (tied ranks), the test statistic needs an adjustment, which can reduce its power and reliability.
- While the test effectively detects median shifts, it does not provide direct insight into effect size, often necessitating supplementary measures such as Cliff’s Delta or rank-biserial correlation for effect size interpretation.

3.6 ANALYSIS OF THE DATA

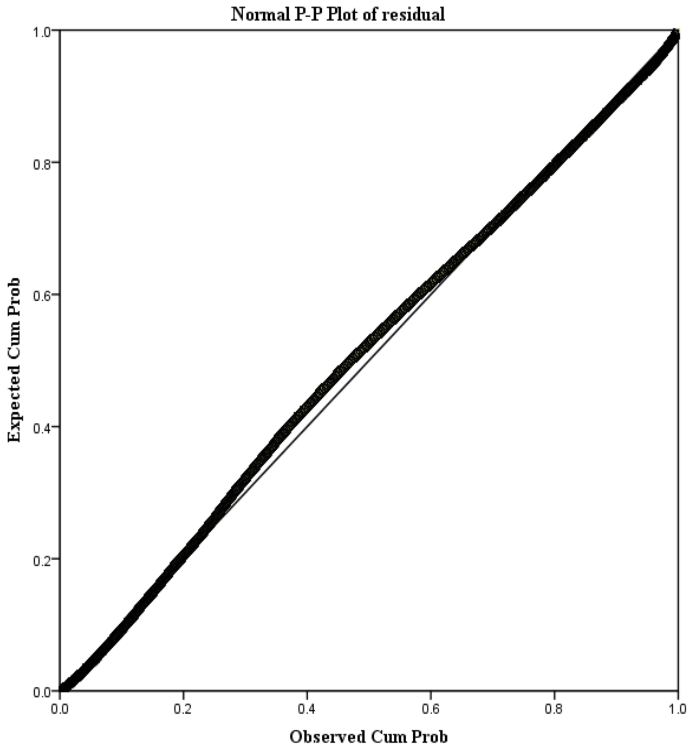
The hypothesis under consideration is as follows:

H_0: There is no significant difference in the mean CD4 count between baseline and 20 weeks.

H_1: There is a significant difference in the mean CD4 count between baseline and 20 weeks.

Since the variables under consideration are continuous, a paired t-test shall be used for this objective if the residuals are normally distributed. Check for normality of the residuals with the help of a P-P plot.

A probability-probability (P-P) plot is used in SPSS to assess whether a dataset follows a specified distribution, most commonly the normal distribution. If the points vary significantly from the line, the residuals are not normally distributed.



Upon careful observation of the graph, we conclude that the residuals of the two variables under consideration are not normal. Hence, move on to the corresponding non parametric test, the Wilcoxon signed rank test.

Ranks

		N	Mean Rank	Sum of Ranks
cd420 - cd40	Negative Ranks	11516[a]	16055.74	184897880.00
	Positive Ranks	38379[b]	27616.21	1059882580.00
	Ties	105[c]		
	Total	50000		

a. cd420 < cd40

b. cd420 > cd40

c. cd420 = cd40

Test Statistics[a]

	cd420 - cd40
Z	-135.978[b]
Asymp. Sig. (2-tailed)	.000

a. Wilcoxon Signed Ranks Test

b. Based on negative ranks.

Here, the p-value is 0.000, which is less than 0.05. Therefore, reject the null hypothesis.

3.7 CONCLUSION

There is a significant difference in mean between CD4 count at baseline and after 20 weeks.

CHAPTER – IV

KRUSKAL-WALLIS TEST

OBJECTIVE: To compare the time to liver failure among people taking different treatments.

4.1 INTRODUCTION

In the realm of non-parametric statistical inference, the Kruskal-Wallis test emerges as a robust alternative to the one-way ANOVA, particularly when the assumptions of normality and homogeneity of variances are untenable. Developed by William Kruskal and W. Allen Wallis in 1952, this rank-based test facilitates the comparison of three or more independent groups, determining whether their underlying population distributions differ significantly. Unlike parametric counterparts, the Kruskal-Wallis test operates on the ordinal properties of the data, making it particularly valuable in the analysis of skewed distributions, ordinal variables, or datasets contaminated by outliers.

Fundamentally, the test evaluates the null hypothesis that all samples originate from identical populations, with medians that are statistically indistinguishable. By converting raw data into ranks and analysing the distribution of these ranks across groups, the test circumvents the need for normally distributed data and preserves analytical integrity in scenarios where classical techniques may fail. Its application spans a multitude of disciplines, from behavioural science to biomedical research, offering a rigorous methodological framework for discerning statistically significant group-level disparities without presuming parametric regularity.

The Kruskal-Wallis test serves as a critical inferential tool, enabling the validation of hypothesized associations within the data structure under conditions of minimal statistical assumptions. The integration of such a non-parametric approach underscores a commitment to methodological rigor and analytical robustness in the presence of real-world data complexities.

4.2 COMPUTATIONAL PROCEDURE OF KRUSKAL-WALLIS TEST

The Kruskal-Wallis test is a non-parametric method that is widely used to examine whether there are statistically significant differences between the medians of three or more independent groups. It is an extension of the Mann-Whitney U test and is used when the assumptions of one-way ANOVA (like normality and equal variances) are not satisfied.

The null hypothesis is H_0: There is no significant difference in medians between the groups.

Computational Procedure:

Gather data from different groups that are to be compared. These groups should be independent. Combine all the observations from all the groups into a single list. Then rank all the values together in ascending order, assigning rank 1 to the smallest value, rank 2 to the next smallest, and so on. If two or more values are the same (called ties), give each one the average of the ranks they would have gotten if they were different. Once all values are ranked, divide the ranked data back into their original groups. For each group, compute the sum of the ranks (denoted as R_i for group *i*)

The Kruskal-Wallis test statistic is given by,

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{\left[R_i - \left(\frac{n_i(N+1)}{2}\right)\right]^2}{n_i}$$

Where:

- H is the Kruskal-Wallis test statistic.
- N denotes the total number of observations across all groups.
- k is the number of groups.
- R_i is the total of ranks for the i^{th} group.
- n_i is the total number of observations in the i^{th} group.

For computation purposes, we use

$$H = \left(\frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i}\right) - 3(N+1)$$

When there are tied ranks in the data, the Kruskal-Wallis test employs a correction factor. The calculated test statistic (H) is modified using this correction to take into consideration the bias caused by tied values. The H statistic may be somewhat overestimated without this adjustment, especially in cases where there are numerous ties.

The correction factor is given by

$$c.f = 1 - \frac{\sum u^3 - \sum u}{N(N^2-1)}$$

where u is the number of tied observations in each group.

The H statistic is then modified using the equation,

$$corrected\ H = \frac{H}{c.f}$$

The degrees of freedom of the Kruskal-Wallis test are calculated as: $df = k - 1$, where k is the number of groups. Compare H with a critical value from the χ^2distribution table with k−1 degrees of freedom at the chosen level of significance. If the calculated H value is greater than the χ^2 critical value, then the null hypothesis is rejected. As an alternative, the p-value linked to the H value is calculated and contrasted with the significance level (e.g., 0.05). The outcome is statistically significant if $p<0.05$.

If the test is significant, it means that at least one group differs in terms of its median. However, it does not tell which groups are different. Dunn's test with Bonferroni correction is a post-hoc statistical procedure used after a significant Kruskal-Wallis test to identify which specific groups differ from one another. Dunn's test addresses this by performing pairwise comparisons between groups using rank sums.

In Dunn's test, the data from all groups are combined and ranked. The average rank for each group is then calculated, and the differences between these average ranks are tested for statistical significance. The test computes a z-statistic for each pairwise comparison based on the difference in average ranks, adjusted for the number of observations in each group and the overall sample variance. Since multiple comparisons are being made, the risk of type I error (false positives) increases. This is where the Bonferroni correction comes into play.

The Bonferroni correction adjusts the significance level to account for the number of comparisons. Specifically, it divides the desired overall alpha level (commonly 0.05) by the number of pairwise tests being conducted. With four groups, there are six possible pairwise comparisons. Therefore, the corrected significance level for each individual comparison would be 0.05 divided by 6, which equals approximately 0.0083. Any p-value below this adjusted threshold is considered statistically significant.

4.3 ASSUMPTIONS OF THE KRUSKAL-WALLIS TEST

- The groups being compared must be independent.
- The data should be at least ordinal (can be ranked) or continuous.
- Observations within each group should be randomly selected from the population.
- The independent variable should consist of two or more categorical groups.

4.4 APPLICATIONS OF THE KRUSKAL-WALLIS TEST

- For comparing the effects of different treatments when the outcome data are skewed or ordinal.
- For comparing student performance across different teaching methods when scores are not normally distributed.
- For studying customer preferences or satisfaction across multiple products or regions.

4.5 LIMITATIONS OF THE KRUSKAL-WALLIS TEST

- It only tells if there is a difference between groups, but not which groups are different.
- Large differences in the number of observations in each group can affect the accuracy of the test.

- If the groups have very different shapes (e.g., some are skewed left and some right), the results may not be reliable.

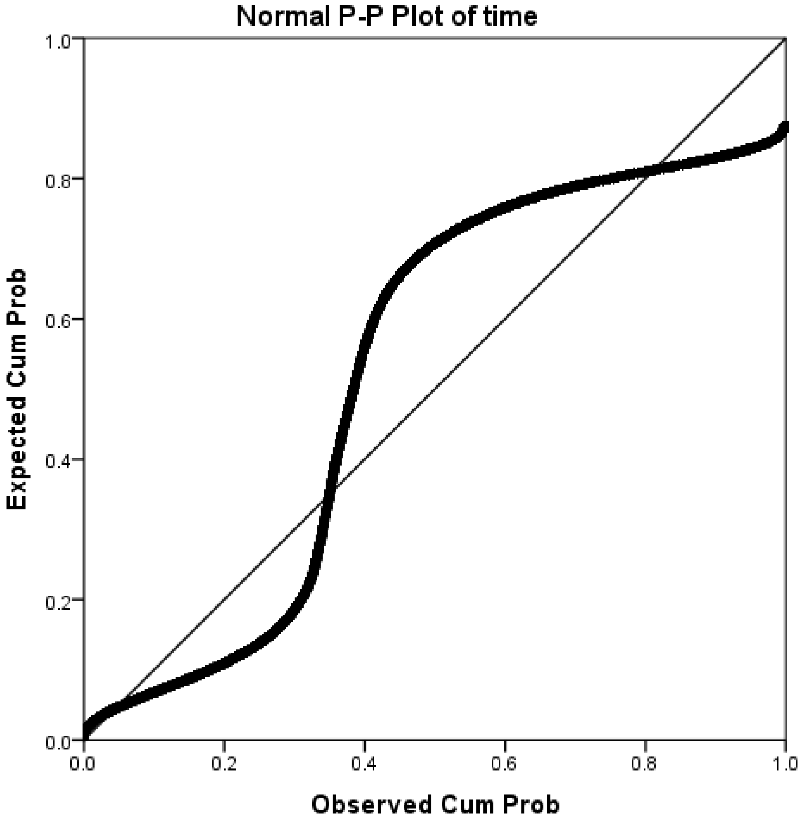
4.6 ANALYSIS OF THE DATA

The hypothesis under consideration is as follows:

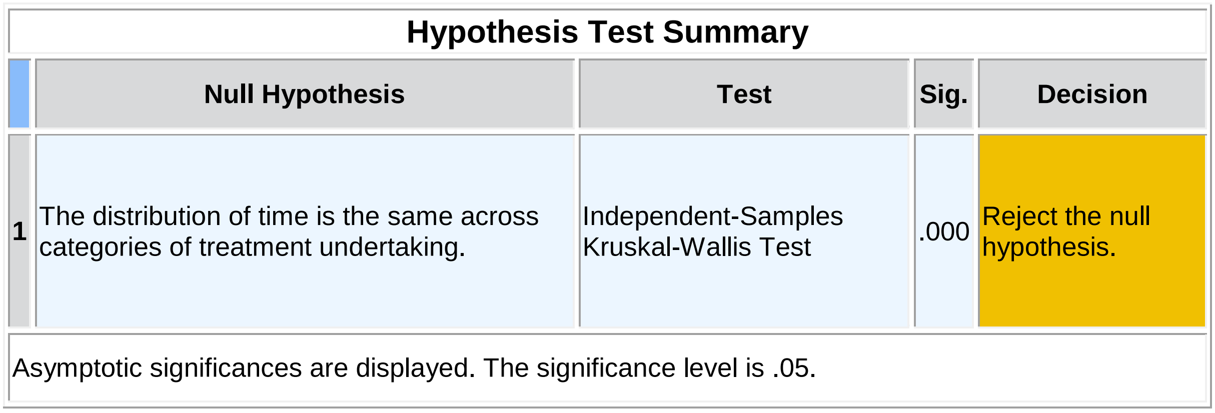
H_0: There is no significant difference in the time to liver failure among people taking different treatments.

Since the dependent variable under consideration is continuous, a one-way ANOVA shall be used for this objective if the dependent variable is normally distributed. Check for normality of the dependent variable with the help of a P-P plot.

A probability-probability (P-P) plot is used in SPSS to assess whether a dataset follows a specified distribution, most commonly the normal distribution. If the points stray significantly from the line, the dependent variable is not normally distributed.



From the graph, we conclude that the variable ‘time to liver failure’ under consideration is not normally distributed. Hence, move on to the corresponding non parametric test, the Kruskal-Wallis test.



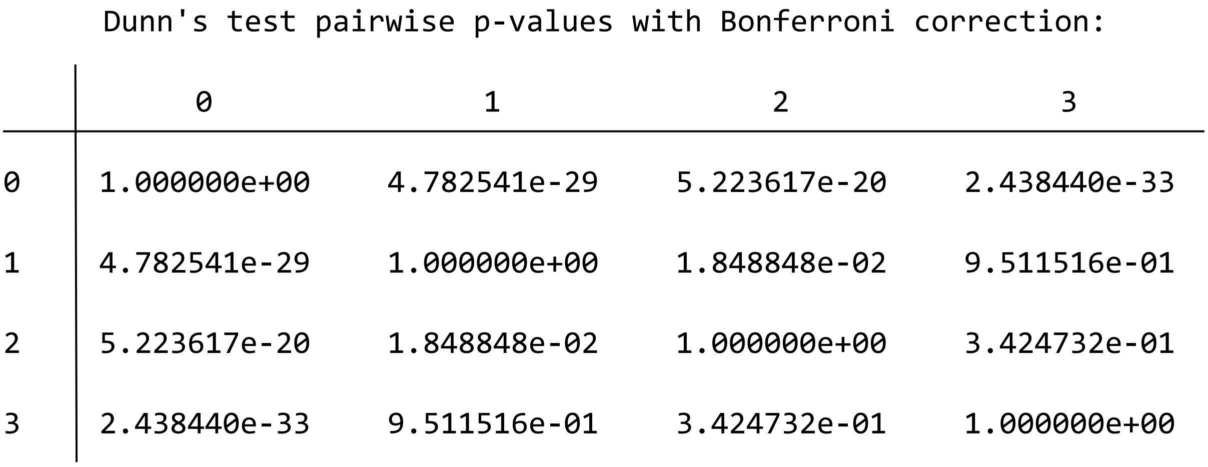
Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of time is the same across categories of treatment undertaking.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Here, the p-value is 0.000, which is less than 0.05. Therefore, reject the null hypothesis.

To assess whether there are significant differences between treatment groups based on the Dunn's test results, each pairwise p-value is compared to the Bonferroni-adjusted significance threshold. With four treatment groups, there are six pairwise comparisons, and the Bonferroni correction adjusts the overall significance level of 0.05 by dividing it by the number of comparisons, resulting in an adjusted alpha level of approximately 0.0083. Pairwise comparisons with p-values below this threshold value are considered statistically significant.



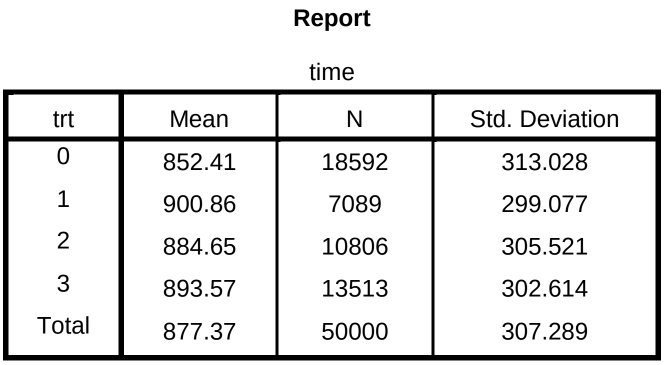
Dunn's test pairwise p-values with Bonferroni correction:

	0	1	2	3
0	1.000000e+00	4.782541e-29	5.223617e-20	2.438440e-33
1	4.782541e-29	1.000000e+00	1.848848e-02	9.511516e-01
2	5.223617e-20	1.848848e-02	1.000000e+00	3.424732e-01
3	2.438440e-33	9.511516e-01	3.424732e-01	1.000000e+00

The analysis revealed

- Treatment group 0 shows a statistically significant difference in time to liver failure compared to all other groups (1, 2, and 3)**.**
- Groups 1, 2, and 3 do not differ significantly from each other.

To find out which treatment works better in delaying or speeding up liver failure, it is important to look at the average (mean) or middle value (median) of the time to liver failure for each treatment group. This helps compare how long patients in each group stay liver failure–free and shows which treatment may be more effective.



Report

time

trt	Mean	N	Std. Deviation
0	852.41	18592	313.028
1	900.86	7089	299.077
2	884.65	10806	305.521
3	893.57	13513	302.614
Total	877.37	50000	307.289

The mean time to liver failure is shortest in group 0 (852.41 days), suggesting this treatment is less effective. Groups 1, 2, and 3 show longer mean times, 900.86, 884.65, and 893.57 days respectively, indicating better outcomes.

These results suggest that treatments 1 and 3 were the most effective in delaying liver failure, while treatment 0 was the least effective.

4.7 CONCLUSION

There is a significant difference in time to liver failure among people taking different treatments, moreover treatment 0 was the least effective in delaying it.

CHAPTER – V

RANDOM FOREST METHOD

OBJECTIVE: To classify patients as high- or low-risk for rapid AIDS progression using CD4/CD8 counts, behavioural indicators, and age.

5.1 INTRODUCTION

A popular and potent machine learning technique that is a member of the ensemble learning technique family is the Random Forest algorithm. In order to produce more accurate and reliable predictions, it constructs a "forest" of numerous decision trees rather than a single one. To put it simply, random forest creates several decision trees using various data points, then votes on each one to determine the outcome. This enhances prediction performance, prevents overfitting, and lowers errors. In technical terms, the algorithm employs bootstrap aggregating, in which a random sample of the dataset is used to train each tree. Additionally, it adds feature randomness, which means that when splitting nodes, each tree takes into account a random subset of features. The ensemble's capacity to generalize is strengthened by the increased model diversity brought about by this randomness. Random forest is very good at both regression (predicting numerical values) and classification (predicting categories).

5.2 TYPES OF RANDOM FOREST METHOD

Random forest comes in two primary varieties:

- When the result is a category (such as yes/no or different kinds of objects), the Random Forest Classifier is employed.
- When the outcome is a numerical value, such as price, age, or temperature, the Random Forest Regressor is employed.

In order to increase the final prediction's accuracy and dependability, both types employ the same concept of creating numerous decision trees and combining their output. The primary distinction is whether a label or a number should be predicted as the response.

1.Classification Using Random Forest

When a category or label is the output to be predicted, this kind of Random forest is employed.

- The algorithm uses various dataset components to create numerous decision trees.
- After examining the data, each tree provides a response (such as "Yes" or "No").
- The random forest selects the most frequently provided response after all trees have provided their responses. We refer to this as majority voting. The prediction's level of confidence increases with the number of trees that concur on the same response.

This variant is frequently referred to as a Random Forest Classifier.

2.Regression using Random Forest

When the output is a number, this type is utilized. It is a regression problem if the objective is to predict a value rather than a label.

- It creates a lot of decision trees, just like in classification.
- Each tree provides a number as its prediction rather than labels.
- The random forest then uses the average of all these numbers as the final prediction. This increases the prediction's accuracy and lessens the impact of any one tree that might be off.

We refer to this variant as a Random Forest Regressor.

5.3 COMPUTATIONAL PROCEDURE OF THE RANDOM FOREST CLASSIFIER METHOD

The Random Forest Classifier is an ensemble machine learning technique that creates numerous decision trees and then mixes their outputs to improve classification task performance and accuracy. The basic idea behind this method is to generate a "forest" of decision trees, each trained on a different section of the data, and then have the trees "vote" on the final forecast. This reduces the risk of overfitting, which is typical when using a single decision tree.

First, the algorithm uses replacement sampling to randomly select a number of smaller datasets from the original dataset. This procedure is called bootstrap sampling, and a decision tree is trained on each of these new datasets. Since some data points may appear more than once in one tree's dataset and not at all in another, the trees get slightly different perspectives on the data.

The algorithm does not utilize every feature when building each tree. Rather, it chooses a subset of features at random to take into account at every split or decision point. This method, known as feature bagging, improves generalization on unseen data by increasing the diversity among the trees.

Recursively, the tree splits. Until a stopping condition is met, the process of selecting a subset of characteristics and determining the optimal split continues along the tree. Common stopping conditions include reaching the maximum tree depth, the minimum number of samples required to divide a node, and ensuring that all instances in a node belong to the same class. When any of these conditions are met, the node transforms into a leaf node with a class label.

The model performs classification once each tree has developed. Each decision tree in the forest receives the input and predicts a new instance's class label. Each tree produces a class prediction. The random forest classifier produces the class with the most votes across all trees as its final output. Individual forecasts are collected, and the final anticipated class is determined by majority vote. In regression problems, the method averages the predictions of all trees.

The data points that are not part of the bootstrap sample, known as OOB samples, can be used as test data for each tree because each tree is trained on a bootstrap sample. Model performance can be assessed without a separate validation set by calculating an overall accuracy estimate for the model based on the predictions made on the OOB samples.

The performance of the model is then evaluated using metrics like accuracy, precision, recall, and F1-score, which can be found in a confusion matrix or classification report.

A confusion matrix is a table that depicts the performance of a classification model. It compares the actual class labels in the dataset to the predicted class labels generated by the model. In binary classification, the confusion matrix is a 2x2 table with four values: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

- True Positives (TP) are cases where the model accurately predicted the positive class.
- True Negatives (TN) are cases where the model accurately predicted the negative class.
- False Positives (FP) occur when a model predicts a positive outcome, but it is actually negative.
- False Negatives (FN) occur when a model predicts a negative value when it is actually positive.

Using these values, important performance metrics such as accuracy (overall correctness), precision, recall, and F1-score can be calculated.

A classification report summarizes the major performance metrics for a classification model. It features precision, recall, F1-score, and support for each class. This report assists in understanding how well the model performs across different classes, particularly when dealing with imbalanced datasets.

- Precision refers to the ratio of correct positive predictions to total anticipated positives. It indicates how accurate the model is at predicting a positive class.
- The recall ratio compares the number of correct positive predictions to the actual positive cases. It demonstrates how successfully the model can identify all positive examples.
- The F1-score represents the harmonic mean of precision and recall. It provides a single score that considers both problems.
- Support is the number of actual occurrences of each class in the dataset.

The classification report provides a more comprehensive and organized evaluation of a model's performance than depending only on accuracy because all factors are considered. For every class, it offers important metrics like precision, recall, and F1-score, which help identify the model's strong points and potential weaknesses. This makes it possible to understand the model's advantages and disadvantages, particularly in cases where the data is not balanced.

5.4 ASSUMPTIONS OF THE RANDOM FOREST CLASSIFIER METHOD

- It is assumed that every data point is unrelated to the others.
- A few input characteristics ought to be pertinent to target prediction.
- There shouldn't be a strong correlation between features.
- To stabilize the predictions, a sufficient number of trees are required.
- Assuming that overfitting will be managed by ensemble averaging, trees are grown to their maximum depth (unless parameters limit them).

5.5 APPLICATIONS OF THE RANDOM FOREST CLASSIFIER METHOD

- Disease prediction: Using medical data, this tool assists physicians in determining whether a patient has a disease.
- Fraud detection: Identifies fraudulent or dubious transactions in online shopping or banking.
- Customer churn: It informs businesses of which customers are most likely to discontinue using their service.

5.6 LIMITATIONS OF THE RANDOM FOREST CLASSIFIER METHOD

- The model's decision-making process is hard to comprehend; it is slow for large data.
- Requires more RAM and storage than more basic models.
- May not be quick enough for real-time predictions because of the large number of trees; if not adjusted appropriately, it may still overfit noisy data.

5.7 PYTHON CODE

Define the input dataset and set the model parameters. The input is a labelled dataset with several features (independent variables) and a categorical target variable (the class label). Before the model begins training, specify key parameters such as the number of trees (`n_estimators`), the number of features to consider when splitting a node (`max_features`), and, optionally, the maximum depth of the trees (`max_depth`). These parameters control the behaviour and complexity of the forest.

```
# Data Processing Packages
import pandas as pd
import numpy as np

# Modelling Packages
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

df=pd.read_csv("AIDS_Classification_50000.csv")

df['cd4_drop'] = df['cd40'] - df['cd420']
df['risk_label'] = df['cd4_drop'].apply(lambda x: 1 if x >= 100 else 0)
#to label patients as: High-risk: Rapid progression (e.g., large CD4 drop)
#Low-risk: Stable/slower progression 1 = High-risk (large CD4 drop) 0 = Low-risk

#Define features and target
features = ['cd40', 'cd420', 'cd80', 'cd820', 'drugs', 'homo', 'hemo', 'age']
X = df[features]
y = df['risk_label']

#standardising
X_scaled = StandardScaler().fit_transform(X)

# Train/Test split
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.3, random_state=42)
#This line splits the dataset into training and testing sets
#test_size=0.3: 30% of the data will be used for testing and 70% for training.
# random_state=42: ensure the split is reproducible.

#X_train, X_test: Features for training and testing.
#y_train, y_test: Target values for training and testing.

# Random Forest model
rf = RandomForestClassifier(n_estimators=100, random_state=42)
#n_estimators=100: to build 100 decision trees in the forest.
rf.fit(X_train, y_train)
#The model builds 100 decision trees on different samples of the data.

#At this point, we have a trained random forest model, but we need to find out whether it makes accurate predictions.
y_pred = rf.predict(X_test)  #make predictions on test data.
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)
```

```
# Evaluate
print("Confusion Matrix:")
print(confusion_matrix(y_test, y_pred))
print("\nClassification Report:")
print(classification_report(y_test, y_pred))
```

5.8 ANALYSIS OF THE DATA

To develop a predictive model for classifying AIDS patients into high-risk or low-risk categories based on their immunological and clinical characteristics. A key component of this analysis was the creation of a new variable, `cd4_drop`, which represents the difference in the CD4 cell count between baseline (`cd40`) and follow-up (`cd420`). This newly formed variable was essential in defining the risk categories. Patients who experienced a significant drop in CD4 count (≥100 units) were labeled as high-risk, indicating rapid disease progression. In contrast, patients with a smaller or no drop in CD4 count were categorized as low-risk, suggesting slower or more stable progression of the disease. This risk categorization serves as the target variable for the classification task.

To predict the risk category of a patient, a random forest classifier was employed, utilizing features such as baseline and follow-up CD4 and CD8 counts, age, and other demographic and clinical indicators, including drug use and conditions like haemophilia. The model was trained on 70% of the dataset and tested on the remaining 30%.

```
Confusion Matrix:
  [[13760   14]
  [52    1174]]

Accuracy: 0.9956
```

The confusion matrix reveals that the model correctly classified 13,760 low-risk patients and 1,174 high-risk patients, with only 14 low-risk and 52 high-risk patients misclassified. This results in a high overall accuracy of 99.56%.

Classification Report

	precision	recall	f1-score	support
0	1.00	1.00	1.00	13774
1	0.99	0.96	0.97	1226
accuracy			1.00	15000
macro avg	0.99	0.98	0.99	15000
weighted avg	1.00	1.00	1.00	15000

The classification report provides additional evidence of the model's excellent performance in differentiating between low-risk and high-risk AIDS patients: for the low-risk class (0), the model obtained perfect scores for all evaluation metrics (precision, recall, and F1-score all equaled 1.00), indicating flawless classification of almost all low-risk cases; for the high-risk class (1), the model showed a precision of 0.99, indicating that 99% of patients predicted as high-risk were in fact high-risk, while the recall was 0.96, indicating that 96% of actual high-risk patients were correctly identified; the F1-score of 0.97 for high-risk patients represents a well-balanced performance in terms of precision and recall. The overall accuracy was reported as 1.00 (or 99.56% when calculated more precisely), with similarly strong macro and weighted averages for all metrics, indicating consistent and reliable performance across both classes. These results confirm that the model is highly effective for clinical risk stratification, especially in identifying patients at risk of rapid disease progression.

Because of the high predictive power of the cd4_drop variable, the classification results indicate that the model is very dependable and successful in differentiating between AIDS patients who are at high risk and those who are not. The model offers a useful evaluation of patient risk by incorporating the rate of decline in CD4 counts, a crucial marker of disease progression. Since high-risk patients are more likely to progress quickly, early detection of these patients may allow clinicians to start more aggressive interventions and treatments, which would improve patient outcomes.

5.9 CONCLUSION

This analysis successfully developed a model to classify AIDS patients as high-risk or low-risk based on CD4/CD8 counts, behavioural indicators, and age. The model effectively identified patients at risk of rapid disease progression, supporting the potential for earlier and more targeted clinical intervention.

CHAPTER – VI

HIERARCHIAL CLUSTERING

OBJECTIVE: To identify the subgroups of variables.

6.1 INTRODUCTION

Hierarchical clustering is an unsupervised machine learning technique used to group similar observations into clusters based on their characteristics. The goal is to build a hierarchy of clusters, where each level of the hierarchy represents a different grouping of the data. This method is particularly useful when the natural number of clusters in the data is unknown or when understanding the nested structure among observations are important.

The process starts by treating each observation as its own cluster and then successively merging the most similar clusters. This continues until all the observations are combined into a single cluster, forming a tree-like structure called a *dendrogram*. The dendrogram visually represents how clusters are formed and merged at different levels of similarity, allowing analysts to determine the appropriate number of clusters by choosing a threshold at which to "cut" the tree.

Hierarchical clustering is widely used in fields such as biology, marketing, and social sciences, where relationships among elements are important to analyse and interpret. Its ability to reveal the underlying structure of data without prior assumptions makes it a powerful tool for exploratory data analysis.

6.2 TYPES OF HIERARCHICAL CLUSTERING

The two types of hierarchical clustering are

1. Agglomerative Hierarchical Clustering (Bottom-Up Approach)
This is the most commonly used type of hierarchical clustering. Every data point is handled as a separate cluster at the start of the process. The two closest clusters are then gradually combined. Until every data point is part of a single cluster, this merging process keeps going. Because it starts from the bottom (individual data points) and moves up (toward one big cluster), it is called the bottom-up approach. The various steps involved are

- Start with n clusters.
- Find the two clusters that are closest together.
- Merge them into one cluster.
- Repeat this process until only one cluster remains.

Agglomerative clustering is widely used because it is easy to understand and interpret, especially when visualized using a dendrogram.

2. Divisive Hierarchical Clustering (Top-Down Approach)

Divisive clustering is the opposite of agglomerative clustering. All of the data points are gathered into a single cluster at the start of this process. The algorithm divides the most dissimilar group into smaller clusters at each stage. Until every data point is in its own distinct cluster, this splitting process keeps going. The different steps that are involved are

- Start with one large cluster containing all data points.
- Split the cluster into two groups based on the largest dissimilarity.
- Continue splitting clusters recursively.
- Stop when each data point becomes its own cluster.

Divisive clustering is less commonly used than agglomerative clustering because it is computationally more complex. However, it can still be useful in some applications where the data is naturally better understood by splitting large groups.

When the clustering algorithm needs to decide which clusters should be combined (or split), it uses a rule called a linkage method. This rule determines how the “distance” between clusters is calculated. The choice of linkage method affects the shape and size of the clusters. Here are the most commonly used linkage methods:

1. Single Linkage (Minimum Linkage)

- Also known as the "nearest neighbour" method.
- The distance between two clusters is defined as the shortest distance between any two points in the clusters.
- Tends to form long, chain-like clusters because it focuses on the closest pair of points.
- Can be sensitive to noise or outliers.

2. Complete Linkage (Maximum Linkage)

- Also called the "farthest neighbour" method.
- The distance between two clusters is the greatest distance between any two points in the clusters.
- Produces compact, round-shaped clusters.
- More robust to outliers than single linkage.

3. Average Linkage

- The distance between two clusters is the average distance between all pairs of points from both clusters.
- It is a balanced approach between single and complete linkage.
- Produces clusters that are relatively well-separated and balanced in size.

4. Ward’s Method

- This method aims to minimize the total within-cluster variance.
- At each step, the pair of clusters that results in the smallest increase in the total variance is merged.
- Produces compact and spherical clusters.
- Often preferred for quantitative data.

6.3 PROCEDURE OF HIERARCHICAL CLUSTERING METHOD

The process of hierarchical clustering involves grouping similar data points step by step until a full hierarchy is built. It begins by measuring how similar or different each pair of data points is. This is done using a distance metric, such as Euclidean distance, which calculates how far apart two points are in the dataset. These distances are used to create a distance matrix, a table that shows how close or far every point is from every other point.

At the start of the clustering process, each data point is treated as its own individual cluster. From there, the algorithm looks for the two most similar clusters, that is, the ones with the smallest distance between them, and merges them into a single cluster. After this merge, the total number of clusters decreases by one.

Once two clusters are merged, the algorithm recalculates the distances between the newly formed cluster and all the remaining clusters. This step is important because the new cluster may have a different position or structure than the individual points that formed it. The process of finding the nearest clusters, merging them, and updating distances continues over and over. With each step, the clusters grow larger and the number of clusters becomes smaller.

As this continues, a tree-like structure begins to form in the background. This is known as a *dendrogram,* which records each merge and the distance at which it happened. The dendrogram becomes a visual summary of the entire clustering process and helps us understand how the data points are grouped at each stage.

The final step in hierarchical clustering is deciding when to stop merging and how many clusters to keep. Since the process continues until all points are eventually grouped into one big cluster, we need to cut the dendrogram at a particular level to identify a meaningful number of clusters. This is usually done by looking for large gaps or jumps in the height of the dendrogram, these gaps indicate that clusters merged at that point are much less similar and may not belong together. Cutting the tree at that level gives us a set of distinct clusters that best represent the structure in the data. In essence, hierarchical clustering builds a natural grouping of data without needing to specify how many groups to make from the beginning. It allows us to explore the relationships between data points at different levels and is especially helpful when the goal is to uncover how groups gradually form or break apart.

6.4 ASSUMPTIONS OF HIERARCHICAL CLUSTERING METHOD

1. Assumes that the chosen distance metric accurately reflects similarity between data points.
2. Data has an inherent hierarchical structure.
3. No need to predefine the number of clusters.
4. All features are treated equally (unless standardized).
5. Clusters are nested within each other.

6.5 APPLICATIONS OF HIERARCHICAL CLUSTERING METHOD

1. Bioinformatics: Used to find gene or sample groupings based on expression levels.
2. Market Segmentation: Groups customers based on behaviours, demographics, etc.
3. Image Segmentation: Groups pixels or objects based on similarity in features like colour or texture.
4. Taxonomy Creation: Builds hierarchical classifications in fields like biology or linguistics.

6.6 LIMITATIONS OF HIERARCHICAL CLUSTERING METHOD

1. High computational complexity.
2. Once a merge or split is made, it cannot be undone.
3. Poor scalability to large datasets.
4. Difficulty interpreting dendrograms for large datasets.
5. Requires manual choice of distance and linkage method.

6.7 ANALYSIS OF THE DATA

In this section, hierarchical clustering was applied to the variables of the dataset to identify groups of variables that behave similarly across the 50,000 observations. Hierarchical clustering was performed using the average linkage method (also known as between-groups linkage) and the Euclidean distance metric. Average linkage measures the distance between clusters as the average of all pairwise distances between elements in each cluster.

The agglomeration schedule records each step of the hierarchical clustering process, documenting how clusters are merged and the dissimilarity coefficients (i.e., distances) at which they are combined. Each row corresponds to a stage where two clusters are joined to form a new cluster. The key columns in the schedule include:

- Cluster 1 and Cluster 2: The clusters which are being merged at each step.
- Coefficients: The dissimilarity measure between the clusters being joined.
- Stage Cluster First Appears: Indicates at which stage each of the two clusters was first formed.

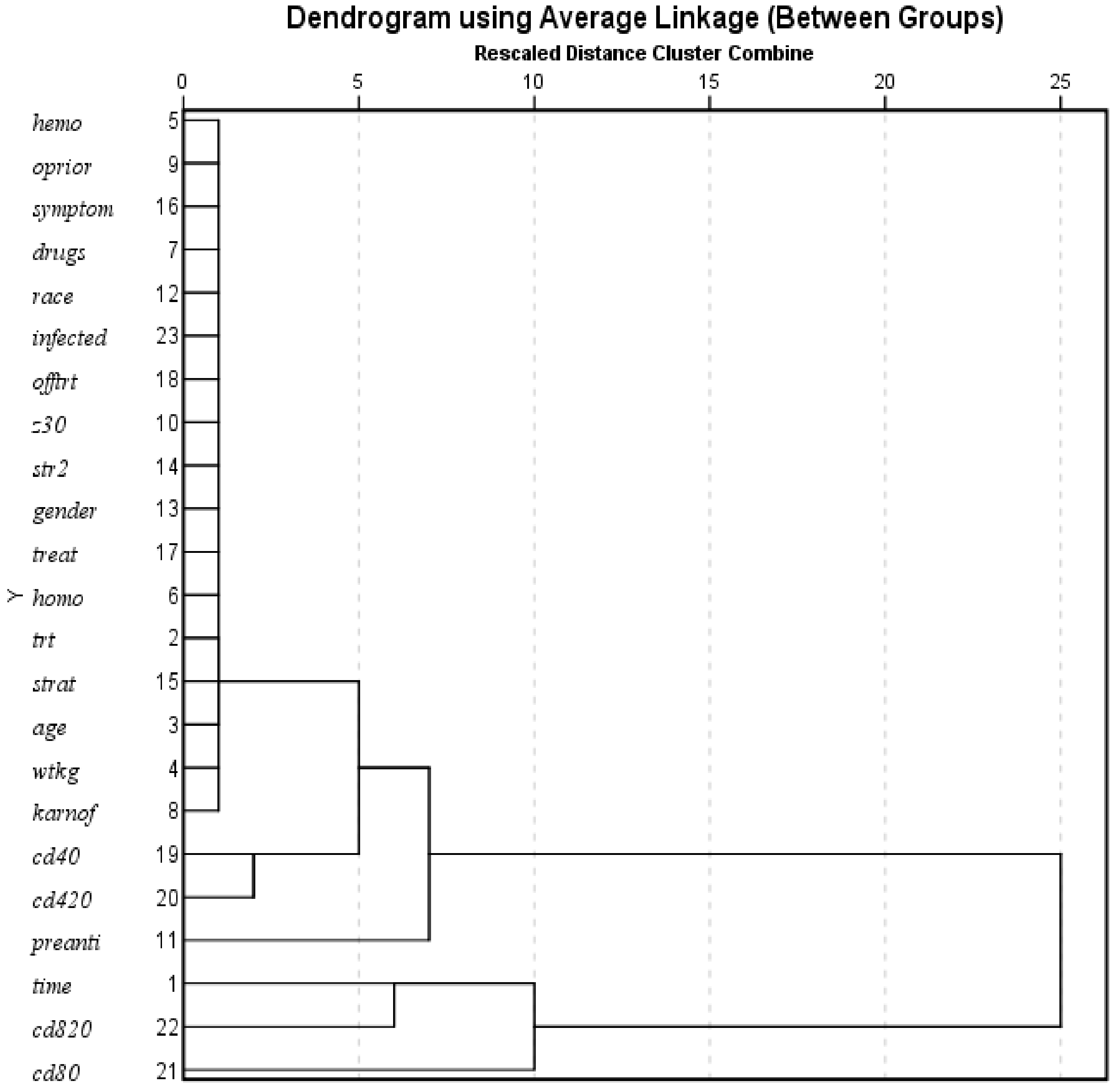
- Next Stage: Shows the subsequent stage in which the new cluster formed will participate.

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	5	9	3597.000	0	0	2
2	5	16	5709.500	1	0	3
3	5	7	8581.667	2	0	6
4	10	14	9580.000	0	0	10
5	13	17	15867.000	0	0	8
6	5	12	16057.250	3	0	7
7	5	23	17736.000	6	0	9
8	6	13	18094.500	0	5	10
9	5	18	19207.333	7	0	11
10	6	10	22437.667	8	4	11
11	5	6	31426.429	9	10	13
12	2	15	130965.000	0	0	13
13	2	5	151504.833	12	11	15
14	4	8	30464181.725	0	0	16
15	2	3	58946334.714	13	0	16
16	2	4	362386081.797	15	14	18
17	19	20	2053049304.000	0	0	18
18	2	19	7700454951.035	16	17	20
19	1	22	10257909769.000	0	0	21
20	2	11	12561704754.866	18	0	22
21	1	21	17941604341.500	19	0	22
22	1	2	47652091595.637	21	20	0

- In stages 1 to 11, the increases in dissimilarity are small, meaning the variables that merged were quite similar.
- At stage 12, there's a noticeable jump, showing that more dissimilar variables are being grouped.
- A big increase happens at stage 14 (~30 million) and continues in stages 15 and 16 (~58 million and ~362 million).
- From stages 17 to 22, the dissimilarity rises sharply, with the final merge reaching over 47 billion, meaning very different clusters are being combined.

The significant jump in coefficients between stage 13 (151,504) & stage 14 (30,464,182) is a strong indicator of a natural break. This suggests that merging beyond stage 13 introduces dissimilar clusters and that an optimal number of clusters might be around 9 to 11. The optimal number of clusters is typically selected by identifying the largest increase in the agglomeration coefficient, which reflects a large gap in similarity.



The dendrogram displays the hierarchical clustering of variables using the average linkage method. Variables that are grouped together at lower rescaled distances are more similar in behaviour across the dataset. The variables cd40, cd420, cd820, cd8, preanti, and time form a tightly connected cluster, suggesting they share similar patterns and could represent related biological markers. Another cluster includes clinical and demographic features such as age, wtkg, strat, and karnoff, indicating similar variability in the dataset. In contrast, variables like hemo, oprior, and symptom are merged at much higher distances, implying they are more distinct and less correlated with other variables.

We determined that 6 clusters were formed by examining the dendrogram.

6.8 CONCLUSION

The hierarchical clustering using the average linkage method and Euclidean distance successfully grouped the variables into approximately 6 distinct clusters.

CONCLUSIONS

Based on the findings from the project entitled **“Statistical Study On HIV Patients’ Dataset,”** the following conclusions have been drawn:

1. There is an association between the type of treatment regimen and the treatment outcome.
2. There is a significant difference in mean between CD4 count at baseline and 20 weeks.
3. There is a significant difference in time to liver failure among people taking different treatments, moreover treatment 0 was the least effective in delaying it.
4. A model was successfully developed which classified AIDS patients as high-risk or low-risk based on CD4/CD8 counts, behavioural indicators, and age.
5. The 23 variables have been grouped into six clusters, based on their similarities.

REFERENCES

1. Agresti, A. (2018). Statistical Methods for the Social Sciences (5th ed.). Pearson.
2. Field, A. (2017). Discovering Statistics Using IBM SPSS Statistics (5th ed.). SAGE Publications.
3. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.
4. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021*)*. An Introduction to Statistical Learning: With Applications in R (2nd ed.). Springer.
5. Kaufman, L., & Rousseeuw, P. J. (2005). Finding Groups in Data*:* An Introduction to Cluster Analysis. Wiley.
6. Sheskin, D. J. (2020). Handbook of Parametric and Nonparametric Statistical Procedures (6th ed.). CRC Press.
7. Tamhane, A. C., & Dunlop, D. D. (2000). Statistics and Data Analysis: From Elementary to Intermediate. Prentice Hall.