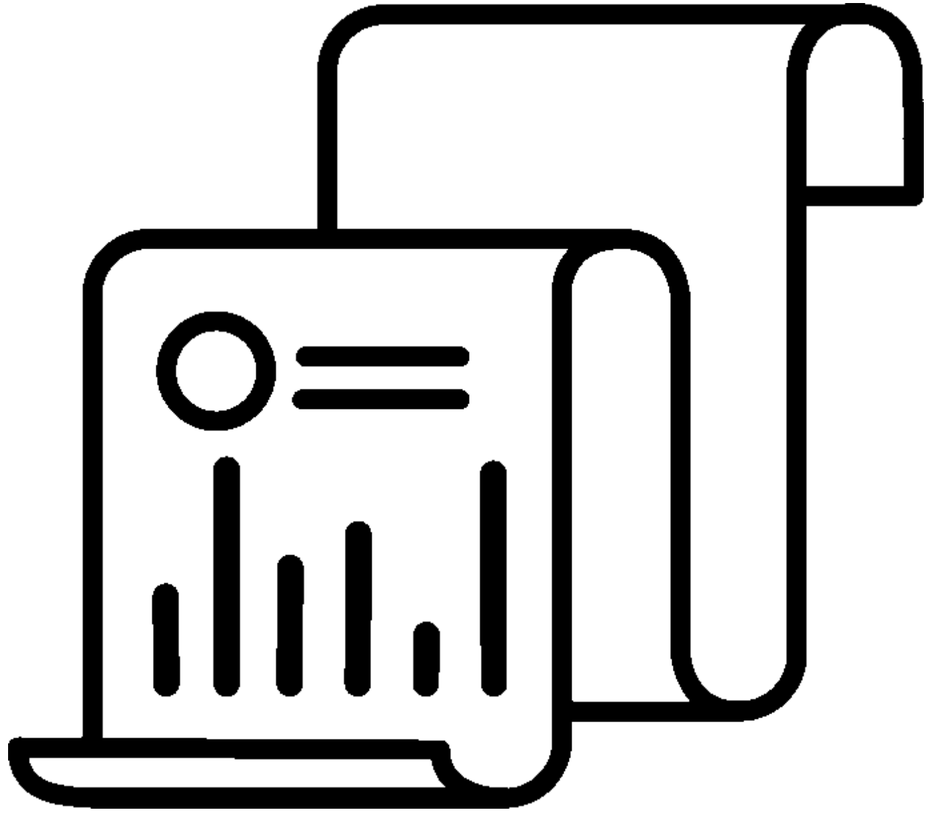
A STATISTICAL STUDY ON HIV PATIENT'S DATASET

PRESENTED BY: ANN LIN DAMI

DATA SOURCE

It is a secondary data collected from ‘Kaggle’. It consists of 23 variables and 50,000 samples on which the analysis is being carried out. Dataset contains healthcare statistics and categorical information about patients who have been diagnosed with HIV. This dataset was initially published in 1996.



VARIABLES

The dataset consists of 23 variables, out of which 9 are continuous and the rest are categorical.
Some of the variables are as follows;

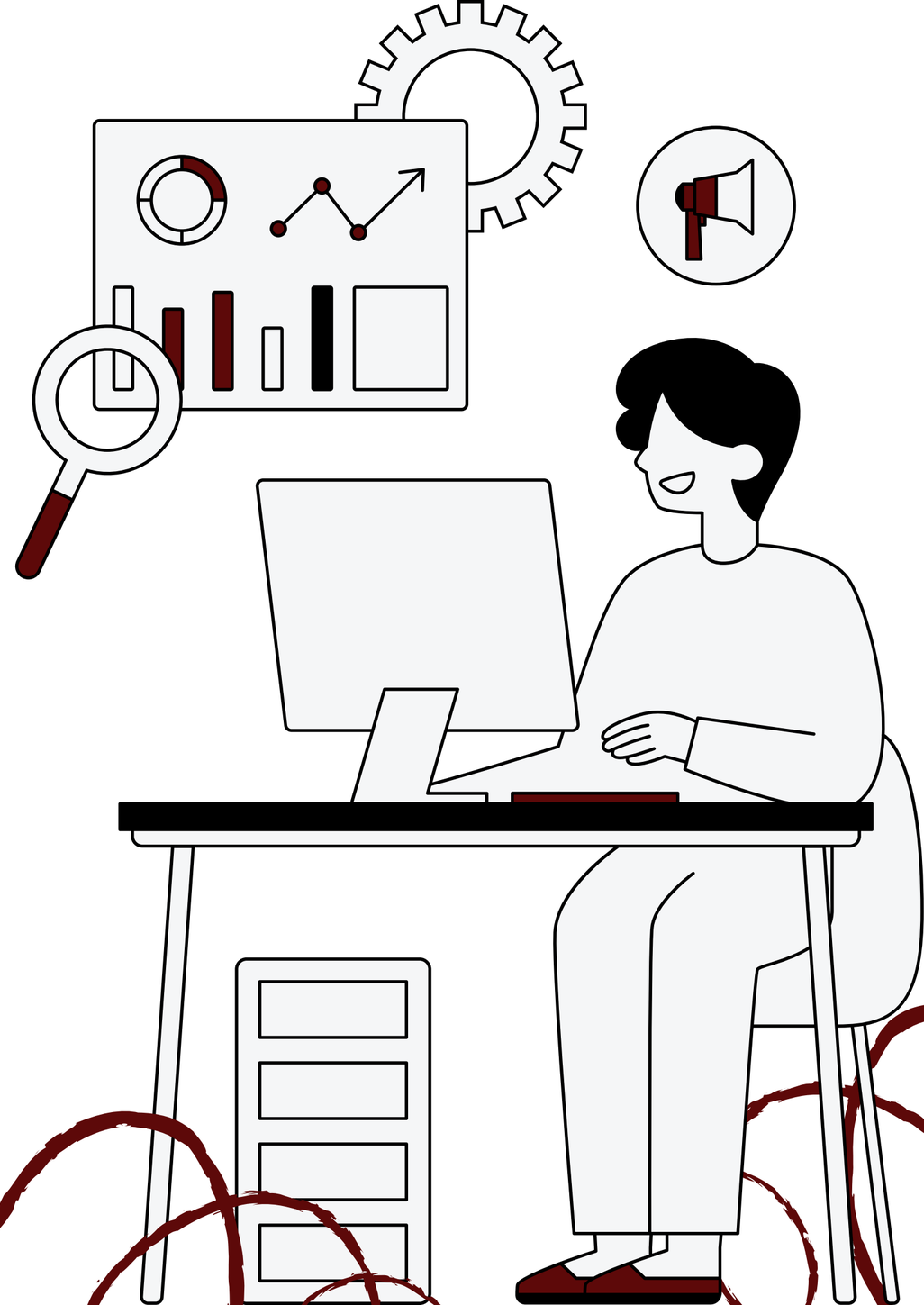
1.Personal information (age, weight, race, gender, sexual activity)
2.Medical history (hemophilia, history of IV drugs)
3.Treatment history (ZDV/non-ZDV treatment history)
4.Lab results (CD4/CD8 counts)

SOFTWARES USED

SPSS

&

PYTHON



OBJECTIVES OF STUDY

01. To examine the association between treatment regimen and treatment outcome.

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

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

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

05. To identify the subgroups of variables.

Statistical Methods USED

01

CHI- SQUARE TEST OF INDEPENDENCE

02

WILCOXON SIGNED RANK TEST

03

KRUSKAL-WALLIS TEST

04

RANDOM FOREST CLASSIFIER METHOD

05

HIERARCHIAL CLASSIFICATION METHOD

01. CHI-SQUARE TEST OF INDEPENDENCE

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

Since both the variables under consideration are categorical,
a Chi-square test for independence shall be used for this objective.

Hypothesis:

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

H1: There is an association between the type of treatment regimen used and the resulting treatment outcome.

ANALYSIS

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.

Here, the p-value is 0.000, which is less than 0.05.

Therefore, we reject the null hypothesis.

CONCLUSION

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

02. WILCOXON SIGNED RANK TEST

Objective:
To compare the CD4 counts of the patients at baseline and after 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.

Hypothesis:

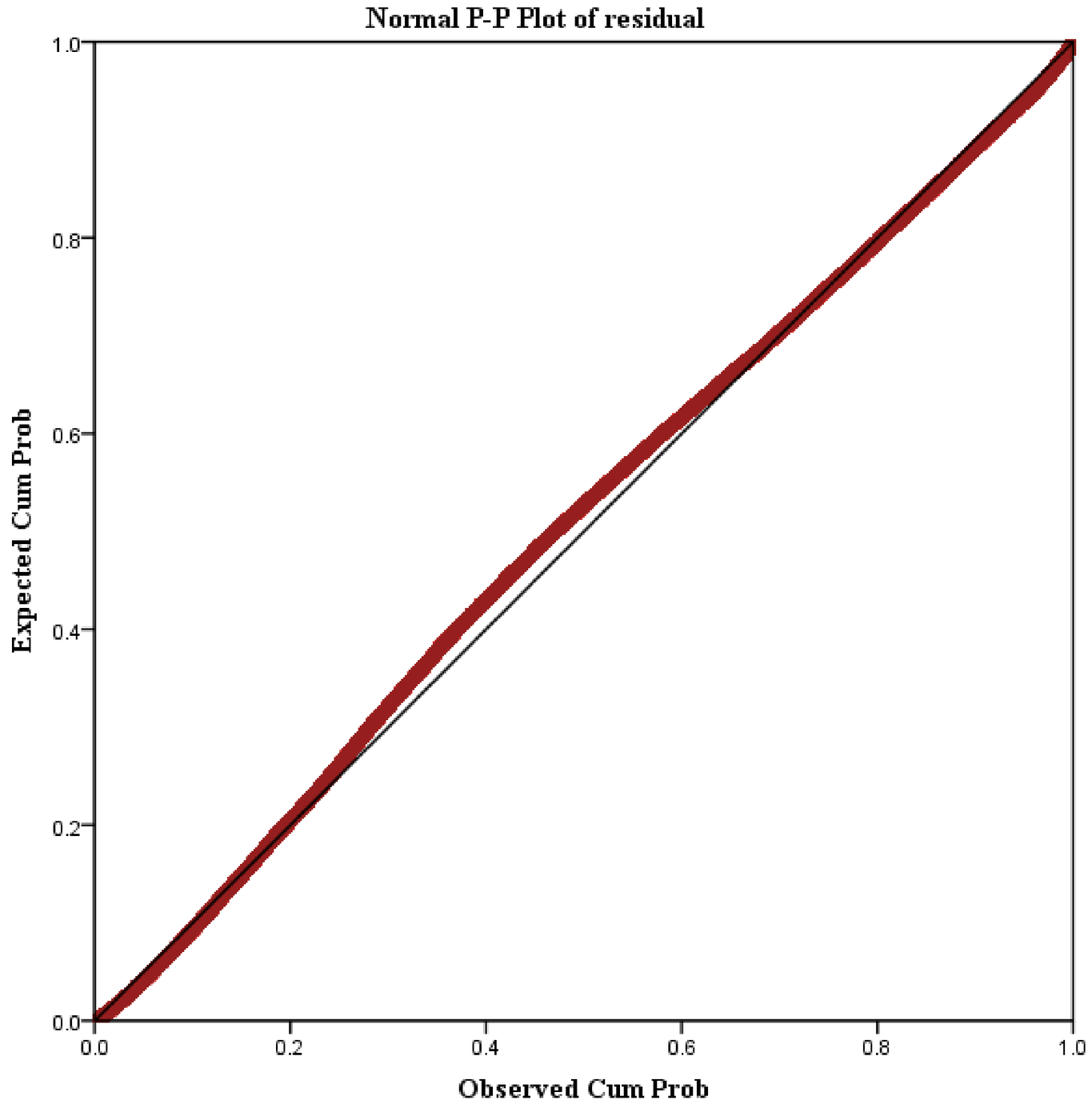
Ho: There is no significant difference in mean between CD4 count at baseline and 20 weeks.

H1: There is a significant difference in mean between CD4 count at baseline and 20 weeks.

ANALYSIS

CHECKING FOR NORAMILTY

Upon careful observation of the P-P plot,
we conclude that the residuals of the two variables
under consideration are not normal.
Hence moving on to the corresponding non parametric test.



THE CORRESPONDING NON PARAMETRIC TEST IS A

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. Hence, we reject the null hypothesis.

CONCLUSION

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

03. KRUSKAL-WALLIS TEST

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

Since the variable under consideration is continuous, a one way ANOVA shall be used for this objective, if the variable is normally distributed.

Hypothesis:

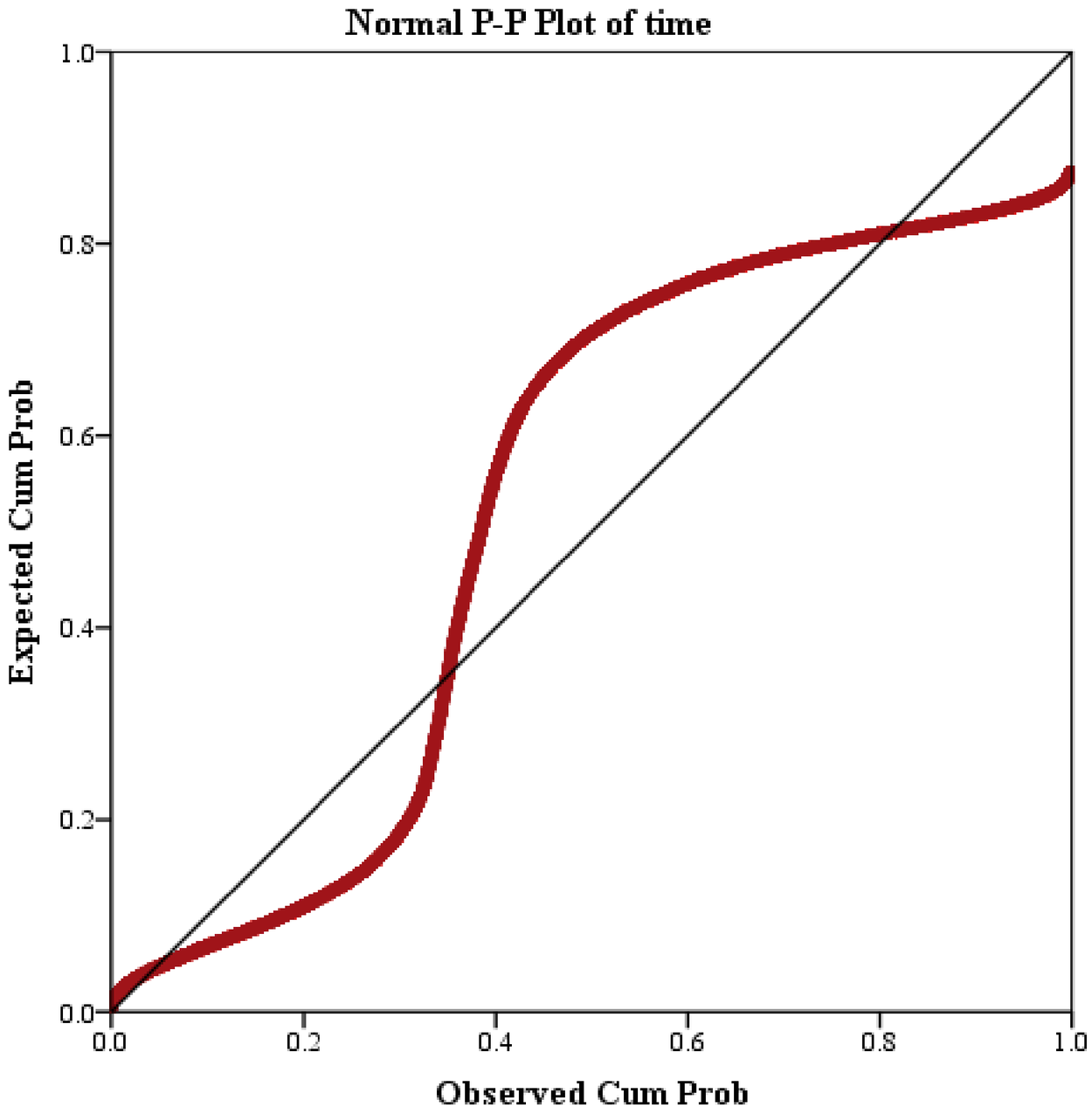
Ho: There is no significant difference in time to liver failure among people taking different treatments.

ANALYSIS

CHECKING FOR NORAMILTY

Upon careful observation of the plot, we conclude that the dependent variable under consideration (time) is not normal.

Hence moving on to the corresponding non parametric test.



the corresponding non parametric test is a

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.

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

A post hoc Dunn's test with Bonferroni correction was conducted to identify specific group differences in time to liver failure across treatment groups.
A p-value < 0.0083 means the two groups are significantly different.
The results indicated that:

- 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 determine which treatment is more effective in delaying or accelerating liver failure, mean time to liver failure for each treatment group is examined.

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

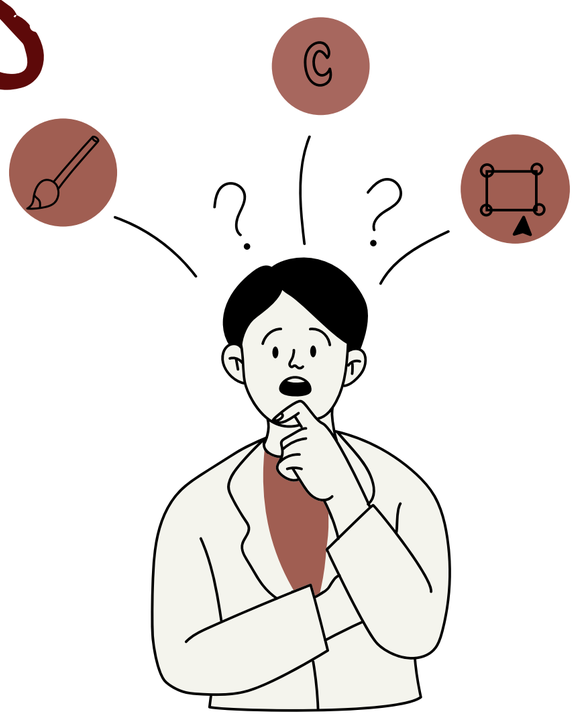
The mean time to liver failure is shortest in Group 0 (852.41 days), suggesting this treatment is less effective. Groups 1, 2, and 3 show longer mean times—900.86, 884.65, and 893.57 days respectively, indicating better outcomes.

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.

04. Random Forest Classifier Method

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



To develop a predictive model for classifying AIDS patients into high- and low-risk categories based on their characteristics, a Random Forest Classifier Method is used, since the output to be predicted is a category or label.

ANALYSIS

- This analysis aimed to classify AIDS patients as high-risk or low-risk for rapid disease progression.
- A new variable, cd4_drop, which represents the difference in the CD4 cell count between baseline (cd40) and follow-up (cd420) was created.
- Patients who experienced a significant drop in CD4 count (≥100 units) were labeled as high-risk, and patients with a smaller or no drop in CD4 count were categorized as low-risk. 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.
- The model was trained on 70% of the dataset and tested on the remaining 30%
- Model performance was evaluated using appropriate classification metrics.

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

For the low-risk class (0),—precision, recall, and F1-score all equal to 1.00 —indicating flawless classification of low-risk cases.

For the high-risk class (1), the model demonstrated

- a precision of 0.99, meaning that 99% of patients predicted as high-risk were indeed high-risk.
- The recall was 0.96, showing that 96% of actual high-risk patients were correctly identified.

The overall accuracy was reported as 1.00. These results confirm that the model is highly effective for clinical risk stratification, especially in identifying patients at risk of rapid disease progression.

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.

05. Hierarchial Clustering Method

Objective:
To identify the subgroups of variables

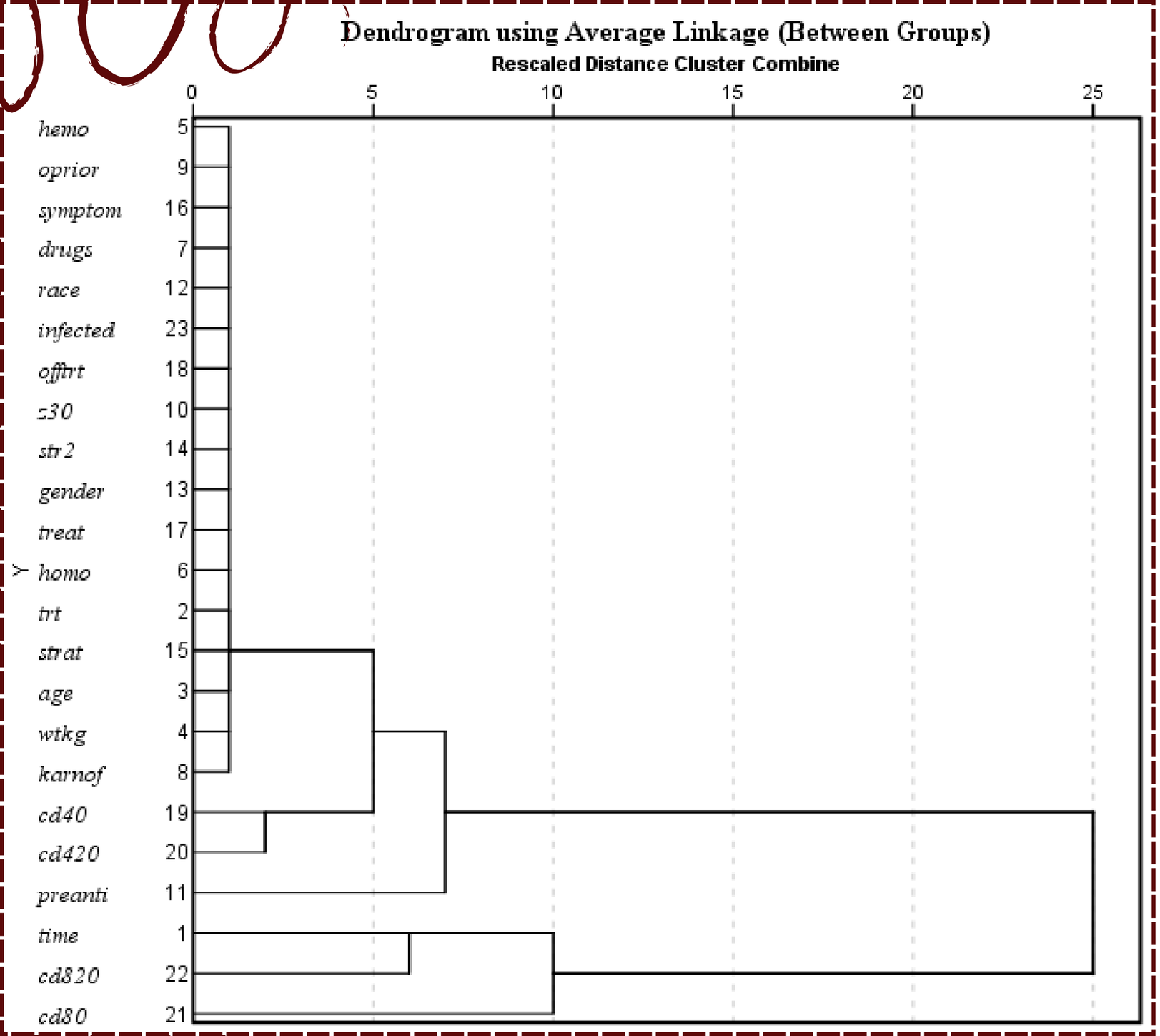
The goal is to find groups of variables that are closely related. To do this, hierarchical clustering is used to group variables that behave in a similar way.

Agglomeration Schedule						
Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	5	9	3597	0	0	2
2	5	16	5709	1	0	3
3	5	7	8581	2	0	6
4	10	14	9580	0	0	10
5	13	17	15867	0	0	8
6	5	12	16057	3	0	7
7	5	23	17736	6	0	9
8	6	13	18094	0	5	10
9	5	18	19207	7	0	11
10	6	10	22437	8	4	11
11	5	6	31426	9	10	13
12	2	15	130965	0	0	13
13	2	5	151504	12	11	15
14	4	8	30464181	0	0	16
15	2	3	58946334	13	0	16
16	2	4	362386081	15	14	18
17	19	20	2053049304	0	0	18
18	2	19	7700454951	16	17	20
19	1	22	10257909769	0	0	21
20	2	11	12561704754	18	0	22
21	1	21	17941604341	19	0	22
22	1	2	47652091595	21	20	0

ANALYSIS

The agglomeration schedule records each step of the hierarchical clustering process.

- In stages 1 to 11, the increases in dissimilarity are small, meaning the variables being grouped were very similar.
- At stage 12, there's a noticeable jump, and the increases become much larger from stage 14 onward, showing that less similar groups are being combined.
- This sharp rise in dissimilarity from stages 17 to 22 suggests that very different clusters are being merged.



The dendrogram shows the hierarchical clustering of variables using the average linkage method.

Based on the dendrogram, six clusters were identified.

CONCLUSION

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

Final CONCLUSIONS

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.
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 to six clusters, based on their similarities.

Bibliography

1. Alan Agresti – Statistical Methods for the Social Sciences – 2017 (5th edition)
2. Jiawei Han, Micheline Kamber, Jian Pei – Data Mining: Concepts and Techniques – 2011 (3rd edition)
3. Myles Hollander, Douglas A. Wolfe, Eric Chicken – Nonparametric Statistical Methods – 2013 (3rd edition)
4. Trevor Hastie, Robert Tibshirani, Jerome Friedman – The Elements of Statistical Learning – 2009 (2nd edition)
5. W. J. Conover – Practical Nonparametric Statistics – 1998 (3rd edition)

THANK YOU