BACS3013 DATA SCIENCE: Assignment Documentation

Title:AIDS Clinical Trials

TABLE OF CONTENT

1.0 Business Understanding

1.1 Business Background

HopeAlive Medical Hospital is a hospital. It was founded in Kuala Lumpur, Setapak on 1 July 1995. HopeAlive Medical Hospital has been dedicated to treating HIV/AIDS patients. Initially a modest clinic, it has grown into a Aids medical facility . Located strategically for accessibility, the hospital serves not only the local community but also patients from surrounding areas. Actively collaborating with government agencies, non-profits, and advocacy groups, HopeAlive Medical Hospital contributes to awareness, prevention, and research efforts in the fight against HIV/AIDS. Committed to excellence and compassion, it remains a beacon of hope for individuals affected by HIV/AIDS, providing them with the support and dignity they deserve.
For nearly three decades, HopeAlive Medical Hospital has remained unwavering in its mission to combat AIDS, offering not just treatment, but a lifeline of hope to those in need.

1.2 Business Aim

HopeAlive Medical Hospital aims to provide customer with the best treatments,especially for AIDS patients.Through prediction of whether a patient is died within a certain window of time, we are able to find out the best treatment between two different types of AIDS treatments to help the AIDS patients and at the same time prolong their lifetime.The prediction is done through development of an accurate predictive model based on collected dataset.This involves meticulously selecting relevant features and target,advanced data preparation techniques,excellent data understanding,well-prepared data and sufficient knowledge on statistics.Through this predictive model,HopeAlive Medical Hospital hopes to establish itself as the leading institution for HIV/AIDS treatment and support in Malaysia. Besides,through rigorous research and innovative approaches, HopeAlive seeks to improve patient outcomes.

1.3 Business objective

The primary business objective of HopeAlive Medical Hospital is to develop an accurate predictive model to predict whether a patient died within a certain window to determine the most effective AIDS treatment options. This endeavor requires meticulous data selection, advanced data preparation techniques, and the utilization of four different predictive models. The goal is to achieve the highest levels of accuracy, root mean square error (RMSE), and mean square error (MSE) to ensure optimal treatment selection for AIDS patients. By leveraging predictive analytics, the hospital aims to personalize treatment plans and improve patient outcomes significantly. (Miotto et al., 2017)

Furthermore,another key objective for HopeAlive Medical Hospital is to establish itself as the leading institution for HIV/AIDS treatment and support in Malaysia. Through the utilization of the developed predictive model, the hospital seeks to consistently deliver superior treatment outcomes for AIDS patients. By offering the best treatment options based on data-driven predictions, HopeAlive aims to enhance its reputation and attract more patients seeking high-quality care for HIV/AIDS. This objective aligns with the hospital's commitment to becoming a trusted healthcare provider and a beacon of hope for individuals affected by HIV/AIDS.

Lastly,in addition to developing a predictive model and establishing leadership in HIV/AIDS treatment, HopeAlive Medical Hospital aims to foster a culture of continual improvement through rigorous research and innovative approaches. By staying abreast of the latest advancements in HIV/AIDS treatment and leveraging data-driven insights, the hospital seeks to continually enhance patient outcomes. This involves ongoing research initiatives, collaboration with experts in the field, and the implementation of innovative treatment strategies to improve the quality and effectiveness of care provided to AIDS patients. Through a commitment to research and innovation, HopeAlive endeavors to remain at the forefront of HIV/AIDS treatment and contribute to the advancement of medical knowledge in the field.

1.4 Inventory of Resources

The project's human resources consist of both project members and university staff.The project members include Khor Kai Xian(23WMR), Wong Kiong Wei(23WMR), Tan Rou Min(23WMR01883),and Chong Wen Rui(23WMR02309). These individuals are Year 1 Semester 3 students, part of Group 1, pursuing a Bachelor of Computer Science (Honors) in Data Science At Tunku Abdul Rahman University of Management and Technology (TAR UMT). Overseeing The project, Dr Noor Aida Binti Husaini, a dedicated assistant professor and tutor at TAR UMT,provided invaluable guidance to the project members throughout its completion.

At the heart of our exploration lie the rich datasets procured from the esteemed repository of UC Irvine. These datasets, meticulously curated and readily available, serve as the cornerstone of our analytical endeavors. Through them, we unravel patterns, glean insights, and unravel the mysteries hidden within the data. Moreover, our reliance on the ubiquitous Google search engine and the versatile Mozilla Firefox browser amplifies our access to a wealth of knowledge spanning every domain imaginable. These digital gateways usher us into a realm where information knows no bounds, fueling our curiosity and driving our quest for understanding.

In the tangible realm of hardware, our arsenal boasts the stalwart companionship of personal computers, steadfast sentinels in our quest for knowledge. These machines, loyal and unwavering, stand ready to crunch numbers, run simulations, and breathe life into our ideas. Moreover, the communal embrace of school computers extends our computing prowess, providing us with additional firepower to tackle the challenges that lie ahead. Together, they form the backbone of our computational prowess, empowering us to navigate the digital landscape with confidence and dexterity.

In the realm of software, our arsenal is as diverse as it is powerful, each tool meticulously chosen to complement our workflow and enhance our productivity. The venerable Jupyter Notebook stands as a testament to our commitment to interactive and exploratory data analysis. With its intuitive interface and robust capabilities, it serves as our playground for experimentation and discovery. Furthermore, the collaborative synergy of Google Docs and the seamless file management offered by Google Drive streamline our documentation and facilitate seamless collaboration. And in the realm of communication, WhatsApp emerges as our conduit for

collaboration, fostering real-time exchanges and nurturing a sense of camaraderie among team members.

1.5 Requirements, Assumptions and Constraints

To ensure the successful completion of this project, two key requirements must be met. Firstly, unrestricted access to the data source without any copyright issues is essential. Fortunately, the dataset utilized originates from UCIrvine, ensuring accessibility to all users and eliminating copyright concerns. Secondly, the appropriate tools and software must be utilized for task completion. Python and its associated libraries will be employed for coding, with all processes executed within a Jupyter Notebook environment.

Two basic assumptions underlie this project. Firstly, it is assumed that the data has been accurately recorded in ascending date order. Secondly, it is presumed that all necessary tasks are executed correctly, leading to the generation of accurate results. Additionally, the dataset sourced from the UCIrvine website is assumed to consist of unaltered real-world values.

However, two significant limitations were encountered during the project's execution. The first limitation pertained to time constraints, as a strict seven-week deadline governed the project's timeline. Adherence to the project plan was imperative to ensure timely delivery. Secondly, the constraint of incomplete knowledge was acknowledged. Given that this assignment was initiated at the semester's commencement, it was recognized that not all requisite knowledge might be immediately available. Consequently, additional effort was invested in conducting thorough research to facilitate the project's successful completion.

1.6 Risk and Contingencies

The execution of this project carries no inherent risks, given that all the tools and resources employed are open-source and available for free usage. However, the provided dataset could potentially contain missing values. In light of this, project members are required to engage in data cleaning using appropriate techniques to ensure the accuracy and efficacy of the patient mortality prediction process, as well as the performance of two different types of HIV treatment regimens. In terms of potential contingencies, it is important to note that the project timeline is constrained by limited time availability, with the report due by 6th May 2024. To successfully navigate this constraint, project members must adhere rigorously to the predefined schedule in order to achieve project completion within the designated time frame. This entails efficient time management and a steadfast commitment to the established plan.

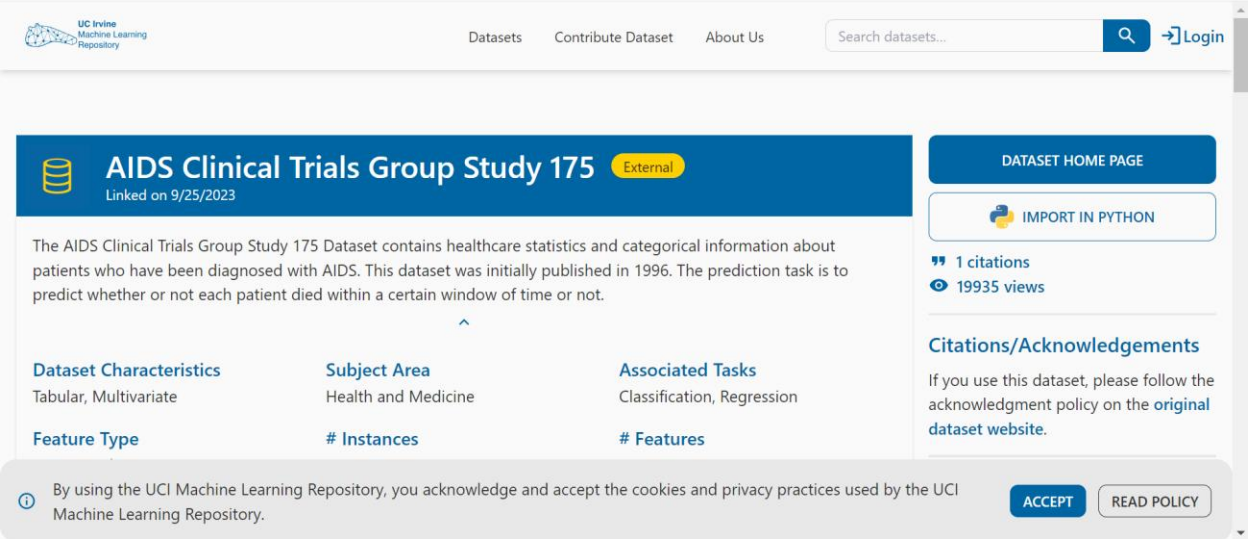
1.7 Project Plan

Project Phases	Activities	Start Time	End Time
Business Understanding	• Understand business background • Determine business aim • Determine business objectives • Discover inventory of resources • Discover requirements, and make assumptions and constraints • Discover risk and contingencies • Define project motivation	20/03/24	25/03/24
Data Understanding	• Collect data • Describe data • Explore and visualise data • Verify data quality	25/03/24	03/04/24
Data Preparation	• Select data • Clean data • Construct new data	30/03/24	15/04/24
Modeling	• Train Test Split • Model Building • Model Evaluation • Hyperparameter Tuning	10/04/24	30/04/24

Evaluation	• Evaluate results • Best Model Determination	01/05/24	03/05/24
Deployment	• Plan deployment • Produce final report • Summarise project	03/05/24	06/05/24

2.0 Data Understanding

2.1 Data Collection



This dataset for predict whether or not each patient died within a certain window of time to determine the best AIDS treatments has been sourced from UC Irvine Machine Learning Repository,a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. It is used by students, educators, and researchers all over the world as a primary source of machine learning data sets. As an indication of the impact of the archive, it has been cited over 1000 times (*UC Irvine Machine Learning Repository | Re3data.org*, n.d.). The dataset is provided in CSV format,or Comma-separated Values, which is an extremely common flat-file format that uses commas as a delimiter between values. Anyone familiar with spreadsheet programs has very likely encountered CSV files before - they're easily consumed by Google Spreadsheet, Microsoft Excel, and countless other applications (*CSV Format | Socrata*, n.d.).The subject area of this dataset is health and medicine.The features type include categorical and integer.Besides, this is a multivariate data, and it is organised in tabular structure. Yet,the problem is a classification problem, because the output value in this project is categorical value,which is 1 and 0.The dataset itself originates from ClinicalTrials.gov.

2.2 Data Description

2.2.1 Data Loading

Read the CSV file named 'data.csv' into a Pandas dataframe.

```
import pandas as pd
df=pd.read_csv('data.csv')
df
```

	pidnum	cid	time	trt	age	wtkg	hemo	homo	drugs	karnof	...	gender	str2	strat	symptom	treat	offtrt	cd40	cd420	cd80	cd820
0	10056	0	948	2	48	89.8128	0	0	0	100	...	0	0	1	0	1	0	422	477	566	324
1	10059	1	1002	3	61	49.4424	0	0	0	90	...	0	1	3	0	1	0	162	218	392	564
2	10089	0	961	3	45	88.4520	0	1	1	90	...	1	1	3	0	1	1	326	274	2063	1893
3	10093	0	1166	3	47	85.2768	0	1	0	100	...	1	1	3	0	1	0	287	394	1590	966
4	10124	0	1090	0	43	66.6792	0	1	0	100	...	1	1	3	0	0	0	504	353	870	782
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2134	990021	0	1091	3	21	53.2980	1	0	0	100	...	1	1	3	0	1	1	152	109	561	720
2135	990026	0	395	0	17	102.9672	1	0	0	100	...	1	1	3	0	0	1	373	218	1759	1030
2136	990030	0	1104	2	53	69.8544	1	1	0	90	...	1	1	3	0	1	0	419	364	1391	1041
2137	990071	1	465	0	14	60.0000	1	0	0	100	...	1	0	1	0	0	0	166	169	999	1838
2138	990077	0	1045	3	45	77.3000	1	0	0	100	...	1	0	1	0	1	0	911	930	885	526

2.2.2 Basic Data Understanding

#	Column	Non-Null Count	Dtype
0	pidnum	2139 non-null	int64
1	cid	2139 non-null	int64
2	time	2139 non-null	int64
3	trt	2139 non-null	int64
4	age	2139 non-null	int64
5	wtkg	2139 non-null	float64
6	hemo	2139 non-null	int64
7	homo	2139 non-null	int64
8	drugs	2139 non-null	int64
9	karnof	2139 non-null	int64
10	oprior	2139 non-null	int64
11	z30	2139 non-null	int64
12	zprior	2139 non-null	int64
13	preanti	2139 non-null	int64
14	race	2139 non-null	int64
15	gender	2139 non-null	int64
16	str2	2139 non-null	int64
17	strat	2139 non-null	int64
18	symptom	2139 non-null	int64
19	treat	2139 non-null	int64
20	offtrt	2139 non-null	int64
21	cd40	2139 non-null	int64
22	cd420	2139 non-null	int64
23	cd80	2139 non-null	int64
24	cd820	2139 non-null	int64

This dataset includes 2139 rows and 25 column data, except the columns called “wtkg” is float typed data, the other 24 column data are integer typed data and in the descriptive form, they only contain values of 1 or 0. All the data is not missing.

The data about Personal information are age, weight, race, gender and sexual activity, data of Medical history are hemophilia, and history of IV drugs, data of treatment history is ZDV or non-ZDV treatment history and the last data of lab results are CD4 and CD8 counts.

The following table summarizes the description of each data column and their respective data type:

Column	Role	Description	Type	Data Type
pidnum	ID	Patient ID	Categorical	Int64
cid	Target	Censoring indicator (1 = failure, 0 = censoring)	Categorical	
time	Features	time to failure or censoring	Continuous	
trt		treatment indicator (0 = ZDV only; 1 = ZDV + ddI, 2 = ZDV + Zal, 3 = ddI only)	Categorica	
age		age (yrs) at baseline	Categorical	
wtkg		weight (kg) at baseline	Continuous	float64
hemo		hemophilia (0=no, 1=yes)	Categorical	Int64
homo		homosexual activity (0=no, 1=yes)	Categorical	
drugs		history of IV drug use (0=no, 1=yes)	Categorical	
karnof		Karnofsky score (on a scale of 0-100)	Continuous	
oprior		Non-ZDV antiretroviral therapy pre-175 (0=no, 1=yes)	Categorical	
z30		ZDV in the 30 days prior to 175 (0=no, 1=yes)	Categorical	
zprior		ZDV prior to 175 (0=no, 1=yes)	Categorical	
prienti		# days pre-175 anti-retroviral therapy	Continuous	
race		race (0=White, 1=non-white)	Categorical	
gender		gender (0=F, 1=M)	Categorical	
str2		antiretroviral history (0=naive, 1=experienced)	Categorical	

strat		antiretroviral history stratification (1='Antiretroviral Naive',2='> 1 but <= 52 weeks of prior antiretroviral therapy',3='> 52 weeks)	Categorica	
symptom		symptomatic indicator (0=asymp, 1=symp)	Categorical	
trest		treatment indicator (0=ZDV only, 1=others)	Categorical	
offtrt		indicator of off-trt before 96+/-5 weeks (0=no,1=yes)	Categorical	
cd40		CD4 at baseline	Continuous	
cd420		CD4 at 20+/-5 weeks	Continuous	
cd80		CD8 at baseline	Continuous	
cd820		CD8 at 20+/-5 weeks	Continuous	

2.2.3 Descriptive Statistics

The goal of descriptive statistics is to provide a clear and concise summary of the data(Simplilearn, 2023) . Descriptive statistics includes measures such as central tendency (e.g., mean, median, mode) and dispersion (e.g., range, variance, standard deviation).(Simplilearn, 2023)

[15]:

	pidnum	cid	time	trt	age	wtkg	hemo	homo	drugs	karnof	...	gender
count	2139.000000	2139.000000	2139.000000	2139.000000	2139.000000	2139.000000	2139.000000	2139.000000	2139.000000	2139.000000	...	2139.000000
mean	248778.252454	0.243572	879.098177	1.520804	35.248247	75.125311	0.084151	0.661057	0.131370	95.446470	...	0.827957
std	234237.289399	0.429338	292.274324	1.127890	8.709026	13.263164	0.277680	0.473461	0.337883	5.900985	...	0.377506
min	10056.000000	0.000000	14.000000	0.000000	12.000000	31.000000	0.000000	0.000000	0.000000	70.000000	...	0.000000
25%	81446.500000	0.000000	727.000000	1.000000	29.000000	66.679200	0.000000	0.000000	0.000000	90.000000	...	1.000000
50%	190566.000000	0.000000	997.000000	2.000000	34.000000	74.390400	0.000000	1.000000	0.000000	100.000000	...	1.000000
75%	280277.000000	0.000000	1091.000000	3.000000	40.000000	82.555200	0.000000	1.000000	0.000000	100.000000	...	1.000000
max	990077.000000	1.000000	1231.000000	3.000000	70.000000	159.939360	1.000000	1.000000	1.000000	100.000000	...	1.000000

str2	strat	symptom	treat	offtrt	cd40	cd420	cd80	cd820
2139.000000	2139.000000	2139.000000	2139.000000	2139.000000	2139.000000	2139.000000	2139.000000	2139.000000
0.585788	1.979897	0.172978	0.751286	0.362786	350.501169	371.307153	986.627396	935.369799
0.492701	0.899053	0.378317	0.432369	0.480916	118.573863	144.634909	480.197750	444.976051
0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	49.000000	40.000000	124.000000
0.000000	1.000000	0.000000	1.000000	0.000000	263.500000	269.000000	654.000000	631.500000
1.000000	2.000000	0.000000	1.000000	0.000000	340.000000	353.000000	893.000000	865.000000
1.000000	3.000000	0.000000	1.000000	1.000000	423.000000	460.000000	1207.000000	1146.500000
1.000000	3.000000	1.000000	1.000000	1.000000	1199.000000	1119.000000	5011.000000	6035.000000

Descriptive statistics offer a concise summary of the dataset's key attributes, aiding in quickly grasping the general nature of the data. This overview can facilitate the subsequent data processing tasks. For instance, when we observe the mean value of the 'age' column, as depicted in Figure 4, we can determine that the typical AIDS patient’s age in the dataset is approximately 35 years. Additionally, by examining the standard deviation, we can conclude that the ages, on average, differ from this mean by approximately 8 years.

2.3 Data Exploration

Data exploration refers to the initial step in data analysis in which data analysts use data visualization and statistical techniques to describe dataset characterizations, such as size, quantity, and accuracy, in order to better understand the nature of the data.(Razavi et al., 2021)

Data exploration techniques include both manual analysis and automated data exploration software solutions that visually explore and identify relationships between different data variables, the structure of the dataset, the presence of outliers, and the distribution of data values in order to reveal patterns and points of interest, enabling data analysts to gain greater insight into the raw data.(Boulos, 2004)

2.3.1 Relationship Exploration Using Chi-squared tests and ANOVA correlation coefficient and Visualisation Using Bar Chart and Contingency Table

To gain insights into the relationships between data columns, chi-squared tests and ANOVA correlation coefficients can be utilized.In this dataset,there are features varying from numerical and categorical.The target variable,cid is a binary data,which is a categorical data.According to Brownlee (2020),for relationship between numerical input and categorical output,the most common techniques are correlation based, although in this case, must take the categorical target into account.Hence, Brownlee (2020) suggested us to use ANOVA correlation coefficient (linear).For relationship between categorical input and categorical output,Brownlee (2020) suggested us using Chi-Squared test (contingency tables).

On the other hand, ANOVA (Analysis of Variance) correlation coefficients can be used to find relationships with a binary variable, such as cid. ANOVA assesses whether the means of a dependent variable differ across levels of an independent variable, providing insights into potential relationships between the variables.

Analysis of variance (ANOVA) is a statistical test used to evaluate the difference between the means of more than two groups. This statistical analysis tool separates the total variability within a data set into two components: random and systematic factors.(Kenton, 2024)

A one-way ANOVA uses one independent variable. A two-way ANOVA uses two independent variables. Analysts use the ANOVA test to determine independent variables' influence on the dependent variable in a regression study.Both of these methods can provide valuable insights into the relationships between data columns and the variable of interest, cid.(Kenton, 2024)

ANOVA Test Table

cuemath
THE MATH EXPERT

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F Value
Between Groups	$SSB = \Sigma\, n_j(\bar{X}_j - \bar{X})^2$	$df_1 = k - 1$	$MSB = SSB / (k - 1)$	$f = MSB / MSE$
Error	$SSE = \Sigma\Sigma(X - \bar{X}_j)^2$	$df_2 = N - k$	$MSE = SSE / (N - k)$	
Total	$SST = SSB + SSE$	$df_3 = N - 1$		

2.3.1.1 Relationship Visualisation

ANOVA F-Value and P-Value

To visualize the relationship between different data columns, a visualization technique known as ANOVA F-Value and P-Value is employed. ANOVA F-Value and P-Value are defined as statistical measures used to assess the significance of the relationship between variables. ANOVA F-Value indicates the extent of variation between groups relative to the variation within groups, while P-Value assesses the statistical significance of this variation. ANOVA F-Value and P-Value use different values or different shades of significance to represent the strength of the relationship and communicate the significance of the variables plotted on the x-axis and y-axis.

Formula

$$F = \frac{MST}{MSE}$$

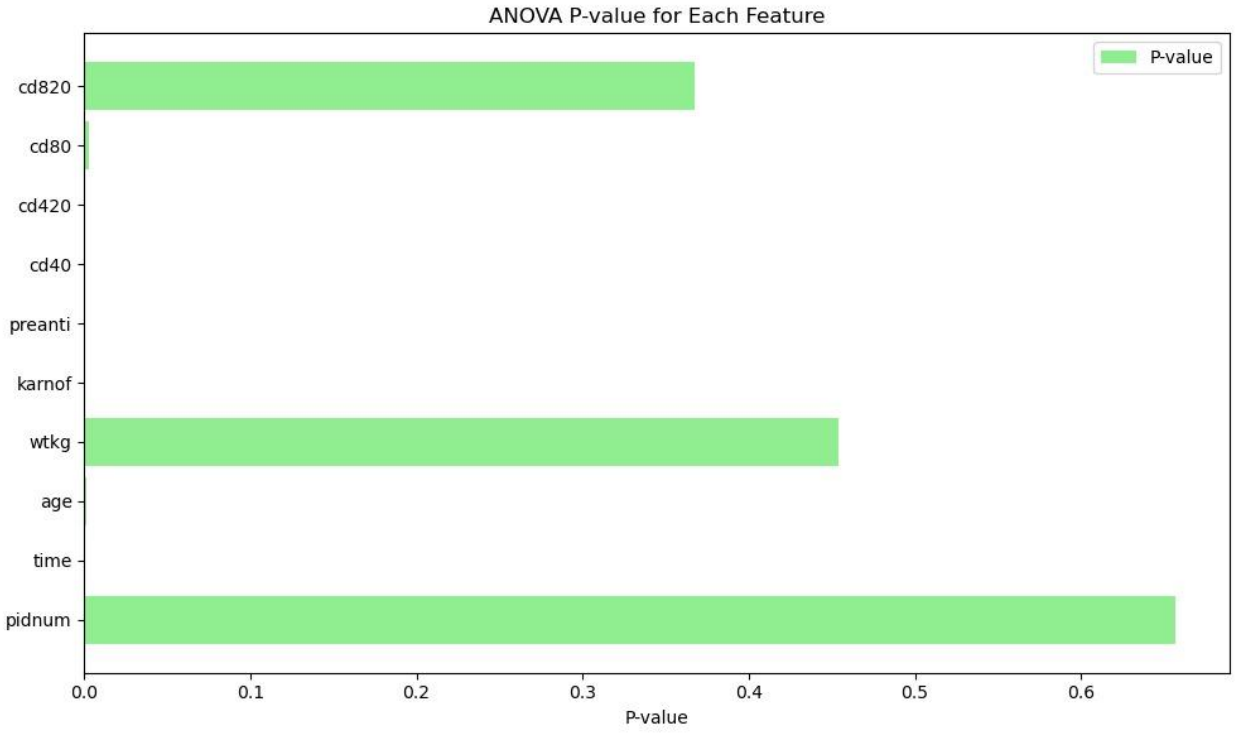
where:

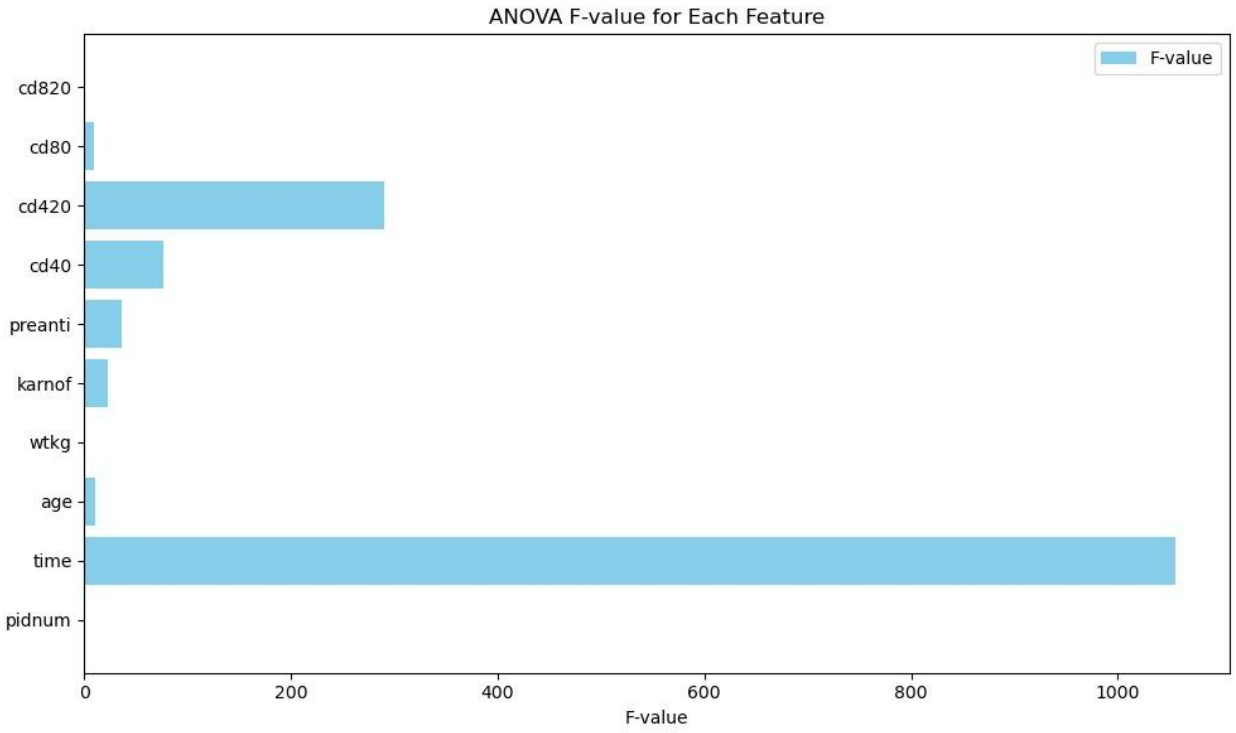
$F = \text{ANOVA coefficient}$

$MST = \text{Mean sum of squares due to treatment}$

$MSE = \text{Mean sum of squares due to error}$

	Feature	F-value	P-value
0	pidnum	0.197711	6.566199e-01
1	time	1055.468458	1.635917e-188
2	age	10.639134	1.124621e-03
3	wtkg	0.561305	4.538172e-01
4	karnof	22.889278	1.833325e-06
5	preanti	35.852420	2.490100e-09
6	cd40	76.279993	4.870459e-18
7	cd420	290.449562	3.628047e-61
8	cd80	9.162161	2.500212e-03
9	cd820	0.812389	3.675165e-01





By using python,we can find out the F-value and P-value of the numerical features with the target variable cid.The F-value is a measure of the difference between the means of the groups. It indicates how much the variance between groups exceeds the variance within groups. Higher F-values suggest a greater difference between group means. In ANOVA, you want to see high F-values, indicating that the variation between groups is larger compared to the variation within groups.The p-value associated with each feature indicates the probability of obtaining the observed F-value (or more extreme) if the null hypothesis is true. In the context of ANOVA, the null hypothesis typically states that there is no significant difference between the means of the groups. A low p-value (typically below a chosen significance level, e.g., 0.05) indicates that the observed differences between group means are unlikely to have occurred by chance alone, leading to the rejection of the null hypothesis. Therefore, lower p-values suggest greater significance.

Chi-squared Tests

Degrees of freedom (*df*)	Significance level (α)							
	.99	.975	.95	.9	.1	.05	.025	.01
1	-------	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892
40	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691
50	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154
60	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379
70	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425
80	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329
100	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116
1000	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807

Critical values of the Chi-square distribution with *d* degrees of freedom

	Probability of exceeding the critical value						
d	0.05	0.01	0.001	*d*	0.05	0.01	0.001
1	3.841	6.635	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820
10	18.307	23.209	29.588	20	31.410	37.566	45.315

INTRODUCTION TO POPULATION GENETICS, Table D.1

Chi-squared tests can be used to find relationships with a categorical variable, such as cid. It measures the association between two categorical variables and determines whether they are independent or related. A chi-square test is a statistical test that is used to compare observed and expected results. The goal of this test is to identify whether a disparity between actual and predicted data is due to chance or to a link between the variables under consideration. As a result, the chi-square test is an ideal choice for aiding in our understanding and interpretation of the connection between our two categorical variables.(Biswal, 2023)

A chi-square test or comparable nonparametric test is required to test a hypothesis regarding the distribution of a categorical variable. Categorical variables, which indicate categories such as animals or countries, can be nominal or ordinal. They cannot have a normal distribution since they can only have a few particular values.(Biswal, 2023)

The Chi-square formula:(Turney, 2023)

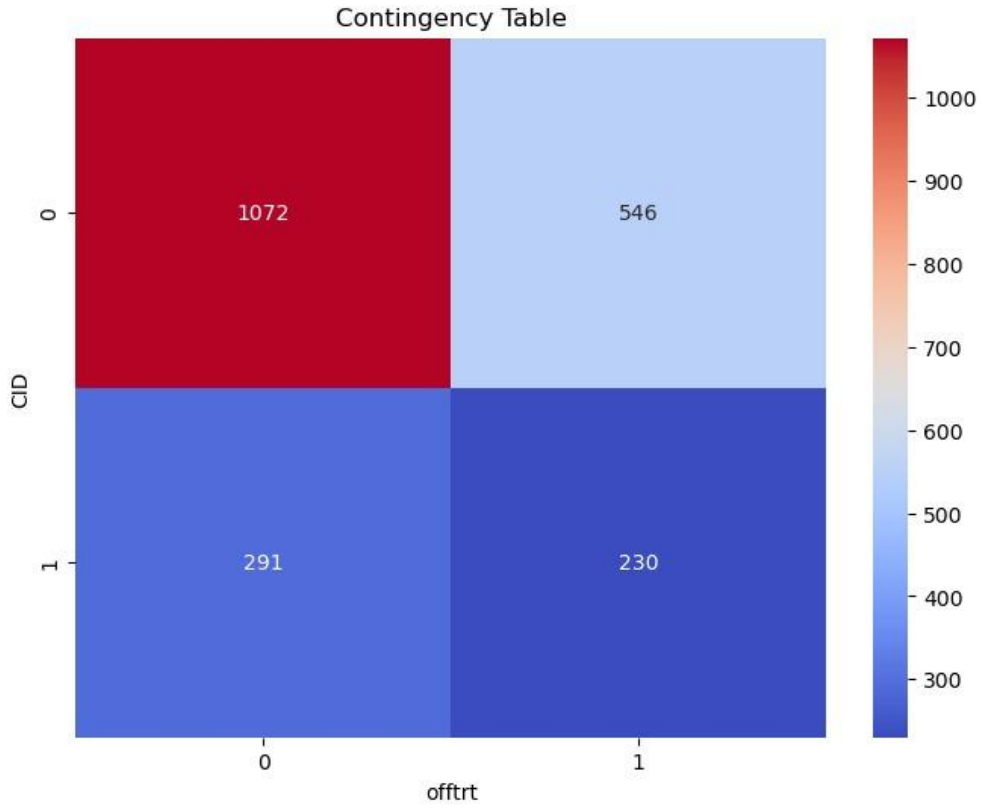
chi-square (X_2):

$$X^2 = \sum \frac{(O - E)^2}{E}$$

Where:

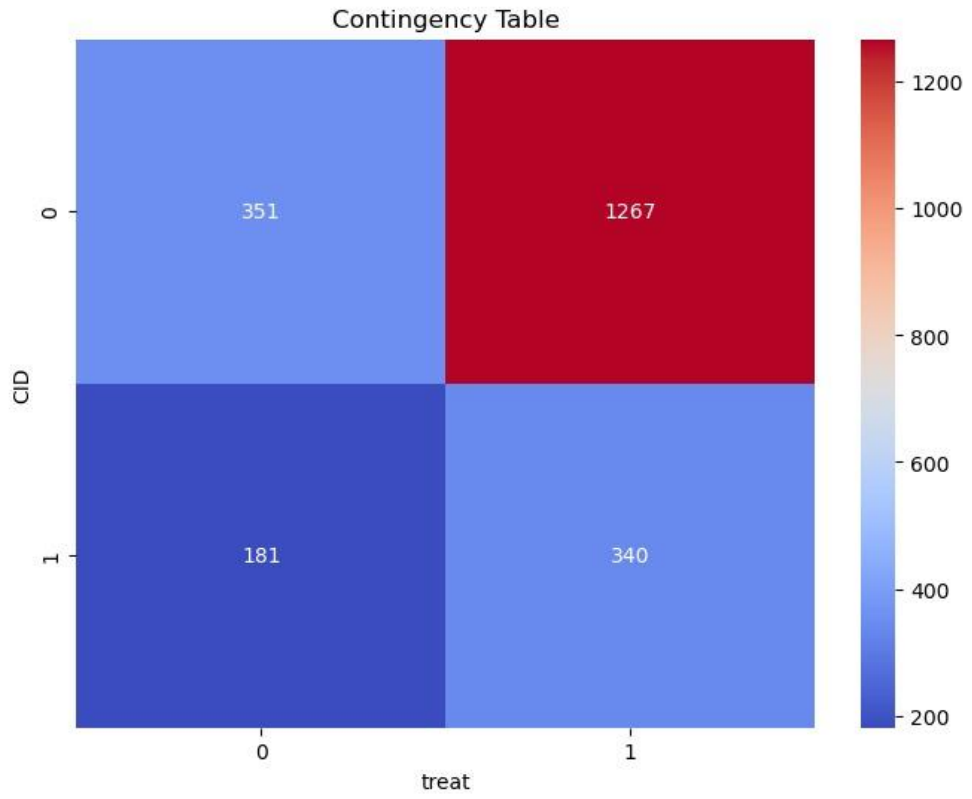
- X^2 is the chi-square test statistic
- Σ is the summation operator (it means “take the sum of”)
- O is the observed frequency
- E is the expected frequency

offtrt
Chi-Squared test results:
Chi-Squared statistic: 17.99359264445338
P-value: 2.2164976511816548e-05
Degrees of Freedom: 1
Expected frequencies:
[[1031.0116877 586.9883123]
 [331.9883123 189.0116877]]
Chi-Squared test is significant between 'cid' and 'offtrt'.



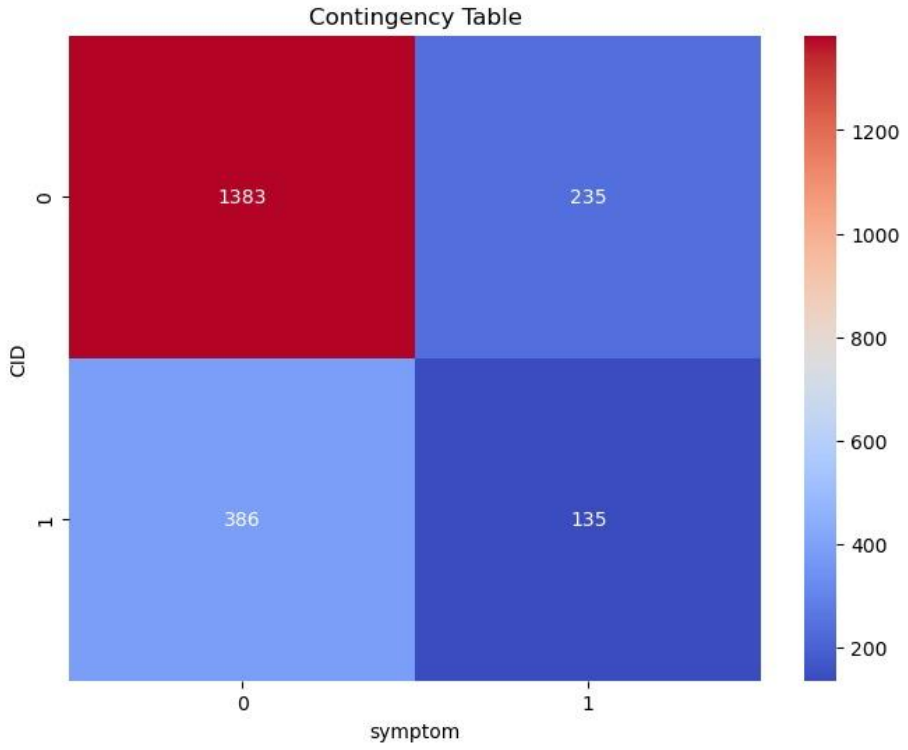
The Chi-Squared test reveals a significant association between the variables 'cid' and 'offtrt'. With a Chi-Squared statistic of 17.99 and a p-value of 2.22e-05 ($p<0.05$), the test suggests that the observed frequencies differ significantly from the expected frequencies. This indicates that 'cid' and 'offtrt' are not independent; their occurrences are related in a statistically significant manner. The degrees of freedom are 1, reflecting the number of variables minus 1. The expected frequencies indicate the anticipated distribution if 'cid' and 'offtrt' were independent, further confirming the significant association.

treat
Chi-Squared test results:
Chi-Squared statistic: 35.209703154948144
P-value: 2.9604482912362604e-09
Degrees of Freedom: 1
Expected frequencies:
[[402.41982235 1215.58017765]
 [129.58017765 391.41982235]]
Chi-Squared test is significant between 'cid' and 'treat'



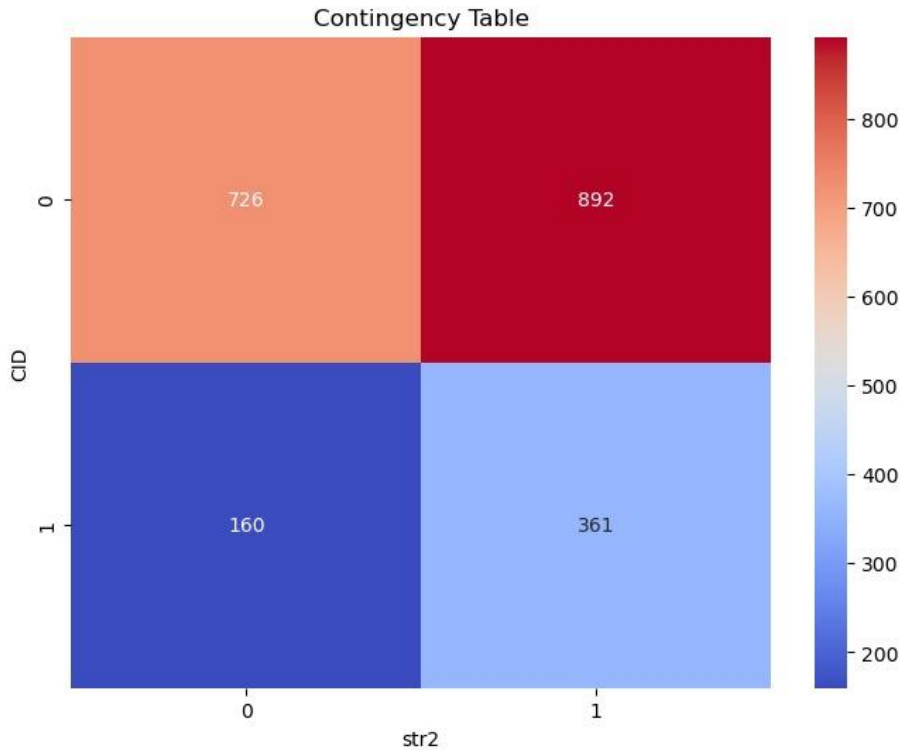
The Chi-Squared test reveals a significant association between 'cid' and 'treat'. With a Chi-Squared statistic of 35.21 and a p-value of 2.96e-09 ($p < 0.05$), the test indicates that the observed frequencies deviate significantly from the expected frequencies. This suggests that 'cid' and 'treat' are not independent; their occurrences are related in a statistically significant manner. The degrees of freedom are 1, reflecting the number of variables minus 1. The expected frequencies show the anticipated distribution if 'cid' and 'treat' were independent, further confirming the significant association.

symptom
Chi-Squared test results:
Chi-Squared statistic: 34.93258896428114
P-value: 3.4131991121101056e-09
Degrees of Freedom: 1
Expected frequencies:
[[1338.12155213 279.87844787]
 [430.87844787   90.12155213]]
Chi-Squared test is significant between 'cid' and 'symptom'.



The Chi-Squared test reveals a significant relationship between 'cid' and 'symptom'. With a Chi-Squared statistic of 34.93 and an extremely low p-value of 3.41e-09 ($p < 0.05$), it indicates that the observed frequencies substantially deviate from the expected frequencies. This suggests that 'cid' and 'symptom' are not independent; their occurrences are related in a statistically significant manner. The degrees of freedom, 1, denote the difference between the number of observed frequencies and the number of variables minus 1. The expected frequencies show the anticipated distribution if 'cid' and 'symptom' were independent, further confirming the significant relationship.

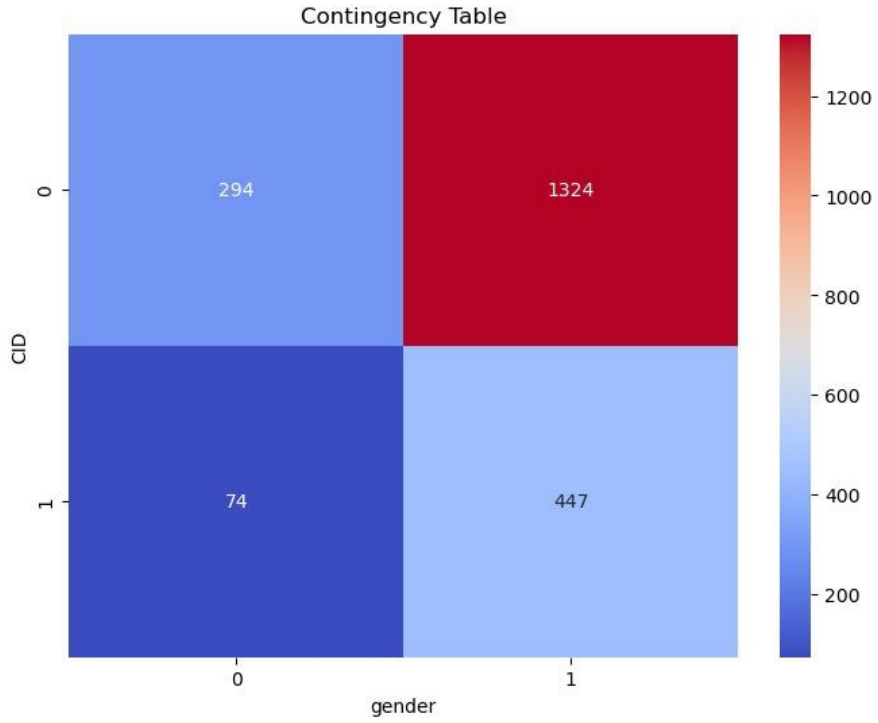
str2
Chi-Squared test results:
Chi-Squared statistic: 31.985529316703058
P-value: 1.5532532507796213e-08
Degrees of Freedom: 1
Expected frequencies:
[[670.19541842 947.80458158]
 [215.80458158 305.19541842]]
Chi-Squared test is significant between 'cid' and 'str2'.



The Chi-Squared test indicates a significant relationship between 'cid' and 'str2'. With a Chi-Squared statistic of 31.99 and an extremely low p-value of 1.55e-08 ($p < 0.05$), it suggests that the observed frequencies differ significantly from the expected frequencies. This implies that 'cid' and 'str2' are not independent; their occurrences are correlated in a statistically significant manner. The degrees of freedom, 1, represent the difference between the number of observed frequencies and the number of variables minus 1. The expected frequencies display the anticipated distribution if 'cid' and 'str2' were independent, further confirming the significant relationship.

gender
Chi-Squared test results:

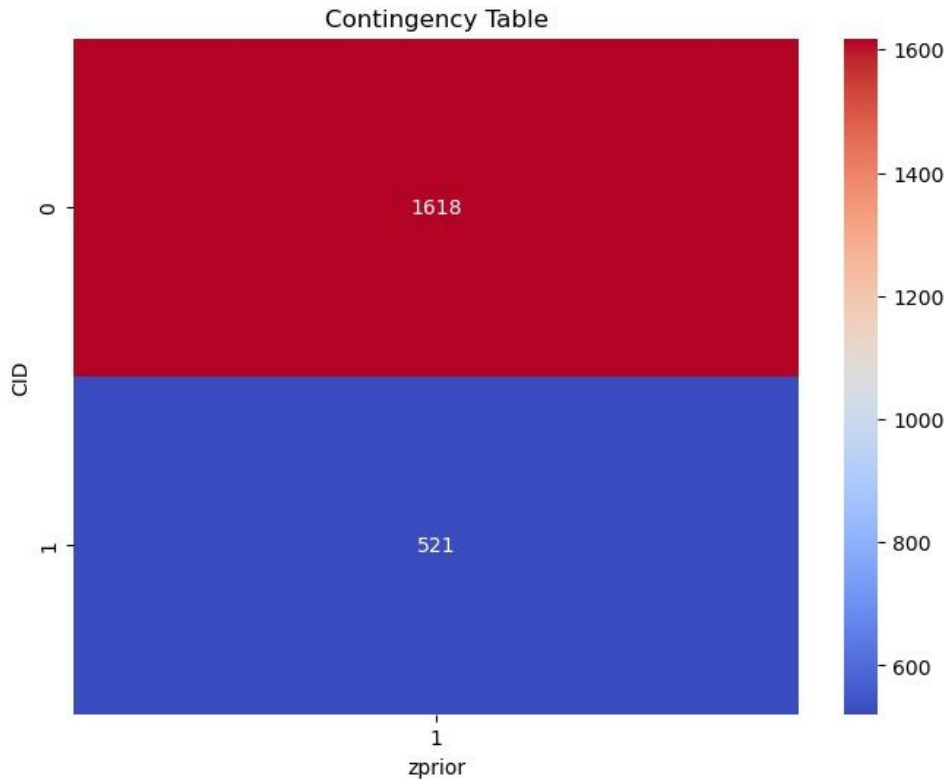
Chi-Squared statistic: 4.080192406029964
P-value: 0.04338871786278083
Degrees of Freedom: 1
Expected frequencies:
[[278.3655914 1339.6344086]
 [89.6344086 431.3655914]]
Chi-Squared test is significant between 'cid' and 'gender'.



The Chi-Squared test suggests a significant relationship between 'cid' and 'gender'. With a Chi-Squared statistic of 4.08 and a p-value of 0.043 ($p < 0.05$), it indicates that observed frequencies deviate from expected frequencies. This implies 'cid' and 'gender' are not independent; their occurrences are linked in a statistically significant manner. The degrees of freedom, 1, reflect the difference between observed frequencies and variables minus 1. Expected frequencies depict the anticipated distribution if 'cid' and 'gender' were independent, further confirming the significant relationship between these variables.

zprior
Chi-Squared test results:
Chi-Squared statistic: 0.0

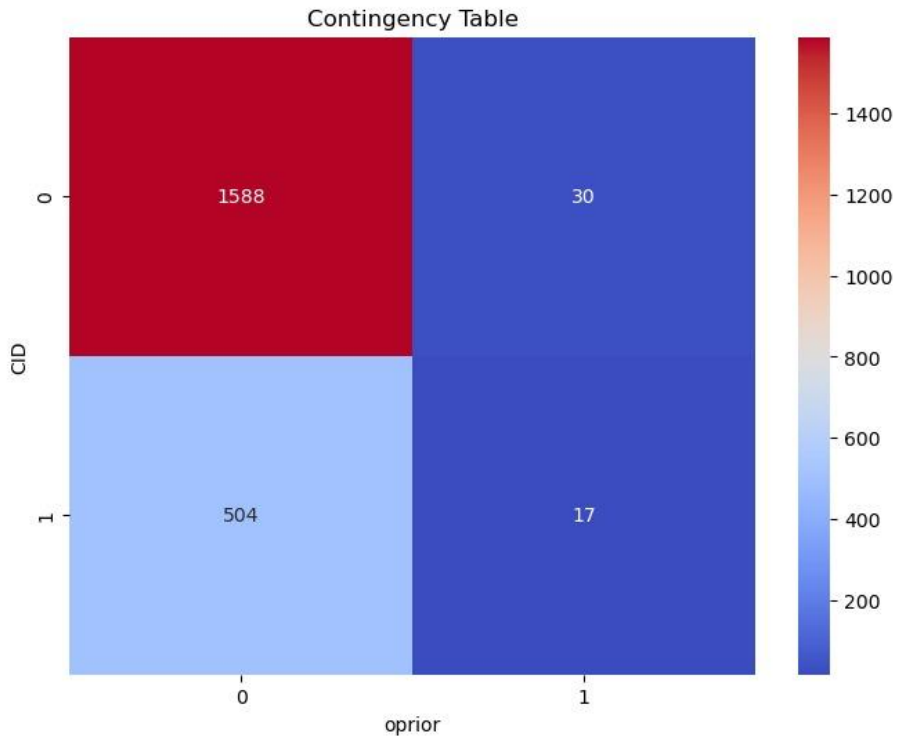
P-value: 1.0
Degrees of Freedom: 0
Expected frequencies:
[[1618.]
 [521.]]
Chi-Squared test is not significant between 'cid' and 'zprior'.



The Chi-Squared test results indicate that there is no significant relationship between 'cid' and 'zprior'. With a Chi-Squared statistic of 0 and a p-value of 1.0, the test suggests that observed frequencies match expected frequencies perfectly. Additionally, having 0 degrees of freedom indicates that there's no variability in the data beyond what is expected. Therefore, 'cid' and 'zprior' appear to be independent variables, with their occurrences not correlated in a statistically significant manner. This lack of significance suggests that any observed differences between 'cid' and 'zprior' frequencies are likely due to chance rather than a meaningful relationship.

oprior
Chi-Squared test results:
Chi-Squared statistic: 3.0137354139386625

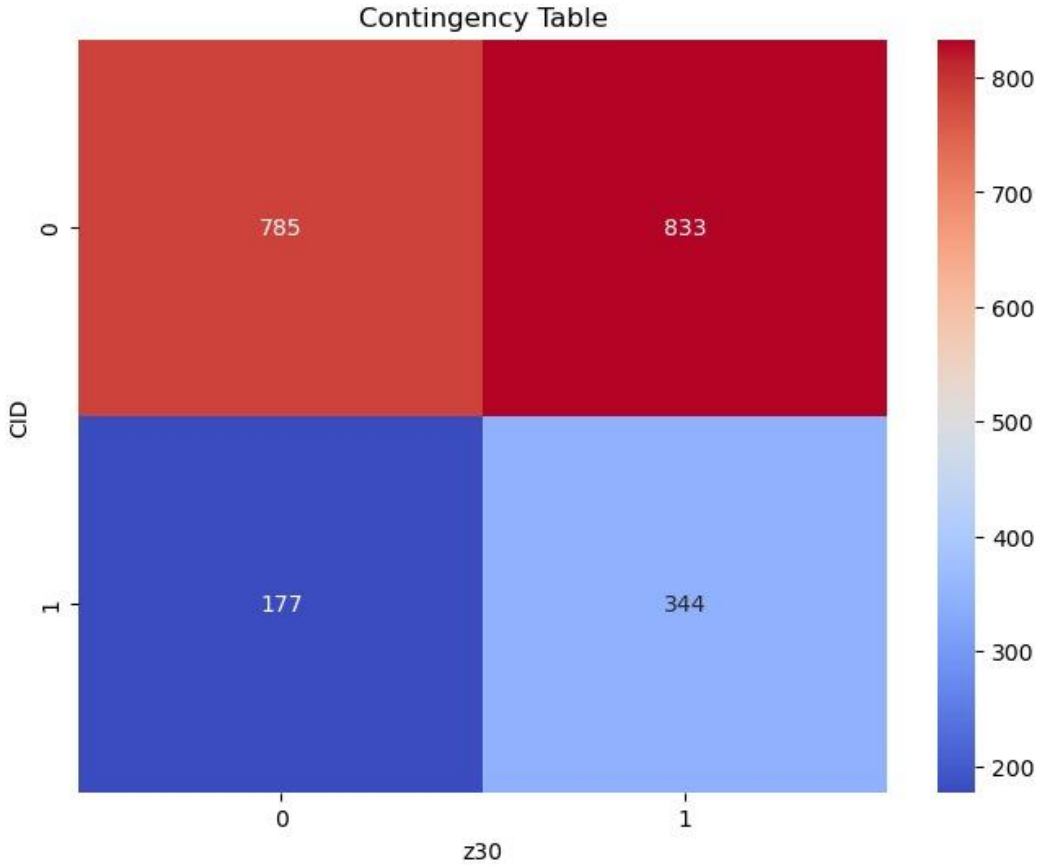
P-value: 0.08256182737180419
Degrees of Freedom: 1
Expected frequencies:
[[1582.44787284 35.55212716]
 [509.55212716 11.44787284]]
Chi-Squared test is not significant between 'cid' and 'oprior'.



The Chi-Squared test indicates that there is no significant relationship between 'cid' and 'oprior'. With a Chi-Squared statistic of 3.01 and a p-value of 0.083 (p>0.05), the test suggests that observed frequencies do not significantly differ from expected frequencies. This implies that 'cid' and 'oprior' are likely independent variables, with their occurrences not correlated in a statistically significant manner. The degrees of freedom, 1, reflect the difference between observed frequencies and variables minus 1. Therefore, any observed disparities between 'cid' and 'oprior' frequencies may be due to random chance rather than a meaningful relationship.

z30
Chi-Squared test results:
Chi-Squared statistic: 33.098356648747504

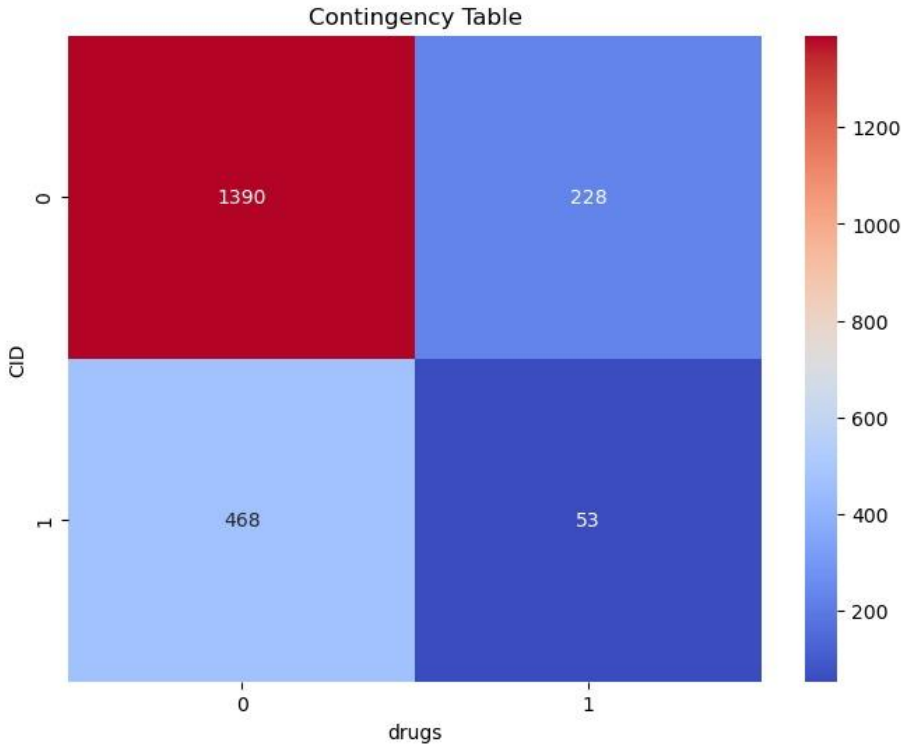
P-value: 8.761273941595931e-09
Degrees of Freedom: 1
Expected frequencies:
[[727.68396447 890.31603553]
 [234.31603553 286.68396447]]
Chi-Squared test is significant between 'cid' and 'z30'.



The Chi-Squared test reveals a significant relationship between 'cid' and 'z30'. With a Chi-Squared statistic of 33.10 and an extremely low p-value of 8.76e-09 ($p < 0.05$), the test indicates that observed frequencies significantly deviate from expected frequencies. This suggests that 'cid' and 'z30' are not independent; their occurrences are correlated in a statistically significant manner. The degrees of freedom, 1, represent the difference between observed frequencies and variables minus 1. The expected frequencies show the anticipated distribution if 'cid' and 'z30' were independent, further confirming the significant relationship between these variables.

Drugs
Chi-Squared test results:
Chi-Squared statistic: 4.965675610901002
P-value: 0.025855212743143514

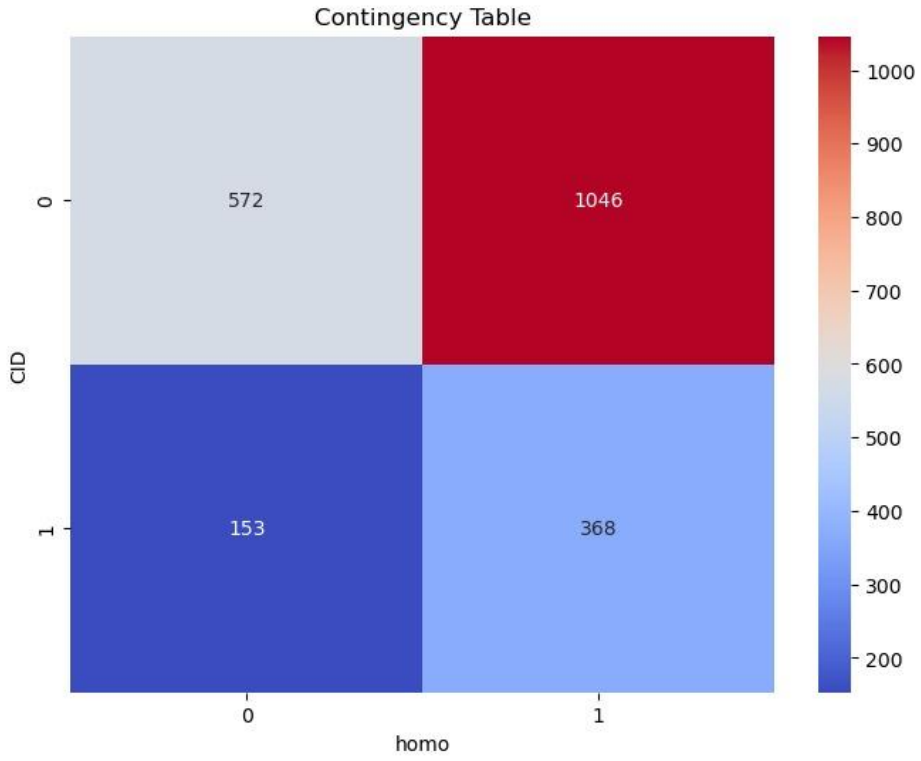
Degrees of Freedom: 1
Expected frequencies:
[[1405.44366526 212.55633474]
 [452.55633474   68.44366526]]
Chi-Squared test is significant between 'cid' and 'drugs'.



The Chi-Squared test reveals a significant relationship between 'cid' and 'drugs' ($p < 0.05$). Observed frequencies significantly differ from expected frequencies, suggesting non-independence. 'cid' and 'drugs' are likely correlated, indicating that the occurrence of drug usage varies significantly across different values of 'cid'. This suggests a potential association between the use of drugs and the 'cid' variable, warranting further investigation to understand the underlying factors driving this relationship and its implications.

homo
Chi-Squared test results:
Chi-Squared statistic: 6.037523030250535
P-value: 0.01400492038961362

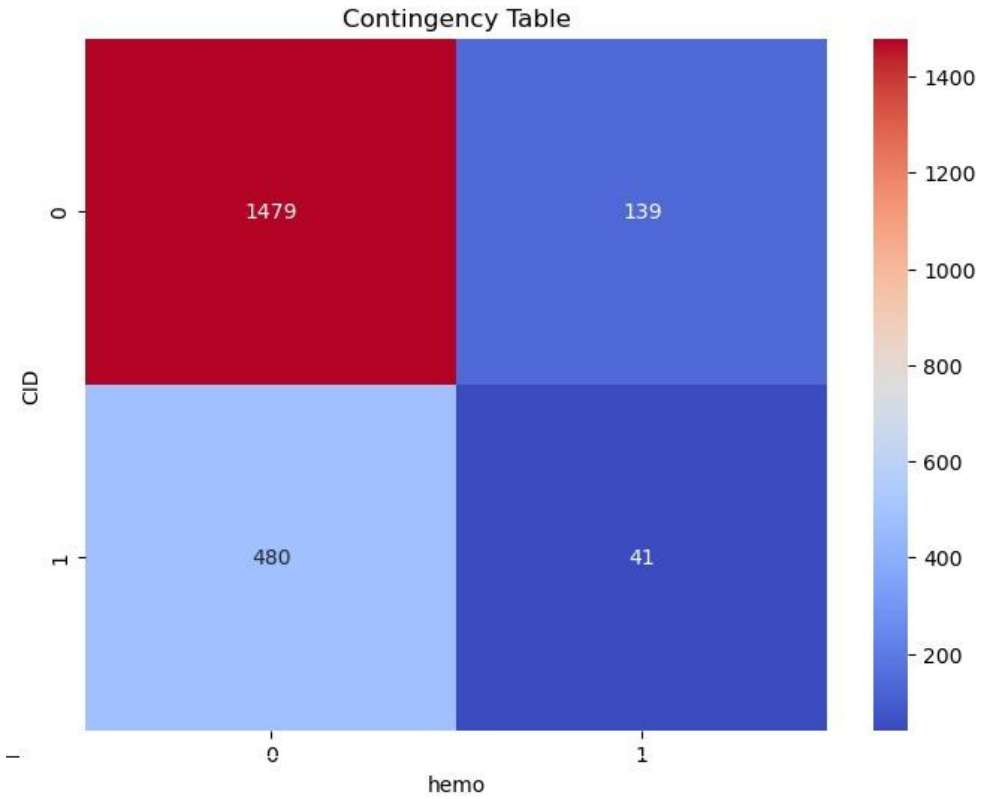
Degrees of Freedom: 1
Expected frequencies:
[[548.41047218 1069.58952782]
 [176.58952782 344.41047218]]
Chi-Squared test is significant between 'cid' and 'homo'.



The Chi-Squared test indicates a significant relationship between 'cid' and 'homo' ($p < 0.05$). Observed frequencies significantly deviate from expected frequencies, suggesting non-independence. This implies that the occurrence of 'homo' varies significantly across different values of 'cid'. Further investigation is warranted to understand the underlying factors driving this relationship. The significant association between 'cid' and 'homo' may indicate potential correlations or influences between these variables, prompting researchers to explore the implications of these findings and potentially adjust analytical approaches to account for this relationship.

hemo
Chi-Squared test results:
Chi-Squared statistic: 0.18072701363136065
P-value: 0.6707491995814665

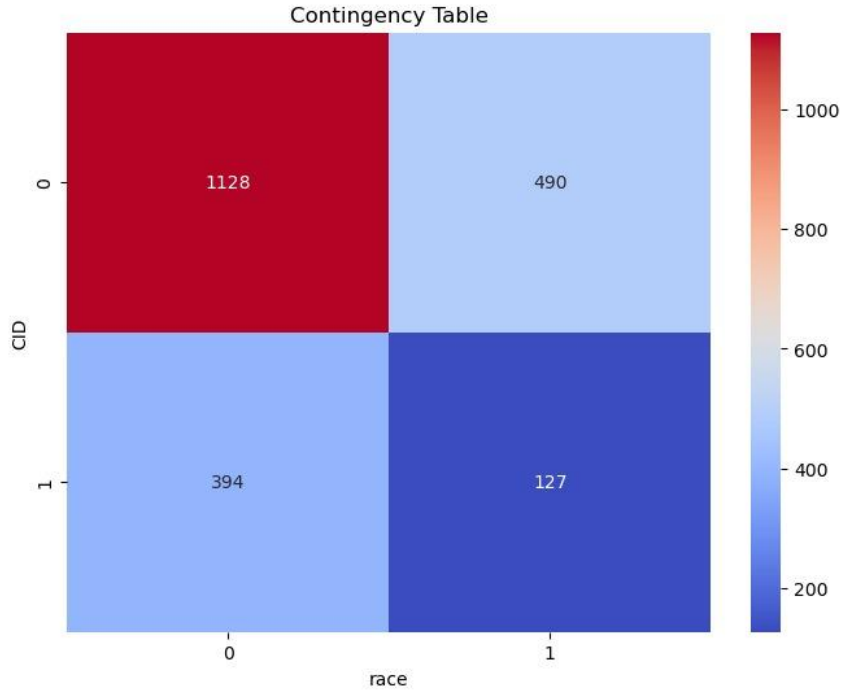
Degrees of Freedom: 1
Expected frequencies:
[[1481.84291725 136.15708275]
 [477.15708275   43.84291725]]
Chi-Squared test is not significant between 'cid' and 'hemo'.



The Chi-Squared test indicates a lack of significant association between 'cid' and 'hemo' ($p = 0.67$), with one degree of freedom. The observed frequencies closely align with the expected frequencies, suggesting that any differences observed between the two variables are likely due to random chance rather than a systematic relationship. This implies that the occurrence of 'hemo' does not significantly vary across different values of 'cid'. Consequently, 'cid' does not appear to influence the occurrence of 'hemo', and further investigation into other factors affecting 'hemo' may be warranted to better understand its distribution within the dataset.

race
Chi-Squared test results:
Chi-Squared statistic: 6.417527774510406
P-value: 0.011299936224922476

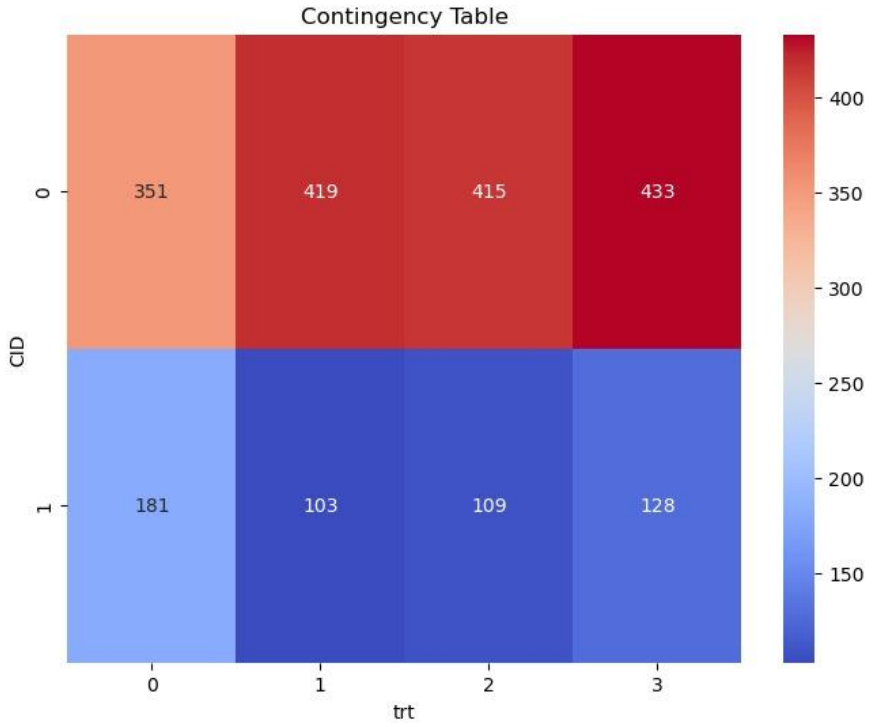
Degrees of Freedom: 1
Expected frequencies:
[[1151.28377747 466.71622253]
 [370.71622253 150.28377747]]
Chi-Squared test is significant between 'cid' and 'race'.



The Chi-Squared test reveals a significant relationship between 'cid' and 'race' ($p < 0.05$). Observed frequencies significantly differ from expected frequencies, indicating non-independence. This suggests that the distribution of 'race' varies significantly across different values of 'cid'. Further investigation is needed to discern the underlying reasons for this association. The significant correlation between 'cid' and 'race' may imply potential influences or interactions between these variables, underscoring the importance of considering race-related factors in analyzing the outcomes or characteristics associated with 'cid' values.

trt
Chi-Squared test results:
Chi-Squared statistic: 37.354420872172085
P-value: 3.871626196387447e-08
Degrees of Freedom: 3

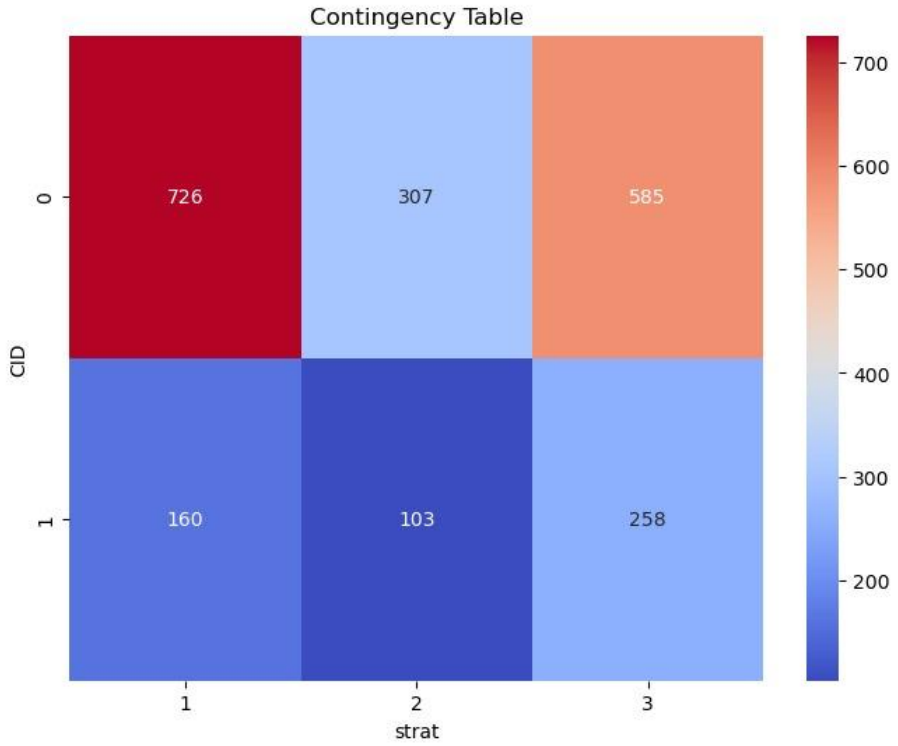
Expected frequencies:
[[402.41982235 394.85553997 396.36839645 424.35624123]
 [129.58017765 127.14446003 127.63160355 136.64375877]]
Chi-Squared test is significant between 'cid' and 'trt'.



The Chi-Squared test indicates a significant relationship between 'cid' and 'trt' ($p < 0.05$). With four degrees of freedom, the observed frequencies significantly diverge from the expected frequencies, signifying non-independence. This suggests that the distribution of 'trt' varies significantly across different values of 'cid'. Further exploration is essential to understand the underlying factors contributing to this association. The significant correlation between 'cid' and 'trt' may imply complex interactions or influences between these variables, underscoring the importance of considering treatment-related factors in analyzing the outcomes or characteristics associated with different 'cid' values.

strat
Chi-Squared test results:
Chi-Squared statistic: 37.0674760299078
P-value: 8.930995099479137e-09
Degrees of Freedom: 2

Expected frequencies:
[[670.19541842 310.13557737 637.66900421]
 [215.80458158 99.86442263 205.33099579]]
Chi-Squared test is significant between 'cid' and 'strat'.



The Chi-Squared test reveals a significant relationship between 'cid' and 'strat' ($p < 0.05$) with two degrees of freedom. The observed frequencies significantly differ from the expected frequencies, indicating a non-random association between the two variables. This suggests that the distribution of 'strat' varies significantly across different values of 'cid'. Further investigation is warranted to discern the underlying factors driving this association. The significant correlation between 'cid' and 'strat' underscores potential complexities in the stratification process or inherent differences in patient characteristics across different 'cid' groups, necessitating careful consideration in subsequent analyses or interpretations.

2.3.1.2 Relationship Exploration

The results indicated the strength of the relationship between each feature variable and the target variable using ANOVA and Chi-squared tests. A very strong relationship with the target

variable was indicated by their respective effect size measures (η^2 for ANOVA and Cramer's V or Phi coefficient for Chi-squared).

Feature Variable	F value	P value	Chi-squared tests
pidnum	0.197711	6.566199e-01	-
cid	-	-	-
time	1055.468458	1.635917e-188	-
trt	-	-	significant
age	10.639134	1.124621e-03	-
wtkg	0.561305	4.538172e-01	-
hemo	-	-	Not significant
homo	-	-	significant
drugs	-	-	significant
karnof	22.889278	1.833325e-06	-
oprior	-	-	Not significant
zprior	-	-	Not significant
z30	-	-	significant
preanti	35.852420	2.490100e-09	-
race	-	-	significant
gender`	-	-	significant
str2	-	-	significant
strat	-	-	significant
symptom	-	-	significant
treat	-	-	significant
offtrt	-	-	significant
cd40	76.279993	4.870459e-18	-

cd420	290.449562	3.628047e-61	-
cd80	9.162161	2.500212e-03	-
cd820	0.812389	3.675165e-01	-

Table 4: Summary of value of F value,P value and Chi-squared test result of features with target with their respective r value

From the table above,we have found out the relationship between features and target variable by using ANOVA correlation coefficient and Chi-squared test.We have use two different types of method for numerical features and categorical features.For numerical features,we use ANOVA correlation coefficient to find their relationship with the target variables.For categorical features,we use Chi-squared test to find their relationship with the target variable.

By interpreting the results obtained,we can find out that ‘time’, ‘age’, ‘karnof’, ‘preanti’, ‘cd40’, ‘cd420’, ‘cd80’, ‘offtrt’, ‘treat’, ‘symptom’, ‘str2’, ‘gender’, ‘z30’, ‘drugs’ , ‘homo’ , ‘race’, ‘trt’ and ‘strat’ have significant association with the target variable ‘cid’.For more details on how we interpret the results,we will discuss further in 3.1 Data Selection.

2.4 Data Quality Verification

During the data quality assessment phase, the primary goal is to detect anomalies. Anomalies are data points that deviate significantly from the norm within the dataset, indicating

potential errors or unusual occurrences. Detecting anomalies is essential for identifying and addressing erroneous data points, which can distort the insights derived from the dataset. By flagging anomalies, further investigation can be conducted, and if necessary, corrective actions can be taken during the data cleaning process. This is crucial because anomalies can disproportionately impact the training of machine learning models. If not addressed, models may prioritize learning from these anomalies rather than capturing the underlying patterns within the majority of the data. Consequently, such models may struggle to generalize well and perform accurately on new, unseen data. The three primary categories of anomalies include missing data, duplicated data, and outliers.

2.4.1 Null Data Detection

Null data, also known as missing data, refers to instances in the dataset where information for a particular variable or observation is unavailable or not recorded. Missing data may arise due to various factors, including improper maintenance or data corruption in the past. In Pandas, missing values are typically represented by 'NaN,' which stands for 'Not a Number'

```
[54]: df.isnull().any().any()

[54]: False
```

Figure 48: Null Data Detection

According to the output in Figure 48, each column in the dataset is populated with complete data, and there are no missing values present. This could be a result of careful and thorough data collection processes to ensure that all necessary information was collected without any omissions.

2.4.2 Duplicate Data Detection

```python
import pandas as pd

# Load the CSV file into a DataFrame
df = pd.read_csv('data.csv')

# Define the list of columns to check for duplicates
columns_to_check = ['time', 'age', 'karnof', 'preanti', 'cd40', 'cd420', 'cd80', 'offtrt',
                    'treat', 'symptom', 'str2', 'gender', 'z30', 'drugs', 'homo', 'race',
                    'trt', 'strat']

def is_binary(series):
    return len(series.unique()) == 2 and 0 in series.unique() and 1 in series.unique()

# Check for duplicates in non-binary columns
non_binary_columns = [col for col in columns_to_check if not is_binary(df[col])]
duplicate_rows = df.duplicated(subset=non_binary_columns, keep='first')

if duplicate_rows.any():
    print("Duplicate rows found in the DataFrame (excluding binary columns).")
    print("Duplicate rows:")
    print(df[duplicate_rows])
else:
    print("No duplicate rows found in the DataFrame (excluding binary columns).")
```

```
No duplicate rows found in the DataFrame (excluding binary columns).
```

Duplicate data records are data entries that have identical or very similar values for one or more attributes, such as name, email, phone number, or ID. They can occur due to various reasons, such as data entry errors, data integration issues, data scraping methods, or data manipulation techniques (*How Do You Manage Duplicate Data Records?*, 2023).

Data deduplication is the process of reducing the amount of duplicate data stored on a system. By identifying and removing duplicate data, you can reduce the amount of storage space needed to store your data. This is especially useful for large datasets that contain a lot of redundant data, such as backups (*What Is Data Deduplication? How to Improve Data Uniqueness*, n.d.-b).

In this dataset, it is expected that each data is unique because data is collected from different patients that are individually unique, meaning that the same data should not appear more than once. Therefore, every column is examined to determine if there are any duplicate records.

Referring to the figure, the output indicates that there are no duplicate rows found in the dataset. Based on this information, it can be concluded that there are no duplicated data entries in the dataset. This may be due to the fact that the database where this dataset was stored may have implemented measures to ensure that duplicate data insertion is prohibited.(Syed, 2023)

2.4.3 Data Outlier

An outlier in a dataset represents a data point that deviates notably from the majority of observations, lying outside the anticipated or typical range for a given variable. These values significantly differ from the bulk of the dataset, potentially skewing statistical analyses and interpretations. (Syed, 2023)Outliers can arise due to various reasons, such as measurement errors, sampling variability, or genuine anomalies in the data. Identifying and understanding outliers is crucial in data analysis as they can impact the validity of conclusions drawn from the dataset, requiring careful consideration and potentially specialized treatment during data processing and analysis. (Syed, 2023)

We will further discuss the outliers of our dataset and how we handle these outliers in 3.3 Outliers Detection.

2.4.4 Zero Values Detection

```
|: import pandas as pd

# Load the CSV file into a DataFrame
df = pd.read_csv('data.csv')

# Define the list of columns to check for zero values
columns_to_check = ['time', 'age', 'karnof', 'preanti', 'cd40', 'cd420', 'cd80', 'offtrt',
                    'treat', 'symptom', 'str2', 'gender', 'z30', 'drugs', 'homo', 'race',
                    'trt', 'strat']

def is_binary(series):
    return len(series.unique()) == 2 and 0 in series.unique() and 1 in series.unique()

# Check if any columns contain zero values (excluding binary columns)
zero_values_columns = [col for col in columns_to_check if col not in df.columns[df.isnull().any()]]
zero_values_columns = [col for col in zero_values_columns if not is_binary(df[col]) and (df[col] == 0).any()]

if len(zero_values_columns) > 0:
    print("Columns containing zero values (excluding binary columns):")
    print(zero_values_columns)
else:
    print("No columns contain zero values (excluding binary columns).")
```

```
Columns containing zero values (excluding binary columns):
['preanti', 'cd40', 'trt']
```

Zero Values Detection refers to the identification and interpretation of data points within a dataset where specific features or variables exhibit a value of zero. In various contexts, a zero value can carry significant meaning or implications. For instance, in the context of treatment-related features such as "trt," "cd40," and "preanti," zero values often convey crucial information about the absence of treatment, baseline conditions, or specific measurements.

In the "trt" feature, a value of zero typically signifies either the absence of treatment or the inclusion in a baseline or control group within a study or experiment.

Similarly, in the "cd40" feature, a zero value indicates the absence or severe depletion of CD4 positive T cells at the baseline measurement, often associated with conditions like advanced HIV/AIDS or severe immunodeficiency.

In the "preanti" feature, a zero value suggests that there were no days of pre-175 anti-retroviral therapy for the corresponding data point or observation.

It's important to note that while zero values may be indicative of meaningful conditions or circumstances in a dataset, not all zero values are necessarily informative. Therefore, careful consideration and domain expertise are required to determine which zero values should be retained and which should be treated as missing or irrelevant.

3.0 Data Preparation

Data preparation in data science involves cleaning and transforming raw data before analysis(Talend, n.d.). It ensures data quality by addressing issues like missing values, outliers, and inconsistencies. By refining the data, researchers enhance its reliability and suitability for accurate analysis, leading to more meaningful insights and decision-making.

The Data preparation steps are Gather data, Discover and assess data, Cleanse and validate data, Transform and enrich data and Store data. Cleaning up the data is traditionally the most time-consuming part of the data preparation process, but it's crucial for removing faulty data and filling in gaps(Talend, n.d.).

3.1 Data Selection

Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection. The data selection stage significantly relies on feature selection. The main idea of feature selection is to choose a subset of input variables by eliminating features with little or no predictive information.(Strauss, 1987) Feature selection implies some degree of cardinality reduction to impose a cutoff on the number of attributes that can be considered when building a model. This may improve the quality of the model, and also makes the process of modeling more efficient .(Strauss, 1987)

For tabular datasets with a binary target variable, two popular feature selection techniques for numerical input data are ANOVA (Analysis of Variance) and Chi-squared tests. The main difference between ANOVA and Chi-squared tests is that ANOVA assesses the relationship between numerical input features and a categorical target variable, while Chi-squared tests evaluate the association between categorical input features and a categorical target variable.

3.1.1 Data Selection by ANOVA and Chi-Square

In the data understanding phase, ANOVA and Chi-squared tests analysis has been conducted. Based on the computed ANOVA and Chi-squared tests(F value and P value) , the following features have a very strong relationship with the target variable was indicated by their respective effect size measures (η^2 for ANOVA and Cramer's V or Phi coefficient for Chi-squared). The very strong relationship is according to F values and a very small P Values ($p<0.05$) (Bobbitt, 2021b) .

the following features should be chosen:

Feature	F value	P value	Chi-squared tests
time	1055.468458	1.635917e-188	-
age	10.639134	1.124621e-03	Significant
karnof	22.889278	1.833325e-06	-
preanti	35.852420	2.490100e-09	-
cd40	76.279993	4.870459e-18	-
cd420	290.449562	3.628047e-61	-
cd80	9.162161	2.500212e-03	-
offtrt	18.583723	1.699882e-05	Significant
treat	36.483399	1.811097e-09	Significant
symptom	36.296981	1.989682e-09	Significant
str2	33.039069	1.032750e-08	Significant
gender	4.359045	3.693109e-02	Significant
z30	34.190385	5.766692e-09	Significant
drugs	5.311739	2.127764e-02	Significant
homo	6.314549	1.204820e-02	Significant
race	6.717070	9.614473e-03	Significant
trt	15.317075	9.373532e-05	Significant

strat	37.569729	1.047388e-09	Significant

Base on the very strong relationship, the selected features are time, age, karnof, preanti, cd40, cd420, cd80, afftrt, treat, symptom, str2, gender, z30, drugs, homo, race, trt, strat.

3.1.2 Summary of Selected Data

Feature	Type
time	Integer
age	Integer
karnof	Integer
preanti	Integer
cd40	Integer
cd420	Integer
cd80	Integer
offtrt	Binary
treat	Binary
symptom	Binary
str2	Binary
gender	Binary
z30	Binary
drugs	Binary
homo	Binary
race	Integer
trt	Integer
strat	Integer

3.2 Data Cleaning

Data cleaning, also referred to as data cleansing or data scrubbing, is the process of fixing incorrect, incomplete, duplicate or otherwise erroneous data in a data set (Stedman, 2022). In the data understanding phase, the anomalies have been identified. Outliers and Zero Values are the only anomalies detected in the given datasat. Hence, the main focus in this stage is to handle these outliers.

3.2.1 Outliers Detection

Outliers are extreme values that are different from most other data points in the data set. They can have a significant impact on your statistical analysis and distort the results of any hypothesis testing.(Bhandari, 2024)

Outlier with time

```
[111]: (array([ 190,  194,  200,  310,  361,  364,  437,  454,  493,  634,  645,
                701,  745,  777,  783,  817,  822,  854,  871,  896,  928, 1113,
               1127, 1162, 1169, 1196, 1254, 1295, 1313, 1316, 1362, 1439, 1529,
               1562, 1578, 1590, 1604, 1691, 1703, 1790, 1893, 1957, 2001, 2017,
               2029, 2030, 2057, 2115, 2123], dtype=int64),)
Outlier Values: 190     159
194     137
200      55
310     162
361     125
364      69
437     148
454     154
493     147
634     135
645      62
701     140
745     105
777      54
783     163
817     154
822     174
854     146
871     147
896     150
928     138
1113    169
1127    147
1162    145
1169     50
1196    140
1254     33
1295    126
1313    140
1316     14
1362    111
1439    169
1529    142
1562    167
1578    146
1590     45
1604     96
1691    139
1703    171
1790    143
1893    158
1957    134
2001    158
2017    133
2029    148
2030    133
2057    109
2115    150
2123    154
Name: time, dtype: int64
```

The outlier in this dataset has a time-to-death of 701 days, which is significantly longer than the other data points. There are several possible reasons for this outlier: The patient may have been young at the time of diagnosis. Younger patients tend to have a longer life expectancy after

diagnosis with HIV than older patients. The patient may have contracted HIV through sexual transmission rather than intravenous drug use. People who contract HIV through intravenous drug use tend to progress to AIDS more quickly than those who contract it through sexual transmission.

The patient may have been diagnosed with HIV in the early stages of the disease. Early diagnosis and treatment with ART can significantly prolong life expectancy.The patient may have received highly effective ART treatment. ART can dramatically slow the progression of HIV and AIDS, and can help people with HIV live long and healthy lives.

It is important to note that these are just possible reasons for the outlier. Without more information about the patient, it is impossible to say for sure what caused them to live so much longer than the other patients in the dataset.

Since this represents the time to failure or censoring, outliers may not necessarily be erroneous data points but rather indicate extreme durations. Removing outliers from this feature could depend on the specific context of your analysis and whether extreme durations are expected or plausible.

Outlier with trt

```
(array([], dtype=int64),)
```

The absence of outliers in "trt" data usually means that there are no unusual observations in the data set that deviate from the normal pattern. Outliers refer to observations that have significantly different values or characteristics than most observations in the data set.

Outlier with age

```
[117]: (array([   1,   49,  100,  104,  170,  217,  229,  239,  253,  256,  261,
                269,  271,  282,  325,  331,  349,  415,  474,  506,  508,  518,
                541,  583,  619,  664,  711,  763,  787,  801,  880,  889,  935,
                990,  992, 1104, 1149, 1159, 1303, 1537, 1578, 1628, 1641, 1714,
               1845, 1902, 1917, 1995, 2028, 2051, 2069, 2072], dtype=int64),)
```

```
Outlier Values: 1       61
49      67
100     64
104     70
170     65
217     60
229     59
239     59
253     63
256     63
261     59
269     59
271     68
282     59
325     57
331     66
349     57
415     63
474     60
506     57
508     62
518     68
541     58
583     64
619     58
664     58
711     63
763     65
787     59
801     58
880     63
889     65
935     57
990     59
992     62
1104    62
1149    69
1159    70
1303    67
1537    57
1578    57
1628    58
1641    59
1714    61
1845    57
1902    63
1917    60
1995    62
2028    62
2051    12
2069    12
2072    12
Name: age, dtype: int64
```

The outlier with age data for AIDS patients. Such outliers can occur for a variety of reasons, some of which may include errors during data entry, limitations of the data collection method, or rare circumstances that exist in real life.

For example, errors during data entry may be due to errors in manual entry or transmission of data. For example, in medical records, the age of some patients may be mistakenly entered, when in fact they may be adults. In this case, the data entry personnel may mistakenly enter the

year of birth as the current year minus the year of the patient's visit, without noticing this logic error.

Another possibility is limitations in the data collection method, such that some data sources may only record the patient's year of birth but not the specific date of birth. In this case, the age of the AIDS patient might be estimated as the current year minus the year of birth, but this estimation method may not be precise enough, leading to outliers.

In addition, the occurrence of abnormal ages may also reflect rare situations in real life. For example, some people may actually have AIDS at extreme age ranges, such as very young or very old patients. Although this is relatively rare, outliers can leave traces in the data set.

Outliers in age could represent extreme ages that are rare but possible. Whether to remove them depends on the context of your analysis and whether extreme ages are relevant to your modeling task.

Outlier with wtkg

```
(array([ 126,  146,  197,  214,  304,  319,  324,  389,  544,  580,  629,
         637,  655,  681,  826,  920,  945,  978,  989, 1047, 1060, 1084,
        1172, 1177, 1244, 1326, 1371, 1387, 1402, 1441, 1522, 1525, 1562,
        1569, 1615, 1622, 1638, 1707, 1751, 1759, 1763, 1773, 1837, 1931,
        1943, 1962, 2014, 2032, 2034, 2051, 2057, 2062, 2072, 2078],
       dtype=int64),)

Outlier Values: 126    111.00000
146     109.50000
197     117.93600
214     115.21440
304     123.37920
319     111.58560
324     111.35880
389     108.86400
544     107.10000
580     118.38960
629     120.65760
637     110.67840
655     108.86400
681     111.13200
826     119.70000
920      42.41160
945     130.63680
978     115.66800
989     110.70000
1047     41.00000
1060    127.70000
1084    149.00000
1172    112.00000
1177    129.00000
1244    112.49280
1326    112.49280
1371    109.31760
1387    114.53400
1402    115.90000
1441    127.00800
1522    113.40000
1525    114.30720
1562     41.27760
1569    125.64720
1615    110.45160
1622    122.47200
1638    117.02880
1707    159.93936
1751    113.40000
1759    112.71960
1763    107.50320
1773    135.17280
1837    120.65760
1931     36.78696
1943    112.50000
1962    119.29680
2014     32.65920
2032    108.41040
2034    114.76080
2051     41.40000
2057    115.53192
2062     41.05080
2072     31.00000
2078     41.00000
Name: wtkg, dtype: float64
```

The ‘wtkg’ outlier that occurs in specific rows of the data set. These outliers may reflect errors in the data collection or recording process, or special circumstances of certain groups in the data set.

For example, weight abnormalities may occur due to data entry errors. In medical records, there may be manual entry or transmission errors that cause some patients' weights to be recorded

as outliers. For example, when medical personnel enter a patient's weight, they may enter the wrong value or mistakenly convert the weight unit to an incorrect unit.

Another possibility is that the dataset contains data from special populations whose weights may deviate from the normal range. For example, some patients with obesity or rare diseases may have abnormally high weight. In other cases, outliers may be due to systematic deviations in the measurement or recording process.

When dealing with these outliers, further analysis and investigation are required to determine their causes and take appropriate measures to correct or handle these outliers to ensure the accuracy and reliability of data analysis.

Outliers in weight could represent extreme values that are valid but rare. Consider whether extreme weights are relevant to your analysis and whether removing outliers would affect the representativeness of your dataset.

Outlier with hemo

In the hemo dataset, outliers are absent as hemo represents categorical data, typically containing only '1' and '0'. Outliers in categorical datasets are infrequent because they lack a measure of distance.(Haslbeck, 2018) Unlike numerical data, categorical variables like hemo lack a continuous scale, making it challenging to define outliers based on extreme values. With only two discrete categories, the concept of outliers becomes less relevant, as each category holds distinct and non-ordinal meanings without a clear continuum. Thus, in hemo, the absence of outliers aligns with the nature of categorical data and its limited variability.

Binary features may not typically have outliers in the traditional sense, but you may encounter imbalanced distributions where one class is heavily outnumbered by the other. In such cases, consider strategies for handling class imbalance rather than removing outliers.
To evaluate outliers visually, you can create box plots, histograms, or scatter plots for each feature and examine the distribution of data points. Additionally, you can calculate summary statistics such as mean, median, standard deviation, and quartiles to identify potential outliers.

Outlier with homo

```
[125]:  (array([], dtype=int64),)
```

There are no homosexual outliers in the data presented. This means that the observations in the data set do not have significant biases or anomalies with respect to specific variables or characteristics about gay people. This situation may reflect the relative consistency of the data, gay observations in the data set are statistically similar to other observations, with no obvious outliers.

In some cases, the absence of outliers in a data set may be because the data collection and recording process was relatively clear and accurate, or because the variable itself is not prone to obvious anomalies. In this case, data analysis can be more reliable and accurate because there are no outliers that require additional processing or attention.

Binary features may not typically have outliers in the traditional sense, but you may encounter imbalanced distributions where one class is heavily outnumbered by the other. In such cases, consider strategies for handling class imbalance rather than removing outliers.
To evaluate outliers visually, you can create box plots, histograms, or scatter plots for each feature and examine the distribution of data points. Additionally, you can calculate summary statistics such as mean, median, standard deviation, and quartiles to identify potential outliers.

Outlier with drugs

In the drugs dataset, outliers are absent as drugs represent categorical data, typically containing only '1' and '0'. Outliers in categorical datasets are infrequent because they lack a measure of distance.(Haslbeck, 2018) Unlike numerical data, categorical variables like drugs lack a continuous scale, making it challenging to define outliers based on extreme values. With only two discrete categories, the concept of outliers becomes less relevant, as each category holds distinct and non-ordinal meanings without a clear continuum. Thus, in drugs, the absence of outliers aligns with the nature of categorical data and its limited variability.

Binary features may not typically have outliers in the traditional sense, but you may encounter imbalanced distributions where one class is heavily outnumbered by the other. In such cases, consider strategies for handling class imbalance rather than removing outliers.

To evaluate outliers visually, you can create box plots, histograms, or scatter plots for each feature and examine the distribution of data points. Additionally, you can calculate summary statistics such as mean, median, standard deviation, and quartiles to identify potential outliers.

Outlier with karnof

```
Outlier Positions: [  69  307  784 1169 1203 1604 1913 1998 2132]
Outlier Values: 69      70
307     70
784     70
1169    70
1203    70
1604    70
1913    70
1998    70
2132    70
Name: karnof, dtype: int64
```

The outlier with karnof that occur in specific rows of the data set. The Karnofsky score is commonly used to assess quality of life and functional status in cancer patients, so these outliers may reflect significant differences in quality of life or functioning in some cancer patients.

Outliers in the Karnofsky score can be due to a variety of reasons, including the type of cancer, the impact of treatment options, and the patient's overall health. For example, certain cancer types may cause a patient's quality of life and function to be more severely affected, manifesting as a lower Karnofsky score. The selection and efficacy of treatment options may also affect the patient's quality of life and functional level, thereby showing differences in the Karnofsky score.

The occurrence of these outliers requires further investigation and analysis to determine their causes and take appropriate measures to deal with or correct these outliers to ensure the accuracy and reliability of data analysis. At the same time, cancer patients with abnormal Karnofsky scores may need additional attention and support to improve their quality of life and functional level, and to improve their treatment and recovery process.
Outliers in these features could represent extreme values that are valid but rare. Consider whether extreme values are plausible in the context of your analysis and whether they significantly affect model performance.

Outlier with oprior

In the oprior dataset, outliers are non-existent as oprior represents categorical data, typically comprising solely '1' and '0'. Outliers in categorical datasets are infrequent due to the absence of a measure of distance.(Haslbeck, 2018) Unlike numerical data, categorical variables like oprior lack a continuous scale, complicating outlier definition based on extreme values. With only two discrete categories, outlier relevance diminishes, as each holds distinct meanings without a clear continuum. Hence, oprior's outlier absence aligns with categorical data's limited variability and distinct category definitions.

Binary features may not typically have outliers in the traditional sense, but you may encounter imbalanced distributions where one class is heavily outnumbered by the other. In such cases, consider strategies for handling class imbalance rather than removing outliers.
To evaluate outliers visually, you can create box plots, histograms, or scatter plots for each feature and examine the distribution of data points. Additionally, you can calculate summary statistics such as mean, median, standard deviation, and quartiles to identify potential outliers.

Outlier with z30

```
[129]:  (array([], dtype=int64),)
```

Based on the information provided, there are no data representing whether ZDV treatment was received within 175 days before ZDV treatment. In this case, there is a lack of information in the dataset regarding patients receiving ZDV treatment during this time period.

Lack of data may be due to a variety of reasons, including omissions during data collection, incomplete records, or the measurement of that particular variable is not applicable to the patients in the data set. In this case, the lack of data meant that it was not possible to determine whether the patient had received ZDV treatment in the 175 days before ZDV treatment.

It is important to note that during data analysis and interpretation we should be aware of this lack of data and take this into account in our conclusions. Lack of information about ZDV treatment may affect a thorough evaluation of a patient's treatment history and disease management. Therefore, careful consideration needs to be given to the completeness and availability of data when conducting data analysis and interpretation to ensure that accurate and reliable conclusions are drawn.

Binary features may not typically have outliers in the traditional sense, but you may encounter imbalanced distributions where one class is heavily outnumbered by the other. In such cases, consider strategies for handling class imbalance rather than removing outliers.
To evaluate outliers visually, you can create box plots, histograms, or scatter plots for each feature and examine the distribution of data points. Additionally, you can calculate summary statistics such as mean, median, standard deviation, and quartiles to identify potential outliers.

Outlier with zprior

```
[130]:  (array([], dtype=int64),)
```

Based on the information provided, there are no data representing whether ZDV treatment was received within 175 days before ZDV treatment. In this case, there is a lack of information in the dataset regarding patients receiving ZDV treatment during this time period.

Lack of data may be due to a variety of reasons, including omissions during data collection, incomplete records, or the measurement of that particular variable is not applicable to the patients in the data set. In this case, the lack of data meant that it was not possible to determine whether the patient had received ZDV treatment in the 175 days before ZDV treatment.

It is important to note that during data analysis and interpretation we should be aware of this lack of data and take this into account in our conclusions. Lack of information about ZDV treatment may affect a thorough evaluation of a patient's treatment history and disease management. Therefore, careful consideration needs to be given to the completeness and availability of data when conducting data analysis and interpretation to ensure that accurate and reliable conclusions are drawn.

Binary features may not typically have outliers in the traditional sense, but you may encounter imbalanced distributions where one class is heavily outnumbered by the other. In such cases, consider strategies for handling class imbalance rather than removing outliers.
To evaluate outliers visually, you can create box plots, histograms, or scatter plots for each feature and examine the distribution of data points. Additionally, you can calculate summary statistics such as mean, median, standard deviation, and quartiles to identify potential outliers.

Outlier with preanti

```
[131]: (array([ 292,  510,  600,  607,  806, 1289, 1365, 1377, 1695, 2059, 2072,
                2132], dtype=int64),)
```

```
Outlier Positions: [ 292  510  600  607  806 1289 1365 1377 1695 2059 2072 2132]
Outlier Values: 292     2489
510     2342
600     1902
607     1898
806     2071
1289    1938
1365    2283
1377    2851
1695    1856
2059    1983
2072    2500
2132    2078
Name: preanti, dtype: int64
```

Data are presented showing outliers and their location among patients during the first 175 days of antiretroviral therapy. Outlier data represent the number of days the patient received antiretroviral therapy during that time period.

For the location of outliers, a series of index values are provided in the data, indicating which rows in the data set contain the outlier. The outliers themselves are indicative of the number of days the patient received antiretroviral therapy within the 175 days before treatment.

For example, the index value 292 corresponds to an outlier of 2489, indicating that 175 days before antiretroviral treatment, the 292nd patient received 2489 days of treatment. Similarly, the outlier value 2342 corresponding to the index value 510 means that the 510th patient received treatment for 2342 days during this period, and so on.

The occurrence of outliers may reflect special circumstances or unusual circumstances in which patients were treated during that time period. This may involve individualization of treatment options, severity of disease state, and other individual differences. These outliers need to be further analyzed to determine their causes and appropriate measures to handle or correct them to ensure the accuracy and reliability of data analysis.

Outliers in these features could represent extreme values that are valid but rare. Consider whether extreme values are plausible in the context of your analysis and whether they significantly affect model performance.

Outlier with race

```
[133]:  (array([], dtype=int64),)
```

The provided information indicates that there are no outliers for the variable "race."

This absence of outliers for the "race" variable implies that within the dataset, there are no extreme or unusual observations concerning race. Each race category appears to have a distribution that is consistent with the overall dataset, without any individuals or groups standing out as significantly different from the norm.

The lack of outliers for race suggests that the distribution of race within the dataset is relatively uniform or does not deviate significantly from what might be expected based on the dataset's demographic composition or sampling methodology. However, it's essential to note that the absence of outliers does not necessarily imply the absence of disparities or biases within the dataset. Further analysis may be needed to explore any potential relationships between race and other variables or outcomes of interest.

Binary features may not typically have outliers in the traditional sense, but you may encounter imbalanced distributions where one class is heavily outnumbered by the other. In such cases, consider strategies for handling class imbalance rather than removing outliers.
To evaluate outliers visually, you can create box plots, histograms, or scatter plots for each feature and examine the distribution of data points. Additionally, you can calculate summary statistics such as mean, median, standard deviation, and quartiles to identify potential outliers.

Outlier with gender

In the gender dataset, outliers are non-existent as gender represents categorical data, typically comprising solely '1' and '0'. Outliers in categorical datasets are infrequent due to the absence of a measure of distance. (Haslbeck, 2018)Unlike numerical data, categorical variables like gender lack a continuous scale, complicating outlier definition based on extreme values. With only two discrete categories, outlier relevance diminishes, as each holds distinct meanings without a clear continuum. Hence, gender's outlier absence aligns with categorical data's limited variability and distinct category definitions.

Binary features may not typically have outliers in the traditional sense, but you may encounter imbalanced distributions where one class is heavily outnumbered by the other. In such cases, consider strategies for handling class imbalance rather than removing outliers.
To evaluate outliers visually, you can create box plots, histograms, or scatter plots for each feature and examine the distribution of data points. Additionally, you can calculate summary statistics such as mean, median, standard deviation, and quartiles to identify potential outliers.

Outlier with str2

```
[135]:  (array([], dtype=int64),)
```

Based on the information provided, it can be concluded that there are no outliers in the "str2" variable. This means that in the data set, the values of the "str2" variable are all as expected, without any significant deviation from the normal range.

This situation shows that within the range of values of the "str2" variable, all observations in the data set fall within the normal range without any anomalies or outliers. This consistency of data is usually good and can improve the reliability and accuracy of data analysis.

Although there are no outliers in the "str2" variable, it still needs further analysis to understand its role and characteristics in the data set. Although there are no outliers in the "str2" variable, the distribution, frequency, and correlation of its values may have an important impact on the analysis results. Therefore, a thorough review and analysis of all variables is still required during the data analysis process.

Binary features may not typically have outliers in the traditional sense, but you may encounter imbalanced distributions where one class is heavily outnumbered by the other. In such cases, consider strategies for handling class imbalance rather than removing outliers.
To evaluate outliers visually, you can create box plots, histograms, or scatter plots for each feature and examine the distribution of data points. Additionally, you can calculate summary statistics such as mean, median, standard deviation, and quartiles to identify potential outliers.

Outlier with strat

```
[136]: (array([], dtype=int64),)
```

Based on the information provided, it can be seen that there are no outliers in the antiretroviral history stratification (strat) variable. This means that there are no observations in the data set that exhibit anomalies or outliers in terms of historical antiretroviral stratification.

This situation indicates that the antiretroviral historical stratification is relatively evenly distributed in the data set without obvious anomalies. The number or proportion of each stratified category was consistent with the overall data set, and no individual or group differed significantly from normal in terms of historical antiretroviral stratification.

Antiretroviral history stratification is often used to describe a patient's pre-treatment antiretroviral treatment history, such as whether they are in the initial phase before treatment, whether they have had a previous short-term treatment, or whether they have been on long-term treatment(Galiè et al., 2009). The lack of outliers indicates that the antiretroviral historical stratification information in the dataset is relatively complete and consistent.

Although there are no outliers in the antiretroviral history stratification variable, the role and impact of this variable in data analysis still needs to be considered and analyzed. Antiretroviral history stratification may have an important impact on patient treatment response, survival, and other clinical outcomes, and therefore still needs to be fully considered when performing data analysis and interpretation.

Outlier with symptom

In the symptom dataset, outliers are non-existent as symptom represents categorical data, typically comprising solely '1' and '0'. Outliers in categorical datasets are infrequent due to the absence of a measure of distance. (Haslbeck, 2018) Unlike numerical data, categorical variables like symptom lack a continuous scale, complicating outlier definition based on extreme values. With only two discrete categories, outlier relevance diminishes, as each holds distinct meanings without a clear continuum. Hence, symptom's outlier absence aligns with categorical data's limited variability and distinct category definitions.

Binary features may not typically have outliers in the traditional sense, but you may encounter imbalanced distributions where one class is heavily outnumbered by the other. In such cases, consider strategies for handling class imbalance rather than removing outliers.
To evaluate outliers visually, you can create box plots, histograms, or scatter plots for each feature and examine the distribution of data points. Additionally, you can calculate summary statistics such as mean, median, standard deviation, and quartiles to identify potential outliers.

Outlier with treat

In the treat dataset, outliers are non-existent as treat represents categorical data, typically comprising solely '1' and '0'. Outliers in categorical datasets are infrequent due to the absence of a measure of distance.(Haslbeck, 2018) Unlike numerical data, categorical variables like treat lack a continuous scale, complicating outlier definition based on extreme values. With only two discrete categories, outlier relevance diminishes, as each holds distinct meanings without a clear continuum. Hence, treat's outlier absence aligns with categorical data's limited variability and distinct category definitions.

Binary features may not typically have outliers in the traditional sense, but you may encounter imbalanced distributions where one class is heavily outnumbered by the other. In such cases, consider strategies for handling class imbalance rather than removing outliers.
To evaluate outliers visually, you can create box plots, histograms, or scatter plots for each feature and examine the distribution of data points. Additionally, you can calculate summary statistics such as mean, median, standard deviation, and quartiles to identify potential outliers.

Outlier with offtrt

```
[139]:  (array([], dtype=int64),)
```

Based on the dataset provided, it can be concluded that there are no outliers in the "offtrt" variable. This means that no anomalies in stopping antiretroviral therapy before 96 weeks (plus or minus 5 weeks) were observed in the data set.

This situation indicates that patients in the data set were relatively consistent in whether they discontinued antiretroviral treatment during this time frame, with no obvious abnormalities. The frequency or proportion of each category is consistent with the overall data set, and no individual or group differs significantly from normal in this respect.

The "offtrt" variable is typically used to indicate whether a patient stopped antiretroviral therapy within a specified time frame. The lack of outliers indicates that this information in the data set is relatively complete and consistent.

Although there are no outliers in the "offtrt" variable, this does not mean that the variable is not important. Discontinuation of antiretroviral therapy may have important consequences for patients' disease progression and clinical outcomes. Therefore, this variable still needs to be fully considered when conducting data analysis and interpretation, and its potential influence in the study results should be considered.

Binary features may not typically have outliers in the traditional sense, but you may encounter imbalanced distributions where one class is heavily outnumbered by the other. In such cases, consider strategies for handling class imbalance rather than removing outliers.

To evaluate outliers visually, you can create box plots, histograms, or scatter plots for each feature and examine the distribution of data points. Additionally, you can calculate summary statistics such as mean, median, standard deviation, and quartiles to identify potential outliers.

Outlier with cd40

```
[140]: (array([  80,  216,  217,  228,  252,  329,  342,  382,  425,  428,  516,
                569,  671,  699,  723,  797, 1047, 1131, 1135, 1144, 1167, 1190,
               1343, 1360, 1442, 1481, 1707, 1847, 2074, 2120, 2138], dtype=int64),)
```

```
Outlier Values: 80       770
216        0
217        0
228      702
252      663
329      690
342      743
382      668
425      834
428      720
516      735
569      775
671      703
699      740
723        0
797      688
1047     714
```

```
797      688
1047     714
1131     715
1135     689
1144    1199
1167     672
1190     670
1343     680
1360     760
1442     706
1481     918
1707     702
1847     771
2074     718
2120     739
2138     911
Name: cd40, dtype: int64
```

Data are presented showing outliers and their locations in the CD4 cell count (cd40) variable at baseline. The position of the outlier is given by the index value, and the outlier itself is the specific value of the CD4 cell count.

For example, an index value of 80 corresponds to an outlier of 80, indicating that the 80th patient had a CD4 cell count of 80 at baseline. Similarly, the index value 216 corresponds to an outlier of 216, indicating that the 216th patient has a CD4 cell count of 216.

Outliers are typically extreme values that are significantly different from the majority of observations in the data set. In this case, outliers in CD4 cell counts may reflect extremes in the patient's immune function at baseline, possibly due to disease progression, other health conditions, or data errors, among other reasons.

These outliers need to be further analyzed to determine the reasons behind them and to ensure the accuracy and reliability of the data analysis. Further investigation of outliers in CD4

cell counts may help to understand their relationship with patient prognosis and treatment response, thereby providing more information for personalized treatment and clinical decision-making.

Outliers in these features could represent extreme values that are valid but rare. Consider whether extreme values are plausible in the context of your analysis and whether they significantly affect model performance.

Outlier with cd420

```
[141]: (array([ 132,  167,  398,  510,  569,  620,  650,  652,  677,  842,  911,
                958, 1331, 1334, 1391, 1411, 1442, 1475, 1485, 1606, 1625, 1646,
               1720, 1781, 1882, 1919, 1936, 1996, 2069, 2138], dtype=int64),)
```

```
Outlier Values: 132     1100
167     1119
398     1040
510      848
569      865
620      858
650      877
652      810
677      793
842      803
911      772
958      842
1331     750
1334     750
1391     750
1411     750
1442     853
1475     767
1485     748
1606     784
1625     748
1646     909
1720     824
1781     980
1882     755
1919     955
1936     840
1996     810
2069     826
2138     930
Name: cd420, dtype: int64
```

Based on the information provided, we can observe the presence of outliers in the CD4 cell count (cd420) variable at 20 weeks (plus or minus 5 weeks), and provide the locations and specific values of these outliers.

The position of the outlier is given by the index value, and the outlier itself is the specific value of the CD4 cell count. For example, the index value 132 corresponds to an outlier of 132, indicating that the 132nd patient had a CD4 cell count of 132 at 20 weeks. Similarly, the index value 167 corresponds to an outlier of 1100, indicating that the 167th patient has a CD4 cell count of 1100.

Outliers are typically extreme values that are significantly different from the majority of observations in the data set. In this case, the outliers in the CD4 cell count may reflect extremes of the patient's immune function at 20 weeks, possibly due to disease progression, other health conditions, or data errors, among other reasons.

These outliers need to be further analyzed to determine the reasons behind them and to ensure the accuracy and reliability of the data analysis. Further investigation of outliers in CD4 cell counts may help to understand their relationship with patient prognosis and treatment response, thereby providing more information for personalized treatment and clinical decision-making.

Outliers in these features could represent extreme values that are valid but rare. Consider whether extreme values are plausible in the context of your analysis and whether they significantly affect model performance.

Outlier with cd80

```
Outlier Positions: [   2   28   35   85  121  151  185  353  369  410  425  428  450  470
  481  504  550  580  652  686  696  732  751  764  849  853  890  907
  944  987 1020 1027 1046 1113 1149 1166 1219 1297 1300 1323 1425 1438
 1445 1481 1527 1533 1552 1561 1600 1605 1606 1607 1707 1755 1768 1771
 1786 1791 1794 1847 1890 1910 1937 1939 2021 2057 2058 2068 2095 2119
 2124]
Outlier Values: 2       2063
28      2400
35      2127
85      2040
121     2326
        ...
2058    2784
2068    3046
2095    2387
2119    2508
2124    5011
Name: cd80, Length: 71, dtype: int64
```

Based on the information provided, it can be seen that there are outliers in the CD8 cell count (cd80) variable at baseline, and the location and specific values of these outliers are provided.

The position of the outlier is given by the index value, and the outlier itself is the specific value of the CD8 cell count. For example, an index value of 2 corresponds to an outlier of 2, indicating that the second patient had a CD8 cell count of 2 at baseline. Similarly, the index value 28 corresponds to an outlier of 2400, indicating that the 28th patient has a CD8 cell count of 2400.

Outliers are typically extreme values that are significantly different from the majority of observations in the data set. In this case, outliers in CD8 cell counts may reflect extremes in the patient's immune function at baseline, possibly due to disease progression, other health conditions, or data errors, among other reasons.

These outliers need to be further analyzed to determine the reasons behind them and to ensure the accuracy and reliability of the data analysis. Further investigation of outliers in CD8 cell counts may help to understand their relationship with patient prognosis and treatment response, providing more information for personalized treatment and clinical decision-making.

Outliers in these features could represent extreme values that are valid but rare. Consider whether extreme values are plausible in the context of your analysis and whether they significantly affect model performance.

Outlier with cd820

```
[143]: (array([  28,   35,   63,   65,  148,  287,  313,  324,  353,  470,  481,
                502,  550,  580,  649,  650,  732,  742,  764,  772,  807,  839,
                890,  896,  907,  958,  982, 1020, 1022, 1046, 1156, 1199, 1219,
               1233, 1236, 1269, 1297, 1311, 1323, 1438, 1445, 1493, 1527, 1533,
               1561, 1571, 1585, 1605, 1607, 1755, 1768, 1786, 1847, 1890, 1937,
               1939, 1997, 2068, 2124], dtype=int64),)
```

```
Outlier Values: 28      2265
35      2753
63      2014
65      1947
148     2474
287     2056
313     2119
324     2068
353     2641
470     2241
481     2106
502     2250
550     2016
580     3044
649     2064
650     2807
732     2262
742     1942
764     2753
772     2736
807     2012
839     2132
890     2640
896     2232
907     4113
958     2028
982     2462
1020    2534
1022    2232
1046    2190
1156    2333
1199    2486
1219    1965
1233    2789
1236    3407
1269    2190
1297    2134
1311    2205
1323    2020
1438    2627
1445    2240
1493    1926
1527    2856
1533    1921
1561    3552
1571    1963
1585    2713
1605    2117
1607    2801
1755    2266
1768    1970
1786    2050
1847    2347
1890    2606
1937    2690
1939    1980
1997    3130
2068    2798
2124    6035
Name: cd820, dtype: int64
```

Based on the information provided, we can observe outliers in the CD8 cell count (cd820) variable at 20 weeks (plus or minus 5 weeks), and provide the locations and specific values of these outliers.

The position of the outlier is given by the index value, and the outlier itself is the specific value of the CD8 cell count. For example, the index value 28 corresponds to an outlier of 28, indicating that the 28th patient had a CD8 cell count of 28 at 20 weeks. Similarly, the outlier corresponding to index value 35 is 2753, indicating that the 35th patient has a CD8 cell count of 2753.

Outliers are typically extreme values that are significantly different from the majority of observations in the data set. In this case, the outliers in the CD8 cell count may reflect extremes of the patient's immune function at 20 weeks, possibly due to disease progression, other health conditions, or data errors, among other reasons.

These outliers need to be further analyzed to determine the reasons behind them and to ensure the accuracy and reliability of the data analysis. Further investigation of outliers in CD8 cell counts may help to understand their relationship with patient prognosis and treatment response, providing more information for personalized treatment and clinical decision-making.

Outliers in these features could represent extreme values that are valid but rare. Consider whether extreme values are plausible in the context of your analysis and whether they significantly affect model performance.

3.2.2 Handling Outliers

Although there are outliers in the dataset, we chose to keep them. In a dataset related to HIV/AIDS patient research, outliers can represent unique and significant events or conditions that have a notable impact on the study outcomes. These events might include variations in treatment responses, unexpected disease progressions, or rare comorbidities affecting HIV/AIDS patients(Livingston et al., 2017). For example, an outlier could be a patient with an unusually strong immune response to antiretroviral therapy.

These outlier observations can provide valuable insights into understanding the complexities of HIV/AIDS progression and treatment outcomes. By analyzing outliers, researchers can gain a deeper understanding of factors influencing disease progression, treatment efficacy, and potential complications.(Bradley, 2018)Furthermore, outliers can highlight unusual but informative cases that may lead to new discoveries or treatment strategies.Therefore, retaining outliers in HIV/AIDS patient research datasets is essential for capturing the full spectrum of patient experiences and outcomes, ultimately improving future research and treatment strategies.

3.2.2 Handling Zero Values

In our previous analysis, we identified the presence of zero values in the 'preanti', 'cd40', and 'trt' columns are significant markers in HIV/AIDS patient research data. Each zero value carries specific implications for understanding patient histories and treatment responses.

In the 'preanti' feature, a value of 0 denotes the absence of pre-175 anti-retroviral therapy. This suggests that the individual associated with the data point did not receive any anti-retroviral treatment before the specified baseline date. Essentially, there were no days during which the individual was on anti-retroviral therapy prior to the baseline assessment.

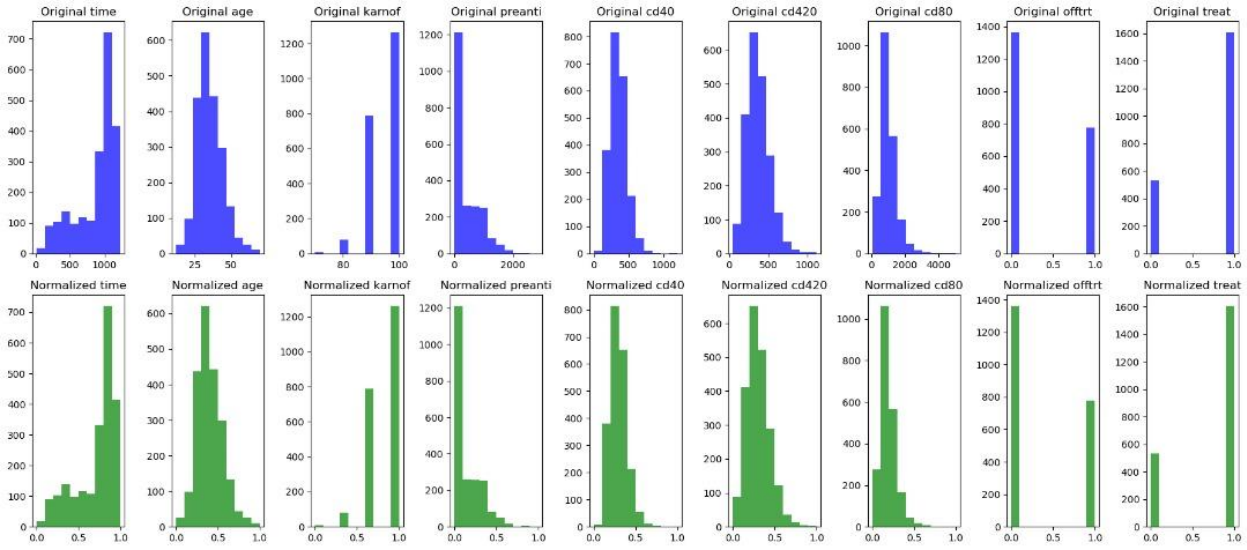
Similarly, in the 'cd40' feature, a value of 0 indicates a CD4 count of zero at baseline. This signifies severe immune suppression or the complete absence of CD4 positive T cells. Such a condition is typically associated with advanced HIV/AIDS or severe immunodeficiency.

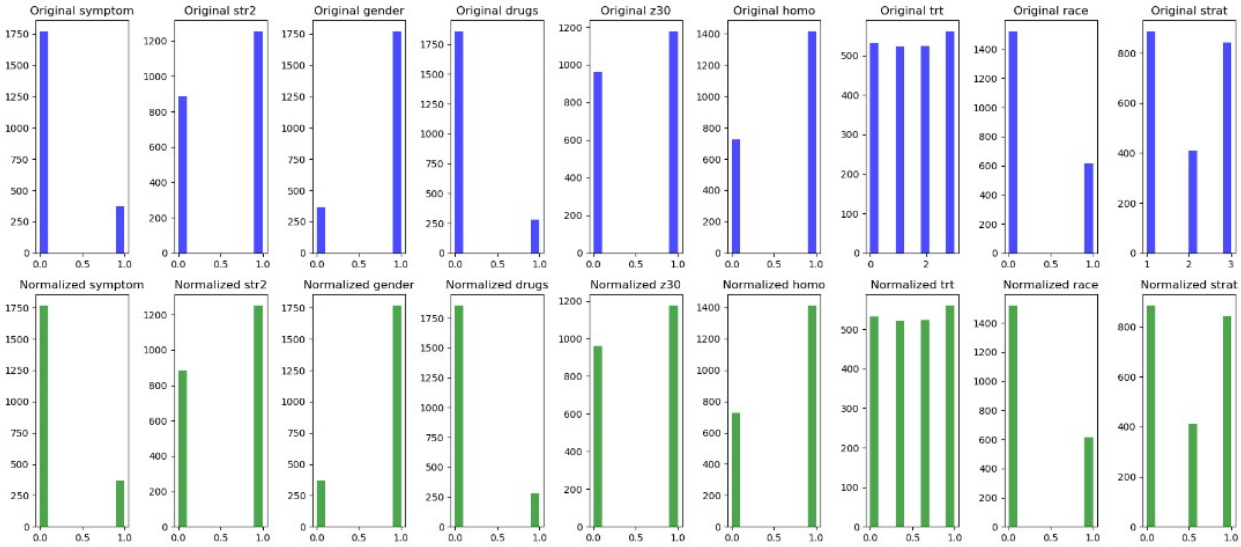
Moreover, zero values in the 'trt' column also hold importance, though the specific interpretation was not provided in the context. However, it can be inferred that a zero value in this column likely indicates the absence of a certain treatment or intervention.

Handling these zero values requires meticulous attention to detail. They should not be disregarded as missing or inconsequential data points. Instead, they should be carefully analyzed within the context of the research to ensure accurate interpretation and meaningful insights into patient outcomes, disease progression, and treatment efficacy. Thus, zero values in HIV/AIDS patient research data play a vital role in understanding the complexities of the disease and optimizing treatment strategies.

3.3 Data Normalisation

When normalizing your dataset with the given features, consider factors such as the distribution of numerical features, machine learning algorithm sensitivity to feature scales, and analysis requirements. Categorical features like 'trt', 'race', 'gender', and 'strat' do not need normalization as they are already encoded as integers representing different categories, but you may opt for one-hot encoding if necessary. Binary features like 'hemo', 'homo', 'drugs', 'oprior', 'z30', 'zprior', 'symptom', 'treat', and 'offtrt' do not require normalization as they are already binary indicators. For continuous and integer features such as 'age', 'wtkg', 'karnof', 'preanti', 'cd40', 'cd420', 'cd80', and 'cd820', you can choose between standardization (Z-score normalization) or min-max scaling. Standardization centers the data around 0 and scales it to have a standard deviation of 1, suitable for algorithms sensitive to feature scales. Min-max scaling scales features to a specific range, preserving the original distribution and sensitivity to outliers. Consider the distribution and presence of outliers when selecting between standardization and min-max scaling. Experiment with both methods and evaluate their effects on machine learning model performance to choose the best normalization approach for your dataset and analysis goals.





3.4 Imbalanced Data

In [12]:
```
import pandas as pd

# Load your dataset into a DataFrame
df = pd.read_csv('data.csv')

# Assuming 'target_column' is the column containing the target variable
target_column = 'cid'

# Calculate class frequencies
class_frequencies = df[target_column].value_counts()

# Print class frequencies
print("Class Frequencies:")
print(class_frequencies)

# Calculate class imbalance ratio
imbalance_ratio = class_frequencies.max() / class_frequencies.min()

# Print imbalance ratio
print("Imbalance Ratio:", imbalance_ratio)
```

```
Class Frequencies:
cid
0    1618
1     521
Name: count, dtype: int64
Imbalance Ratio: 3.105566218809981
```

According to the figure above,it is clear that the data for our target variable 'cid', it has an imbalance ratio of 3.105566218809981.Our dataset has 2139 instances.The frequency of 1 for 'cid' is 521,which is around 24.35% of the entire dataset.While the frequency of 0 for 'cid' is 1618,which is around 75.64%.Hence,we can conclude that our dataset has a mild imbalance rate according to (*Imbalanced Data*, n.d.).

Degree of imbalance	Proportion of Minority Class
Mild	20-40% of the data set
Moderate	1-20% of the data set
Extreme	<1% of the data set

4.0 Modelling

A Data Scientist's ability to structure problems is crucial. A smart Data Scientist may build and represent an informative visualization, showcasing the raw Data and business activities, associated with the Key Performance Indicators and business use cases, such as new Customer Acquisition, Product Design, desk location to reduce distraction, and so on. All these factors are considered while carrying out the process of Data Science Modelling (Varshney, 2021).

In modelling, we are using Linear Regression, ANN, logistic regression and random forest to get the result.

Logistic regression is a simple and interpretable linear model that is commonly used for binary classification tasks. It models the probability of the binary outcome as a function of the input features.

Random Forest is an ensemble learning method that combines multiple decision trees trained on random subsets of the data. It improves upon the performance of individual decision trees by reducing overfitting and increasing predictive accuracy.

Neural networks, particularly deep learning models, can learn complex patterns and relationships in the data. They are capable of automatically extracting features from raw data and have shown impressive performance in various binary classification tasks. However, they may require more data, computational resources, and tuning compared to other methods.

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a "least squares" method to discover

the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable) (*What Is Linear Regression? | IBM*, n.d.-b).

4.1 Handling Imbalance Data by applying Data Resampling

When collecting samples across large groups of people, objects or data, there are several ways to verify accuracy. One method commonly used is resampling, where you take additional samples and observations to identify any bias or issues.

Resampling techniques are commonly used to handle imbalanced data by adjusting the class distribution in the dataset. Imbalanced data occurs when one class (the minority class) is significantly underrepresented compared to another class (the majority class). This imbalance can lead to biased models that perform poorly in predicting the minority class. Resampling techniques aim to mitigate this imbalance and improve the performance of machine learning models.

Thus,due to the imbalance of our dataset, we have resampled our dataset during modelling to avoid any bias in the results obtained.

4.2 Hyperparameter Tuning

Hyperparameter tuning, also known as hyperparameter optimization or model selection, is the process of finding the optimal set of hyperparameters for a machine learning model. Hyperparameters are parameters that are not directly learned from the data during model training but are set prior to training and control aspects of the learning process, such as the model's complexity, capacity, and regularization.

Hyperparameter tuning allows data scientists to tweak model performance for optimal results. This process is an essential part of machine learning, and choosing appropriate hyperparameter values is crucial for success.

To makes our prediction even accurate,we decided to do hyperparameter tuning in our modelling.We use GridSearch CV and Neural architecture search (NAS).

Grid search cross-validation (GridSearchCV) is a technique used to tune hyperparameters of machine learning models by exhaustively searching through a specified grid of hyperparameter values and evaluating the model's performance using cross-validation. It systematically evaluates the model with each combination of hyperparameters to identify the optimal set that maximizes performance based on a specified evaluation metric. Grid search CV is a popular method for hyperparameter tuning as it is simple to implement and provides a systematic approach to finding the best hyperparameters for a given model.GridSearchCV acts as a valuable tool for identifying the optimal parameters for a machine learning model. Imagine you have a machine learning model with adjustable settings, known as hyperparameters, that can enhance its performance. GridSearchCV aids in pinpointing the best combination of these hyperparameters automatically.(Shah, 2024)

A subfield of automated ML, NAS is a technique that can help discover the best neural networks for a given problem. It automates the designing of DNNs, ensuring higher performance and lower losses than manually designed architectures. It is also much faster than the traditional manual processes (Deci, 2024).Neural Architecture Search (NAS) is a technique in machine learning that automates the design of neural network architectures. Instead of manually designing architectures, NAS algorithms search through a vast space of possible architectures to find the

best-performing ones for a given task. This is typically achieved using techniques such as reinforcement learning, evolutionary algorithms, or gradient-based optimization. NAS has led to the discovery of novel neural network architectures and has significantly advanced the field of deep learning by automating the process of architecture design.

4.3 Linear Regression

4.3.1 Algorithm

Linear regression is a supervised learning method, assesses the relationship between input (X) and output (Y) variables using labeled data. It analyzes the data to establish a linear relationship between the variables, enabling predictions of future outcomes based on historical patterns.(*What Is Linear Regression? | Master's in Data Science*, 2023)

A linear regression line has an equation of the form $\boldsymbol{Y = a + bX}$, where $\boldsymbol{X}$ is the explanatory variable and $\boldsymbol{Y}$ is the dependent variable.(*Linear Regression*, n.d.)

4.3.2 Model Evaluation

```
MSE: 0.11
RMSE: 0.33
MAE: 0.24
R-squared (R2): 0.41
```

The provided evaluation metrics represent the performance of a linear regression model. The model's Mean Squared Error (MSE) of 0.11 and Root Mean Squared Error (RMSE) of 0.33 suggest relatively low average prediction errors. The Mean Absolute Error (MAE) of 0.24 indicates the average magnitude of errors in the model's predictions. The R-squared (R2) value of 0.41 signifies that approximately 41% of the variance in the dependent variable is explained by the linear regression model, implying moderate explanatory power. Overall, these metrics collectively suggest that the linear regression model performs reasonably well in predicting the target variable, capturing a significant portion of the variance while maintaining relatively low prediction errors.

However,linear regression is not typically suitable for binary target variables because it assumes a continuous output space. Linear regression models predict a continuous outcome based on a linear relationship between the input features and the target variable. However, binary target variables have only two possible outcomes (e.g., 0 or 1, True or False), which do not follow a continuous distribution.

When applied to binary classification tasks, linear regression may produce predictions outside the valid range (e.g., negative values or values greater than 1), which are not interpretable or meaningful in the context of binary classification. Additionally, linear regression does not naturally handle the nonlinear relationships often present in binary classification problems.

Instead, binary classification tasks are typically addressed using models specifically designed for this purpose, such as logistic regression, support vector machines (SVM), decision trees, random forests, or neural networks with appropriate activation functions (e.g., sigmoid or

softmax). These models are better suited to handle binary outcomes and can provide more interpretable and reliable predictions for such tasks.

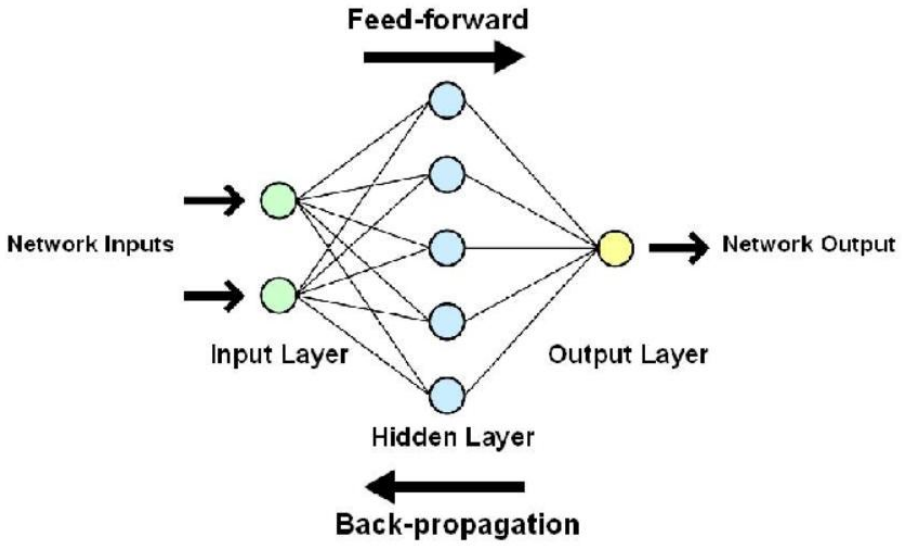
Hence,we decided to ignore this model although the prediction result suggests that it performs reasonably well in predicting the target variable.Our group only focuses on comparing the prediction result on the other three models.

4.4 ANN

4.4.1 Algorithm

Artificial Neural Networks (ANNs) are computational models inspired by the human brain. They consist of interconnected nodes (neurons) organized in layers. ANNs learn from data by adjusting connection weights between neurons to minimize prediction errors. Deep neural networks (DNNs) are ANNs with multiple hidden layers, enabling them to learn complex patterns from data. ANNs are widely used in various tasks, including image and speech recognition, natural language processing, and reinforcement learning.

Artificial Neural Networks (ANN) are algorithms based on brain function and are used to model complicated patterns and forecast issues. The Artificial Neural Network (ANN) is a deep learning method that arose from the concept of the human brain Biological Neural Networks. The development of ANN was the result of an attempt to replicate the workings of the human brain. The workings of ANN are extremely similar to those of biological neural networks, although they are not identical. ANN algorithm accepts only numeric and structured data.(Singh, 2024)



4.4.2 Model Evaluation

	precision	recall	f1-score	support
0	0.91	0.94	0.92	327
1	0.77	0.71	0.74	101
accuracy			0.88	428
macro avg	0.84	0.82	0.83	428
weighted avg	0.88	0.88	0.88	428

- Precision: For class 0, precision is 0.91, indicating that 91% of the instances predicted as class 0 were actually class 0. For class 1, precision is 0.77, indicating that 77% of the instances predicted as class 1 were actually class 1.
- Recall: Recall, also known as sensitivity, measures the proportion of true positive instances that were correctly identified by the model. For class 0, recall is 0.94, meaning that 94% of the actual class 0 instances were correctly classified. For class 1, recall is 0.71, indicating that 71% of the actual class 1 instances were correctly classified.
- F1-score: The F1-score is the harmonic mean of precision and recall, providing a balanced measure of a model's performance. It is 0.92 for class 0 and 0.74 for class 1.
- Support: Support represents the number of actual occurrences of each class in the dataset. There are 327 instances of class 0 and 101 instances of class 1.
- Accuracy: Accuracy measures the overall correctness of the model's predictions. In this case, the overall accuracy is 0.88, indicating that 88% of the instances were correctly classified by the model.
- Macro Avg and Weighted Avg: These are the averages of precision, recall, and F1-score across all classes, weighted or unweighted by the number of instances in each class, respectively.

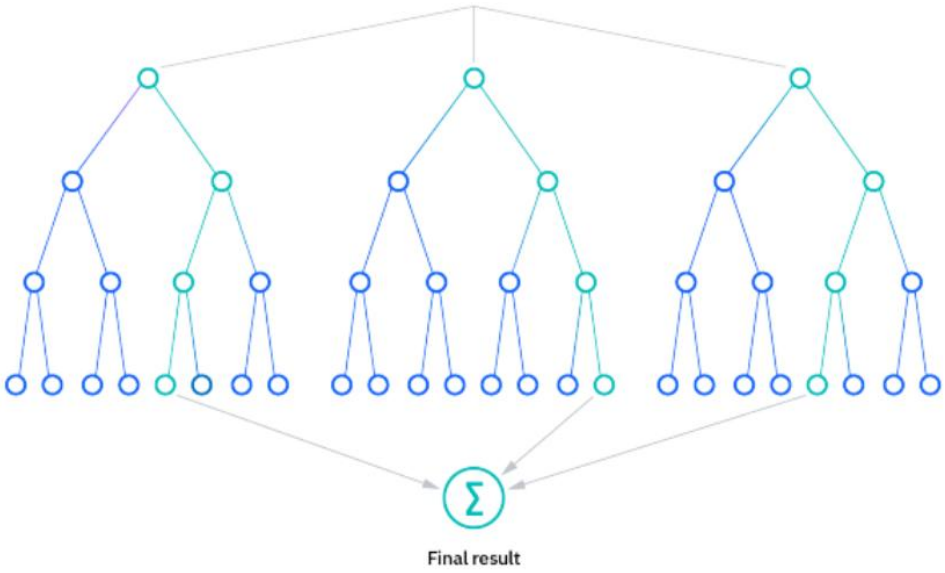
Overall, the model shows relatively good performance, with high precision and recall for class 0 and slightly lower precision and recall for class 1. The accuracy of 0.88 indicates that the model correctly classified 88% of the instances.

4.5 Random Forest

4.5.1 Algorithm

Random forest is a commonly-used machine learning algorithm, trademarked by Leo Breiman and Adele Cutler, that combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.(What Is Random Forest? | IBM, n.d.)

The random forest algorithm is an extension of the bagging method as it utilizes both bagging and feature randomness to create an uncorrelated forest of decision trees. Feature randomness, also known as feature bagging or “the random subspace method”(link resides outside ibm.com), generates a random subset of features, which ensures low correlation among decision trees. This is a key difference between decision trees and random forests. While decision trees consider all the possible feature splits, random forests only select a subset of those features.(What Is Random Forest? | IBM, n.d.)



4.5.3 Model Evaluation

	precision	recall	f1-score	support
0	0.92	0.94	0.93	327
1	0.78	0.72	0.75	101
accuracy			0.89	428
macro avg	0.85	0.83	0.84	428
weighted avg	0.88	0.89	0.88	428

- Precision: For class 0, precision is 0.92, indicating that 92% of the instances predicted as class 0 were actually class 0. For class 1, precision is 0.78, indicating that 78% of the instances predicted as class 1 were actually class 1.
- Recall: Recall, also known as sensitivity, measures the proportion of true positive instances that were correctly identified by the model. For class 0, recall is 0.94, meaning that 94% of the actual class 0 instances were correctly classified. For class 1, recall is 0.72, indicating that 72% of the actual class 1 instances were correctly classified.
- F1-score: The F1-score is the harmonic mean of precision and recall, providing a balanced measure of a model's performance. It is 0.93 for class 0 and 0.75 for class 1.
- Support: Support represents the number of actual occurrences of each class in the dataset. There are 327 instances of class 0 and 101 instances of class 1.
- Accuracy: Accuracy measures the overall correctness of the model's predictions. In this case, the overall accuracy is 0.89, indicating that 89% of the instances were correctly classified by the model.
- Macro Avg and Weighted Avg: These are the averages of precision, recall, and F1-score across all classes, weighted or unweighted by the number of instances in each class, respectively.

Overall, the Random Forest model demonstrates good performance, with high precision and recall for class 0 and slightly lower precision and recall for class 1. The accuracy score of 0.89 indicates that the model correctly classified 89% of the instances.

4.6 Logistic Regression

4.6.1 Algorithm

Logistic regression estimates the probability of an event occurring, such as voted or didn't vote, based on a given data set of independent variables.

This type of statistical model (also known as logit model) is often used for classification and predictive analytics. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a logit transformation is applied on the odds—that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds, and this logistic function is represented by the following formulas:

$$\text{Logit}(pi) = 1/(1+ \exp(-pi))$$

$$\ln(pi/(1-pi)) = \text{Beta}\_0 + \text{Beta}\_1*X\_1 + \ldots + B\_k*K\_k$$

In this logistic regression equation, logit(pi) is the dependent or response variable and x is the independent variable. The beta parameter, or coefficient, in this model is commonly estimated via maximum likelihood estimation (MLE). This method tests different values of beta through multiple iterations to optimize for the best fit of log odds. All of these iterations produce the log likelihood function, and logistic regression seeks to maximize this function to find the best parameter estimate. Once the optimal coefficient (or coefficients if there is more than one independent variable) is found, the conditional probabilities for each observation can be calculated, logged, and summed together to yield a predicted probability. For binary classification, a probability less than .5 will predict 0 while a probability greater than 0 will predict 1. After the model has been computed, it's best practice to evaluate the how well the model predicts the dependent variable, which is called goodness of fit. The Hosmer–Lemeshow test is a popular method to assess model fit (*What Is Logistic Regression? | IBM*, n.d.).

4.6.2 Model Evaluation

	precision	recall	f1-score	support
0	0.93	0.88	0.91	327
1	0.67	0.79	0.73	101
accuracy			0.86	428
macro avg	0.80	0.84	0.82	428
weighted avg	0.87	0.86	0.86	428

- Precision: For class 0, precision is 0.93, indicating that 93% of the instances predicted as class 0 were actually class 0. For class 1, precision is 0.67, indicating that 67% of the instances predicted as class 1 were actually class 1.
- Recall: Recall, also known as sensitivity, measures the proportion of true positive instances that were correctly identified by the model. For class 0, recall is 0.88, meaning that 88% of the actual class 0 instances were correctly classified. For class 1, recall is 0.79, indicating that 79% of the actual class 1 instances were correctly classified.
- F1-score: The F1-score is the harmonic mean of precision and recall, providing a balanced measure of a model's performance. It is 0.91 for class 0 and 0.73 for class 1.
- Support: Support represents the number of actual occurrences of each class in the dataset. There are 327 instances of class 0 and 101 instances of class 1.
- Accuracy: Accuracy measures the overall correctness of the model's predictions. In this case, the overall accuracy is 0.86, indicating that 86% of the instances were correctly classified by the model.
- Macro Avg and Weighted Avg: These are the averages of precision, recall, and F1-score across all classes, weighted or unweighted by the number of instances in each class, respectively.

Overall, the Logistic Regression model demonstrates good performance, with high precision and recall for class 0 and slightly lower precision and recall for class 1. The accuracy score of 0.86 indicates that 86% of the instances were correctly classified by the model.

5.0 Evaluation

5.1 Evaluation Metrics

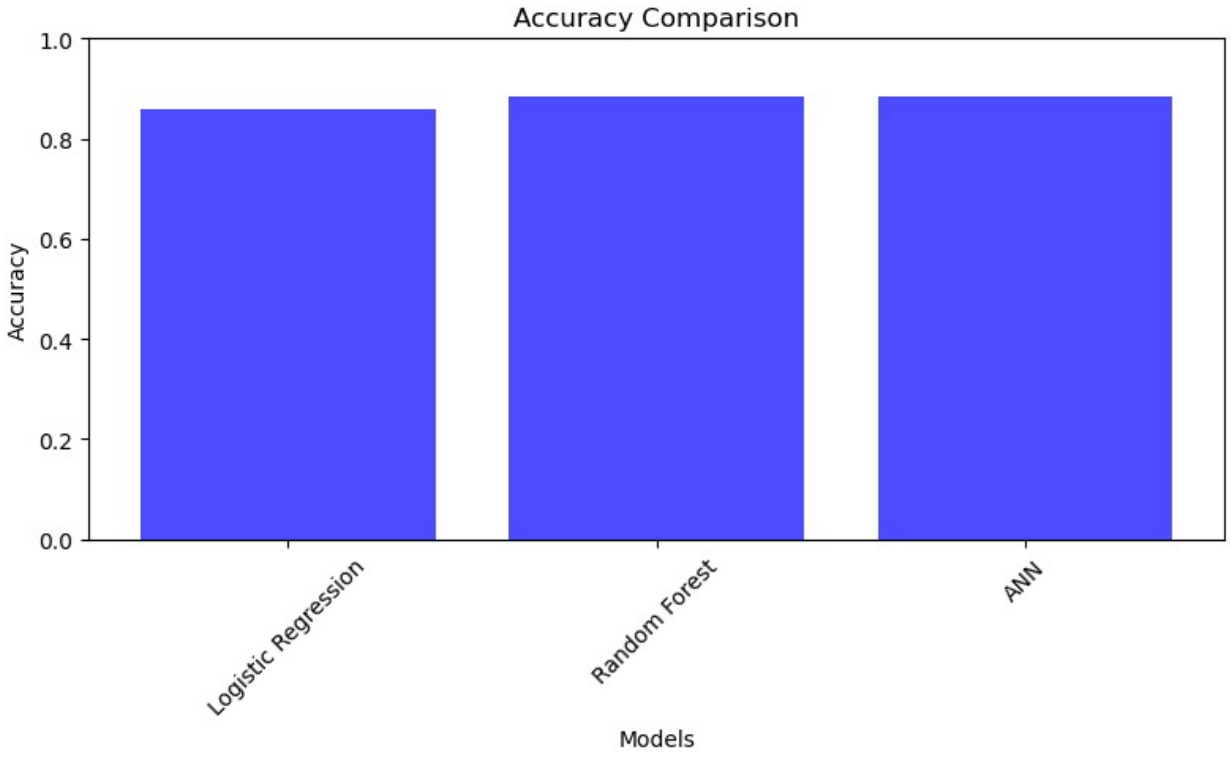
Evaluation metrics are tools used to measure how good a statistical or machine learning model is. They help us understand if the model can predict well and work on different kinds of data. These metrics also let us compare different models to see which one is better. The choice of metrics depends on what problem we're solving, the type of data we have, and what we want to achieve.(Srivastava, 2024) for the Evaluation Metrics, we use accuracy, recall, f1 score, precision and Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared Score for Linear Regression.

5.1.1 Accuracy

Accuracy is a measure that determines the proportion of correct predictions made by a model out of the total number of predictions. (D, 2021)The formula for accuracy is:

The formula for Accuracy

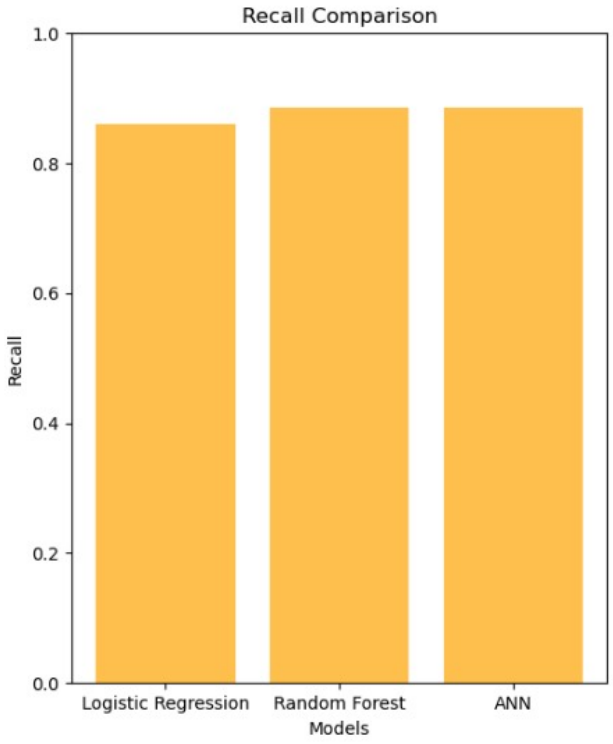
$$Accuracy = \frac{TrueNegatives + TruePositive}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$



5.1.2 Recall

Recall calculates the percentage of actual positives a model correctly identifies (True Positive). We should use recall when the cost of a false negative is high.(D, 2021) The formula for recall is below:

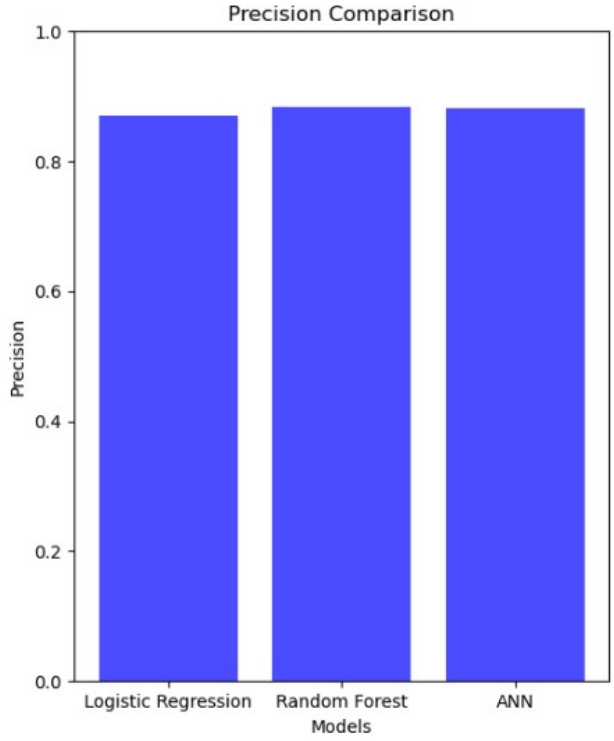
$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$



5.1.3 Precision

Precision measures a model's accuracy in predicting positive labels. It assesses the proportion of correct positive predictions out of all positive predictions made by the model. Precision indicates the relevance of the model's positive predictions.(D, 2021) The formula for precision is:

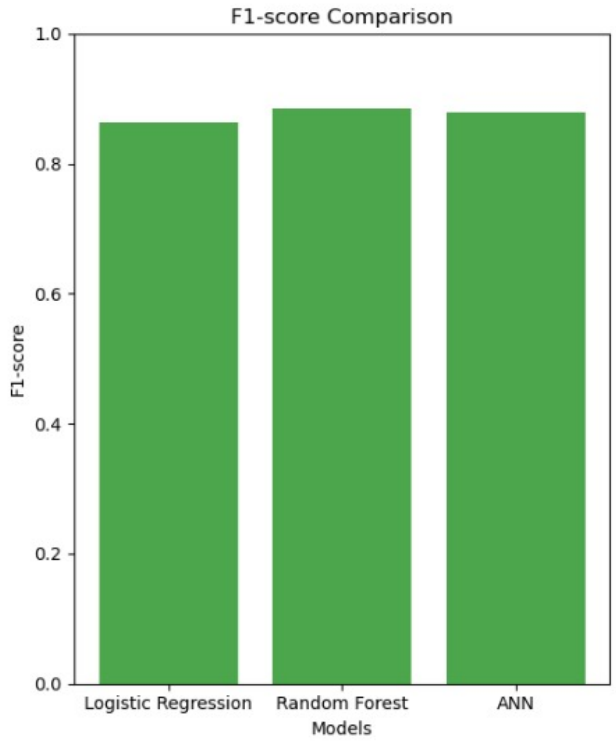
$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$



5.1.4 F1 Scores

F1 score is a machine learning evaluation metric that combines precision and recall scores. Learn how and when to use it to measure model accuracy effectively.(Kundu, 2024)

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$



5.1.5 Evaluation for Linear Regression

Mean Squared Error (MSE)

Mean Squared Error represents the average of the squared difference between the original and predicted values in the data set. It measures the variance of the residuals.(Chugh, 2024)

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

The output:

```
MSE: 0.11
```

Root Mean Squared Error is the square root of Mean Squared error. It measures the standard deviation of residuals.(Chugh, 2024)

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

The output:

```
RMSE: 0.33
```

Mean Absolute Error (MAE)

The Mean absolute error represents the average of the absolute difference between the actual and predicted values in the dataset. It measures the average of the residuals in the dataset.(Chugh, 2024)

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

The output:

```
MAE: 0.24
```

R-squared Score

The coefficient of determination or R-squared represents the proportion of the variance in the dependent variable which is explained by the linear regression model. It is a scale-free score i.e. irrespective of the values being small or large, the value of R square will be less than one.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

The output:

```
R-squared (R2): 0.41
```

5.2 Determination of Model with Best Performance

Evaluation	Logistic Regression	Random Forest	Artificial Neural Network (ANN)
Accuracy	86% (3)	89% (1)	88% (2)
Precision (0)	93% (1)	92% (2)	91% (3)
Precision (1)	67% (3)	78% (1)	77% (2)
Recall (0)	88% (2)	94% (1)	94% (1)
Recall (1)	79% (1)	72% (2)	71% (3)
F1 - Score (0)	91% (3)	93% (1)	92% (2)
F1 - Score (1)	73% (3)	75% (1)	74% (2)

Overall,the result suggests that Random Forest modelling is the best model because it has the best result compared to the other models,which are Logistic Regression and Artificial Neural Network.Therefore,we select Random Forest as the best model for our deployment in 6.0.

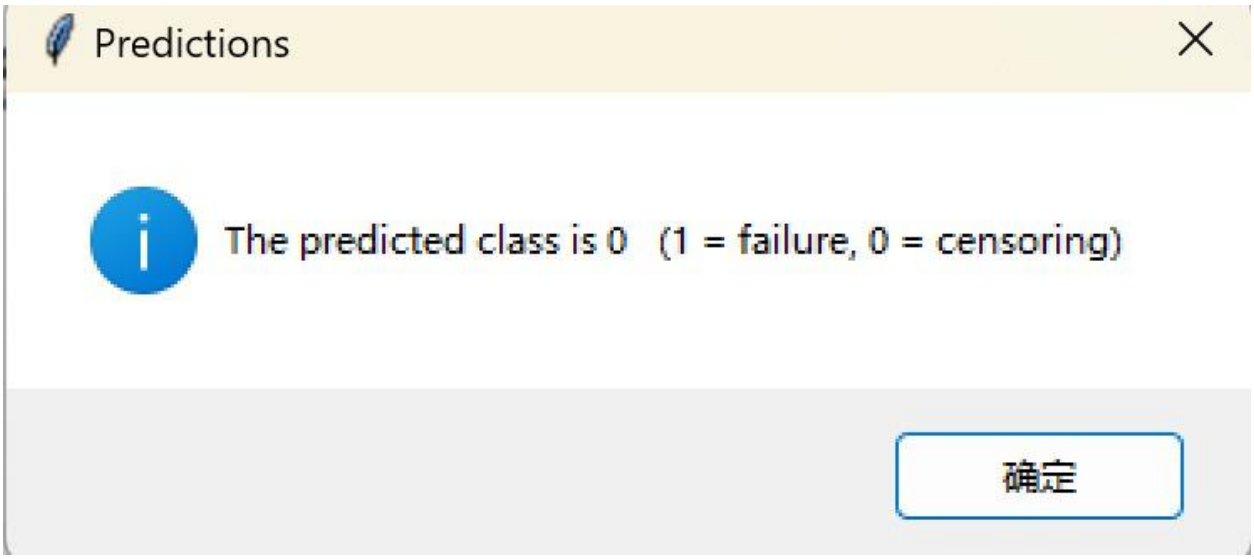
6.0 Deployment

Model deployment in machine learning is the process of integrating your model into an existing production environment where it can take in an input and return an output. The goal is to make the predictions from your trained machine learning model available to others.(Shin, 2023b)

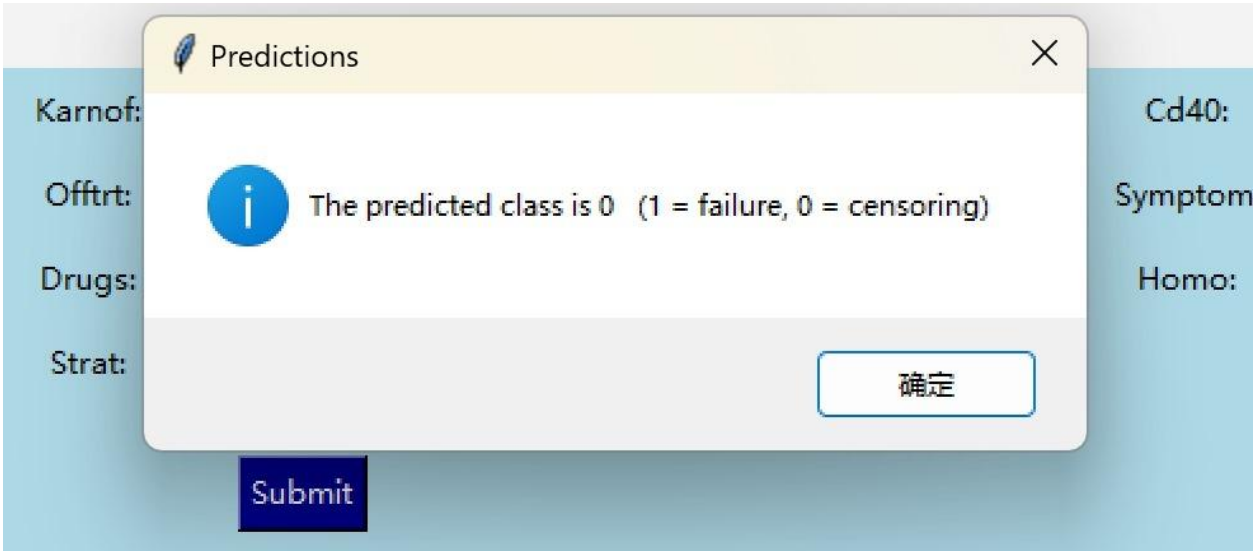
Model deployment refers to the process of making a machine learning model available for use in a production environment where it can make predictions or perform tasks on new, unseen data. During deployment, the trained model is integrated into an application or system where it can be accessed by end-users or other software components. This typically involves deploying the model to a server or cloud platform, setting up APIs or endpoints for communication, and ensuring scalability, reliability, and security. Model deployment is a critical step in the machine learning lifecycle, as it enables organizations to leverage the insights gained from their models to drive real-world decision-making and automation.

6.1 Model Deployment

For model deployment, we utilised the model that has demonstrated the highest level of performance, specifically the Random Forest.



The photo above shows our user interface for our prediction model.The user required to input data for the selected features.Then,the model will run and give the user the prediction made.



The photo above clearly shows that the model successfully predicted the target variable which is ‘cid’ to the user.

7.0 Conclusion

In conclusion, the project has been completed in a limited time, and we have effectively applied all the data science concepts covered in our lectures and practical sessions.

In the Data Understanding phase, we employed various visualisation tools and techniques to explore our dataset. We have used ANOVA correlation coefficient and Chi-squared test to find the relationship between features and target variable.Besides,we also use contingency table and bar chart to visualize the relationship between features and target variable.

Moving on to the Data Preparation phase, we use the result of ANOVA correlation coefficient and Chi-squared test to assist in data selection. For data cleaning, we implemented imputation methods to detect zero values and using IQR to detect outliers.We have made the decision to retain outliers and zero values. Additionally,we normalise the data using the min-max scaling technique.

In the Modeling phase, we constructed four models: Linear Regression,Artificial Neural Network,Logistic Regression and Random Forest. To improve our predictions, we fine-tuned the hyperparameters and resampled our data to balance our dataset. However,we have found out that Linear Regression is not suitable for our dataset.Hence,we only evaluate among the other three models.Our model evaluations were based on key metrics such as accuracy, precision, recall and F1-Score.

During the Evaluation phase, we conducted a comprehensive comparison of our models by examining their respective evaluation metric results. To facilitate this comparison, we created bar charts to visualise the performance distinctions. Our findings revealed that all models exhibited outstanding performance but Random Forest stood out as the top performer with an accuracy of 89%.

In the Deployment phase, we developed a user-friendly function that allows users to input data for the selected features.The selected model for deployment, which is Random Forest, efficiently produced predictions for users.

We are pleased to report that we achieved our data mining goal by successfully developing models with accuracy exceeding 80%,indicating highly accurate predictions closely aligned with actual values.

Throughout this assignment, we had the opportunity to apply the knowledge gained from lectures and practical sessions, gaining a deeper understanding of the entire data science process.Although we faced a lot of challenges in the early stages due to lack of knowledge and expertise, we managed to overcome these obstacles and completed the assignment in limited time.

We would like to express our sincere appreciation to our lecturer and tutor, Dr Noor Aida, for providing invaluable guidance and support throughout this journey. Looking forward, we are committed to working diligently to further enhance our expertise in the field of data science.

References

https://github.com/scikit-learn/scikit-learn/issues/26768
UC Irvine Machine Learning Repository | Re3data.org. (n.d.). https://www.re3data.org/repository/r3d100010960

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2017). Deep learning for healthcare: review, opportunities and challenges. Briefings in Bioinformatics, 19(6), 1236–1246. https://doi.org/10.1093/bib/bbx044

Simplilearn. (2023, October 19). *What is Descriptive Statistics: Definition, Types, Applications, and Examples*. Simplilearn.com. https://www.simplilearn.com/what-is-descriptive-statistics-article

Razavi, S., Jakeman, A., Saltelli, A., Prieur, C., Iooss, B., Borgonovo, E., Plischke, E., Lo Piano, S., Iwanaga, T., Becker, W. E., Tarantola, S., Guillaume, J. H. A., Jakeman, J. D., Gupta, H., Melillo, N., Rabitti, G., Chabridon, V., Duan, Q., Sun, X., . . . Maier, H. R. (2021). The Future of Sensitivity Analysis: An essential discipline for systems modeling and policy support. *Environmental Modelling & Software*, *137*, 104954. https://doi.org/10.1016/j.envsoft.2020.104954

CSV Format | Socrata. (n.d.). https://dev.socrata.com/docs/formats/csv.html

Bhandari, P. (2024, January 17). *How to Find Outliers | 4 Ways with Examples & Explanation*. Scribbr. https://www.scribbr.com/statistics/outliers/

Biswal, A. (2023, October 11). *What is a Chi-Square Test? Formula, Examples & Application*. Simplilearn.com. https://www.simplilearn.com/tutorials/statistics-tutorial/chi-square-test#:~:text=A%20chi%2Dsquare%20test%20is,between%20the%20variables%20under%20consideration.

Simplilearn. (2023, October 19). *What is Descriptive Statistics: Definition, Types, Applications, and Examples*. Simplilearn.com. https://www.simplilearn.com/what-is-descriptive-statistics-article

NCI Dictionary of Cancer Terms. (n.d.). Cancer.gov. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/karnofsky-performance-status

Talend. (n.d.). *What is Data Preparation? Processes and Example*. Talend - a Leader in Data Integration & Data Integrity. https://www.talend.com/resources/what-is-data-preparation/

How do you manage duplicate data records? (2023, December 9). www.linkedin.com. https://www.linkedin.com/advice/1/how-do-you-manage-duplicate-data-records-skills-data-analysis

What is Data Deduplication? How to Improve Data Uniqueness. (n.d.-b). https://www.sagacitysolutions.co.uk/about/news-and-blog/data-deduplication/

Kenton, W. (2024, April 19). *What is analysis of variance (ANOVA)?* Investopedia. https://www.investopedia.com/terms/a/anova.asp

Talend. (n.d.). *What is Data Preparation? Processes and Example*. Talend - a Leader in Data Integration & Data Integrity. https://www.talend.com/resources/what-is-data-preparation/

Brownlee, J. (2020, August 20). *How to choose a feature selection method for machine learning*. MachineLearningMastery.com. https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/#:~:text=The%20most%20common%20techniques%20are,Pearson's%20correlation%20coefficient%20(linear)

Turney, S. (2023, June 22). *Chi-Square (X2) Tests | Types, Formula & Examples*. Scribbr. https://www.scribbr.com/statistics/chi-square-tests/

Strauss, A. L. (1987). *Qualitative analysis for social scientists*. https://doi.org/10.1017/cbo9780511557842

Friedman, J. H., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *Annals of Statistics*, *28*(2). https://doi.org/10.1214/aos/1016218223

Yermack, D. (2017). Corporate governance and blockchains. *Review of Finance*, rfw074. https://doi.org/10.1093/rof/rfw074

Pendergast, M. (2006). Teaching Introductory Programming to IS Students: Java Problems and Pitfalls. *Journal of Information Technology Education*, *5*, 491–515. https://doi.org/10.28945/261

Galiè, N., Hoeper, M. M., Humbert, M., Torbicki, A., Vachiéry, J., Barberá, J., Beghetti, M., Corris, P. A., Gaine, S., Js, G., Ma, G., Jondeau, G., Klepetko, W., Opitz, C., Peacock, A., Rubin, L. J., Zellweger, M., & Simonneau, G. (2009). Guidelines for the diagnosis and treatment of pulmonary hypertension: The Task Force for the Diagnosis and Treatment of Pulmonary Hypertension of the European Society of Cardiology (ESC) and the European Respiratory Society (ERS), endorsed by the International Society of Heart and Lung Transplantation (ISHLT). *European Heart Journal*, *30*(20), 2493–2537. https://doi.org/10.1093/eurheartj/ehp297

Kenton, W. (2024, April 19). *What is analysis of variance (ANOVA)?* Investopedia. https://www.investopedia.com/terms/a/anova.asp

Syed, A. H. (2023, April 20). Dealing with Outliers in Data Science: Techniques and Best Practices. *Medium*. https://syedabis98.medium.com/dealing-with-outliers-in-data-science-techniques-and-best-practices-a08172643b7a

Bradley, L. (2018). *HIV+ women's reproductive decision-making: perceiving reproductive choice*. https://doi.org/10.22215/etd/2005-08489

Livingston, G., Sommerlad, A., Orgeta, V., Costafreda, S. G., Huntley, J., Ames, D., Ballard, C., Banerjee, S., Burns, A., Cohen-Mansfield, J., Cooper, C., Fox, N. C., Gitlin, L. N., Howard, R., Kales, H. C., Larson, E. B., Ritchie, K., Rockwood, K., Sampson, E. L., . . . Mukadam, N. (2017). Dementia prevention, intervention, and care. *Lancet*, *390*(10113), 2673–2734. https://doi.org/10.1016/s0140-6736(17)31363-6

Bobbitt, Z. (2021b, December 7). *How to interpret the F-Value and P-Value in ANOVA*. Statology. https://www.statology.org/anova-f-value-p-value/

What is linear regression? | Master's in Data Science. (2023, December 15). CORP-MIDS1 (MDS). https://www.mastersindatascience.org/learning/machine-learning-algorithms/linear-regression/

\Srivastava, T. (2024, January 8). *12 Important model evaluation Metrics for Machine Learning Everyone should know (Updated 2023)*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/

Linear regression. (n.d.). http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm#:~:text=A%20linear%20regression%20line%20has,Y%20is%20the%20dependent%20variable.

Varshney, H. (2021, July 19). *Data Science Modelling: 8 easy steps*. Learn | Hevo. https://hevodata.com/learn/data-science-modelling/#udsm

What is linear regression? | IBM. (n.d.-b). https://www.ibm.com/topics/linear-regression#:~:text=IBM-,What%20is%20linear%20regression%3F,is%20called%20the%20independent%20variable.

Imbalanced data. (n.d.). Google for Developers. https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data

D, E. (2021, December 12). Accuracy, Recall & Precision - Erika D - Medium. *Medium*. https://medium.com/@erika.dauria/accuracy-recall-precision-80a5b6cbd28d

What is Hyperparameter Tuning? - Hyperparameter Tuning Methods Explained - AWS. (n.d.). Amazon Web Services, Inc. https://aws.amazon.com/what-is/hyperparameter-tuning/#:~:text=Hyperparameter%20tuning%20allows%20data%20scientists,the%20model%20as%20a%20hyperparameter.

Kundu, R. (2024, April 10). F1 Score in Machine Learning: Intro & Calculation. *V7*. https://www.v7labs.com/blog/f1-score-guide

Chugh, A. (2024, January 18). MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better? *Medium*. https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e

Shah, R. (2024, April 16). *Tune Hyperparameters with GridSearchCV*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/tune-hyperparameters-with-gridsearchcv/#:~:text=GridSearchCV%20acts%20as%20a%20valuable,that%20can%20enhance%20its%20performance.

Deci. (2024, February 20). *Neural Architecture Search: Everything You need to know | DECI*. https://deci.ai/neural-architecture-search/

Singh, G. (2024, April 5). *Introduction to artificial neural networks*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/09/introduction-to-artificial-neural-networks/

What is Random Forest? | IBM. (n.d.). https://www.ibm.com/topics/random-forest

What is logistic regression? | IBM. (n.d.). https://www.ibm.com/topics/logistic-regression

Shin, T. (2023b, November 7). *What is model deployment in machine learning?* Built In. https://builtin.com/machine-learning/model-deployment