PREDICTIVE MODELING PROJECT

BUSINESS REPORT

Prepared By : Sahid
COURSE : DSBA
DATE : 04-07-2024

1. Introduction

1.1 Project Overview

1.2 Objectives

2. Problem 1: Predicting 'usr' (CPU User Mode Time)

2.1. Problem Definition & Exploratory Data Analysis (EDA)

- **Data Description and Statistical Summaries**
- **Univariate Analysis** & **Multivariate Analysis**

2.2 Data Preprocessing

- **Missing Values & Outliers**: Detection and treatment.
- **Data Encoding**: Methods for encoding categorical variables.
- **Train-Test Split**: Ratio and summary statistics.

2.3 Model Building

- **Linear Regression**: Using Sklearn and Statsmodels.
- **Performance Metrics**: R^2, Adj R^2, RMSE for train and test sets.
- **Model Comparison**: Evaluate different models and their performance.

2.4 Business Insights & Recommendations

- **Regression Equation**: Final model and impact of key variables.
- **Actionable Insights**: Recommendations for system optimization.

3. Problem 2: Predicting Contraceptive Method Choice

3.1 Problem Definition & Exploratory Data Analysis (EDA)

- **Data Description and Statistical Summaries**
- **Univariate Analysis** & **Multivariate Analysis**

3.2 Data Preprocessing

- **Missing Values & Outliers**: Detection and treatment.
- **Data Encoding**: Methods for encoding categorical variables.
- **Train-Test Split**: Ratio and summary statistics.

3.3 Model Building

- **Models**: Logistic Regression, Linear Discriminant Analysis, CART.
- **Hyperparameter Tuning**: GridSearch for CART.
- **Performance Metrics**: Accuracy, Precision, Recall, F1-score.
- **Model Comparison**: Evaluate and select the best model.

3.4 Business Insights & Recommendations

- **Feature Importance**: Key features from the best model.
- **Actionable Insights**: Recommendations for healthcare policy.

4. Conclusion

- **Summary of Findings**: Key insights from both problems.
- **Final Recommendations**

5. References

List of Tables

List of Figures

1. Introduction:

1.1 Project Overview

Problem 1: Predicting 'usr' (CPU User Mode Time)

"In today's digitally-driven world, the performance of computer systems is paramount, especially in multi-user environments. One critical performance metric is the percentage of time CPUs operate in user mode, known as 'usr'. The usage of CPUs in user mode can significantly affect the overall system performance, responsiveness, and resource allocation. Understanding and predicting the 'usr' metric can help in optimizing system performance and ensuring efficient resource utilization."

Problem 2: Predicting Contraceptive Method Choice

"Effective family planning is a crucial element in public health policymaking, especially in developing countries. The Republic of Indonesia has been actively conducting surveys to understand contraceptive prevalence among married women. Accurately predicting whether a woman opts for a contraceptive method, given her demographic and socio-economic background, can significantly aid in designing targeted family planning programs. This analysis focuses on understanding the factors influencing contraceptive choices among married women in Indonesia."

1.2 Objectives

Problem 1: Predicting 'usr' (CPU User Mode Time)

"The primary objective of this analysis is to establish a linear regression model to predict the 'usr' metric based on various system attributes. By achieving this, we aim to understand the influence of these attributes on CPU usage in user mode and provide actionable insights for system optimization and management. Specific objectives include conducting exploratory data analysis to identify patterns, building and evaluating multiple linear regression models, and drawing business insights from the final model."

Problem 2: Predicting Contraceptive Method Choice

"The primary objective of this analysis is to develop a predictive model that determines whether a woman will choose to use a contraceptive method based on her demographic and socio-economic attributes. This involves performing exploratory data analysis to uncover patterns, building and comparing various classification models (such as Logistic Regression, Linear Discriminant Analysis, and CART), and deriving the most effective model. Additionally, the study aims to provide actionable recommendations based on the best-performing model to support public health initiatives.**"**

2. Problem 1: Predicting 'usr' (CPU User Mode Time)

2.1. Problem Definition & Exploratory Data Analysis (EDA)

Problem Definition

Context: The comp-activ database comprises activity measures of computer systems. Data was gathered from a Sun Sparcstation 20/712 with 128 Mbytes of memory, operating in a multi-user university department. Users engaged in diverse tasks, such as internet access, file editing, and CPU-intensive programs. Being an aspiring data scientist, you aim to establish a linear equation for predicting 'usr' (the percentage of time CPUs operate in user mode). Your goal is to analyze various system attributes to understand their influence on the system's 'usr' mode.

Exploratory Data Analysis (EDA):

Data Description:

System measures used:

- lread - Reads (transfers per second) between system memory and user memory
- lwrite - writes (transfers per second) between system memory and user memory
- scall - Number of system calls of all types per second
- sread - Number of system read calls per second .
- swrite - Number of system write calls per second .
- fork - Number of system fork calls per second.
- exec - Number of system exec calls per second.
- rchar - Number of characters transferred per second by system read calls
- wchar - Number of characters transfreed per second by system write calls
- pgout - Number of page out requests per second
- ppgout - Number of pages, paged out per second
- pgfree - Number of pages per second placed on the free list.
- pgscan - Number of pages checked if they can be freed per second
- atch - Number of page attaches (satisfying a page fault by reclaiming a page in memory) per second
- pgin - Number of page-in requests per second
- ppgin - Number of pages paged in per second
- pflt - Number of page faults caused by protection errors (copy-on-writes).
- vflt - Number of page faults caused by address translation .
- runqsz - Process run queue size (The number of kernel threads in memory that are waiting for a CPU to run. Typically, this value should be less than 2. Consistently higher values mean that the system might be CPU-bound.)
- freemem - Number of memory pages available to user processes
- freeswap - Number of disk blocks available for page swapping.

- usr - Portion of time (%) that cpus run in user mode

Data Overview :

The dataset has 8192 rows and 22 columns. It is always a good practice to view a sample of the rows. A simple way to do that is to use head() function.

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgscan	atch	pgin	ppgin	pflt	vflt	runqsz	freemem	freeswap	usr
0	1	0	2147	79	68	0.2	0.2	40671.0	53995.0	0.0	...	0.0	0.0	1.6	2.6	16.00	26.40	CPU_Bound	4670	1730946	95
1	0	0	170	18	21	0.2	0.2	448.0	8385.0	0.0	...	0.0	0.0	0.0	0.0	15.63	16.83	Not_CPU_Bound	7278	1869002	97
2	15	3	2162	159	119	2.0	2.4	NaN	31950.0	0.0	...	0.0	1.2	6.0	9.4	150.20	220.20	Not_CPU_Bound	702	1021237	87
3	0	0	160	12	16	0.2	0.2	NaN	8670.0	0.0	...	0.0	0.0	0.2	0.2	15.60	16.80	Not_CPU_Bound	7248	1863704	98
4	5	1	330	39	38	0.4	0.4	NaN	12185.0	0.0	...	0.0	0.0	1.0	1.2	37.80	47.60	Not_CPU_Bound	633	1760253	90

(Table 1: Top five rows of the dataset for Problem1)

Descriptive Statistics :

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgfree	pgscan	atch	pgin	ppgin	pflt	vflt	freemem	freeswap	usr
count	8192	8192	8192	8192	8192	8192	8192	8.09E+03	8.18E+03	8192	...	8192	8192	8192	8192	8192	8192	8192	8192	8.19E+03	8192
mean	19.55	13.1	2306.31	210.47	150.05	1.88	2.791998	1.97E+05	9.59E+04	2.285317	...	11.91971	21.52684	1.127505	8.27796	12.38858	109.7937	185.3157	1763.456	1.33E+06	83.96887
std	53.35	29.89	1633.61	198.98	160.47	2.47	5.212456	2.40E+05	1.41E+05	5.307038	...	32.36352	71.14134	5.708347	13.87497	22.28131	114.4192	191.0006	2482.104	4.22E+05	18.40190
min	0	0	109	6	7	0	0	2.78E+02	1.50E+03	0	...	0	0	0	0	0	0	0.2	55	2.00E+00	0
25%	2	0	1012	86	63	0.4	0.2	3.41E+04	2.29E+04	0	...	0	0	0	0.6	0.6	25	45.4	231	1.04E+06	81
50%	7	1	2051.5	166	117	0.8	1.2	1.25E+05	4.66E+04	0	...	0	0	0	2.8	3.8	63.8	120.4	579	1.29E+06	89
75%	20	10	3317.25	279	185	2.2	2.8	2.68E+05	1.06E+05	2.4	...	5	0	0.6	9.765	13.8	159.6	251.8	2002.25	1.73E+06	94
max	1845	575	12493	5318	5456	20.12	59.56	2.53E+06	1.80E+06	81.44	...	523	1237	211.58	141.2	292.61	899.8	1365	12027	2.24E+06	99

(Table 2: Descriptive statistics of the dataset for Problem1)

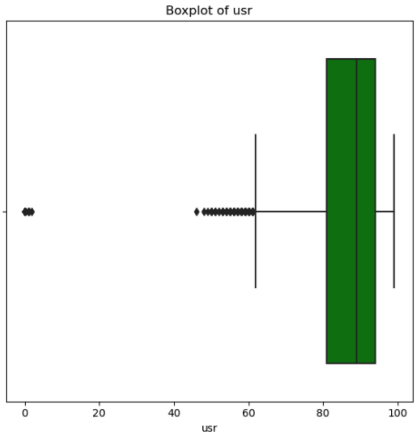
Observations:

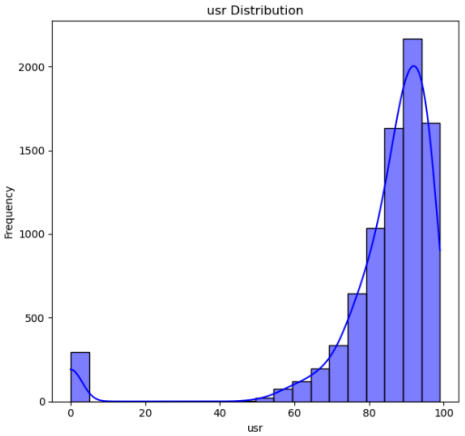
- High Variability: Most variables show significant variability (e.g., lread, lwrite).
- Outliers: Presence of extreme values in several metrics like lread, scall, and sread.
- Skewness: Many variables exhibit right skewness with medians lower than means.
- Zero Values: Numerous variables have minimum values of zero, indicating periods of inactivity.
- Target Variable (usr): usr has a high mean (~84%) with a standard deviation of 18.4, indicating CPUs frequently operate in user mode.
- Potential Correlations: Variables like rchar and wchar have high variability, possibly impacting usr

Univariate Analysis

For performing Univariate analysis we will take a look at the Boxplots and Histograms of variables such as scall, fork, pgout, lread, and runqsz are typically highly relevant for predicting usr due to their direct impact on CPU operations and memory management.

Observations on usr





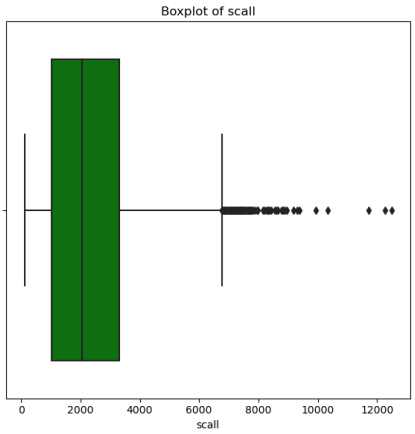
(Fig1 : Plot showing the boxplot and distribution for usr)

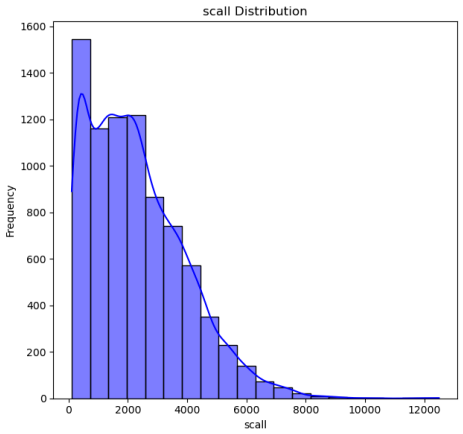
Observations:

● The Distribution of usr is left skewed.

● There are a lot of outliers in this variable.

Observations on scall





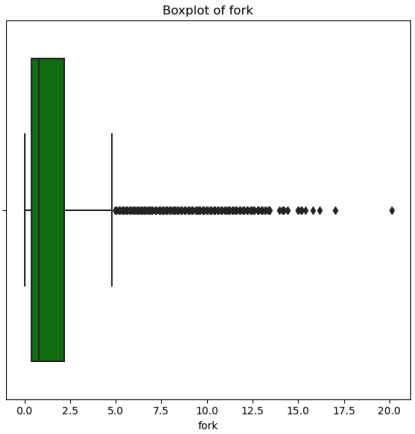
(Fig2 : Plot showing the boxplot and distribution for scall)

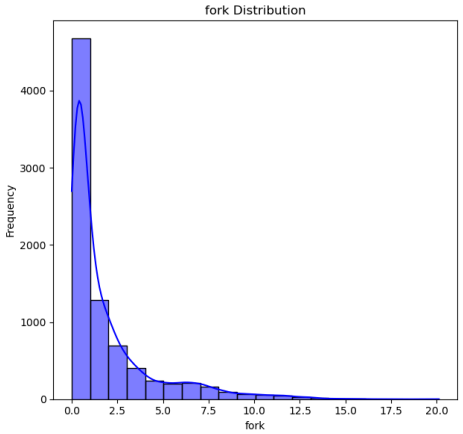
Observations:

● The Distribution of scall is right skewed.

● There are a lot of outliers in this variable.

Observations on fork





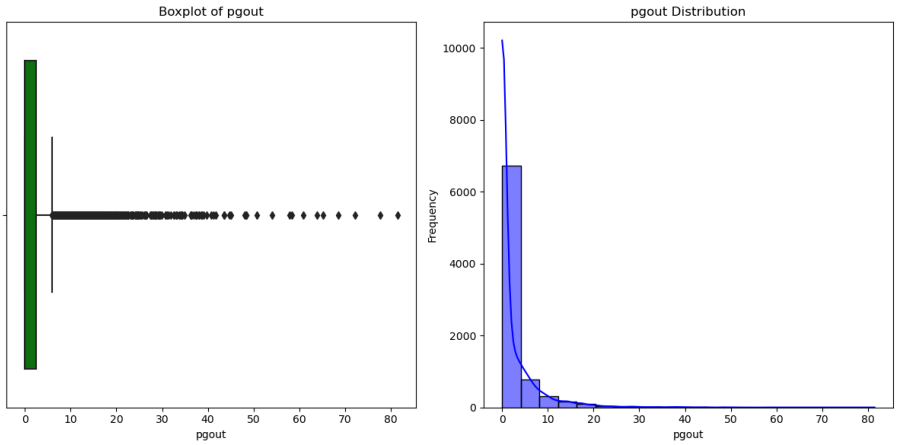
(Fig3 : Plot showing the boxplot and distribution for fork)

Observations:

● The Distribution of fork is right skewed.

● There are a lot of outliers in this variable.

Observations on pgout

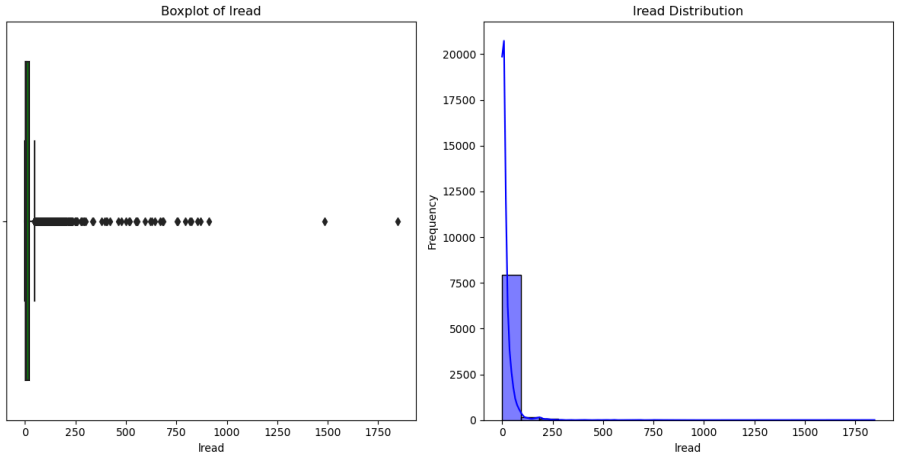


(Fig4 : Plot showing the boxplot and distribution for pgout)

Observations:

- The Distribution of pgoutis heavily right skewed.
- There are a lot of outliers in this variable

Observations on lread

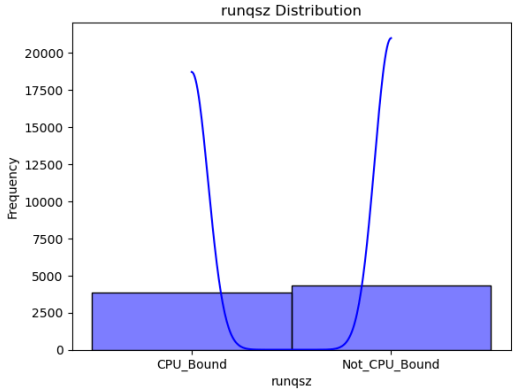


(Fig5 : Plot showing the boxplot and distribution for lread)

Observations:

- The Distribution of lread is heavily right skewed.
- There are a lot of outliers in this variable

Observations on runqsz

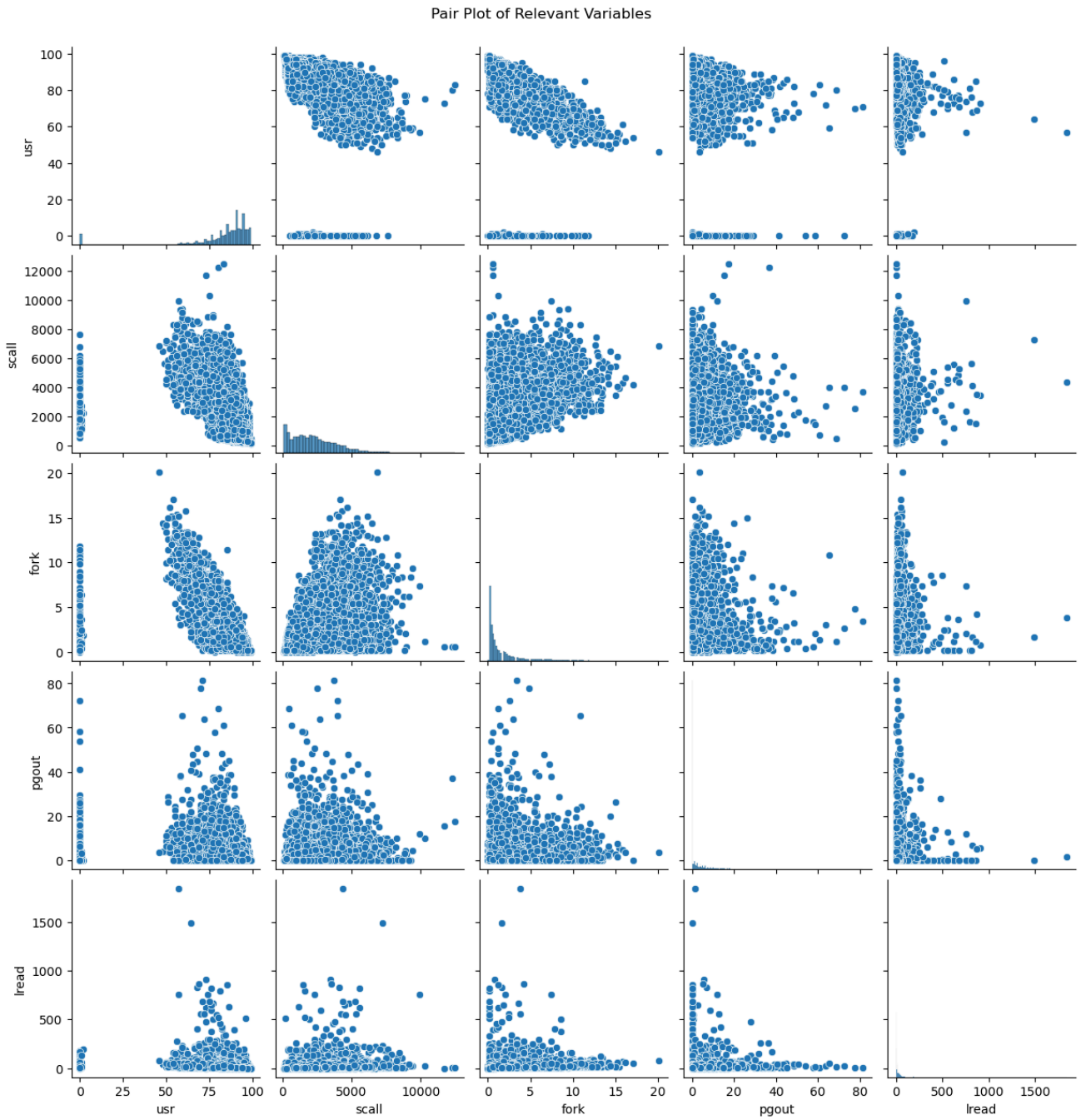


(Fig6 : Plot showing the distribution for runqsz)

Observations:

- The Distribution of runqsz is heavily right skewed.
- There are a lot of outliers in this variable

Multivariate analysis

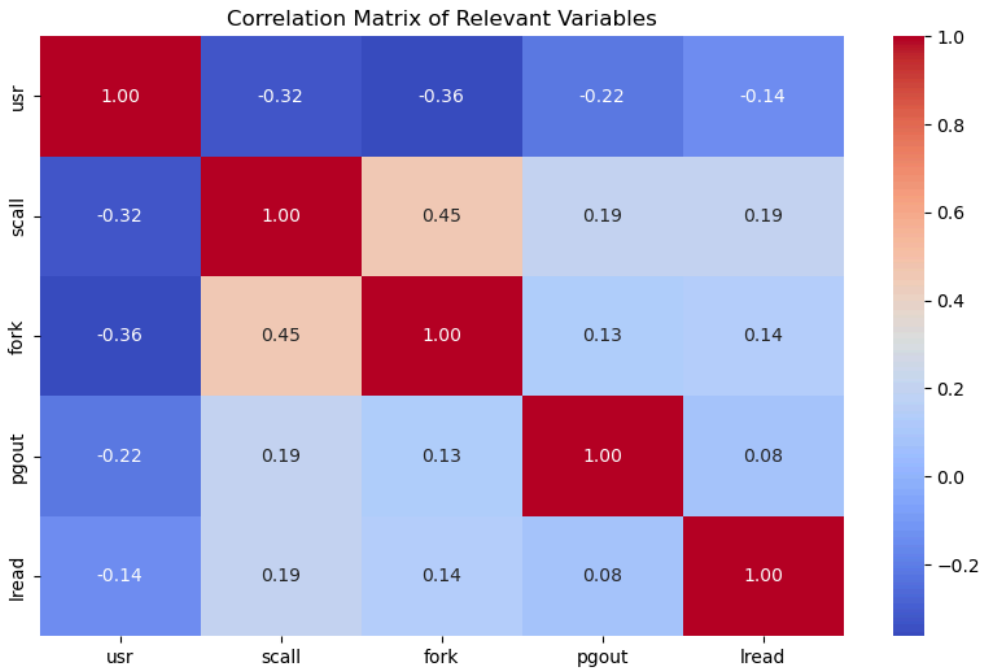


(Fig7 : Pair plot showing relevant variables for predicting 'usr'.)

Observations:

Scatter Plots: There are noticeable patterns in the relationships between 'usr' and other variables such as 'scall', 'fork', 'pgout', and 'lread'.

Correlation of Numerical Variables



(fig 8 showing heatmap of Correlation of relevant numerical variables)

Observations

usr Correlations:

- scall: Moderate negative correlation (-0.323).
- fork: Moderate negative correlation (-0.363)
- pgout: Weak negative correlation (-0.222).
- lread: Weak negative correlation (-0.141).

scall Correlations:

- fork: Moderate positive correlation (0.447).
- pgout: Weak positive correlation (0.195).
- lread: Weak positive correlation (0.191).

fork Correlations:

- pgout: Weak positive correlation (0.130).
- lread: Weak positive correlation (0.140).

pgout and lread: Very weak positive correlation (0.082).

Summary

- usr decreases with increases in scall, fork, pgout, and lread.
- scall and fork are significantly related (0.447).
- Focus on scall and fork as they show notable impact on usr

2.2 Data Preprocessing

Missing Values & Outliers: Detection and treatment.

Check for missing values : There are two variables “rchar” and “wchar” that has missing values of 104 and 15 respectively

To ensure the integrity of our analysis and model performance, we addressed the missing values in the dataset through the following method:

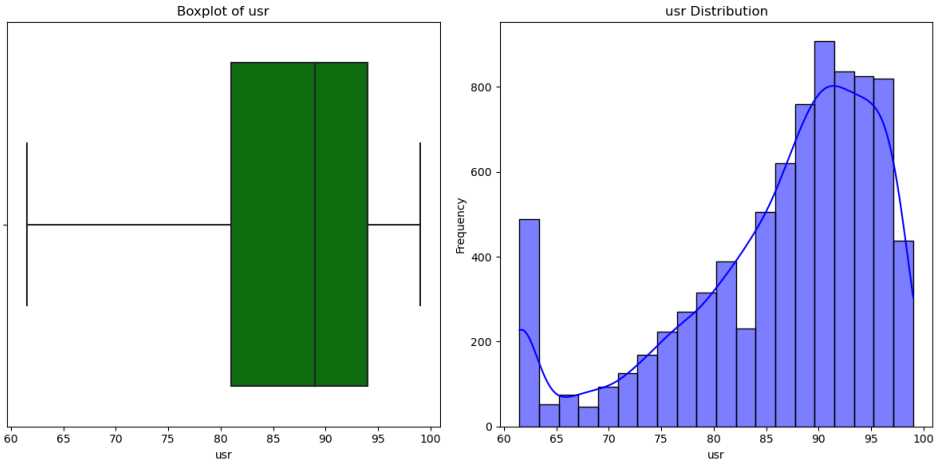
- **Imputation with Mean:** For each column with missing values, we filled the missing entries with the mean value of that column. This method helps to maintain the overall distribution of the data and prevents the loss of valuable information.

Outlier Detection and Treatment

To ensure the accuracy and reliability of our analysis, we performed outlier detection and treatment on the numerical columns in our dataset. This process helps in mitigating the potential skewing effect of extreme values, ensuring a more robust model.

Methodology: Interquartile Range (IQR)

- **Interquartile Range Calculation**: We computed the first quartile (Q1) and the third quartile (Q3) for each numerical column in the dataset. The IQR is then determined as the difference between Q3 and Q1.
- **Identifying Outliers**: Using the IQR, we established the lower and upper bounds for detecting outliers. The lower bound is calculated as Q1−1.5×IQRQ1 - 1.5 \times IQRQ1−1.5×IQR and the upper bound as Q3+1.5×IQRQ3 + 1.5 \times IQRQ3+1.5×IQR.
- **Treating Outliers**: For each numerical column, values below the lower bound were set to the lower bound, and values above the upper bound were set to the upper bound. This method, known as capping, ensures that extreme values are brought within a reasonable range without entirely discarding the data**.**



(fig 9 : Plot showing the boxplot and distribution for usr after outlier treatment)

Data Encoding

To prepare the dataset for modeling, it was necessary to convert categorical variables into a numeric format suitable for analysis. Specifically, the 'runqsz' column, which originally contained categorical values, was encoded into binary numeric values.

- **Conversion Process:** The 'runqsz' column, which had categories 'CPU_Bound' and 'Not_CPU_Bound', was converted into binary numeric values. The encoding was as follows:
    - 'CPU_Bound' was mapped to 0.
    - 'Not_CPU_Bound' was mapped to 1.
- **Validation:** After the conversion, the unique values in the 'runqsz' column were checked to confirm the successful encoding process.

This transformation ensures that the categorical data is in a format compatible with our analytical and modeling techniques, enhancing the overall effectiveness of the analysis.

Train-Test Split: Ratio and summary statistics.

Data Splitting

To prepare the data for model building, the dataset was divided into features and the target variable, followed by a train-test split:

- **Feature and Target Selection:**
    - **Features (X):** All columns except 'usr'.
    - **Target (Y):** The 'usr' column, representing the percentage of time CPUs operate in user mode.

- **Train-Test Split:**
    - The dataset was split into training and testing sets to evaluate model performance.
    - **Training Set:** 80% of the data, used for training the models.
    - **Testing Set:** 20% of the data, used for testing and validating the models.
    - A random state of 42 was used to ensure reproducibility of the results.

This splitting ensures that the models can be trained on one portion of the data and tested on an unseen portion, providing a robust evaluation of their performance

(Table 3: Summary statistics for the training set for Problem1)

	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
	Training Set	Training Set	Training Set	Training Set	Training Set	Training Set	Training Set	Training Set	Testing Set	Testing Set	Testing Set	Testing Set	Testing Set	Testing Set	Testing Set	Testing Set
lread	6553	1.35E+01	15.16981	0	2	7	20	4.70E+01	1.64E+03	1.33E+01	15.12303	0	2	6	19	4.70E+01
lwrite	6553	6.67E+00	9.289463	0	0	1	11	2.50E+01	1.64E+03	6.59E+00	9.304422	0	0	1	9	2.50E+01
scall	6553	2.29E+03	1593.526065	109	1012	2047	3305	6.78E+03	1.64E+03	2.30E+03	1591.84577	128	1020	2074	3349	6.78E+03
sread	6553	1.98E+02	145.410875	6	86	165	276	5.69E+02	1.64E+03	2.05E+02	151.949357	6	84.5	167	295.5	5.69E+02
swrite	6553	1.37E+02	96.104416	7	62	116	184	3.68E+02	1.64E+03	1.43E+02	101.075899	9	64.5	120	191	3.68E+02
fork	6553	1.56E+00	1.594736	0	0.4	0.8	2.4	4.90E+00	1.64E+03	1.55E+00	1.57754	0	0.4	0.8	2.2	4.90E+00
exec	6553	1.95E+00	2.045013	0	0.2	1.2	2.8	6.70E+00	1.64E+03	1.87E+00	1.959087	0	0.2	1.2	2.6	6.70E+00
rchar	6553	1.79E+05	173650.6614	278	35708	128515	263192	6.11E+05	1.64E+03	1.82E+05	177874.826	419	33715.5	126309	274447	6.11E+05
wchar	6553	7.56E+04	70852.07201	1498	22916	46726	106369	2.31E+05	1639	7.64E+04	72873.4635	1522	23167.5	45752	103802.5	2.31E+05
pgout	6553	1.43E+00	2.201995	0	0	0	2.4	6.00E+00	1639	1.40E+00	2.193754	0	0	0	2.2	6.00E+00
ppgout	6553	2.58E+00	4.051437	0	0	0	4.2	1.05E+01	1639	2.48E+00	3.980508	0	0	0	3.8	1.05E+01
pgfree	6553	3.19E+00	5.005648	0	0	0	5.2	1.25E+01	1639	3.05E+00	4.893017	0	0	0	4.4	1.25E+01
pgscan	6553	0.00E+00	0	0	0	0	0	0.00E+00	1639	0.00E+00	0	0	0	0	0	0.00E+00
atch	6553	3.86E-01	0.560653	0	0	0	0.6	1.50E+00	1639	3.96E-01	0.572088	0	0	0	0.8	1.50E+00
pgin	6553	6.37E+00	7.663443	0	0.6	2.8	9.6	2.35E+01	1639	6.44E+00	7.769845	0	0.6	2.8	9.8	2.35E+01
ppgin	6553	9.10E+00	11.108803	0	0.6	3.8	13.8	3.36E+01	1639	9.32E+00	11.36869	0	0.6	3.8	13.9	3.36E+01
pflt	6553	1.06E+02	101.713188	0	24.8	64	160.2	3.62E+02	1639	1.05E+02	100.914027	0.8	25.4	63.07	153	3.62E+02
vflt	6553	1.76E+02	163.178312	0.2	45.4	120.4	251.8	5.61E+02	1639	1.74E+02	159.787856	3.6	46.6	120.76	252.05	5.61E+02
runqsz	6553	5.26E-01	0.499376	0	0	1	1	1.00E+00	1639	5.41E-01	0.498503	0	0	1	1	1.00E+00
freemem	6553	1.40E+03	1613.175558	55	230	582	2018	4.66E+03	1639	1.35E+03	1575.77737	62	235	572	1926	4.66E+03
freeswap	6553	1.33E+06	420519.1454	10989.5	1043112	1300490	1731210	2.24E+06	1639	1.32E+06	421791.889	10989.5	1038995.5	1128165	1724438.5	1.89E+06

Model Building

Model Overview: A linear regression model was trained to predict the percentage of time CPUs operate in user mode (usr) using the system attributes. The model was evaluated using the test set to assess its performance.

Key Model Parameters:

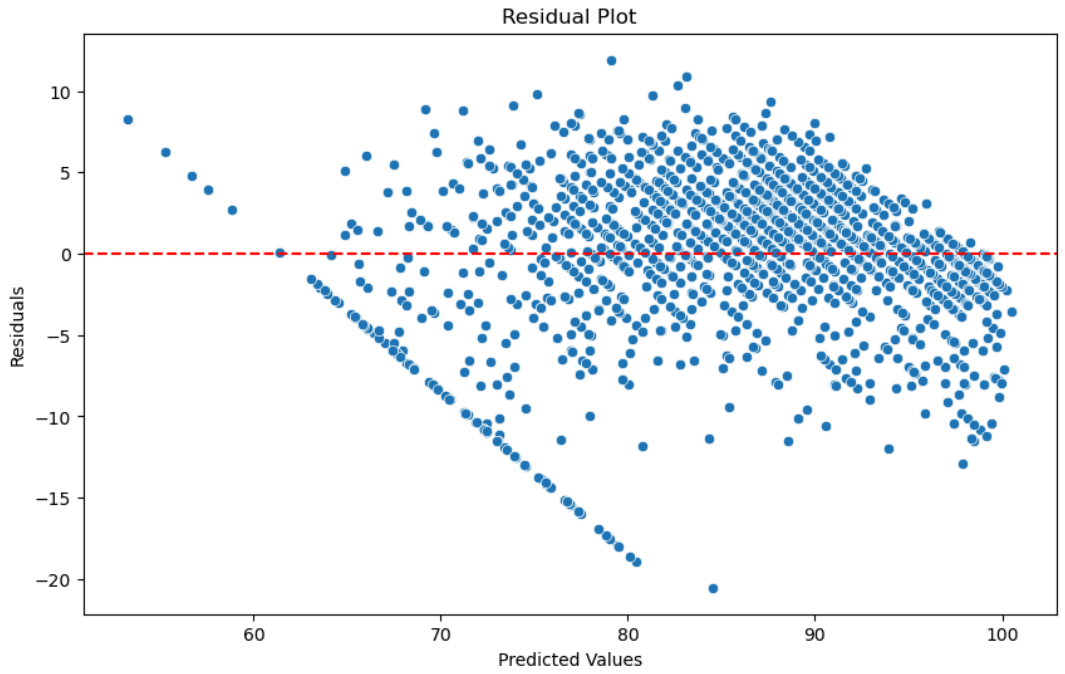
- Coefficients: The model's coefficients for each feature are as follows:
  - Feature 1: -0.0558
  - Feature 2: 0.0412
  - Feature 3: -0.0007
  - Feature 4: 0.0023
  - Feature 5: -0.0060
  - Feature 6: -0.0838
  - Feature 7: -0.2670
  - Feature 8: -0.000005
  - Feature 9: -0.000005
  - Feature 10: -0.4823
  - Feature 11: -0.0504
  - Feature 12: 0.1125
  - Feature 13: -0.0000
  - Feature 14: 0.6205
  - Feature 15: 0.0384
  - Feature 16: -0.0684

    - Feature 17: -0.0320
    - Feature 18: -0.0063
    - Feature 19: 1.8131
    - Feature 20: -0.0005
    - Feature 21: 0.000009
- Intercept: The model's intercept is 82.98.

Performance Metrics:

- Mean Squared Error (MSE): The mean squared error of the model is 20.75, indicating the average squared difference between the observed actual outcomes and the predictions.
- Coefficient of Determination (R^2): The R^2 value is 0.78, suggesting that the model explains 78% of the variance in the target variable usr.

These results reflect the model's effectiveness in capturing the relationship between system attributes and CPU usage in user mode. The relatively high R^2 value indicates a good fit to the data, while the MSE provides insight into the model's accuracy.



(fig 10 : Plot showing the residual plot created to assess the fit of the regression model)

Model Summary : R-squared: 0.789: This indicates that approximately 78.9% of the variability in usr can be explained by the model. This is a high value, suggesting the model fits the data well.

Adj. R-squared: 0.788: This adjusts the R-squared value for the number of predictors in the model. It's very close to the R-squared value, indicating that the model's complexity is appropriate.

2.4 Business Insights & Recommendations

Coefficients

Each coefficient represents the estimated change in the dependent variable (usr) for a one-unit change in the corresponding independent variable, holding all other variables constant.

Intercept (82.97526974636537): This is the expected value of usr when all independent variables are zero. It represents the baseline usr value.

Individual Coefficients:

The coefficients are listed in the same order as the features in the model. Here's a general interpretation for each

* lread: -0.0558: For each unit increase in lread, usr is expected to decrease by approximately 0.0558 units.

* lwrite: 0.0412: For each unit increase in lwrite, usr is expected to increase by approximately 0.0412 units.

* scall: -0.0007: For each additional system call, usr decreases by 0.0007 units.

* sread: 0.0023: For each unit increase in sread, usr increases by approximately 0.0023 units.

* swrite: -0.0060: For each unit increase in swrite, usr decreases by 0.0060 units.

* fork: -0.0838: For each additional fork, usr decreases by approximately 0.0838 units.

* exec: -0.2670: For each additional exec, usr decreases by approximately 0.2670 units.

* rchar: -5.182e-06: For each unit increase in rchar, usr decreases by a very small amount.

* wchar: -4.966e-06: For each unit increase in wchar, usr decreases by a very small amount.

* pgout: -0.4823: For each unit increase in pgout, usr decreases by approximately 0.4823 units.

* ppgout: -0.0504: For each unit increase in ppgout, usr decreases by approximately 0.0504 units.

* pgfree: 0.1125: For each unit increase in pgfree, usr increases by approximately 0.1125 units.

* pgscan: -1.110e-16: This coefficient is extremely close to zero, indicating negligible impact.

* atch: 0.6205: For each unit increase in atch, usr increases by approximately 0.6205 units.

* pgin: 0.0384: For each unit increase in pgin, usr increases by approximately 0.0384 units.

* ppgin: -0.0684: For each unit increase in ppgin, usr decreases by approximately 0.0684 units.

* pflt: -0.0320: For each unit increase in pflt, usr decreases by approximately 0.0320 units.

* vflt: -0.0063: For each unit increase in vflt, usr decreases by approximately 0.0063 units.

* runqsz: 1.8131: For each unit increase in runqsz, usr increases by approximately 1.8131 units.

* freemem: -0.0005: For each unit increase in freemem, usr decreases by approximately 0.0005 units.

* freeswap: 9.402e-06: For each unit increase in freeswap, usr increases by a very small amount.

Model Evaluation Metrics

- Mean Squared Error (MSE): 20.75: This represents the average squared difference between the observed actual outcomes and the predictions. Lower values indicate a better fit, with 20.75 being the average squared error in the predictions.

- Coefficient of Determination (R^2): 0.78: This value indicates that approximately 78% of the variance in usr can be explained by the model. A value of 0.78 suggests a good fit, indicating the model explains a substantial portion of the variability in the data.

I**nterpretation and Considerations**

The positive and negative signs of the coefficients indicate the direction of the relationship between each predictor and usr.

The magnitude of each coefficient indicates the strength of the effect, with larger absolute values indicating stronger relationships.

Some coefficients are very small, suggesting minimal impact on usr.

The model explains a significant portion of the variance in usr, but there is still 22% unexplained variability, which might be due to factors not included in the model, measurement error, or random noise.

3. Problem 2: Predicting Contraceptive Method Choice

3.1 Problem Definition & Exploratory Data Analysis (EDA)

Problem Definition

Context

As a statistician at the Republic of Indonesia Ministry of Health, you are tasked with analyzing a dataset from the Contraceptive Prevalence Survey. This dataset includes detailed information from 1,473 married females who were either not pregnant or unsure of their pregnancy status at the time of the survey.The primary objective of this analysis is to develop a predictive model that

can forecast whether these women will choose to use a contraceptive method. This prediction will be based on a thorough examination of their demographic and socio-economic attributes, including factors such as age, education level, number of children, religion, employment status, and exposure to media.

By leveraging statistical and machine learning techniques, the goal is to identify patterns and relationships within the data that can inform policy decisions and health interventions aimed at improving contraceptive uptake and family planning efforts.

Exploratory Data Analysis (EDA):

Data Description:

- 1. Wife's age (numerical)
- 2. Wife's education (categorical) 1=uneducated, 2, 3, 4=tertiary
- 3. Husband's education (categorical) 1=uneducated, 2, 3, 4=tertiary
- 4. Number of children ever born (numerical)
- 5. Wife's religion (binary) Non-Scientology, Scientology
- 6. Wife's now working? (binary) Yes, No
- 7. Husband's occupation (categorical) 1, 2, 3, 4(random)
- 8. Standard-of-living index (categorical) 1=verlow, 2, 3, 4=high
- 9. Media exposure (binary) Good, Not good
- 10. Contraceptive method used (class attribute) No,Yes

Data Overview :

The dataset has 1473 rows and 10 columns. It is always a good practice to view a sample of the rows. A simple way to do that is to use head() function.

	Wife_age	Wife_ education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_	Media_exposure	Contraceptive_method_used
0	24	Primary	Secondary	3	Scientology	No	2	High	Exposed	No
1	45	Uneducated	Secondary	10	Scientology	No	3	Very High	Exposed	No
2	43	Primary	Secondary	7	Scientology	No	3	Very High	Exposed	No
3	42	Secondary	Primary	9	Scientology	No	3	High	Exposed	No
4	36	Secondary	Secondary	8	Scientology	No	3	Low	Exposed	No

(Table 4: Top five rows of the dataset for Problem2)

Descriptive Statistics :

	Wife_age	No_of_children_born	Husband_Occupation
count	1402.000000	1452.000000	1473.000000
mean	32.606277	3.254132	2.137814
std	8.274927	2.365212	0.864857
min	16.000000	0.000000	1.000000
25%	26.000000	1.000000	1.000000
50%	32.000000	3.000000	2.000000
75%	39.000000	4.000000	3.000000
max	49.000000	16.000000	4.000000

(Table 5: Descriptive statistics of the dataset for Problem2)

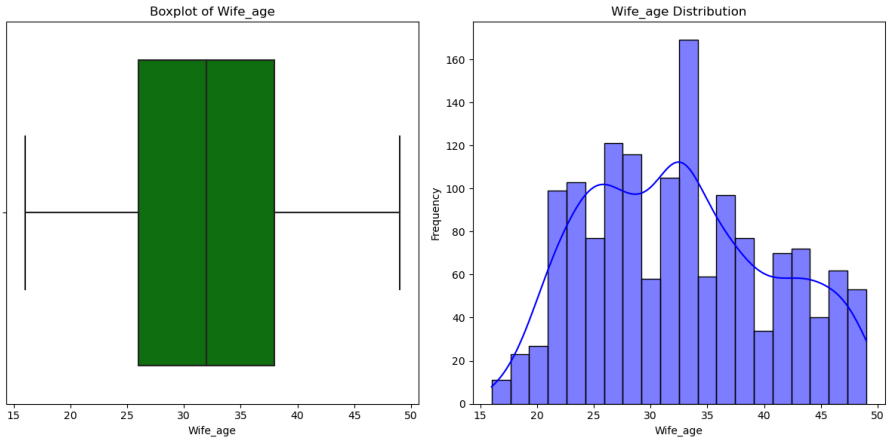
Observations:

- **Wife's Age**
  - **Average Age:** The mean age of the wives is approximately 32.6 years, indicating a mature sample group.
  - **Age Range:** The dataset includes women ranging from 16 to 49 years old, which covers a broad span of reproductive age.
  - **Age Distribution:** The concentration of ages between 26 and 39 years suggests that most women fall within their prime reproductive years, but there is also significant representation across both younger and older age groups.
- **Number of Children Ever Born**
  - **Average Number:** The average number of children ever born is 3.25, reflecting a moderately large family size.
  - **Wide Range:** The range of 0 to 16 children shows considerable variability, with some women having no children and others having a large number of children.
  - **Central Tendency:** The majority of women have between 1 and 4 children, indicating that while many have smaller families, there are also some with significantly larger families.
- **Husband's Occupation**
  - **Average Occupation Category:** The average occupation category is around 2.14, suggesting a mix of occupations primarily in the lower to mid-level categories.
  - **Occupation Distribution:** The range from 1 to 4 shows diversity in the types of occupations held by husbands, with a balanced representation across various occupation levels.

Univariate Analysis

For performing Univariate analysis we will take a look at the Boxplots and Histograms to get a better understanding of the distributions.

Observations on Wife_age



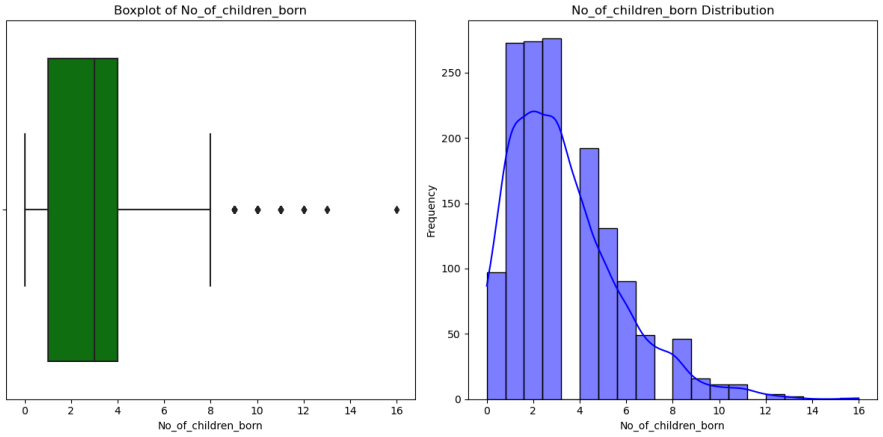
(fig 11 : Plot showing the boxplot and distribution for Wife Age for Problem 2)

Observations:

● The Distribution of Wife Age is normally distributed.

- There are no outliers in this variable.

Observations on No_of_children_born

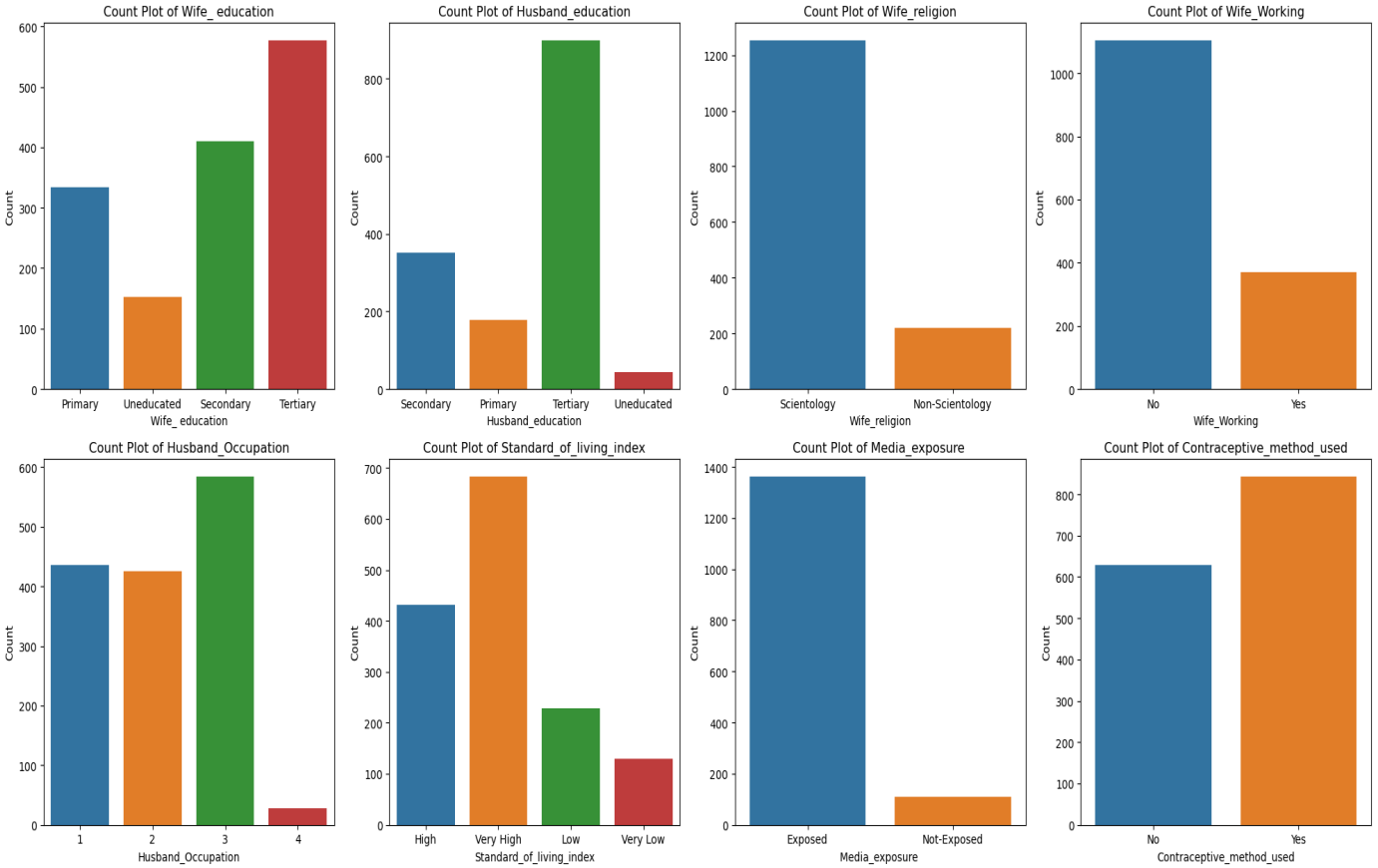


(fig 12 : Plot showing the boxplot and distribution for No_of_children_born for Problem 2)

Observations:

- The Distribution of Wife Age is skewed to the right
- There are af ew outliers in this variable.

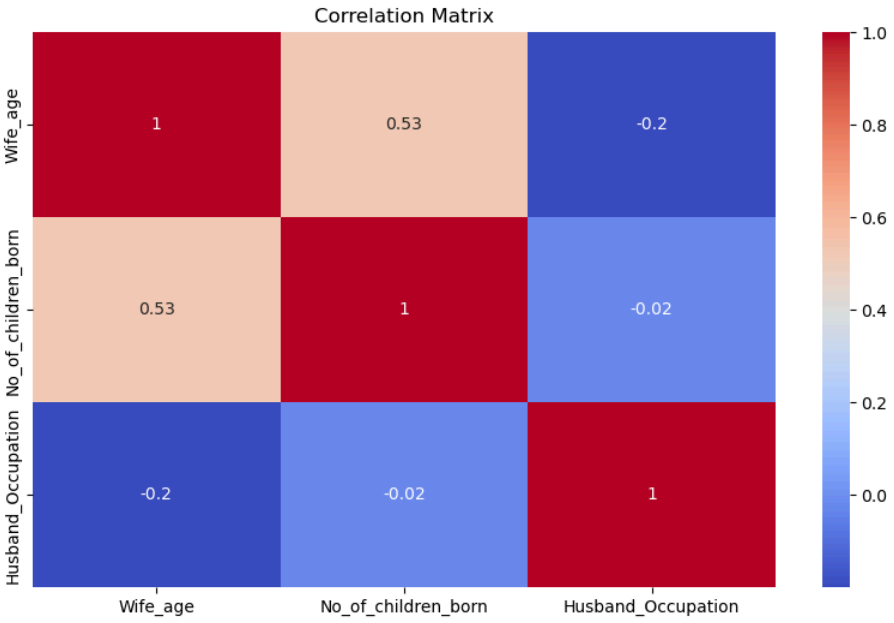
Count plots for categorical variables



(fig 13 : Plot showing the count plot of the categorical varables for Problem 2)

Multivariate Analysis

Correlation of Numerical Variables



(fig 14: showing heatmap of Correlation of numerical variables for Problem 2)

Observations

- Wife_age and No_of_children_born have a moderate positive correlation of 0.527. This suggests that as the age of the wife increases, the number of children born tends to increase as well. This could indicate that older wives might have had more time to have children.
- Wife_age and Husband_Occupation have a slight negative correlation of -0.199. This implies a weak inverse relationship between the wife's age and the husband's occupation, meaning that as the wife's age increases, the occupation of the husband has a slight tendency to be lower, though the relationship is weak.
- No_of_children_born and Husband_Occupation have a very weak negative correlation of -0.020. This indicates that there is almost no linear relationship between the number of children born and the husband’s occupation.

3.2 Data Preprocessing

Missing Values & Outliers: Detection and treatment.

Check for missing values： There are two variables “Wife_age” and “No_of_children_born” that has missing values of 71 and 21 respectively

To ensure the integrity of the dataset and maintain the accuracy of our analysis, missing values in key numerical variables were addressed as follows:

- **Wife's Age**: Missing values in the 'Wife_age' column were imputed with the mean value of the non-missing data. This approach assumes that the missing data is missing completely at random and the mean provides a reasonable estimate of the missing values.
- **Number of Children Born**: Missing values in the 'No_of_children_born' column were imputed with the median value of the non-missing data. Using the median is preferred in this case as it is less sensitive to outliers and provides a more robust estimate of central tendency.

Outlier Detection and Treatment

During the analysis, it was observed that:

- **Wife's Age**: There were no significant outliers detected in this variable. Therefore, no adjustments were necessary for this feature.
- **Number of Children Born**: While some values may appear extreme, they are not deemed significant outliers. As a result, no treatment for outliers was applied.

The decision to retain the existing data without adjustments ensures that the analysis remains accurate and reflective of the original dataset's characteristics.

Data Encoding: Methods for encoding categorical variables.

Encoding of Categorical Variables

To prepare the dataset for analysis, the following encoding techniques were applied to convert categorical variables into a numerical format:

Binary Variables: Binary categorical variables, including Wife_religion, Wife_Working, Media_exposure, and Contraceptive_method_used, were label-encoded. This transformation converts the binary categories into numerical values (0 and 1), allowing them to be used in the modeling process.

Categorical Variables with Multiple Categories: For categorical variables with more than two categories, such as Wife_education, Husband_education, Husband_Occupation, and Standard_of_living_index, one-hot encoding was applied. This method creates new binary variables for each category, enabling the model to interpret and utilize these categorical features effectively.

By applying these encoding methods, the dataset was prepared for subsequent analysis and modeling, ensuring that all variables are appropriately represented in numerical form.

Train-Test Split

Data Splitting

To evaluate the performance of the predictive model, the dataset was divided into training and testing subsets. This was achieved through the following steps:

Feature and Target Variable Separation: The dataset was split into features (X) and the target variable (Contraceptive_method_used), which indicates whether a contraceptive method was used.

Training and Testing Split: The features and target variable were further divided into training and testing sets using an 80-20 split. This means that 80% of the data was allocated for training the model, while the remaining 20% was reserved for testing and evaluating the model's performance.

Random Seed: A random seed of 42 was set to ensure reproducibility of the results, allowing for consistent and comparable results across different runs.

This approach ensures that the model is trained on a substantial portion of the data and evaluated on an independent subset, providing a robust assessment of its predictive accuracy.

	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
	Training Set	Training Set	Training Set	Training Set	Training Set	Training Set	Training Set	Training Set	Testing Set	Testing Set	Testing Set	Testing Set	Testing Set	Testing Set	Testing Set	Testing Set
Wife_age	1178	3.29E+01	8.165499	16	26	32.606277	39	4.90E+01	2.95E+02	3.14E+01	7.592324	16	25	31	36	4.90E+01
No_of_children_born	1178	3.33E+00	2.416528	0	2	3	5	1.60E+01	2.95E+02	2.93E+00	2.025651	0	1	3	4	1.10E+01
Wife_religion	1178	8.54E-01	0.353266	0	1	1	1	1.00E+00	2.95E+02	8.37E-01	0.36973	0	1	1	1	1.00E+00
Wife_Working	1178	2.56E-01	0.436337	0	0	0	1	1.00E+00	2.95E+02	2.31E-01	0.421874	0	0	0	0	1.00E+00
Media_exposure	1178	7.98E-02	0.271093	0	0	0	0	1.00E+00	2.95E+02	5.08E-02	0.220059	0	0	0	0	1.00E+00

(Table 6 : Summary statistics for the training set for Problem2)

3.3 Model Building

Model Evaluation for Logistic Regression

Accuracy: The logistic regression model achieved an accuracy of approximately 65.4% on the test set. This means the model correctly predicted the outcome for about 65.4% of the instances.

Confusion Matrix:

- True Negatives (TN): 61
- False Positives (FP): 69
- False Negatives (FN): 33
- True Positives (TP): 132

Classification Report:

- **Precision**:
  - Class 0: 0.65
  - Class 1: 0.66

Precision indicates the proportion of true positive predictions among all positive predictions. The model is slightly better at predicting class 1 correctly compared to class 0.

- **Recall**:
  - Class 0: 0.47
  - Class 1: 0.80

Recall reflects the proportion of actual positives that were correctly identified. The model has higher recall for class 1, suggesting it identifies more instances of this class accurately.

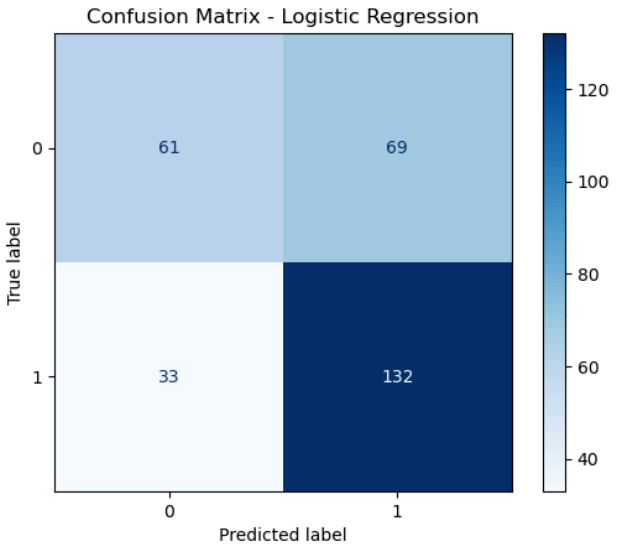
- **F1-Score**:
  - Class 0: 0.54
  - Class 1: 0.72

The F1-score, which balances precision and recall, is higher for class 1, indicating better overall performance in classifying this category.

Overall Performance:

- The model's performance is moderate with a balance between precision and recall. It performs better at predicting class 1 (which could represent a positive outcome) compared to class 0 (which could represent a negative outcome).
- The confusion matrix shows a significant number of false positives and false negatives, suggesting that while the model has a decent accuracy, there is room for improvement in distinguishing between the two classes.

This evaluation indicates that while the logistic regression model is effective, further tuning or alternative models might improve classification performance.



(fig 15 showing confusion matrix of Logistic Regression for Problem 2)

Model Evaluation for Linear Discriminant Analysis (LDA)

Accuracy: The LDA model achieved an accuracy of approximately 66.1% on the test set. This indicates that the model correctly predicted the outcome for about 66.1% of the instances.

Confusion Matrix:

- True Negatives (TN): 62
- False Positives (FP): 68
- False Negatives (FN): 32
- True Positives (TP): 133

Classification Report:

- **Precision:**
  - Class 0: 0.66
  - Class 1: 0.66

Precision measures the accuracy of the positive predictions. The LDA model has similar precision for both classes, suggesting it is equally effective at predicting each class.

- **Recall**:
    - Class 0: 0.48
    - Class 1: 0.81

Recall reflects the model's ability to identify all actual positives. The model shows higher recall for class 1, meaning it identifies more instances of this class correctly compared to class 0.

- F1-Score:
    - Class 0: 0.55
    - Class 1: 0.73

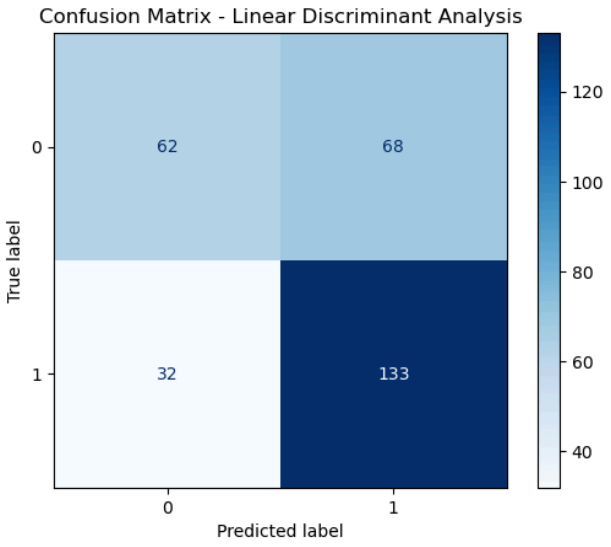
The F1-score balances precision and recall, and the LDA model performs better in classifying class 1, demonstrating improved overall performance in identifying this class.

Overall Performance:

The LDA model has a slightly higher accuracy compared to the logistic regression model. It performs better in classifying class 1 while showing improved precision and recall for this class.

The confusion matrix indicates that the model has fewer false positives and false negatives compared to the logistic regression model, which suggests better performance in distinguishing between classes.

In summary, LDA provides a modest improvement in classification accuracy and performance metrics over logistic regression, especially for class 1. This could indicate that LDA's approach to dimensionality reduction and class separation is beneficial for this dataset.



(fig 16: showing confusion matrix of Linear Discriminant Analysis for Problem 2)

Model Evaluation for Decision Tree Classifier (CART)

Accuracy: The CART model achieved an accuracy of approximately 64.1% on the test set. This means the model correctly predicted the outcome for around 64.1% of the instances.

Confusion Matrix:

- True Negatives (TN): 79
- False Positives (FP): 51
- False Negatives (FN): 55
- True Positives (TP): 110

Classification Report:

- **Precision**:
  - Class 0: 0.59
  - Class 1: 0.68

Precision measures how many of the positive predictions were correct. The CART model has better precision for class 1, meaning it is more accurate when predicting class 1 compared to class 0.

- **Recall**:
  - Class 0: 0.61
  - Class 1: 0.67

Recall indicates the model's ability to identify all actual positives. The CART model has slightly higher recall for class 1, showing it is better at identifying instances of class 1 compared to class 0.

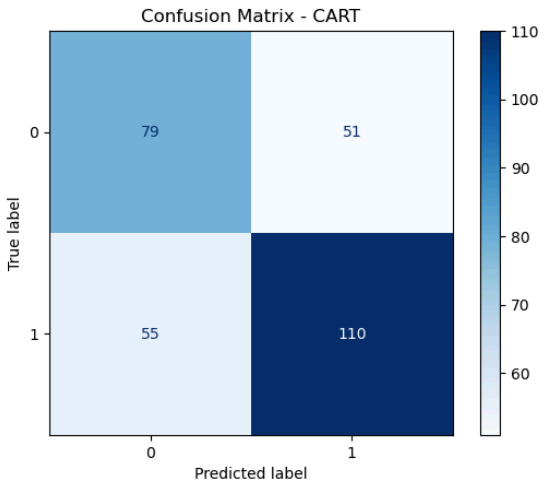
- **F1-Score**:
  - Class 0: 0.60
  - Class 1: 0.67

The F1-score balances precision and recall. The CART model performs better for class 1, reflecting an improved ability to classify this class effectively.

Overall Performance:

- The CART model has a lower accuracy compared to the LDA model but is comparable to the logistic regression model.
- It performs well in classifying class 1 with higher precision and recall, suggesting that the decision tree's approach to splitting based on feature values is beneficial for identifying class 1.
- The confusion matrix shows the model has a balance between false positives and false negatives, indicating reasonable performance across both classes.

In summary, the Decision Tree Classifier (CART) provides a competitive classification performance, particularly excelling in predicting class 1, though its overall accuracy is slightly lower compared to LDA.



(fig 17: showing confusion matrix of CART for Problem 2)

Optimized Decision Tree Classifier (CART) Performance

Accuracy: The optimized CART model achieved an accuracy of approximately 69.8% on the test set, showing an improvement over the initial model.

Confusion Matrix:

- True Negatives (TN): 53
- False Positives (FP): 77
- False Negatives (FN): 12
- True Positives (TP): 153

Classification Report:

- **Precision**:
    - Class 0: 0.82
    - Class 1: 0.67

Precision for class 0 improved significantly with the optimized model, indicating that the model is now better at correctly predicting class 0.

- **Recall**:
    - Class 0: 0.41
    - Class 1: 0.93

Recall for class 1 improved substantially, suggesting the model is more effective at identifying class 1 instances. However, recall for class 0 decreased, indicating that while the model is better at predicting class 1, it may miss some class 0 cases.

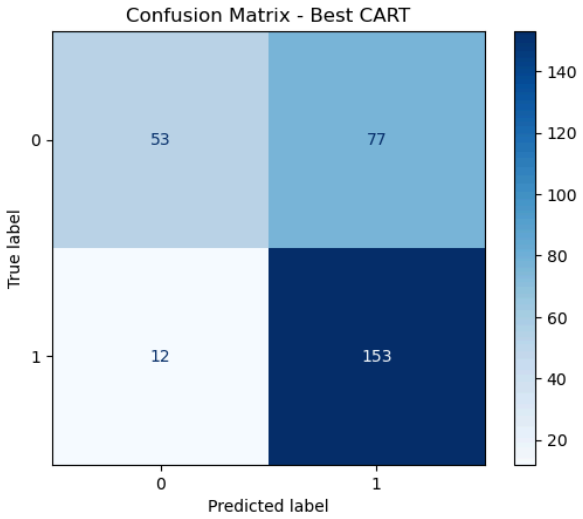
- **F1-Score**:
    - Class 0: 0.54
    - Class 1: 0.77

The F1-score, which balances precision and recall, reflects improved performance for class 1, though performance for class 0 is relatively lower.

Overall Performance:

- The optimized CART model shows improved accuracy and better classification performance for class 1 compared to the initial model.
- Precision for class 0 has increased, but recall for class 0 has decreased, indicating a trade-off where the model now has a higher likelihood of missing class 0 instances.
- The optimization process has yielded a model that provides a more balanced trade-off between precision and recall for class 1, but attention is needed to address the reduced recall for class 0.

In summary, the Grid Search optimization has led to a significant improvement in the Decision Tree Classifier's performance, especially in predicting class 1, although there is a trade-off with class 0 performance that should be considered.



(fig 18: showing confusion matrix of Best CART for Problem 2)

3.4 Business Insights & Recommendations

Key Takeaways

Age and Number of Children: These are typically important features in predicting contraceptive use, as observed in the feature importance ranking. Younger women or those with fewer children may be less likely to use contraceptive methods.

Education: Both wife's and husband's education levels are significant predictors. Higher education levels are often associated with increased awareness and use of contraceptive methods.

Standard of Living and Media Exposure: A higher standard of living and good media exposure are crucial. These factors indicate better access to information and resources related to contraceptive methods.

Employment Status: The working status of the wife plays a role. Employed women might have different family planning needs and access to contraceptive methods.

Religion: While the dataset may have binary encoding, cultural and religious beliefs significantly influence contraceptive use.

Actionable Insights and Recommendations

Targeted Awareness Programs: Develop and implement targeted awareness programs focusing on younger women and those with lower education levels. Use educational campaigns to highlight the benefits and availability of contraceptive methods.

Improve Accessibility: Enhance the accessibility of contraceptive methods in areas with lower standards of living. Subsidize contraceptive methods or provide them for free to low-income families.¶

Leverage Media Exposure: Use media channels effectively to disseminate information about family planning and contraceptive use. Partner with popular media outlets to run campaigns.

Workplace Initiatives: Introduce workplace family planning programs and resources. Encourage employers to provide information and access to contraceptive methods for working women.

Cultural Sensitivity: Develop culturally sensitive educational materials and programs that respect religious beliefs while promoting contraceptive use.

4. Conclusion

CPU User Mode Prediction

Model Performance: The linear regression model explained approximately 78.9% of the variance in 'usr', indicating a strong fit. The model's R-squared value and Adjusted R-squared are close, suggesting appropriate model complexity.

Key Insights:

Positive Influences: Variables like lwrite, sread, pgfree, atch, pgin, and runqsz positively impact 'usr'.

Negative Influences: Variables such as lread, swrite, fork, exec, pgout, and ppgout negatively affect 'usr'.

Negligible Effects: Some variables have minimal impact on 'usr', indicating they contribute less to explaining the variability in CPU user mode time.

Recommendation: The model effectively captures major influences on CPU user mode time. Further improvements could involve incorporating additional features or addressing potential measurement errors to capture the remaining unexplained variance.

Contraceptive Method Prediction

Model Performance: The Logistic Regression model achieved an accuracy of 65.4%, while the Linear Discriminant Analysis (LDA) model had a slightly higher accuracy of 66.1%. The Decision Tree Classifier (CART) initially performed at 64.1% accuracy but improved to 69.8% after hyperparameter tuning with GridSearchCV.

Key Metrics:

Logistic Regression: Precision, recall, and F1-scores varied, with lower recall for the non-contraceptive class and higher recall for the contraceptive class.

LDA: Showed better balance with higher recall for the contraceptive class and improved performance over logistic regression.

CART (Tuned): Provided the best performance with higher accuracy and better balance in precision and recall, especially for the contraceptive class.
Recommendation: The tuned CART model is recommended for its superior accuracy and balanced performance. Further analysis might focus on refining the model or exploring additional features to enhance predictive capability.

5. References

- **Previous jupyter notebooks**
- **Google**
- **Course Resources**
