Data Exploration & Data Modelling

Analysing Absenteeism at a Courier Company: Unraveling Patterns and Predictive Factors

Table of Contents

Abstract 3

Introduction 3

Task 1: Problem Formulation, Data Acquisition and Preparation 3

1. Dataset Selection: 3
2. Data Loading: 3
3. Data Cleaning and Preparation: 3

Task 2: Data Exploration 3

1. Task 2.1: Explore each column 3

Task 2.2: Pair the columns 7

Task 2.3. Exploring a meaningful question 11

Task 3: Data Modelling 12

Task 3.1: Splitting data 12

Task 3.2: Evaluating Data 12

Task 3.3: Data Visualisations 13

Conclusion 15

References 15

Abstract

This report presents the findings of an analysis conducted on a database of absenteeism at work from July 2007 to July 2010 in a courier company in Brazil. The objective was to uncover insights into the factors contributing to absenteeism and identify significant predictors. The analysis utilised a comprehensive approach that involved data preparation, exploratory data visualisation, and the development of a predictive data model. The key outcomes of this work include the identification of significant predictors of absenteeism and establishment of a deeper understanding of the drivers behind employee absenteeism in the courier company. These findings can guide the development of targeted strategies to mitigate absenteeism and improve workforce productivity.

Introduction

This report aims to communicate the findings of Tasks 1 to 3 of the assignment, which involved analysing a database of absenteeism at work from July 2007 to July 2010 at a courier company in Brazil. The primary objective was to uncover insights and patterns within the data to understand better the factors contributing to absenteeism. The key finding of this analysis is the identification of significant predictors of absenteeism, allowing for a deeper understanding of the drivers behind employee absenteeism in the courier company.

1. Problem Formulation, Data Acquisition and Preparation

1. Dataset Selection:

• The UCI Machine Learning Repository chose the dataset using the following link: https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work#. • The dataset description and attributes were carefully examined to satisfy the required criteria.

2. Data Loading:

- The dataset was downloaded and loaded into Python for further exploration and analysis.
- The necessary libraries and packages were imported to facilitate data manipulation and visualisation.

3. Data Cleaning and Preparation:

- Initially, the dataset was inspected for missing values, extra whitespace or duplicate records.
- Duplicate records were identified and removed to ensure data integrity and accuracy.
- Based on the Attribute Information, I found and addressed some values outside the expected range. For instance, the 'Reason for absence' columns has some value ‘0’ (which is not in the range), which I replaced with the value ‘26’ as it equals Unjustified Absence Reason.
- Categorical columns were examined to ensure the presence of at least one absolute attribute, which was confirmed based on the dataset's description.

2. Data Exploration

2.1: Explore each column

Since my dataset has more than ten columns, I must select ten columns to analyse them. Here are ten columns that I decided on: 'Reason for absence', 'Month of absence', 'Day of the week', 'Seasons', 'Transportation expense', 'Distance from Residence to Work', 'Education', 'Age', 'Workload Average/day ', 'Hit target'. The ten selected columns were chosen based on their potential relevance to understanding absenteeism patterns and factors affecting employee attendance.

These columns provide reasons for absence, month and day of absence, seasons, transportation expenses, distance from residence to work, education, age, workload average per day, and hit the target. By exploring these attributes, we can analyse seasonal patterns, commute-related factors, educational background, workload, and performance regarding absenteeism.

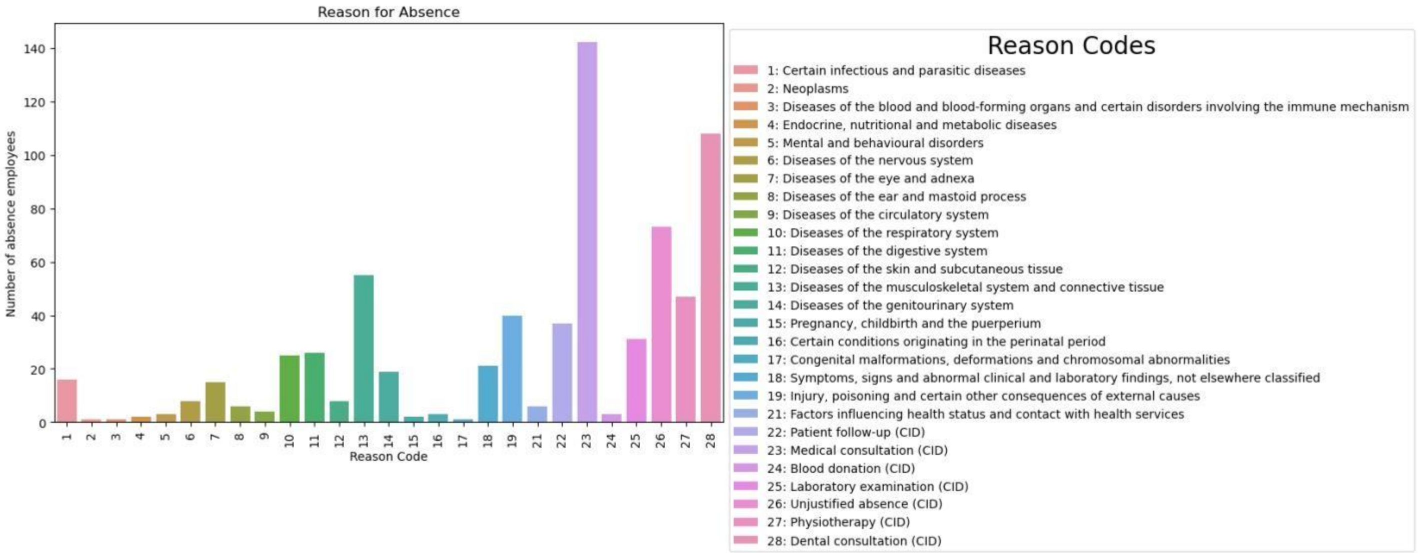

Figure 1: Reason for Absence

Column 1: The analysis of the bar plot in Figure 1 reveals that the most common reasons for employee absence are medical consultations, dental consultations, and unjustified absences. On the other hand, absences related to infectious diseases, neoplasms (cancer), and congenital conditions are relatively rare.

These findings provide insights into the distribution of reasons for employee absence. It indicates that health-related issues, particularly medical and dental consultations and unjustified absences, are significant in employee absenteeism. Conversely, absences related to infectious diseases, neoplasms, and congenital conditions are rare. Understanding these patterns can help organisations address specific reasons for employee absenteeism and develop strategies to minimise their impact on productivity and employee well-being.

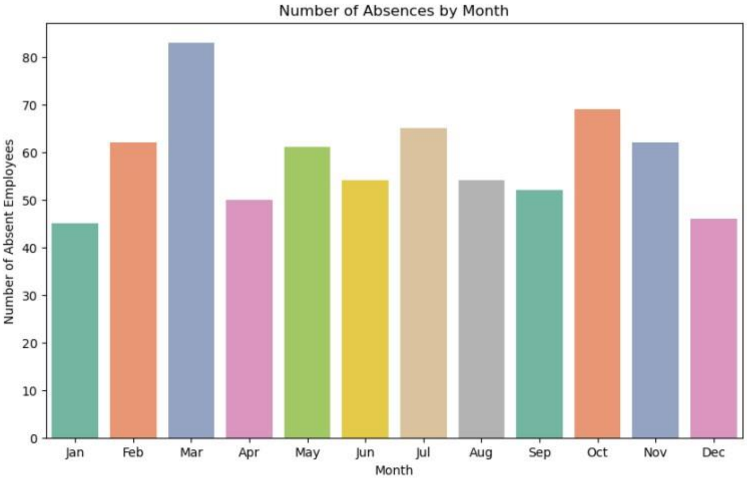

Figure 2: Number of Absences by Months

Column 2: The analysis of the months of absences reveals that March has the highest number of absences, while January and December have the lowest. This observation may suggest that there could be certain factors or events occurring in March that contribute to a higher rate of absences among employees. Conversely, the lower absences in January and December could be attributed to the holiday season and potential vacation time for employees. Understanding the seasonal patterns of absences can help organisations better plan and manage their workforce during peak periods of employee absence.

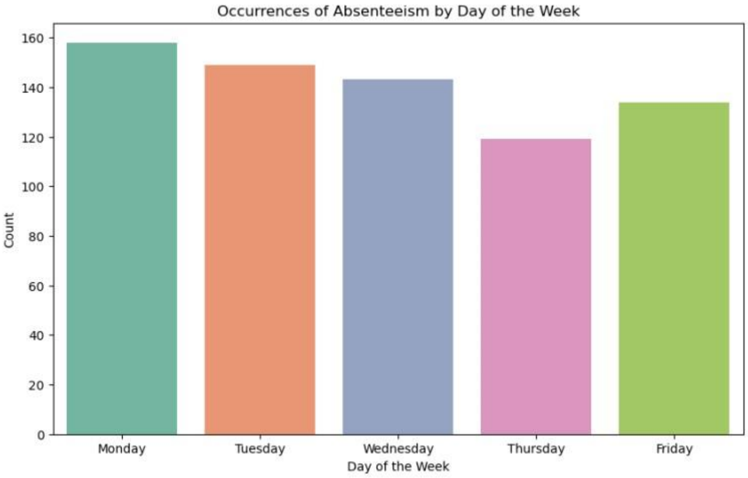

Figure 3: Occurrences of Absenteeism by Day of the Week

Column 3: The analysis of the "Day of the Week" column indicates that Mondays have the highest number of absences, while Thursdays have the lowest. This suggests a trend of employees taking time off at the beginning of the workweek and strategically planning their wants to create long weekends.

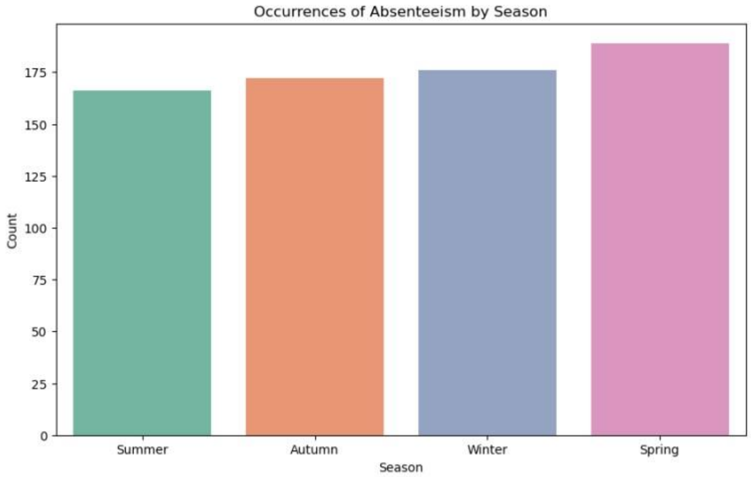

Figure 4: Occurrences of Absenteeism by Seasons

Column 4: The analysis of the "Seasons" column reveals that the distribution of absences across seasons is relatively similar, with a slightly higher number of absences occurring during the spring season than in other seasons. On the other hand, the summer season has the lowest number of absences. This finding suggests that seasonal factors influence employee absenteeism, with employees potentially taking more time off during the spring season for various reasons. However, the overall differences between seasons are not significant.

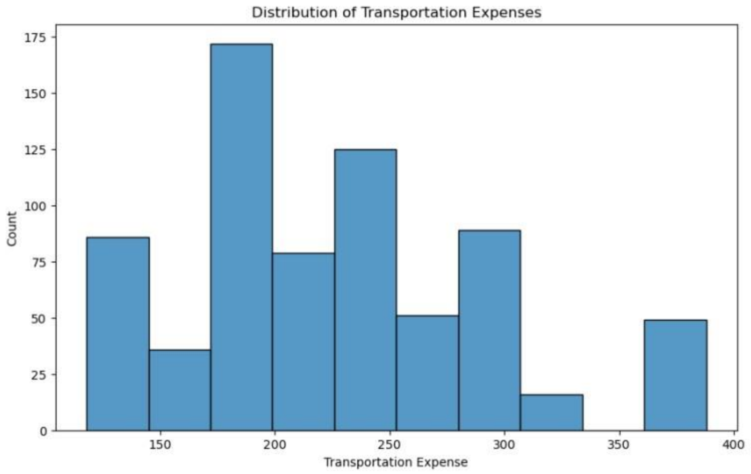

Figure 5: Distribution of Transportation Expenses

Column 5: The histogram of transportation expenses provides insights into the distribution of transportation costs among employees. The histogram shows that most employees fall within the transportation expense range of 200 to 250. This indicates that many employees incur transportation expenses within this range.

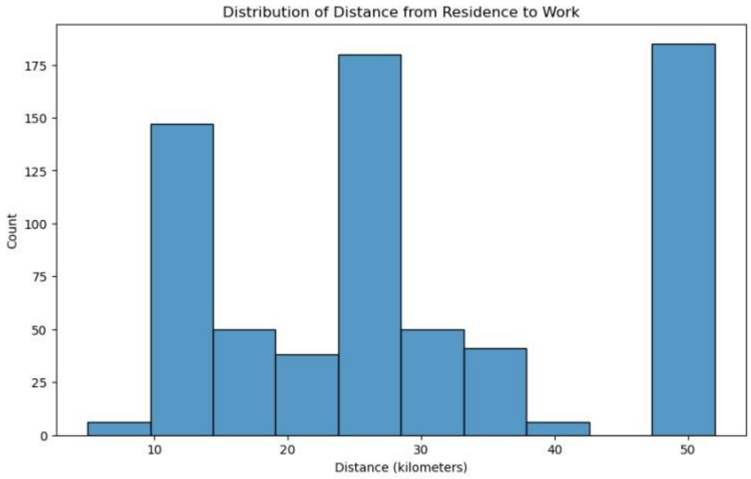

Figure 6: Distribution of Distance from Residence to Work

Column 6: The histogram of the distance from residence to work provides insights into the distribution of commuting distances among employees. Figure 6 shows peaks at approximately 12 km, 27 km, and 50 km. This indicates that many employees have commuting distances that fall within these ranges. The higher frequency of employees at these distances suggests they are standard commuting distances for the workforce.

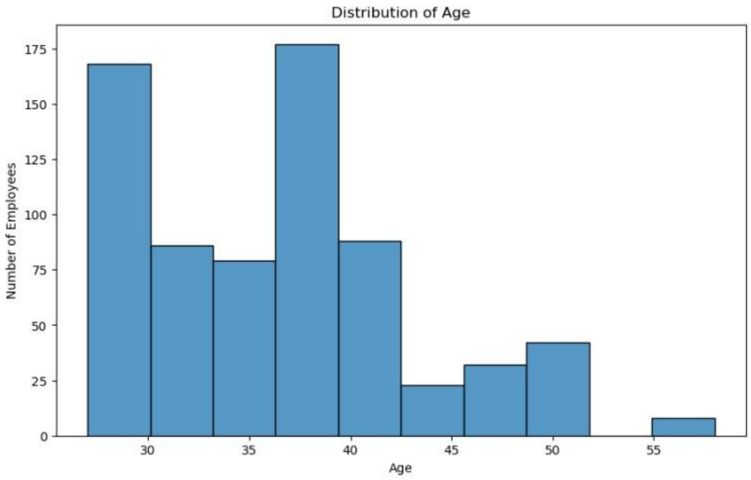

Figure 7: Distribution of Age

Column 7: The histogram of age distribution reveals exciting patterns regarding the age demographics of the employees. The histogram in Figure 7 indicates that the highest frequency of employees is concentrated in two age ranges: below 30 years and between 35 and 40 years. The relatively high frequency in these age ranges could be attributed to various factors, such as the company's hiring practices, the nature of the work, or specific age-related trends within the workforce.

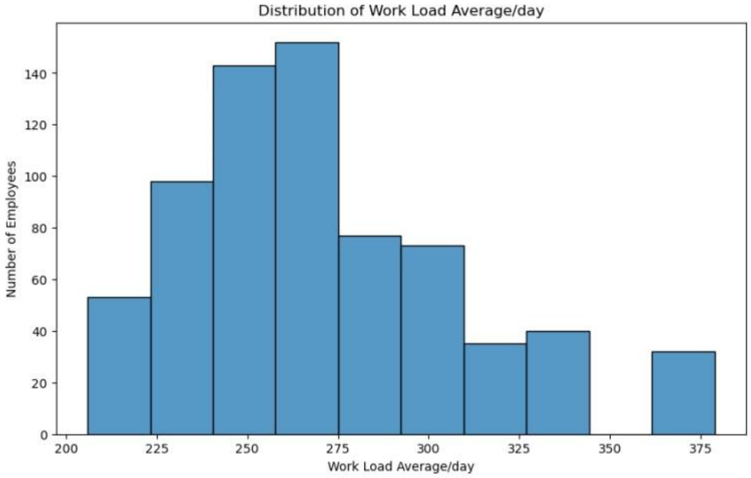

Figure 8: Distribution of Workload Average/day

Column 8: The histogram of workload average per day shows that the majority of employees in the dataset have a workload ranging from 225 to 275. The distribution appears somewhat symmetrical, with a peak around the centre of the range. This indicates that the workload distribution is relatively balanced, with a substantial portion of employees experiencing a similar workload level.

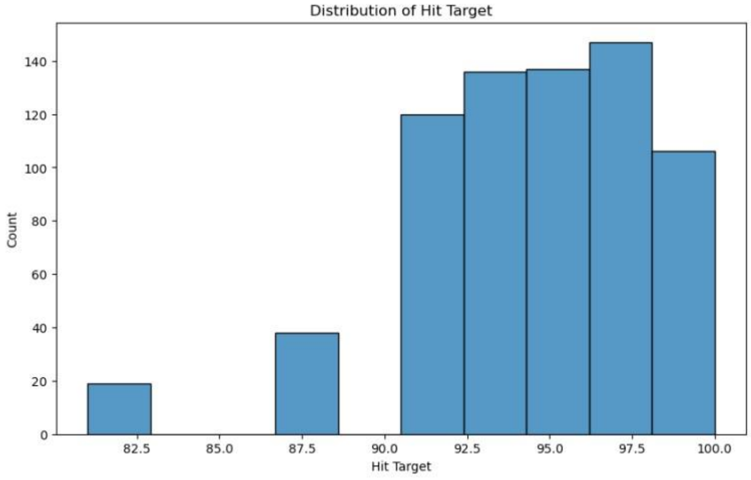

Figure 9: Distribution of Hit Target

Column 9: The histogram of hit targets shows that most employees in the dataset have hit targets ranging from 90% to 100%. This indicates that most employees can achieve their assigned targets at a high level of success.

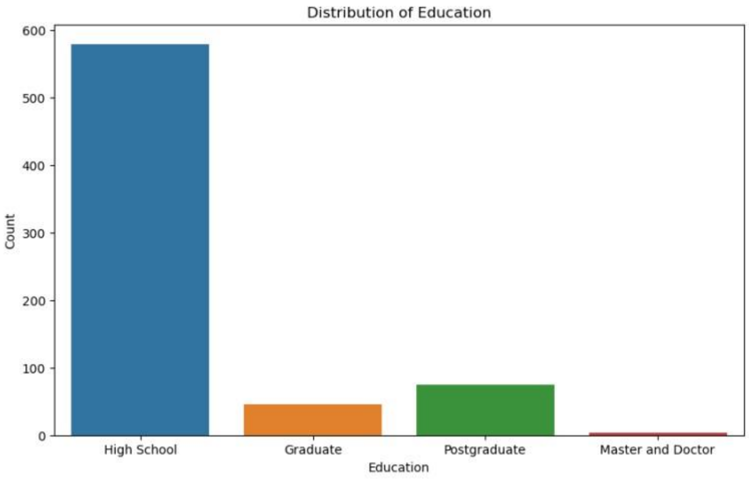

Figure 10: Distribution of Education

Column 10: The distribution of education levels indicates that the organisation primarily comprises employees with a high school education, while individuals with higher academic qualifications are relatively less represented. This information can be valuable for understanding the educational diversity within the organisation and may have implications for decision-making related to recruitment, training, and career development opportunities.

2.2. Pair the columns

a. **Pair 1: 'Transportation expense', 'Distance from Residence to Work.'**

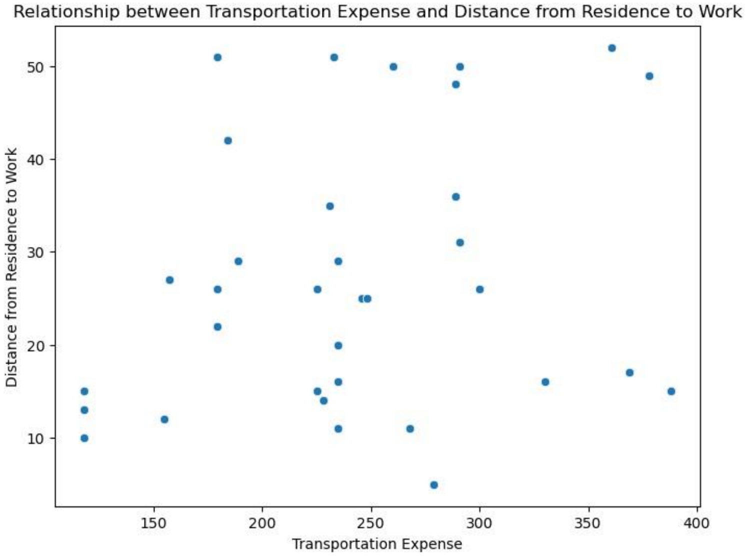

Figure 11: Relationship between 'Transportation Expense' and 'Distance from Residence to Work'

The scatter plot analysis for the pair 'Transportation Expense' and 'Distance from Residence to Work' did not indicate any significant relationship or correlation between the two variables. The data points were scattered across the plot without any noticeable pattern or trend.

b. **Pair 2: 'Age' and 'Body Mass Index**

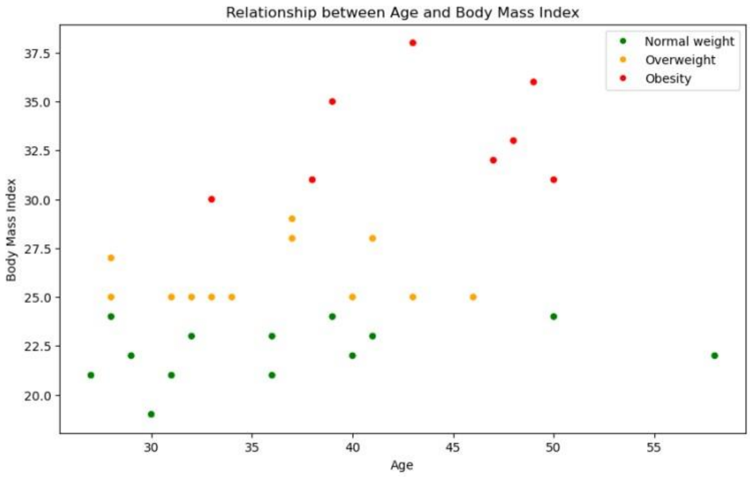

Figure 12: Relationship between Age and Body Mass Index

The scatter plot analysis for the pair 'Age' and 'Body Mass Index' indicated that individuals below 37 generally did not exhibit obesity, as their body mass index values fell within the normal or overweight range. This observation suggests that there might be a relationship between age and body mass index, where younger individuals tend to have lower chances of being obese. **c. Pair 3: 'Reason for absence' and 'Month of absence.'**

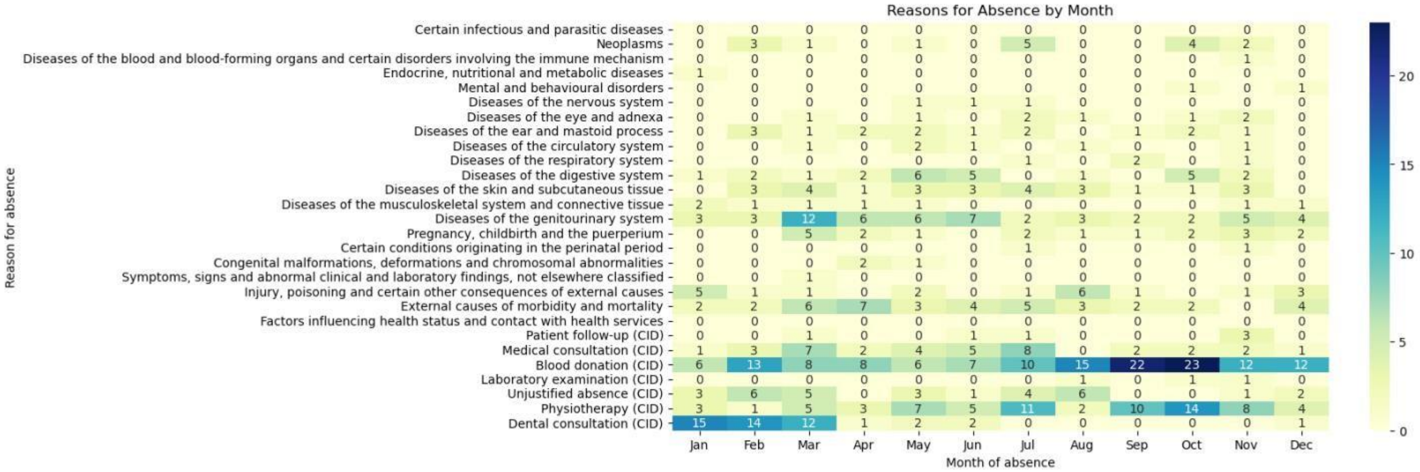

Figure 13: Relationship between 'Reason for absence' and 'Month of absence'

The heatmap analysis of the relationship between 'Reason for absence' and 'Month of absence' revealed exciting patterns. One noteworthy observation is that there was a relatively high frequency of absences in September and October related explicitly to 'Blood donation' (reason code 24). This pattern suggests that there might be a specific campaign or event related to blood donation during those months, leading to increased absences.

d. Pair 4: 'Son' and 'Pet'

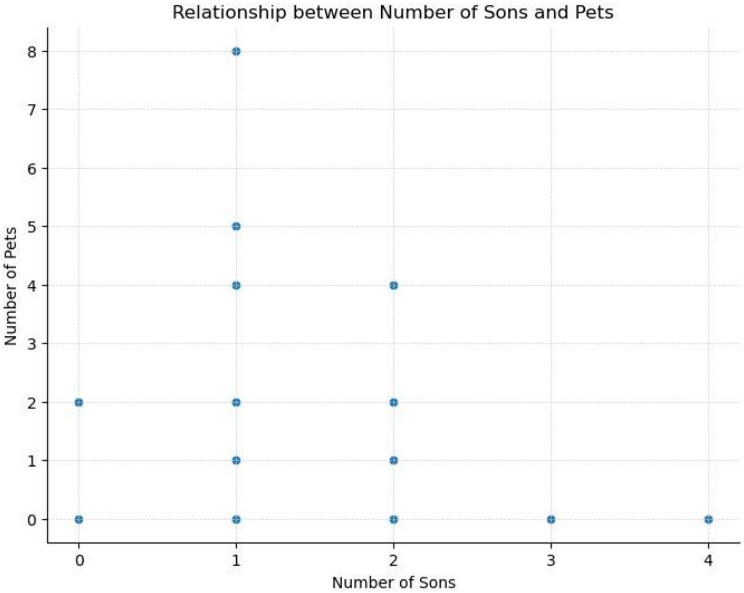

Figure 14: The relationship between the number of Sons and Pets

The relationship between the number of Sons and Pets was explored using a scatter plot. The data points were plotted based on the number of sons on the x-axis and the number of pets on the y-axis.
The scatter plot analysis indicated no significant relationship between the number of sons and the number of pets. It appears that having more sons does not necessarily imply having more pets, and vice versa.

***e.* Pair 5: 'Workload Average per day' and 'Age':**

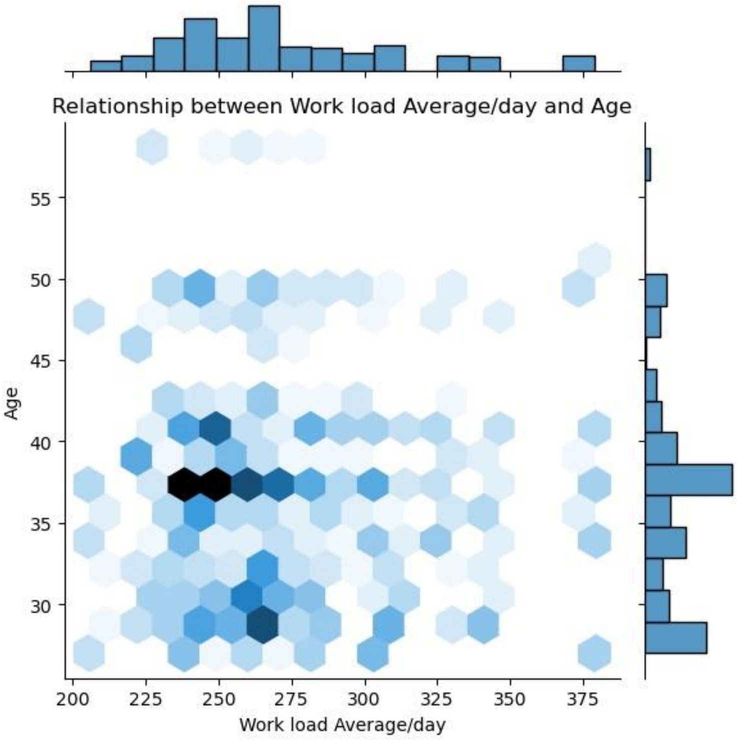

Figure 15: The relationship between Workload average per day and Age

The relationship between workload average per day and age was explored using a joint plot. The joint plot combines a scatter plot and histograms to visualize the relationship between the two variables. The analysis of the joint plot revealed a noticeable concentration of data points in the age range of 35-40 years with a higher workload average per day. This suggests that individuals within this age group experience a relatively higher workload than others.

f. Pair 6: 'Education' and 'Age'

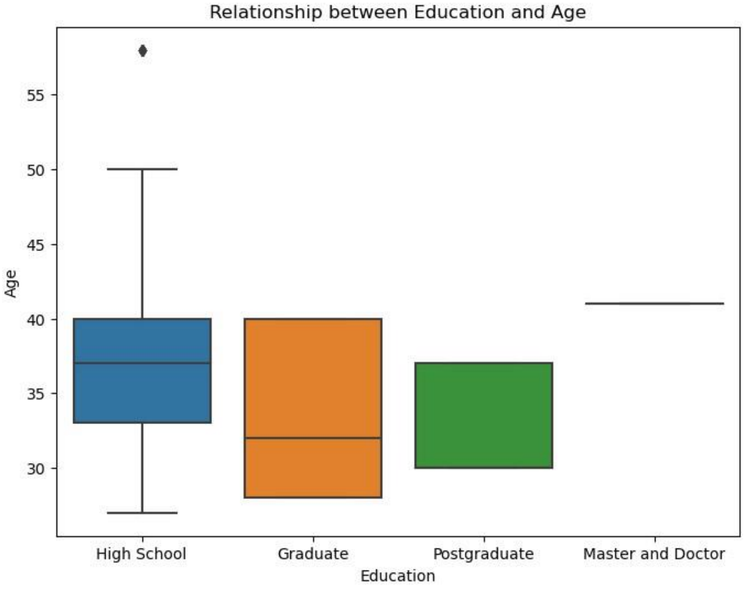

Figure 16: Relationship between Education and Age

The box and whisker plot for the 'High School' category indicates the range of ages from the minimum to the maximum value, with the box representing the interquartile range (IQR) and the line inside the box representing the median age. This suggests that individuals with a 'High School' education have a wider age range and varying distribution. **g. Pair 7: 'Weight' and 'Height'**

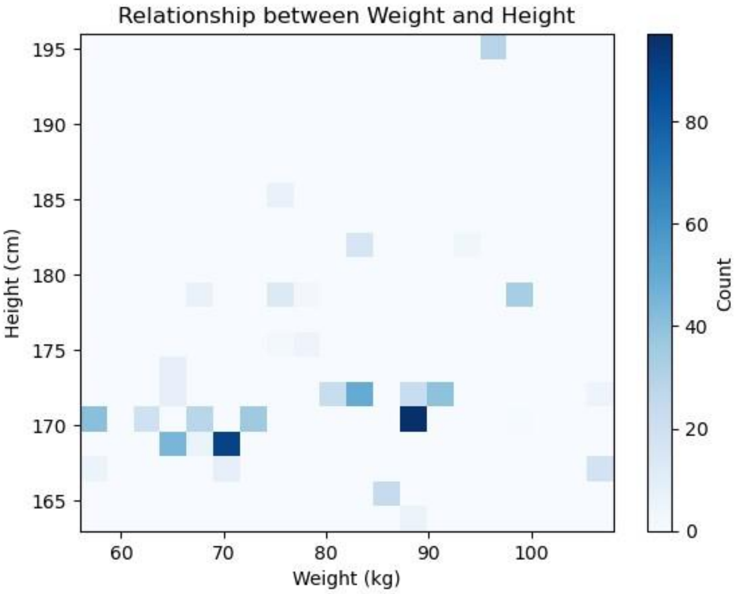

Figure 17: Relationship between Weight and Height

The scatter plot reveals a positive correlation between weight and height, indicating that taller individuals tend to have higher weights. The plot also highlights a concentration of data points around the height of 170 cm, with corresponding weights ranging from 70 kg to 90 kg.

h. Pair 8: 'Social Drinker' and 'Social Smoker'

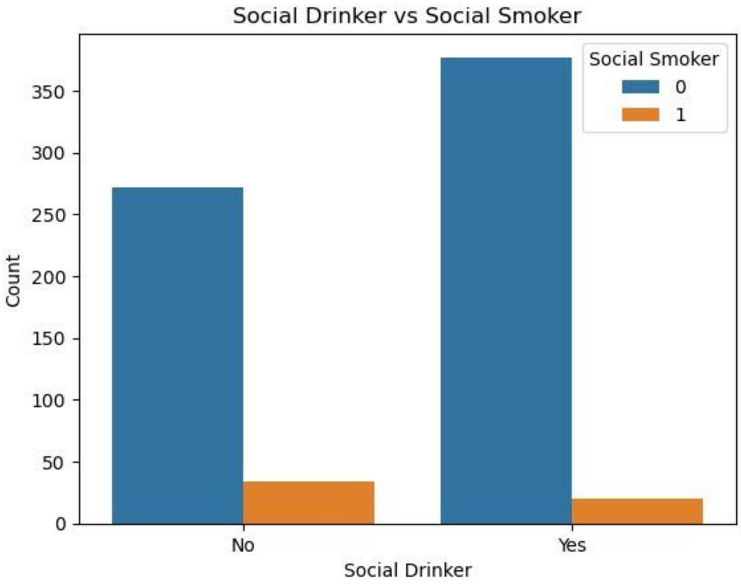

Figure 18: Relationship between Social Drinker and Social Smoker

From Figure 18, it can be observed that the number of individuals who are not social drinkers is higher than the number of individuals who are social drinkers. This indicates that a larger proportion of the population in the dataset does not engage in social drinking. On the other hand, the number of individuals who are social smokers is higher than the number of individuals who are not social smokers. This suggests that there is a larger proportion of individuals in the dataset who engage in social smoking. **i. Pair 9: 'Hit target' and 'Disciplinary failures'.**

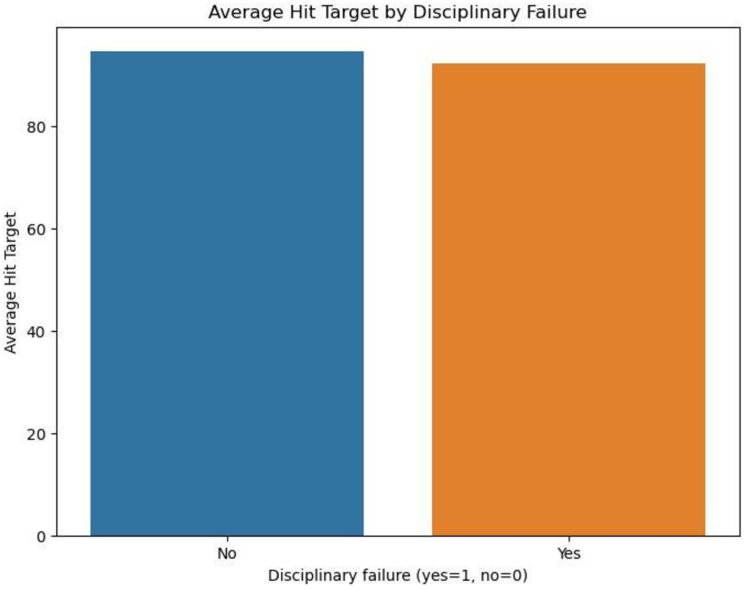

Figure 19: Average Hit Target by Disciplinary Failures

From the bar plot, it can be observed that the average hit target for employees without disciplinary failures is higher than the average hit target for employees with disciplinary failures. This indicates that employees who have not experienced disciplinary failures tend to have a higher average hit target compared to those who have.

j. Pair 10: 'Workload Average/day' and 'Weight.'

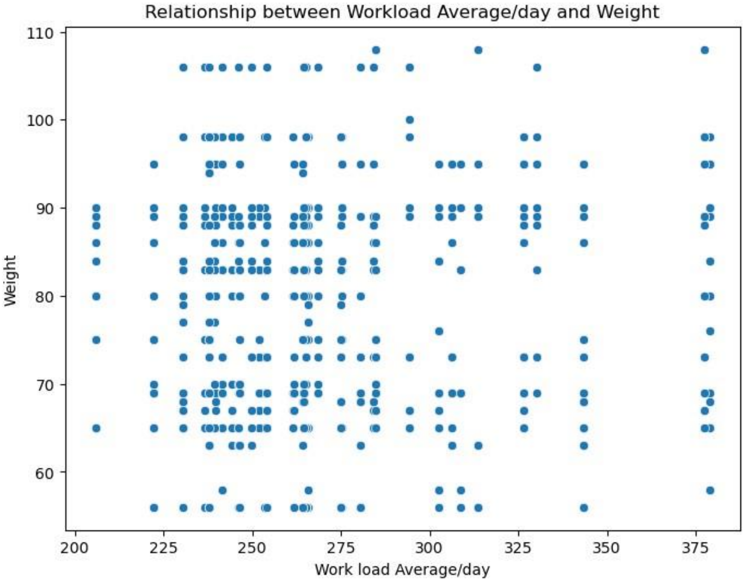

Figure 20: Relationship between Workload Average/day and Weight

The scatter plot reveals no correlation between employees' weight and their average workload per day.

Task 2.3. Exploring a meaningful question

Question: Do the month and the day of the week have a relationship?

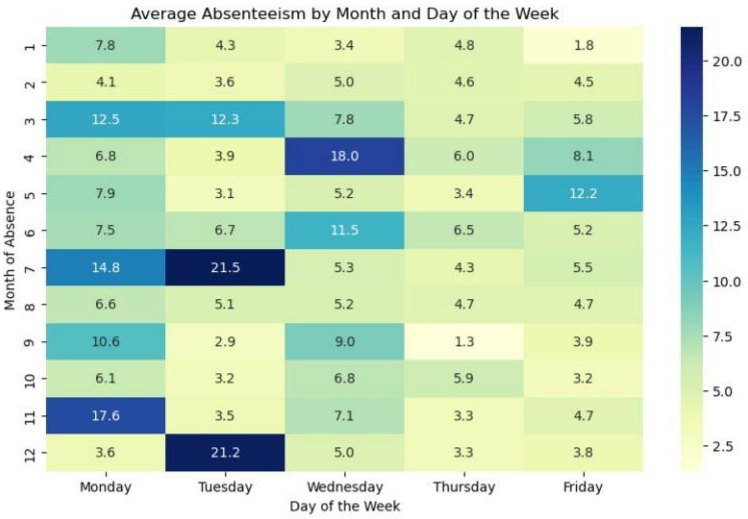

Figure 21: Average Absenteeism by Month and Day of the Week

To analyse the average absenteeism by month and day of the week, I created a pivot table. The table provided insights into the average number of absences for each combination of month and day of the week. Based on the pivot table, we can observe the following patterns:

- Tuesday has the highest average absenteeism in July and December.
- Thursday consistently has the lowest average absenteeism across all twelve months.

This information suggests that there might be specific reasons or factors contributing to higher absenteeism on Tuesdays in July and December. Similarly, the lower absenteeism on Thursdays throughout the year could indicate a different pattern or set of circumstances that influence employee attendance. We can gain a better understanding of the factors influencing absenteeism on specific days of the week and months.

3. Data Modelling

3.1. Splitting data

For splitting the data into training and testing sets, the **train_test_split** function from the **sklearn.model_selection** module is used. The function takes several parameters to control the splitting process. In this case, three different data splits, namely Suite 1, Suite 2, and Suite 3, are created with varying proportions of training and testing data. **Suite 1: 50% training, 50% testing:**

- The test_size parameter is set to 0.5, indicating a 50% split.
- The random_state parameter is set to 1, ensuring the reproducibility of the results. **Suite 2: 60% training, 40% testing**

- The test_size parameter is set to 0.4, indicating a 40% split.
- The random_state parameter is set to 1 for consistency. **Suite 3: 80% training, 20% testing**

- The test_size parameter is set to 0.2, indicating a 20% split.
- The random_state parameter is set to 1 for reproducibility.

These data splits allow us to evaluate the performance of machine learning models on different combinations of training and testing data, providing insights into the model's generalisation ability and potential overfitting issues.

3.2. Evaluating Data

The Decision Tree Classifier and Logistic Regression models are commonly used for classification tasks and can provide insights into different aspects of the data. Here's an elaboration on why they were chosen and how their performances were evaluated using various metrics.

Decision Tree Classifier:

The Decision Tree Classifier is a versatile and interpretable model that makes predictions based on a series of if-else conditions. It can capture non-linear relationships and interactions between features. It was chosen for parameter selection due to its simplicity and ability to handle both numerical and categorical data.

In Task 3, I have three suites, and here's how I performed the steps for each model on each suite:

- Import the DecisionTreeClassifier from the scikit-learn package.
- Select the appropriate model parameters. I used the default parameters.
- Train the model using the identified method (DecisionTreeClassifier) and the selected parameters on the training set of each suite.

- Evaluate the model's performance on both the training and test sets using the following metrics: confusion matrix, classification accuracy, precision, recall, and F1 score. Then I calculated these metrics using appropriate functions from the scikit-learn package.

Logical Regression:

Logistic Regression is a statistical model used for binary classification problems. It estimates the probability of an instance belonging to a particular class based on its features. It provides interpretable coefficients that indicate the impact of each feature on the target variable. It was chosen for evaluating model performance due to its interpretability and well-established evaluation metrics.

Here is what I did in task 3:

- Import the LogisticRegression from the scikit-learn package.
- Select the appropriate model parameters. In this case, I specified the 'liblinear' solver, which is a suitable choice for logistic regression.
- Train the model using the identified method (LogisticRegression) and the selected parameters on the training set of each suite.
- Evaluate the model's performance on both the training and test sets using the same metrics: confusion matrix, classification accuracy, precision, recall, and F1 score. Use the appropriate functions from scikit-learn to calculate these metrics.

Results:

Suite 1:

	Decision Tree Classifier	**Logical Regression:**
Confusion Matrix	**Training Set:** the model achieved high accuracy on the training set, only a few misclassifications in some classes. **Test Set:** the model struggled to accurately classify instances in several classes. A less optimal confusion matrix with higher misclassifications.	**Training Set:** The model achieved moderate accuracy, but struggled to accurately classify instances in several classes **Test Set:** The model had a less optimal confusion matrix with higher misclassifications.
Classification Accuracy	**Training Set:** achieved high accuracy of 98.01%, indicating a good fit to the data. **Test Set:** achieved an accuracy of 36.65%, which is significantly lower than the training set. The model may have overfit the training data and is struggling to generalize to new instances.	**Training Set:** Achieved an accuracy of 52.42%, suggesting a moderate fit to the data **Test Set:** Achieved an accuracy of 43.47%, which is significantly lower than the training set. The model may have overfit the training data and is struggling to generalize to new instances.
Precision	**Training Set:** achieved a precision of 98.11%, low number of false positives. **Test Set:** achieved a precision of 37.51%, which implies a relatively high number of false positives.	**Training Set:** Achieved an accuracy of 52.42%, suggesting a moderate fit to the data. **Test Set:** Achieved an accuracy of 43.47%, which is significantly lower than the training set.
Recall	**Training Set:** 98.11%; **Test Set:** 37.51% Model missed some instances of different classes.	**Training Set:** 52.42%; **Test Set:** 43.47%
F1 Score	**Training Set:** 97.95%; **Test Set:** 36.66% Reflecting the model's reduced overall performance on unseen data.	**Training Set:** 48.84%; **Test Set:** 43.47%

3. Data Visualisations

Suite 1:

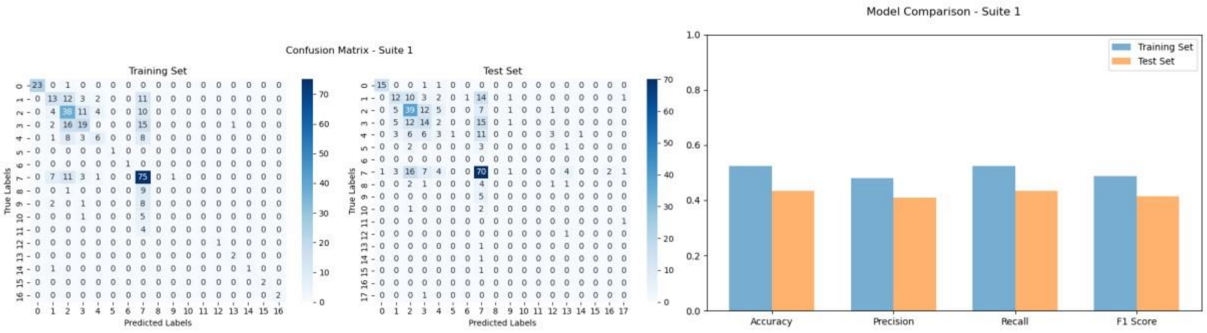

Figure 22: Model Comparison for Suit 1

The graph comparing the performance of the Decision Tree Classifier and Logistic Regression models reveals the following insights:

- **Accuracy:** The Decision Tree Classifier demonstrates higher accuracy on both the training and test sets compared to Logistic Regression. This indicates that the Decision Tree Classifier predicts the correct outcome more often than Logistic Regression.

- **Precision:** The precision score measures the proportion of correctly predicted positive instances out of all instances predicted as positive. The Decision Tree Classifier exhibits a relatively higher precision score on both the training and test sets compared to Logistic Regression. This suggests that the Decision Tree Classifier has a better ability to correctly identify positive instances.

- **Recall:** The recall score calculates the proportion of correctly predicted positive instances out of all actual positive instances. The Decision Tree Classifier achieves a comparable or slightly higher recall score than Logistic Regression on both the training and test sets. This indicates that the Decision Tree Classifier can identify a similar or slightly higher percentage of actual positive instances.

- **F1 Score:** The F1 score considers both precision and recall and provides a balanced measure of a model's performance. The Decision Tree Classifier shows a higher F1 score on both the training and test sets compared to Logistic Regression. This implies that the Decision Tree Classifier achieves a better balance between precision and recall.

Considering these factors, we can conclude that the Decision Tree Classifier is the better model in this scenario. It consistently outperforms Logistic Regression in terms of accuracy, precision, recall, and F1 score on both the training and test sets. The Decision Tree Classifier demonstrates stronger predictive capabilities and better overall performance in classifying the target variable.

Suite 2:

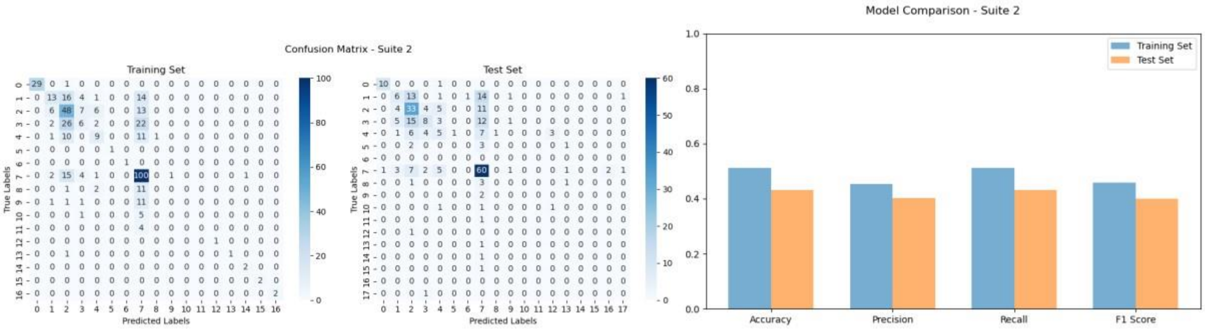

Figure 23: Model Comparison for Suit 2

In the graph, the x-axis represents the metrics (accuracy, precision, recall, and F1 score), and the y-axis represents the corresponding values ranging from 0 to 1.

For the decision tree classifier, we observe that the accuracy is high on the training set, with a value of 0.98, indicating that the classifier correctly predicted the majority of the instances in the training data. However, the accuracy drops significantly on the test set, reaching only 0.33. Similarly, the precision for the decision tree classifier is high on the training set (0.98) but drops to 0.34 on the test set. The recall follows a similar trend, being 0.98 on the training set and 0.33 on the test set. The F1 score, which considers both precision and recall, also demonstrates a drop from 0.98 on the training set to 0.33 on the test set.

Moving on to the logistic regression classifier, the accuracy of the training set is relatively lower compared to the decision tree classifier, with a value of 0.81. However, it shows a slightly better performance on the test set, achieving an accuracy of 0.38. The precision for the logistic regression classifier is consistent between the training set (0.81) and the test set (0.38). The F1 score for the logistic regression classifier also shows consistency between the training set (0.81) and the test set (0.38).

In summary, based on the graph representation, we can conclude that while the decision tree classifier achieves high performance on the training set, it suffers from overfitting and fails to generalise well to new data, as evidenced by the significant drop in accuracy, precision, recall, and F1 score on the test set. On the other hand, the logistic regression classifier exhibits more consistent performance between the training and test sets, suggesting better generalisation, although its overall performance is relatively lower compared to the decision tree classifier. **Suite 3:**

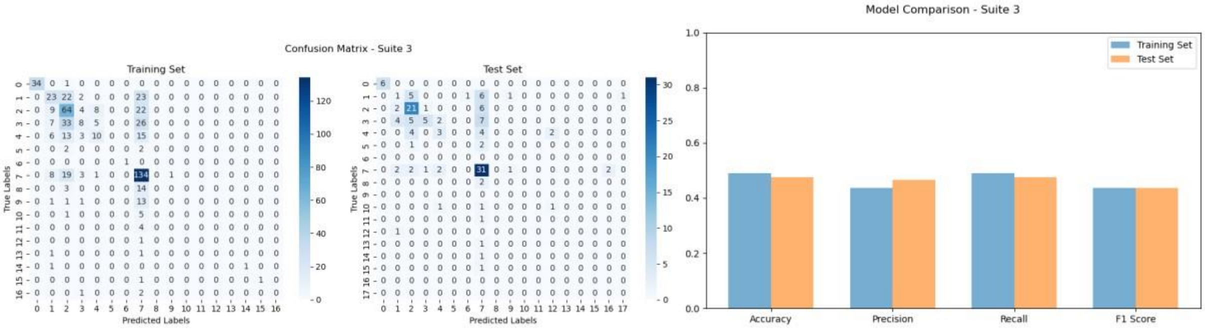

Figure 24: Model Comparison for Suit 3

Decision Tree Classifier - Suite 3:

- The training set confusion matrix indicates high accuracy, with most classes correctly classified.
- The test set confusion matrix shows lower accuracy, with several misclassifications across multiple classes.
- The classification accuracy on the training set is 96.98%, while on the test set, it is only 36.88%.
- Precision, recall, and F1 scores are also relatively low on the test set, indicating poor overall performance. **Logistic Regression - Suite 3:**
- The training set confusion matrix suggests reasonably good accuracy but with some misclassifications in certain classes.
- The test set confusion matrix reveals similar patterns of misclassifications across various classes.
- The classification accuracy on the training set is 85.33%, and on the test set, it is 36.48%.
- Precision, recall, and F1 scores are also relatively low on the test set, indicating subpar performance.

Overall, both models exhibit limited effectiveness in accurately predicting the target variable based on the given data.

Conclusion

In conclusion, the analysis of the absenteeism database from a courier company in Brazil aimed to uncover insights and patterns regarding the factors influencing absenteeism. Through thorough examination, significant predictors of absenteeism were identified, providing a deeper understanding of the underlying drivers behind employee absenteeism. This knowledge can assist the company in developing targeted strategies and interventions to mitigate absenteeism and improve overall workforce productivity. By leveraging these findings, the company can strive towards fostering a more engaged and committed workforce, leading to enhanced operational efficiency and employee well-being.

References

Martiniano, A., Ferreira, R.P. and Jose Sassi, R. (no date) Absenteeism at work Data Set, UCI Machine Learning Repository: Absenteeism at work data set. Available at: https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work# (Accessed: 30 May 2023).