Teamwork Success Predictor

Aliyu Bello

Supervisor: Dr. Aya Salama

Faculty of Computer Science

University of Prince Edward Island, Cairo Campus

July 2023

Acknowledgements

First and foremost, I am profoundly grateful to my loving family for their unwavering support throughout my academic pursuits. Their constant encouragement, understanding, and belief in my abilities have been invaluable. I would like to express my deepest appreciation to Dr. Aya Salama for her guidance and expertise as my project supervisor. Her insightful feedback, constructive criticism, and continuous support have been vital in shaping the direction and quality of my research. Additionally, I would also like to extend my sincere thanks to Omar Afifi and Ola Galal for their support in enhancing my understanding in this research. Their valuable suggestions and clarifications have greatly contributed to the improvement of my work.

Abstract

This thesis presents a comprehensive analysis of predicting team success toward milestone achievement. The research encompasses several stages, including data cleaning and processing, various statistical tests for goodness of fit and normality, feature engineering and selection, and the application of machine learning algorithms. The research has a great deal of potential to be pursued since as can meet practical needs, offer decision support, optimize resource allocation, decrease risks, increase continuous development, contribute to knowledge, and be advantageous in practical situations. To begin, extensive data cleaning and processing to ensure the quality and reliability of the dataset was conducted. Subsequently, employing a range of statistical tests to assess the goodness of fit and normality of the variables enables us to make informed decisions during the subsequent modeling phases. Feature engineering techniques were explored to improve the prediction performance to create additional meaningful features. The process involves transforming the existing variables, deriving new variables, and selecting the most relevant ones for the prediction task. It utilizes various machine-learning approaches to the dataset, such as logistic regression, decision trees, and random forests. A comparative evaluation identified the random forest as the most effective algorithm for predicting team success toward milestone achievement, as it consistently demonstrates superior accuracy across the provided datasets. To optimize the performance of the chosen machine learning model, hyperparameters tunning was performed with the essential parameters. This process ensures that the model is optimized to produce the best possible prediction accuracy. Finally, employment of the interpretability method known as LIME (Local Interpretable Model-Agnostic Explanations) to assess the prediction outcomes of specific instances. By using LIME, gaining insights into how the model makes predictions for selected instances was prioritized,

enhancing the understanding of the underlying factors contributing to team success. This thesis offers a comprehensive analysis of predicting team success toward milestone achievement, covering data cleaning, statistical tests, feature engineering, and machine learning algorithms. The findings highlight the effectiveness of tuned random forest model with an accuracy of 68%. This results to similar outcome from the previous research on the same dataset.

Abbreviations

RF - Random Forest

ML - Machine Learning

DT - Decision Tree

ADA - ADA Boost

LR - Logistic Regression

AI - Artificial Intelligence

TP - True Positives

FP - False Positives

TN - True Negatives

FN - False Negatives

PCA - Principal Component Analysis

LIME - Local Interpretable Model-agnostic Explanations

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and Problem Statement

Success in various fields, including business, academia, and research, has long been seen as significantly influenced by effective teamwork. For complicated objectives to be accomplished and complex issues to be resolved, the ability of teams to collaborate and interact is crucial. In order to reach an agreement and accomplish common goals, teamwork requires coordination of efforts and ongoing, constant negotiation [Reis, 2018].

The amount of study done to predict teamwork performance is inadequate. This offers convincing argument for exploring this area of study as it meets practical demands, provides decision support, optimizes resource allocation, reduces risks, encourages continuous development, adds to knowledge, and can have a beneficial effect in real-world settings. The potential applications of predicting group success using a statistical approach can be diverse, depending on the specific context and goals of the application like career guidance and placement, education policy and planning.

Teamwork and collaboration are commonly practiced in various settings, including workplaces, educational institutions, sports teams, community organizations, and research projects. In a study published in 2016, Belbin et al. examined the effect of teamwork on organizational success in the healthcare industry. The study's findings showed that cooperative teamwork improved healthcare

delivery efficiency and produced positive results. Working as a team means using diverse skills and abilities within a group while sharing information, resources, and responsibilities.

The Random Forest (RF) algorithm's advantages—and the reason [D. Petkovic, 2016] selected it—including its capacity to compute the variable significance measure, Mean Decrease Gini, and other metrics. In two ways, their research demonstrates results. *Predicting factors* generated the best-ranked Team Activity Measures (TAM) variables using Gini measures for the optimal Random Forest RF predictive models (i.e., operating points with maximum accuracy) and looked into whether they have an understandable explanation based on instructors' or any other experience [D. Petkovic, 2016]. These elements can provide practitioners with guidance. The outcomes demonstrate the Gini in random forest's most accurate prediction parameters. *Accuracy of Prediction*: They employed Out of Bag Error (OOB), which measures the typical RF misclassification ratio. In addition to OOB, time intervals were examined for their accuracy estimates; the best accuracy was attained using the R functions ntree and mtry. They concluded more work remains in deeper understanding of why RF works, and more insight into how exactly the top ranked features contribute to RF predictive power.

This research aims study the prediction of success or failure of a team towards a milestone using the hypothesis.

- The number of the team members significantly influences the success of the group towards a milestone positively.

1.2 Thesis Outline

The thesis outline follows: Chapter 2 provides basic definitions and terminologies, showing data explorations and preprocessing, expository data analysis and statistical analysis, Data visualization, feature engineering, model optimization. Chapter 3 discusses development and prediction of the milestones using the datasets. Chapter 4 shows overall performance results and discussions. Chapter 5 includes conclusions and future work.

Chapter 2

Background

2.1 Basic Definitions and Terminologies

This section provides a comprehensive overview of the basic definitions and terminologies related to the approaches and methods employed in the thesis that focuses on the five distinct datasets representing teams at various milestones. Understanding these fundamental concepts is crucial for grasping the subsequent analyses and findings presented in the thesis. Each dataset represents teams having unique milestones achieved by the respective teams. By delving into the definitions and terminologies, readers will gain a solid foundation in the methodology employed throughout the thesis, enabling them to comprehend the intricacies of the subsequent analyses and results. This section is a valuable resource for readers to navigate the thesis effectively and engage with the research presented meaningfully.

2.1.1 Descriptive statistics

A branch of statistics called descriptive statistics deals with summarizing and describing data. It involves using numerical measures and graphical representations to present the main characteristics of a dataset. Measures of central tendency, dispersion, and skewness provide information about the data's average, spread, and shape. Visualizations like histograms and bar charts help to visualize data patterns and outliers. Descriptive statistics is essential in various fields and aids in data

exploration, hypothesis testing, and decision-making. It allows for effective communication of research findings and enables comparisons and generalizations based on the data. Descriptive statistics is used as preliminary data, which can provide the foundation for future research by defining initial problems or identifying essential analyses in more complex investigations [Brian Conner, 2017].

Feature Encoding

The process of turning raw input features, typically categorical or textual data, into a numerical representation that machine learning algorithms may use is feature encoding. For more efficient analysis, feature encoding converts categorical information into numerical characteristics [[Famili and Simoudis, 1997].

2.1.2 Shapiro-Wilk test

The Shapiro-Wilk test calculates a test statistic (W) based on the correlation between the observed data and the corresponding expected values under the assumption of normality. The test statistic is compared to critical values from the Shapiro-Wilk distribution to determine whether the data significantly deviate from normality. [Mayette Saculinggan, 2013]The Shapiro-Wilk Test, which compares measures against the Normal distribution, evaluates the Goodness of Fit test. It can be used to ensure that the data being used for parametric testing follow a normal distribution before executing such tests. Below is the formula of shapiro-wilk[Virenrehal, 2023]

$$W = \frac{(\sum_{i=1}^{n} aixi)^2}{\sum_{i=1}^{n}(x - xbar)^2} \tag{2.1}$$

where:

- x(i) represent the observation.
- xbar is the is the sample mean.
- ai are tabulated coefficient.

2.1.3 Random Forest

This is an ensemble learning method that combines multiple decision trees to make predictions. Random Forest is known for its ability to handle complex datasets and provide robust predictions.

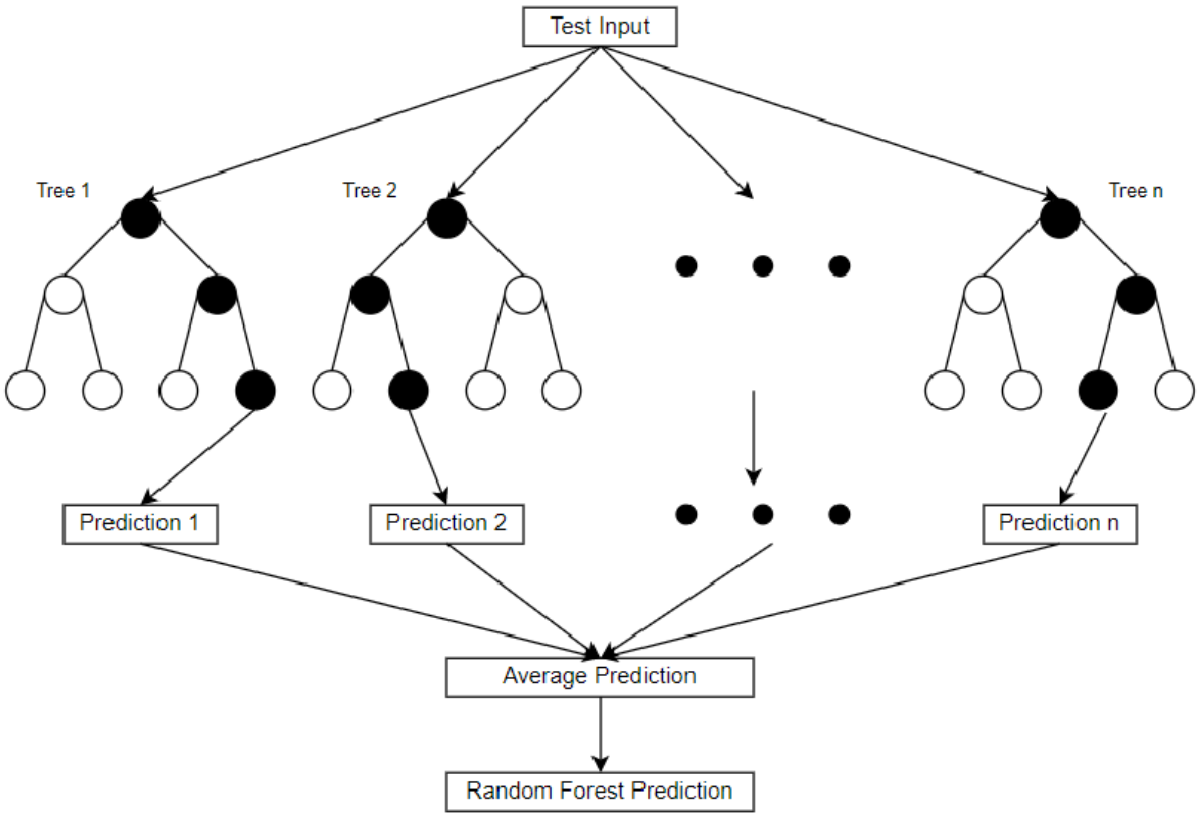


Figure 2.1: Random Forest structure [Jaeho Son, 2022]

The Machine Learning Model Development and Training Module is a vital part of the project, utilizing the Random Forest algorithm for prediction. This module involves developing and training the Random Forest model using labeled data. The model combines multiple decision trees to improve accuracy and handles complex relationships Ángel Hernández García [2018]. The trained Random Forest model becomes a powerful tool for predicting the success of a group towards a milestone and supports decision-making. Continuous monitoring and refinement further enhance the model's accuracy and reliability.

2.1.4 Decision Tree

Decision trees are hierarchical models that offer an organised method for making decisions and resolving issues. Businesses, healthcare, finance, and engineering, have found multiple applications for them. Decision trees enable researchers and practitioners to comprehend intricate relationships within data, classify instances, and make predictions based on learned patterns.[QUINLA, 1996] By using a divide-and-conquer approach, a decision tree can be created from a group of instances. The tree is a leaf with that class labelled if all the instances are members of the same class. Otherwise, a test that yields distinct results for at least two of the cases is selected, and the instances are divided in accordance with this result. The construction of decision trees using different algorithms, such as entropy-based and information gain-based approaches.

$$Entropy = -\sum_{n=1} pi\ log(pi) \tag{2.2}$$

$$GiniImpurity = 1 - \sum_{n=1} pi^2 \tag{2.3}$$

where:

- Entropy represents the entropy of the dataset.
- p(i) is the proportion of instances belonging to class i in the dataset.
- The summation is taken over all distinct classes in the dataset.

The entropy-based approach quantifies the impurity or disorder in a dataset using the concept of entropy. Entropy represents the average amount of information required to classify an instance in a dataset. The information gain-based approach measures the information gain but incorporates a normalization factor. This allows for fair attribute comparison, considering the number of distinct values an attribute can take.

2.1.5 AdaBoost

An ensemble learning method used for both classification and regression tasks is called AdaBoost. [Ying CAO, 2013] AdaBoost can turn a weak learning algorithm into a strong learning algorithm with accuracy that is arbitrarily accurate, improving on random guessing while introducing a novel design approach.

2.1.6 Logistic Regression

Another statistical method using one or more independent variables is logistic regression, it is used to estimate the probability of a binary occurrence, in this context, success or failure. It is a popular and widely used classification algorithm in machine learning and statistics. Finding the best-fitting model that connects the independent variables to the likelihood of the binary outcome is the aim of logistic regression.

The link between the dependent and independent variables has been examined using a binary logistic regression model [Sens, 2004]. In research done by [Bera, 2013] about the prediction of deforestation probability of the Silabati watershed, India, using the binary logistic regression (BLR) model. They made a BLR model to define and prepare probable deforestation areas within the watershed. They used statistical software (SPPSS v17) for Social Science was used for all mathematical calculations. Value of binary logistic regression model using the Arc GIS software and result shows the Silabati watershed's center and medium lower portions will see the most deforestation, according to the factors like distance, forest density, and model-based predicted deforestation map.

2.1.7 Support Vecto Machine (SVM)

Support vector machine, also known as SVM, is an effective supervised machine learning method that is used for classification and regression applications.[Gammermann, 2000] It can be used for solving extremely high-dimensional problems which are infeasible for the previously known learning machines.

2.1.8 Model Evaluation and Performance Metrics

Accuracy

This will measure the overall correctness of the predictions made by each model.

Precision

This will evaluate the model's ability to correctly identify positive instances.

Recall

This will measure its ability to identify all positive instances correctly. If there are statistically significant differences in how the models perform.

AUC

This will provide an assessment of the model's performance in terms of its ability to rank instances correctly.

Significance Testing

This will assist in identifying them. To test significance t-test was used to compare the performance of the ML technique classifier with baseline model. *Baseline accuracy* = 0.5.
p_value < 0.05

Cross-Validation

Cross-validation is a method for assessing and validating the efficiency and performance of machine learning models, including Random Forests. It helps to assess the model's generalization ability and reduces the risk of overfitting. [Berrar, 2013] One of the most popular approaches for resampling data to evaluate true model prediction error and adjust model parameters is cross-validation.

By evaluating these metrics across the different datasets, the module will provide insights into the strengths and weaknesses of the Random Forest, Decision Tree, and Logistic Regression techniques. This information facilitates informed decision-making regarding the selection and application of these models in various ways and assists in optimizing their performance.

Random Search

Random search is a technique used for hyperparameter optimization in machine learning. It involves randomly sampling combinations of hyperparameters from a predefined search space and evaluating the model's performance. For practical reasons relating to the statistical independence of each trial, random experiments are also simpler to conduct than grid experiments [James Bergstra, 2013]. For instance, the experiment can be terminated at any point, and the trials make up the entirety of the experiment.

2.1.9 Model Tunning

In this section, below are some techniques were used to optimize the performance of the model.

Hyperparameters

Hyperparameters are parameters set before the learning process begins and not learned from the data. They are essential in shaping the behavior and performance of a machine learning model. Random search optimization involves randomly sampling hyperparameter combinations from a predefined search space and evaluating the model's performance with each combination. Below are the key hyperparameter that will be used for all the dataset.

Esembled Models

Ensemble models in machine learning are a combination of multiple individual models that work together to make predictions or decisions. Ensemble models have gained popularity for their ability to enhance accuracy, handle complex problems, reduce overfitting, and improve robustness.

The success of these models is attributed to the better feature representation via multi layer processing architectures [M.A. Ganaie, 2022].

Decision tree, Random forest, Logistic regression, ADA boost

2.1.10 Model Interpretation

Feature Importance

Feature importance determines the relative significance of different features in a dataset for predicting the target variable. It helps identify the most influential features that affect the outcome. To understand how trained machine learning models produce their predictions, explainable artificial intelligence (AI) literature has given feature significance techniques much attention [Villani, 2022].

LIME (Local Interpretable Model-agnostic Explanations) Method

LIME (Local Interpretable Model-agnostic Explanations) is an interpretability method widely used to explain the predictions of machine learning models. It offers insights into the variables that affect a particular result, such as a team's success of a particular milestone. With the support of LIME, we could learn more about the key point that influence a team's performance and the elements that determine the success or failure. LIME maintains interpretability in the generated explanations by using an interpretable representation of the data. These explanations are simpler because they demonstrate a closer relationship between the input and prediction [Saumitra Mishra, 2017].

Chapter 3

Methodology

3.1 Methodology Overview

This research presents a sequential analysis approach that encompasses different phases, starting with data exploration and processing, followed by statistical analysis, feature engineering, machine learning (ML) model development, ML model optimization, and interpretation. The research emphasizes the importance of each phase and demonstrates how they contribute to a comprehensive analysis.

Data is thoroughly explored and processed, statistical techniques are applied to establish a solid foundation, features are engineered to enhance predictive performance, ML models are developed and optimized, and interpretation methods are employed to gain insights into the model's predictions. This sequential approach ensures a structured and comprehensive analysis methodology.

Figure 3.1: Methodology overview

3.2 Dataset

he Dataset section explores the nature of the utilized datasets, including their sources, key attributes, and preprocessing steps. Understanding this information is crucial for interpreting our results accurately and ensuring the reliability of the analyses.

3.2.1 Features

The datasets contain information about team projects conducted during a period of time. Each represents a collection of information related to team composition, project management, and performance evaluation over multiple years and semesters. The datasets originate from an educational setting, where teams are formed to work on projects and milestones. The dataset is a structured dataset.

Lists are easy to create:

- Dataset T1 consists of 20 organized variables with 64 corresponding observations. Out of those variables, four of them are categorical variables.
- Dataset T4 consists of 20 organized variables with 63 corresponding observations. Out of those variables, four of them are categorical variables.
- Datasets T2, T3, and T5 consists of 20 organized variables with 74 corresponding observations. Out of those variables, four of them are categorical variables in each.

Certain variables were removed from the dataset because they were determined to be irrelevant to the research topic and the current objectives, ensuring a focused and valuable analysis. In order to speed up the analysis and improve the precision and understandability of the findings, the irrelevant variables were removed. This selection method enabled the research to focus on the most significant variables and have a major effect on the outcomes, resulting in a more thorough and productive analysis.

These are properties that describe each instance or observation in a dataset. They represent the different types of information or measurements that are collected for analysis.

Table 3.1: Datasets features summary.

Variable Name	Variable Type	Entry type
semester	Categorical	Fall / Spring
teamMemberCount	Numerical	0 - 100
femaleTeamMembersPercent	numerical	0 - 100
teamLeadGender	Categorical	M / F
teamDistribution	Categorical	Global / Local
teamMemberResponseCount	Numerical	0 - 100
meetingHoursTotal	Numerical	0 - 100
inPersonMeetingHoursTotal	Numerical	0 - 100
nonCodingDeliverablesHoursTotal	Numerical	0 - 100
codingDeliverablesHoursTotal	Numerical	0 - 100
helpHoursTotal	Numerical	0 - 100
leadAdminHoursResponseCount	Numerical	0 - 100
leadAdminHoursTotal	Numerical	0 - 100
commitCount	Numerical	0 - 100
uniqueCommitMessageCount	Numerical	0 - 100
uniqueCommitMessagePercent	Numerical	0 - 100
commitMessageLengthTotal	Numerical	0 - 1000
issueCount	Numerical	0 - 100
onTimeIssueCount	Numerical	0 - 100
productLetterGrade	Categorical	A / F

Five different datasets, each having a significant milestone, were used for exploration. The researchers were able to study and analyze several aspects of student teamwork by using these diverse datasets. Each milestone was important in understanding and predicting student performance.

Table 3.2: Datasets and Milestones

Dataset	Milestone	Description
T1	M1	High Level requirement and spec
T2	M2	Detailed requirement and spec
T3	M3	Prototype development
T4	M4	Beta launch
T5	M5	Final delivery and demo

3.2.2 Missing Values

Null values were encountered in five datasets, and to effectively represent and manage them, the table was created to document the presence of these missing values in each dataset.

Table 3.3: Missing values

Variable Name	T1	T2	T3	T4	T5
leadAdminHoursTotal	8	3	5	2	7

There are several instances of null values, particularly in the leadAdminHoursTotal variable. Dataset T1 has 8 null values in leadAdminHoursTotal, indicating missing information on the total hours contributed by the lead administrator. Similarly, T2 and T3 have 3 null values each in leadAdminHoursTotal, while T4 has 2 null values in the same variable. Lastly, dataset T5 contains 7 null values, also reflecting unavailable leadAdminHoursTotal data. These null values signify the absence of recorded information for the lead administrator's hours in the respective datasets.

Imputation

Null values were present in multiple datasets (T1, T2, T3, T4, and T5), specifically in the variable leadAdminHoursTotal. The imputation technique of replacing null values with the mean value of the non-null leadAdminHoursTotal values was employed to address this issue. The mean value was calculated for each dataset to impute the respective null values. This approach allowed the datasets to be analyzed without losing important information. The null values were effectively resolved by utilizing imputation with mean, ensuring the datasets were more complete and suitable for subsequent analysis.

3.2.3 Duplicates

In each dataset, it is observed that the total number of observations match the number of distinct records proving there are no duplicated found. This suggests that all datasets have unique observations, ensuring data integrity and accuracy.

Table 3.4: Duplicates

Variable Name	T1	T2	T3	T4	T5
Total records	64	74	74	63	74
Distinct records	64	74	74	63	74

3.2.4 Outliers

Outliers were identified in five datasets, and to systematically analyze them, a table was created to present the outlier observations in each dataset along with their corresponding variable values.

Table 3.5: Outliers in variables

Variable Name	T1	T2	T3	T4	T5
meetingHoursTotal	0	0	0	0	3
inPersonMeetingHoursTotal	0	0	0	0	2
nonCodingDeliverablesHoursTotal	0	0	0	0	0
codingDeliverablesHoursTotal	0	0	0	1	1
leadAdminHoursTotal	0	1	0	1	0
commitCount	2	0	0	0	0
uniqueCommitMessageCount	2	0	0	1	0
uniqueCommitMessagePercent	0	0	0	0	0
commitMessageLengthTotal	2	0	0	1	0

- In the T1 dataset, there were two outliers found in commitMessageLengthTotal variable, one in uniqueCommitMessageCount, and one in commitCount. To fix this issue, the outliers in each respective variable were replaced using the previous highest value that was not considered an outlier. By replacing the outliers with the previous highest non-outlier value, the dataset's overall integrity and consistency were maintained.

- In the T2 dataset, there was a single outlier identified in the variable leadAdminHoursTotal. The outlier in the leadAdminHoursTotal variable was replaced using the previous highest value that was not considered an outlier. By employing this approach, the dataset's integrity and consistency were maintained while effectively handling the outlier observation

- In the T3 dataset, two outliers were identified, one in the variable uniqueCommitMessageCount and another in the variable commitCount. Similar to the approach used in the T1 dataset, the

outliers in uniqueCommitMessageCount and commitCount were replaced with the previous highest value that was not considered an outlier

3.3 Exploratory Data Analysis (EDA) and Statistical Analysis Module

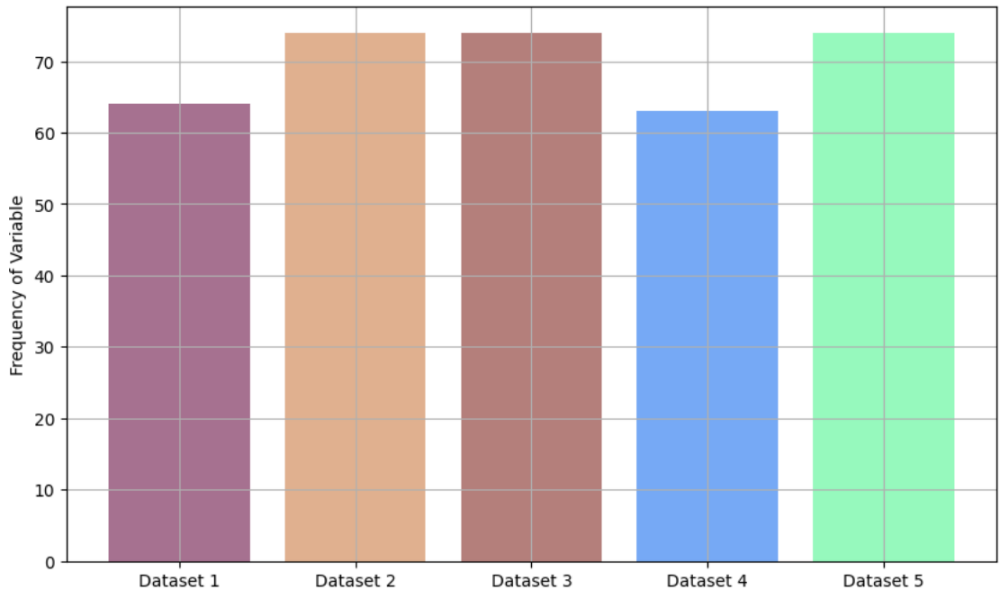
Descriptive Statistics

Descriptive statistics were computed for five separate datasets, and to effectively summarize and compare the characteristics of each dataset, individual tables were created to present the key statistical measures. These tables included commonly used descriptive statistics such as mean, standard deviation, minimum, maximum, and quartiles for each dataset, providing a comprehensive overview of the data and its distributions.

Table 3.6: Dataset T4 Summary statistics.

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
teamMemberCount	63	5.21	1.12	3.00	5.00	5.00	6.00	7.00
femaleTeamMembersPercent	63	0.19	0.17	0.00	0.07	0.17	0.25	0.83
teamMemberResponseCount	63	17.63	7.64	6.00	13.50	15.00	22.50	41.00
meetingHoursTotal	63	42.40	24.63	3.64	23.29	36.29	55.29	107.14
inPersonMeetingHoursTotal	63	32.71	18.33	3.64	18.79	29.07	43.71	84.40
nonCodingDeliverablesHoursTotal	63	25.39	14.33	4.22	15.29	21.43	31.29	67.50
codingDeliverablesHoursTotal	63	76.75	45.00	6.43	38.61	70.29	104.29	172.57
helpHoursTotal	63	21.44	13.63	1.71	9.95	19.57	28.43	70.29
leadAdminHoursResponseCount	63	3.40	1.31	0.00	3.00	3.00	4.00	8.00
leadAdminHoursTotal	61	4.07	2.68	0.00	2.14	3.29	5.43	12.29
commitCount	63	84.76	65.36	0.00	43.50	74.00	125.50	269.00
uniqueCommitCountMessageCount	63	75.46	54.37	0.00	39.00	71.00	116.00	187.00
uniqueCommitCountMessagePercent	63	0.80	0.30	0.00	0.79	0.93	0.97	1.00
commitMessageLengthTotal	63	5022	4044	0.00	1716	4726	7417	15753
issueCount	63	0.70	0.73	0.00	0.00	1.00	1.00	3.00
onTimeIssueCount	63	0.51	0.72	0.00	0.00	0.00	1.00	3.00

Figure 3.2: Frequency of datasets



The datasets T1, T2, T3, T4, and T5 provide valuable insights into a diverse set of 17 quantitative variables, offering a rich pool of information for comprehensive analysis. These datasets are characterized by variations in the number of observations across the variables, with the count ranging from 0 to 74. However, the variable "leadAdminHoursTotal" stands out due to its presence of null values in different datasets, affecting 56, 71, 69, 61, and 67 observations, respectively, in T1 to T5 datasets.

One notable aspect across all datasets is the consistent presence of "femaleTeamMembersPercent" as the variable with the lowest average value. Conversely, the variable with the highest average value varies among the datasets, with "commitCount" or "commitMessageLengthTotal" taking the lead, depending on the specific dataset under consideration. This variation in the highest average value reflects the diverse nature of the datasets and the importance of understanding the particular context of each dataset.

The datasets also exhibit intriguing variations in the variability of the variables. While some variables demonstrate a low variability, with values as small as 0.09 or 0.17, others showcase substantial fluctuations, with values reaching as high as 552.14, 50.17, or even an astonishing 4964. This disparity in variability underscores the need for a nuanced approach while analyzing

the datasets and highlights the potential for uncovering unique patterns and trends within each variable.

Furthermore, several variables in each dataset display a minimum observation of 0, which could imply specific underlying data characteristics or measurement constraints for these particular variables. On the other hand, the maximum values of these variables vary significantly, with some reaching impressive heights like 21925, 15753, 4530, or 488, depending on the dataset and the specific variable under examination. These extreme values could indicate potential outliers or exceptional instances within the data, warranting thorough investigation to ensure accurate interpretations.

It is worth noting that each dataset presents distinct characteristics, as evident from the variations in the number of observations, the presence of null values, average and maximum values, and the overall variability of the variables. These differences offer valuable opportunities for detailed descriptive analysis and facilitate the identification of unique insights, patterns, and trends within each dataset.

The T1, T2, T3, T4, and T5 datasets form a robust collection of quantitative variables with diverse observations and measures. The comprehensive analysis of these datasets can yield valuable information, contributing to a deeper understanding of the underlying phenomena and enriching various fields, such as statistics, data science, and decision-making processes across industries. However, thorough attention to data preprocessing, outlier detection, and context-specific interpretations is crucial to draw meaningful conclusions and unlock the true potential of these datasets.

3.4 Normality Test

The goodness of fit statistical technique evaluates how well a model matches actual data. It helps to decide whether variations between expected and actual values from a model are statistically significant. Chi-squared, Shapiro-Wilk, Kolmogorov-Smirnov, kruskal wallis, and T-test are some tests and techniques for the goodness of fit analysis. The findings help decide by revealing the model's applicability. With the statistical approach of this project, the tables below show the

results of different tests using different datasets. The tables have a variable name and the critical value for the decider for each variable from the datasets T1, T2, T3, T4, and T5.

Shapiro-Wilk Test

The Shapiro-Wilk test was performed on five datasets using a significance threshold of 0.05. The test aimed to determine whether the observed values significantly departed from a normal distribution to evaluate the data's normality. The test results provide insights into the suitability of the datasets for parametric statistical analyses that assume normality.

Variable Name	P value T1	P value T2	P value T3	P value T4	P value T5
teamMemberCount	1.056e-05	1.362e-06	1.362e-06	1.081e-05	1.362e-06
femaleTeamMembersPercent	2.525e-06	3.665e-07	3.665e-07	5.516e-06	3.665e-07
teamMemberResponseCount	0.0001397	0.05424	0.00465	0.0001261	3.367e-09
meetingHoursTotal	0.0005126	0.07833	0.000131	0.00189	8.475e-09
inPersonMeetingHoursTotal	0.0008244	0.00854	8.341e-05	0.003282	4.178e-08
nonCodingDeliverablesHoursTotal	3.836e-05	0.1082	0.008222	8.66e-05	2.607e-10
codingDeliverablesHoursTotal	3.301e-09	2.81e-05	0.0006252	0.007208	3.067e-08
helpHoursTotal	9.451e-05	0.03222	7.961e-06	0.003225	4.59e-08
leadAdminHoursResponseCount	9.676e-08	1.575e-05	3.377e-06	7.346e-07	1.931e-09
leadAdminHoursTotal	1.204e-07	3.934e-08	1.161e-13	1.648e-05	3.276e-13
commitCount	2.738e-15	4.904e-06	4.948e-06	0.003025	5.812e-11
uniqueCommitMessageCount	1.703e-14	1.887e-06	0.0009877	0.01067	5.998e-11
uniqueCommitMessagePercent	1.324e-08	7.097e-05	1.145e-08	2.395e-11	2.317e-10
commitMessageLengthTotal	6.571e-11	2.609e-07	0.001514	0.00125	3.167e-12
issueCount	4.208e-14	2.368e-09	5.084e-07	3.44e-08	2.917e-15
onTimeIssueCount	4.208e-14	3.729e-09	1.495e-06	5.776e-10	5.776e-10

Table 3.7: Shapiro-wilk p-values.

The Shapiro-Wilk test was performed with a significance threshold of 0.05 to assess the normality of the dataset variables. The R output revealed that all the variables had p-values below the threshold, indicating that none passed the normality test. This finding emphasizes the importance of considering non-parametric tests or transformations when analyzing the data, as assumptions of normality may not hold. To counter normality result, transformation of the variable was done, but it results to the datasets not being well organized for further analysis. Alternative approaches are explored to ensure accurate and appropriate analysis.

Kruskal-Wallis Test

The Kruskal Wallis test was performed across the datasets for each variable. The p-values for each variable were obtained based on the Kruskal-Wallis test results for the five datasets. For the variables "teamMemberCount" and "femaleTeamMembersPercent," the p-values were 0.9998 and 0.9224, respectively. These p-values exceed the 0.05 level for significance. Therefore, we fail to

Table 3.8: A table for kruskal Wallis.

Variable Name	Kruskal-Wallis P values of T1 , T2, T3 , T4, T5
teamMemberCount	0.9998
femaleTeamMembersPercent	0.9224
teamMemberResponseCount	2.2e-16
meetingHoursTotal	2.2e-16
inPersonMeetingHoursTotal	2.2e-16
nonCodingDeliverablesHoursTotal	2.2e-16
codingDeliverablesHoursTotal	2.2e-16
helpHoursTotal	2.2e-16
leadAdminHoursResponseCount	2.2e-16
leadAdminHoursTotal	2.2e-16
commitCount	2.2e-16
uniqueCommitCountMessageCount	2.2e-16
uniqueCommitCountMessagePercent	0.0009395
commitMessageLengthTotal	2.2e-16
issueCount	2.2e-16
onTimeIssueCount	2.2e-16

reject the null hypothesis for these variables. This suggests that there is no significant difference in "teamMemberCount" and "femaleTeamMembersPercent" across the five datasets.

On the other hand, it was found that the p-values for the remaining variables were 2.2e-16, which is significantly less than 0.05. Hence, we reject the null hypothesis for these variables. This suggests that these variables significantly differ statistically among the five datasets.

It can be concluded that there are statistically significant differences in the variables with p-values less than 0.05 based on the results of the Kruskal-Wallis test. As their p-values are more than 0.05, the variables "teamMemberCount" and "femaleTeamMembersPercent," however, do not provide enough data to support a conclusion that there are significant differences between the datasets.

3.5 Data Visualization

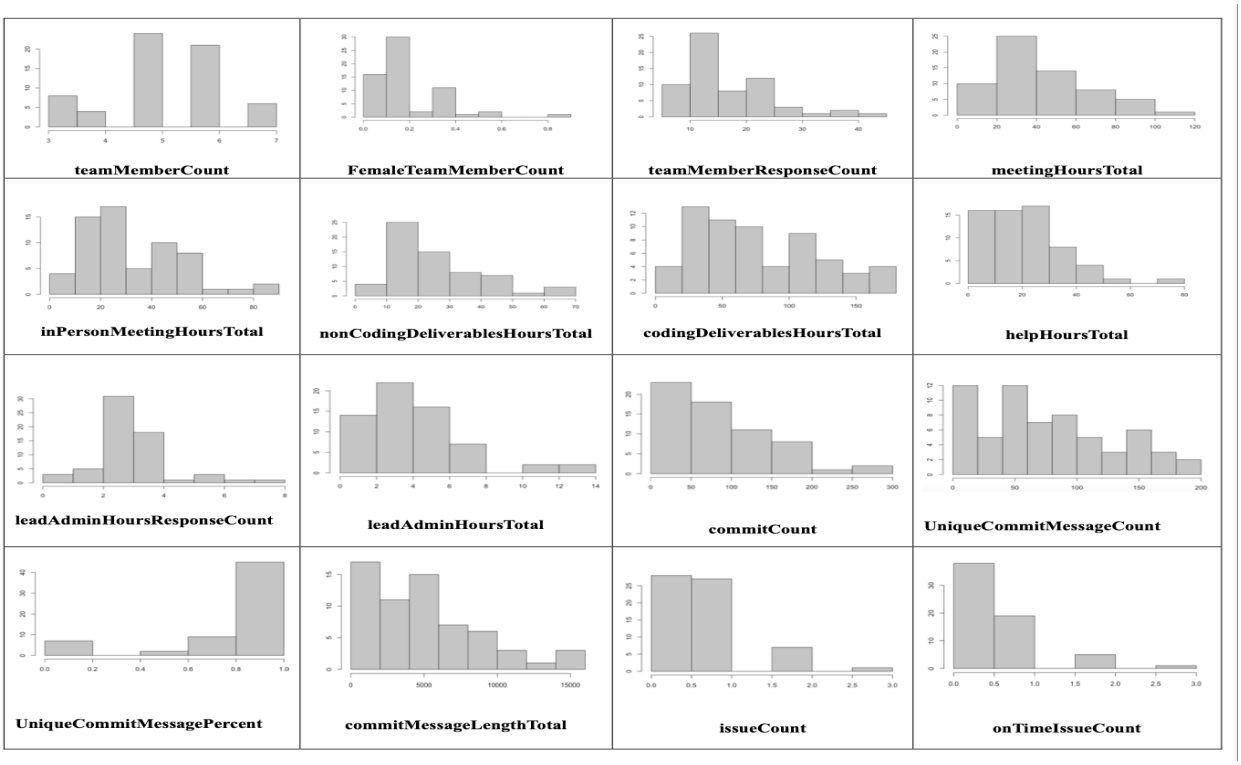
Data visualization plays a crucial role in understanding and communicating information effectively. In this scenario, we have five datasets, each with its milestone for visualization. Histograms are used for quantitative variables like teamMemberCount, commitCount, and meetingHoursTotal, while pie charts are employed for qualitative variables such as teamLeaderGender and semester. The milestones include plotting a histogram for teamMemberCount to analyze teamMemberCount number and distribution and a pie chart for teamLeaderGender to assess the percentages of gender. Visualization lets us gather insights, spot patterns, and base judgments on the data. It allows us to comprehend distributions, identify trends, and convey information more intuitively to stakeholders. Overall, data visualization aids in comprehending complex datasets and facilitates data-driven decision-making.

Histograms

In datasets T1, T2, T3, T4, and T5, the visualization approach focuses on representing data using histograms for quantitative variables and pie charts for qualitative variables. Histograms are employed to display the distribution and frequency of numerical data, showcasing the range

of values and their corresponding frequencies. On the other hand, pie charts are used to present categorical or qualitative data, illustrating the proportion or percentage of each category within the entire dataset.

Figure 3.3: Histograms for dataset T4 quantitative variables

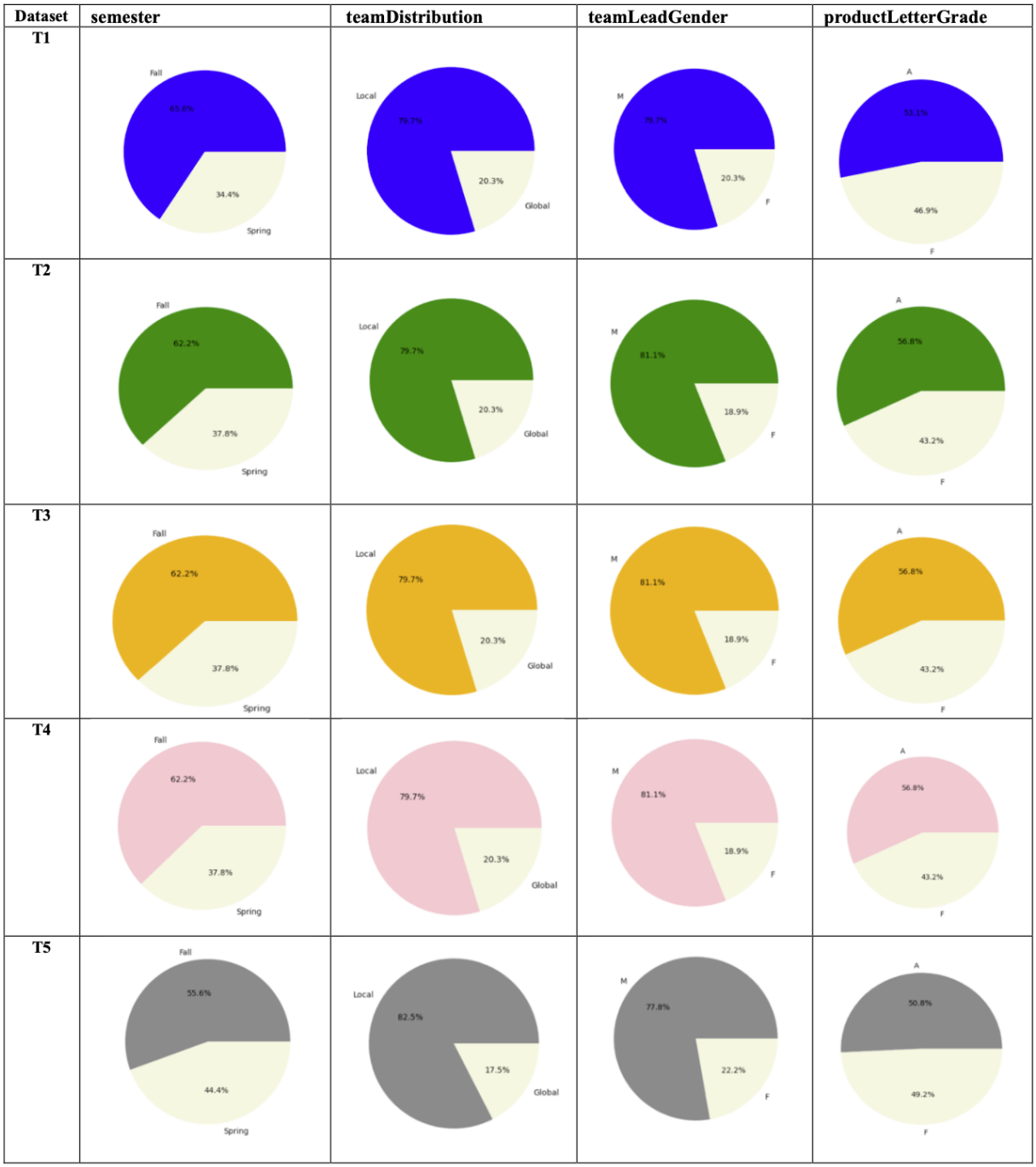


The histogram plots of the quantitative variables in T1, T2, and T3 exhibit a positively skewed distribution. However, in T2, there are also normal distributions and a smaller number of negatively skewed distributions. Similarly, T4 shows histogram plots with normal distributions, positively skewed distributions, and fewer negatively skewed distributions, just like in T2 and T3. In the T5 dataset, the histogram plots of the quantitative variables are primarily positively skewed with a smaller number of negatively skewed distributions, resembling the pattern observed in the T1 dataset.

Pie charts

Below are pie for all the datasets to visualize the qualitative variables.

Figure 3.4: Pie charts for categorical variables on all datasets



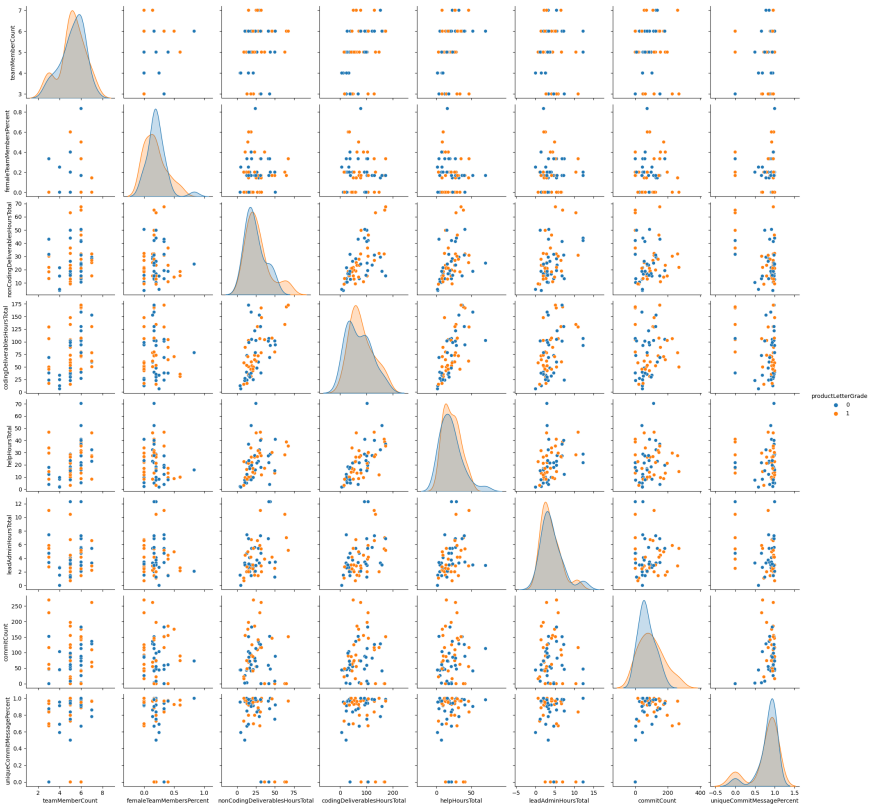
In dataset T1, the qualitative variables reveal that males constitute a higher percentage, the fall

semester enrollment surpasses the spring semester, local students outnumber global students, and the grade distribution is relatively close with only slight differences. Similarly, in datasets T2 and T3, males represent a higher percentage, the fall semester has a higher enrollment than the spring semester, local students are more numerous than global students, and the percentage of A grades exceeds that of F grades. Moving on to dataset T4, we observe that males still constitute a higher percentage, while the fall and spring semesters show a minor difference in enrollment percentages. Local students dominate over global students, and surprisingly, the percentage of F grades is nearly the same as that of A grades. Lastly, in dataset T5, we find that males have a higher representation, the fall semester enrollment outweighs the spring semester, local students are more abundant than global students, and the percentage of A grades surpasses that of F grades.

Scatterplots

Below are scatterplots for dataset T4 to check the relationship between two quantitative variables.

Figure 3.5: Scatterplots of variables from dataset T4

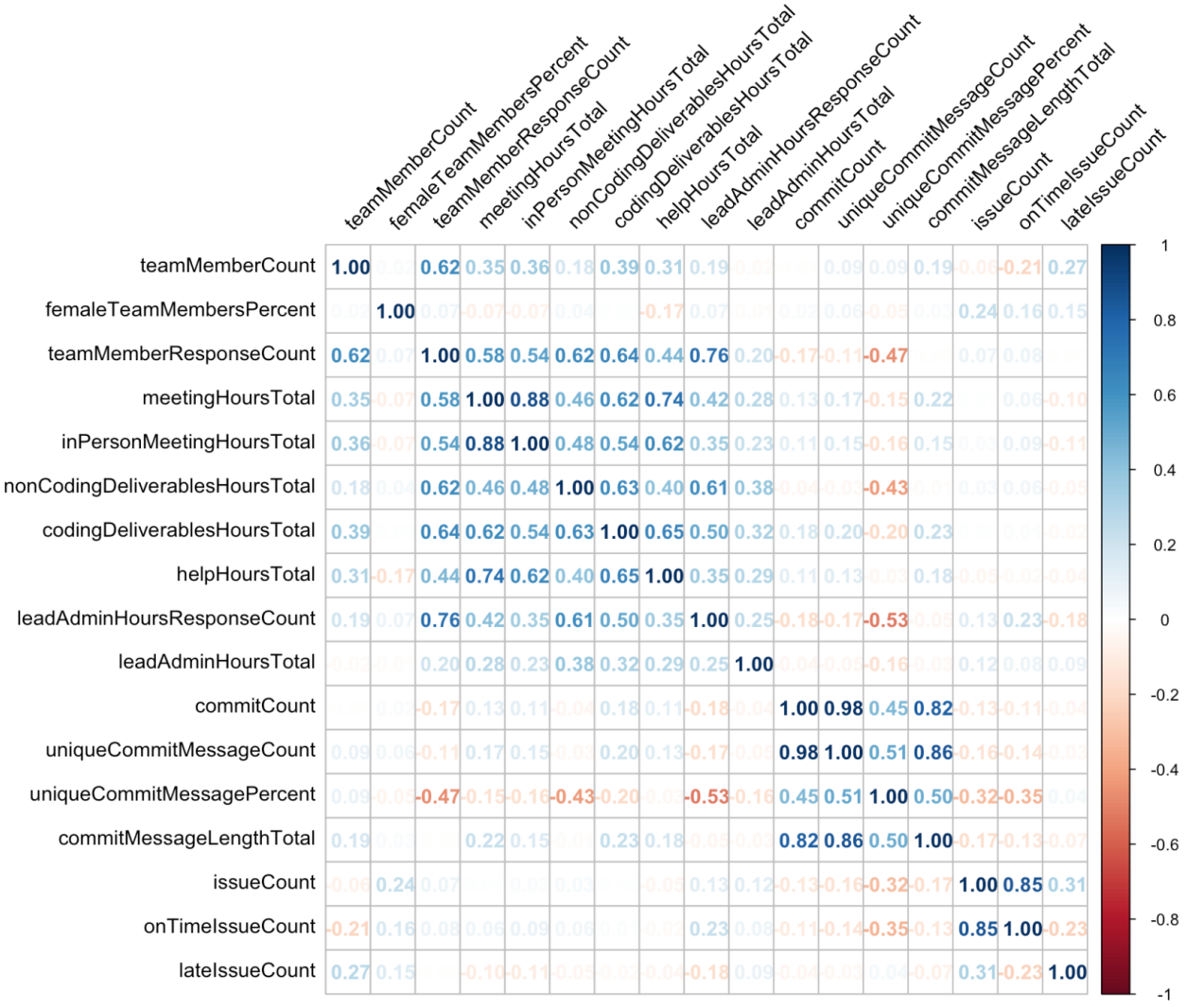


The scatterplots provide valuable insights into the relationships between the variables. For example, It highlights a strong positive correlation between **helphourstotal** and **codingDeliverablesHourstotal** with a trend, while indicating no substantial correlation between **leadAdminHoursTotal** and **teamMemberCount.**

Correlations

Below is a correlation plot for dataset T4.

Figure 3.6: Correlation plot of dataset T4



Using a threshold of 0.75, the Pearson correlation output plot illustrates the correlations between the variables in a dataset. It aids in determining the significance and direction of correlations. A

strong positive correlation is present when the correlation value between two variables exceeds the threshold. Below are some of the observed variables that are strongly correlated.

- *commitMessageLengthTotal* and leadAdminHoursTotal
- *commitMessageLengthTotal* and uniqueCommitMessagePercent

3.6 Feature Engineering

A crucial part of the data processing cycle is feature engineering, which involves transforming raw data into valuable features that can enhance the performance of analysis and machine learning models. It includes a range of methods and procedures to reduce the dimensionality of the feature space, choose the most essential features, and extract relevant details from the data. Predictive modeling involves many steps, one of which is feature engineering. Reducing the modeling error for a specific target involves the transformation of a given feature space, often employing mathematical functions. [Udayan Khurana et al. 2019]

Feature Selection

For feature selection, Pearson correlation was performed, and variables were dropped as a result of strong correlation.

Feature Selection T1, T2, T3, and T4

Using Pearson correlation for feature selection for the dataset T1, T2, T3, and T4 with a threshold of 0.7, the most influential variables were filtered to 8. These variables are similar in all four datasets.

The variables are: *teamMemberCount, femaleTeamMembersPercent, nonCodingDeliverablesHoursTotal,, codingDeliverablesHoursTotal, helpHoursTotal, leadAdminHoursTotal, countCommit, uniqueCommitMessagePercent.*

Feature Selection T5

Using Pearson correlation for feature selection for the dataset T5 with a threshold of 0.7, the most influential variables were filtered to 10.

The variables are: *teamMemberCount, femaleTeamMembersPercent, nonCodingDeliverablesHoursTotal, codingDeliverablesHoursTotal, helpHoursTotal, leadAdminHoursTotal.*

Dimentionality Reduction

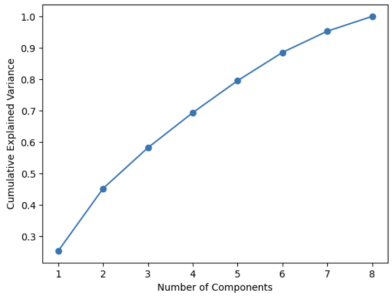
In this analysis, a threshold of 70 percent will be used for the cumulative proportion of variance explained as a criterion for selecting principal components in order to capture a significant amount of information. This threshold allows for a balance between retaining sufficient information while reducing dimensionality and complexity in the dataset. Principal components that cumulatively explain at least 70 percent of the total variance will be kept, Once the treshold is reached and the upcoming component does not add much, it will not be included as in pca it is considered the less componenets the better.

Principal Component Analysis (PCA)

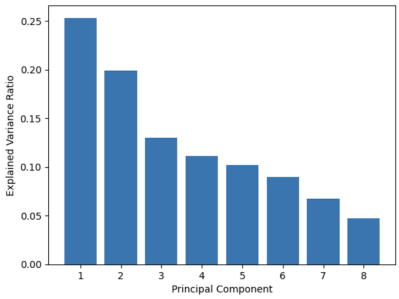
From each dataset, the number of pca components correspond to the number of filtered variables. Each component describes the percentage of variance in the dataset. The cumulative proportion explains the total variance.

- From Figure 3.12, in T1 dataset the PC4 shows about 70 percent, meaning the first four components can almost accurately represent the data.
- From Figure 3.13, in T2 dataset the PC4 shows about 80 percent, meaning the first three components can almost accurately represent the data.
- From Figure 3.14, in T3 dataset the PC4 shows about 79 percent, meaning the first four components can almost accurately represent the data.

Figure 3.7: Plots for T1 PCA

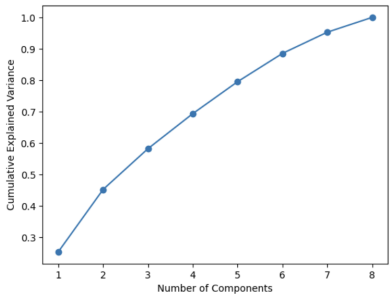


(a) Scree plot for cumulative variance of principal components.

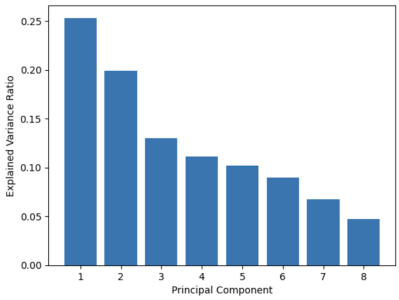


(b) Bar plots for single principal component.

Figure 3.8: Plots for T2 PCA



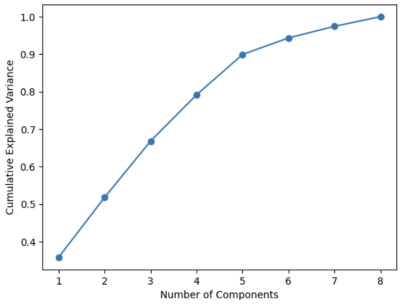
(a) Scree plot for cumulative variance of principal components.



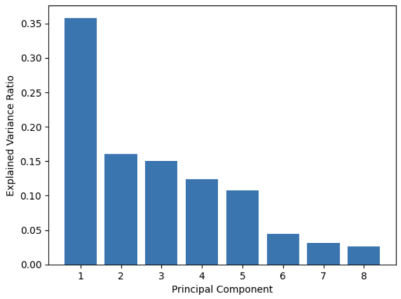
(b) Bar plots for single principal component.

- From Figure 3.15, in T4 dataset the PC4 shows about 75 percent, meaning the first four components can almost accurately represent the data.
- From Figure 3.16, in T5 dataset the PC3 shows about 73 percent, meaning the first three components can almost accurately represent the data.

Figure 3.9: Plots for T3 PCA

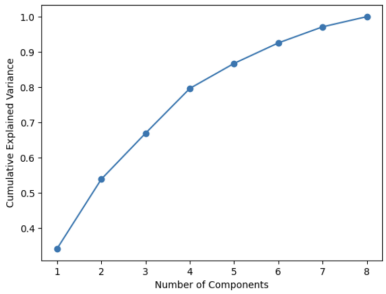


(a) Scree plot for commulative variance of principal component

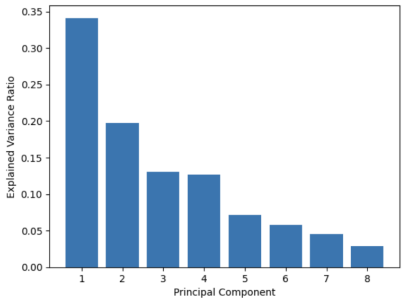


(b) Bar plots for single principal component

Figure 3.10: Plots for T4 PCA

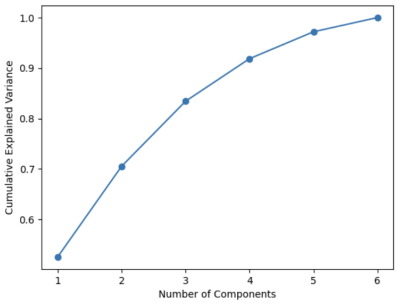


(a) Scree plot for commulative variance of principal component

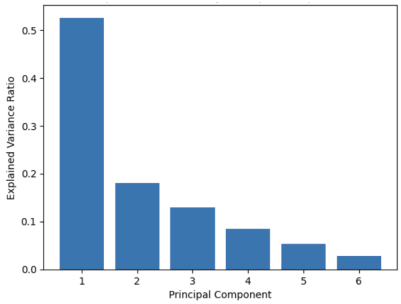


(b) Bar plots for single principal component

Figure 3.11: Plots for T5 PCA



(a) Scree plot for commulative variance of principal component



(b) Bar plots for single principal component

3.7 Machine Learning Models

In the machine learning models section, various machine learning algorithms and techniques are explored and used to solve complex problems and make data-driven predictions.

3.7.1 Exploring the Applicable Models

This section will observe a thorough analysis of the accuracy, precision, recall, and F1-score of the logistic regression, decision tree, and random forest models on the datasets in this section. We will investigate their performances across different domains and analyze their strengths and weaknesses based on empirical results. This analysis will help understand each model's suitability for the prediction and assist in making informed decisions by choosing the best suitable machine learning approach.

Feature Encoding

Table 3.9: Performance of the models.

productLetterGrade	Encoding
A	1
F	0

"A" indicates success, whereas "F" indicates failure. Simple encoding method by giving "A" the value of 1 and "F" the value of 0. This binary representation simplifies the target variable and facilitates its interpretation.

Logistic Regression

The point behind logistic regression lies in transforming the linear regression equation using a logistic function. This transformation ensures that the predicted output remains within the range of 0 to 1, making it suitable for probability estimation.

Decision Tree

The goal of the decision tree method is to maximise information gain or minimise impurity at each split as it builds the tree by recursively separating the data based on several qualities. A stopping requirement, such as reaching a maximum depth or obtaining a minimum number of data points in a leaf node, is then reached, and the attributes with the greatest discriminatory power are chosen as the decision nodes. This was performed across the dataset in order to see the accuracy of the model

Random Forest

This is an ensemble learning method that combines multiple decision trees to make predictions. Random Forest is known for its ability to handle complex datasets and provide robust predictions.

3.8 Model Fine-tuning and Optimization

The Fine-tuning and Optimization module is a critical component of the machine learning pipeline. It involves adjusting pre-trained models for specific tasks and optimizing model parameters to improve performance. This approach examines optimization algorithms, hyperparameter tuning, and other fine-tuning and optimization techniques. By enhancing our understanding of this module, enhancing the efficiency and effectiveness of machine learning models in various domains will be done.

Chapter 4

Results and Discussions

This chapter reveals the outcomes of our comprehensive machine learning tests. The training set is used to train the model, while the testing set is used to evaluate its performance and generalization ability on unseen data. The dataset was split into a training set and a testing set with a ratio of 80% training data and 20% testing data. Below are outputs of the three machine learning test. These metrics are the F1-score, precision, recall, and accuracy as shown in the equations.

Accuracy

$$\textit{Accuracy} = \frac{TP + TrueNegatives(TN)}{TP + TN + FP + FN} \tag{4.1}$$

Precision

$$\textit{Precision} = \frac{TruePositive(TP)}{TP + FalsePositive} \tag{4.2}$$

Recall

$$\textit{Recall} = \frac{TruePositive(TP)}{TP + FalseNegative(FN)} \tag{4.3}$$

Area under the Curve (AUC)

where:

- TP - True Positives
- FP - False Positives
- TN - True Negatives
- FN - False Negative

4.1 Machine Learning performance

Table 4.1: Random Forest evaluation metrics

Evaluation Metrics	T1	T2	T3	T4	T5
Accuracy	0.7	0.5	0.4	0.5	0.6
Precision	0.6	0.7	0.3	0.4	0.8
Recall	1	0.5	0.5	0.6	0.5
AUC	0.7	0.35	0.4	0.5	0.6
Sig Test	Sig	Sig	Not Sig	Sig	Sig

Table 4.2: Decision Tree evaluation metrics

Evaluation Metrics	T1	T2	T3	T4	T5
Accuracy	0.5	0.4	0.4	0.5	0.5
Precision	0.5	0.6	0.4	0.4	0.7
Recall	1	0.4	0.5	0.8	0.5
AUC	0.5	0.3	0.4	0.5	0.5
Sig Test	Sig	Nor Sig	Not Sig	Sig	Sig

Table 4.3: Logistic Regression evaluation metrics

Evaluation Metrics	T1	T2	T3	T4	T5
Accuracy	0.5	0.6	0.5	0.2	0.6
Precision	0.6	0.7	0.5	0.2	0.8
Recall	0.6	0.6	0.7	0.4	0.5
AUC	0.5	0.5	0.5	0.2	0.6
Sig Test	Sig	Sig	Sig	Not Sig	Sig

Random Forest, Decision Tree, and Logistic Regression were evaluated using various performance criteria, including significance testing, accuracy, precision, recall, and AUC. Random Forest was shown to be the best model among the three. The evaluation findings consistently demonstrated the Random Forest model's higher performance across all datasets.

4.2 Random Forest

After conducting the evaluation using various performance metrics, including significance testing, accuracy, precision, recall, and AUC, Random Forest emerged as the best model among the three (Random Forest, Decision Tree, and Logistic Regression). The evaluation results consistently showcased superior performance by the Random Forest model across all datasets. Its high accuracy scores demonstrated its ability to make correct predictions, while it's precision and recall metrics indicated its effectiveness in identifying positive instances accurately. Furthermore, the Random Forest model consistently exhibited strong performance in terms of AUC, highlighting its robustness in ranking instances correctly. Based on these comprehensive evaluations, Random Forest was selected as the top-performing model, surpassing Decision Tree and Logistic Regression, making it the preferred choice for the given task and datasets.

Table 4.4: Performance of the models.

Evaluation metrics	Accuracy	Precision	Recall	AUC	Significance Test
Best Model in datasets	RF T1, T4, T5	RF T1, T2, T4, T5	RF T1, T2, T4, T5	RF T1, T2, T4, T5	RF T1, T2, T4, T5

4.3 Model Evaluation and Performance Metrics

Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. Random Forest is known for its ability to handle complex datasets and provide robust predictions. The Random Forest machine learning approache will all be evaluated in the Model

Table 4.5: Random Forest evaluation metrics

Evaluation metrics	Classes	T1	T2	T3	T4	T5
Accuracy		70%	40%	40%	50%	60%
Precision	Pass	62%	70%	38%	43%	88%
	Fail	80%	20%	43%	67%	43%
Recall	Pass	83%	64%	43%	60%	64%
	Fail	57%	25%	38%	50%	75%
F1 - Score	Pass	75%	22%	40%	43%	50%
	Fail	60%	67%	40%	67%	57%
AUC		71%	50%	40%	55%	60%
Sig Test		Sig	Sig	Not Sig	Sig	Sig

(a) Random Forest evaluation metrics

Evaluation and Performance Metrics chapter. The 5 datasets will be used for this evaluation, each with a milestone. The chapter will employ appropriate evaluation metrics such as significance testing, accuracy, precision, recall, and area under the curve (AUC).

Random Forest For all datasets

- **T1**: The Random Forest model in T1 datasets achieved an accuracy of 0.7, indicating that it correctly predicted 70% of the instances in the dataset. It had a precision that accurately identified 50% of the positive instances. The model demonstrated perfect recall with a score of 1.0, correctly identifying all positive instances. The AUC score of 0.71 suggests a moderate ability to rank instances. Overall, the Random Forest model showed potential for improvement with promising performance in accuracy and recall, while precision and AUC could benefit from further optimization.

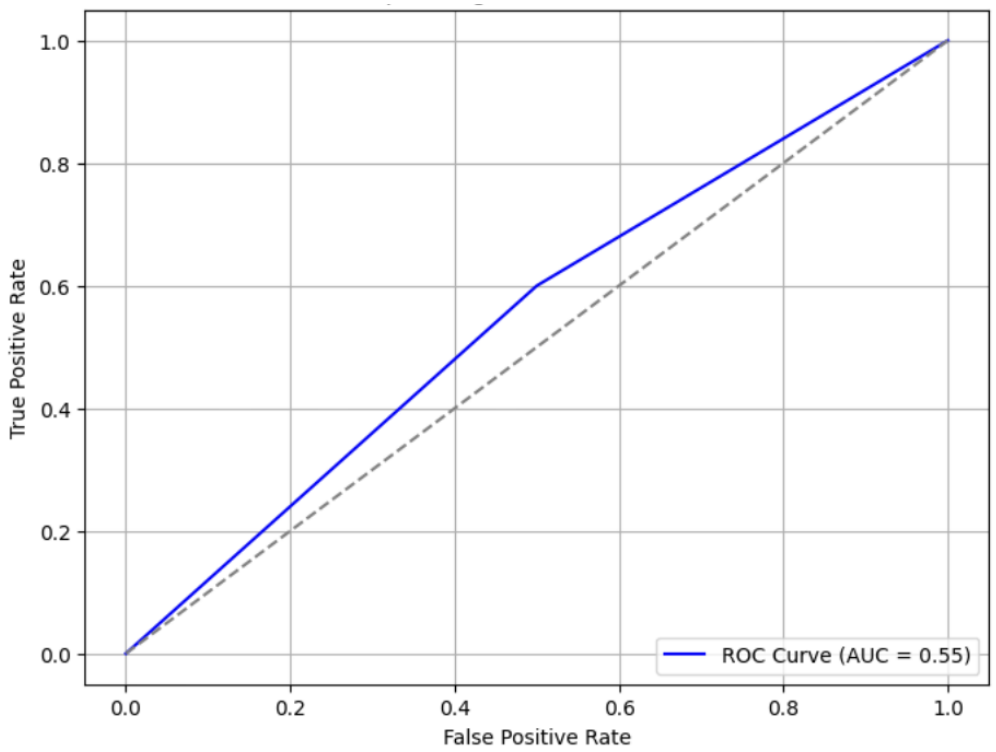
- **T2**: The Random Forest model of T2 achieved an accuracy of 0.5, and precision identified 70% of the positive instances. It captured 60% of the actual positive instances on recall and an AUC of 0.5. The results indicate moderate performance with room for improvement in accuracy, recall, and AUC.

- **T4**: The model's performance on T4 indicates that it correctly predicted 50% of the instances, accurately identified 40% of positive instances, captured 60% of the actual positive instances,

and performed at a chance level in terms of ranking instances. The results suggest opportunities for improvement in precision and AUC metrics while showing moderate performance in accuracy and recall.

- **T5**: The Random Forest model on the T5 dataset achieved an accuracy of 0.6, indicating that it correctly predicted 60%of the instances in the dataset. The precision score of 0.8 suggests that it accurately identified 80% of the positive instances. With a recall score of 0.5, the model captured 50% of the actual positive instances. The AUC score of 0.6 indicates that the model showed moderate performance in ranking instances correctly. Overall, the results indicate relatively good accuracy and precision, while the recall and AUC could be further improved for enhanced performance.

Figure 4.1: AUC curve from dataset T4



The AUC curve of dataset T4 is close to 0.5, which shows that the model's capacity to differentiate between the positive and negative classes is slightly limited.

4.3.1 Cross Validation

This is performed across all datasets, identifying the average of each and getting the grand average for all the datasets. CV

Table 4.6: Cross Validation scores.

Datasets						**Average Score in %**
T1	0.69	0.69	0.53	0.76	0.67	***67%***
T2	0.46	0.4	0.6	0.8	0.5	55%
T3	0.6	0.46	0.4	0.4	0.7	62%
T4	0.6	0.6	0.6	0.25	0.5	53%
T5	0.53	0.6	0.53	0.6	0.5	59%

- **T1**: These scores represent the accuracy achieved by the model on different folds of the data. The average score was 0.67 or approximately 67%. This average score provides an overall evaluation of the model's performance, indicating that it achieved an accuracy of around 67% on average across the different folds of the dataset. This is the highest between the datasets.

- **T2**: The cross-validation scores were [0.46666667, 0.4, 0.6, 0.8, 0.5]. These scores represent the accuracy achieved by the model on different folds of the data. The average score was 0.55 or approximately 55%. This average score provides an overall evaluation of the model's performance, indicating that it achieved an accuracy of around 55% on average across the different folds of the dataset.

- **T3**:The average score was 0.52, roughly 52%. This average score gives a general assessment of the model's effectiveness and shows that it averaged an accuracy of about 52% across the various folds of the dataset.

- **T4**:The average result of T4 was 0.53, or nearly 53%. This average score provides a summary evaluation of the model's performance and reveals that, on average, it had a 53% accuracy across all folds of the dataset.

- **T5**: 59%, was the average score in T5 dataset. The accuracy of the model throughout the several folds of the dataset was on average roughly 59%, according to the average score, which provides a broad evaluation of the model's efficacy.

Grand Average of RF model

$$\frac{0.7 + 0.4 + 0.4 + 0.5 + 0.6}{5} = 0.52$$

Grand Average = 0.52

52%

The Grand average score of approximately 0.57 represents the overall performance of the Random Forest model across five datasets. This score indicates that, on average, the model achieved an accuracy of around 54% when predicting the target variable for different instances in the cross-validation process. It is important to consider additional evaluation metrics and the variability of scores across datasets to fully assess the model's generalization ability and understand its strengths and limitations.

4.4 Model Fine-tuning and Optimization

This section will detail machine learning optimization and fine-tuning principles and procedures for the random forest. How different hyperparameters, optimization, and fine-tuning strategies affect model performance will be investigated.

4.4.1 Random Forest

Some critical components in maximizing the effectiveness of machine learning models are Fine-tuning and Optimization. In this project, fine-tuning a Random Forest model, a popular ensemble learning algorithm known for its robustness and flexibility, was focused on. A technique

was employed to optimize the model: Random Search for hyperparameter tuning. A hyperparameter optimization technique called "Random Search" (RS) randomly explores the hyperparameter space in search of the optimal set of values. Exhaustively searching a wide range of values enables us to discover optimal hyperparameter settings for the Random Forest model. This technique helps improve model performance and avoid overfitting. By running Random Search for hyperparameter tuning, the performance and generalization capabilities of the Random Forest model is enhanced. The Random Forest algorithm's performance is optimized in this project using fine-tuning and optimization techniques.

4.4.2 Random Search RS Hyperparameter Tuning

Random search was used performed to optimize the key hyperparameters. It involves choosing random hyperparameter combinations from an initial search space and assessing the model's performance. Below are the initial accuracy from Random Forest algorithm.

Table 4.7: Random Forest accuracy

Dataset	T1	T2	T3	T4	T5
Accuracy	0.7	0.5	0.4	0.5	0.6

Table 4.8: Table for average value for the datasets.

Best Hyperparameter	Most frequent values
n_estimators	200
max_features	log2
max_depth	10
min_sample_split	10
m_sample_leaf	2

- n_estimators: number of trees in the forest.
- max_features: max number of features considered for splitting a node.
- max_depth: max number of levels in each decision tree.

- min_sample_split: min number of data points placed in a node before the node is split.
- m_sample_leaf: min number of data points allowed in a leaf node. [Koehrsen, 2018].

Table 4.9: Tuned accuracy random search result

Dataset	T1	T2	T3	T4	T5
Tuned Accuracy	0.75	0.71	0.67	0.6	0.7

The outcomes of the random search for model tuning and optimization have greatly enhanced the performance and accuracy of the Random Forest model across five datasets (T1, T2, T3, T4, and T5). For the T1 dataset, the initial accuracy of the Random Forest model was 0.7. After applying the optimization techniques, the accuracy increased to 0.75. This indicates a real improvement in model performance, with a gain of 0.05 in accuracy. Similarly, for the T2 dataset, the initial accuracy of the Random Forest model was 0.5. However, after optimization, the accuracy significantly improved to 0.71. This demonstrates a substantial enhancement, with an increase of 0.21 in accuracy. The Random Forest model's initial accuracy for the T3 dataset was 0.4. Through the optimization process, the accuracy improved to 0.67, representing a remarkable boost of 0.27 in accuracy. The T4 dataset initially yielded an accuracy of 0.5 for the Random Forest model. After optimization, the accuracy slightly improved to 0.6, a 0.1 improvement in accuracy. Finally, for the T5 dataset, the Random Forest model initially achieved an accuracy 0.6. After optimization, the accuracy reached 0.7, exhibiting a moderate improvement of 0.1. Overall, the random search results for model tuning and optimization have proven to be highly effective in enhancing the performance and accuracy of the Random Forest model across all five datasets. Optimization has consistently resulted in significant improvements, leading to more accurate predictions and better model performance.

Esembled Models In ensemble modelling, numerous independent models are combined to provide a more accurate and reliable model. The models used for this are *Decision tree*, *Random forest*, *Logistic regression*, *ADA boost*

Ensemble models combine multiple individual models to create a more robust and accurate model. The accuracy of four independent models—Decision Tree, Random Forest, AdaBoost, and

Table 4.10: Esembled models accuracy.

Models	T1	T2	T3	T4	T5
DT	0.5	0.4	O.4	0.53	0.6
RF	0.6	0.4	0.4	0.38	0.53
LR	0.5	0.6	0.53	0.23	0.6
ADA	0.6	0.5	0.4	0.53	0.4
EM	0.6	0.5	0.46	0.5	0.53

Table 4.11: Optimized esembled models accuracy.

Models	T1	T2	T3	T4	T5
EM Accuracy	0.7	0.66	0.5	0.6	0.6

Logistic Regression—are combined to create the ensemble model in this instance. The ensemble model was created utilizing the accuracy of these separate models for each dataset (T1, T2, T3, T4, and T5).

The ensemble model was optimized, resulting in improved accuracies for each dataset: For dataset T4, the optimized accuracy of the ensemble model increased to 0.6. For dataset T5, the optimized accuracy of the ensemble model increased to 0.6.

- For dataset T1, the optimized accuracy of the ensemble model increased to 0.7.
- For dataset T2, the optimized accuracy of the ensemble model increased to 0.6.
- For dataset T3, the optimized accuracy of the ensemble model increased to 0.5.
- For dataset T4, the optimized accuracy of the ensemble model increased to 0.6.
- For dataset T5, the optimized accuracy of the ensemble model increased to 0.6.

The optimization process resulted in higher accuracies for the ensemble model across all datasets, indicating improved performance and predictive capability. By combining the strengths of different models, the ensemble model leverages their individual accuracies to achieve a more accurate and reliable prediction. The optimization further fine-tunes the ensemble model, leading to enhanced accuracy and performance for each dataset.

Grand Average of Optimized model

$$\frac{0.75 + 0.71 + 0.67 + 0.6 + 0.7}{5} = 0.68$$

Grand Average = 0.68

68%

4.4.3 Support Vector Machine (SVM)

This section shwos the results of SVM model.

Table 4.12: Accuracy output for SVM model.

Model	T1	T2	T3	T4	T5
SVM Accuracy	0.69	0.46	0.46	0.46	0.6

The performance of the SVM algorithm is summarized in **Table 4.12**, showcasing the accuracy achieved on each dataset. It demonstrated good performance on dataset T1, achieving the highest accuracy of 69%. Dataset T5 also showed an accuracy of 60%. However, the remaining datasets yielded comparable results, all at approximately 46%. While the SVM algorithm proves to be effective for certain datasets, its accuracy might not be sufficient for accurately and reliably predicting team success. Further improvements or alternative approaches might be necessary to achieve more accurate and effective predictions in this context.

4.5 Sample Instances And Interpretability methods

Feature Importance Feature importance in Random Forest refers to the measure of how much each feature contributes to the model's predictive power. It is determined by analyzing the reduction in impurity or error when each feature is used in decision-making within the ensemble of trees.

Feature importance of T1

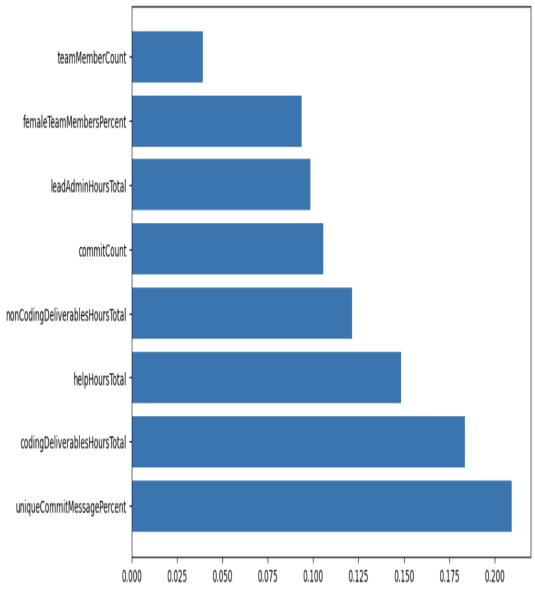


Figure 4.2: T1 Feature Importance Bar Plot

Variables	%
teamMemberCount	0.03
femaleTeamMembersPercent	0.093
leadAdminHoursTotal	0.098
commitCount	0.10
nonCodingDeliverablesHoursTotal	0.12
helpHoursTotal	0.14
codingDeliverablesHoursTotal	0.18
uniqueCommitMessagePercent	0.20

Table 4.13: A table with feature in percentage

In performing a random forest, feature importance measures the relevance of each feature in predicting the target variable. Here's a description of the importance of the features you provided,ranked from highest to lowest:

- codingDeliverablesHoursTotal (0.20): The time spent on coding deliverables is highly important for predictions.
- uniqueCommitMessagePercent (0.18): The percentage of unique commit messages is a strong predictor.
- helpHoursTotal (0.14): The total hours spent on providing help or support is significant.
- nonCodingDeliverablesHoursTotal (0.13): Time spent on non-coding deliverables also contributes significantly.

- leadAdminHoursTotal (0.110118): Hours spent in lead or administrative roles have a reasonable impact.
- femaleTeamMembersPercent (0.098): The percentage of female team members still influences predictions, but less so.
- commitCount (0.093): The total number of commits made contributes, but to a lesser extent.
- teamMemberCount (0.03): The count of team members has the least impact.

Feature importance of T2

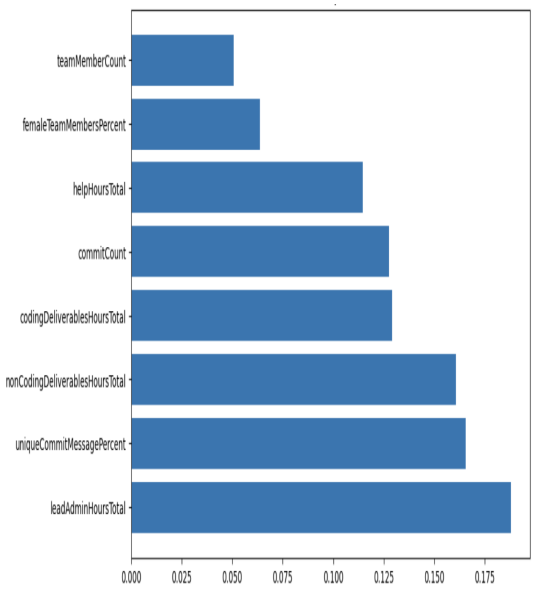


Figure 4.3: T2 Feature Importance Bar Plot

Variables	%
teamMemberCount	0.05
femaleTeamMembersPercent	0.06
helpHoursTotal	0.011
commitCount	0.12
codingDeliverablesHoursTotal	0.125
nonCodingDeliverablesHoursTotal	0.16
uniqueCommitMessagePercent	0.165
leadAdminHoursTotal	0.18

Table 4.14: A table with feature in percentage

For the T2 dataset, feature importance reveals the relevance of each feature in predicting the target variable in random forest. Among the features provided, **leadAdminHoursTotal** (0.18) stands out as highly important, indicating that the total hours spent by team members in a lead or administrative role significantly influence the model's predictions. **UniqueCommitMessagePercent** (0.165) and **nonCodingDeliverablesHoursTotal** (0.125) also have reasonable importance, suggesting their contribution to the predictions. Other features, such as **codingDeliverablesHoursTotal** (0.16), **commitCount** (0.12), **helpHoursTotal** (0.11), **femaleTeamMembersPercent** (0.06), and **teamMemberCount** (0.05) have comparatively lesser importance in the model. Understanding feature importance aids in focusing efforts on the most influential features to enhance the model's accuracy and interpretability.

Feature importance of T3

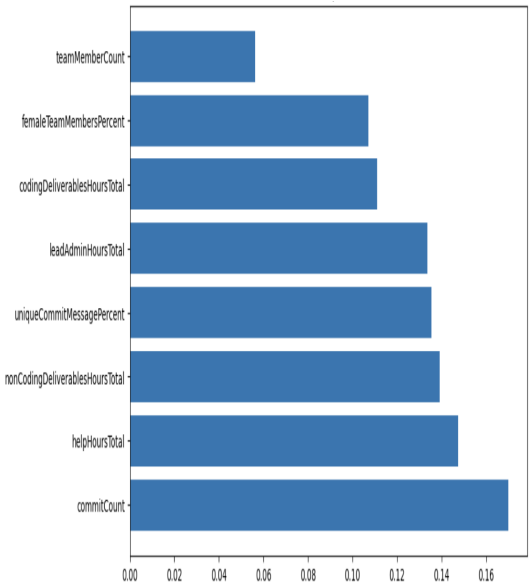


Figure 4.4: T3 Feature Importance Bar Plot

Variables	%
teamMemberCount	0.05
femaleTeamMembersPercent	0.10
codingDeliverablesHoursTotal	0.11
leadAdminHoursTotal	0.13
uniqueCommitMessagePercent	0.135
nonCodingDeliverablesHoursTotal	0.139
helpHoursTotal	0.14
commitCount	0.16

Table 4.15: A table with feature in percentage

In random forest analysis for T3 dataset, feature importance indicates the significance of each feature in predicting the target variable. Among the given features, **commitCount** (0.16) stands out as highly important, indicating its strong influence on the model's predictions. Other moderately important features include **helpHoursTotal** (0.14), **nonCodingDeliverablesHoursTotal** (0.139), and **uniqueCommitMessagePercent** (0.135). On the other hand, features like **leadAdminHoursTotal** (0.13), **codingDeliverablesHoursTotal** (0.11), **femaleTeamMembersPercent** (0.10), and **teamMemberCount** (0.05) have relatively less importance. Recognizing feature importance helps prioritize efforts in enhancing model accuracy and interpretability by focusing on the most influential features.

Feature importance of T4

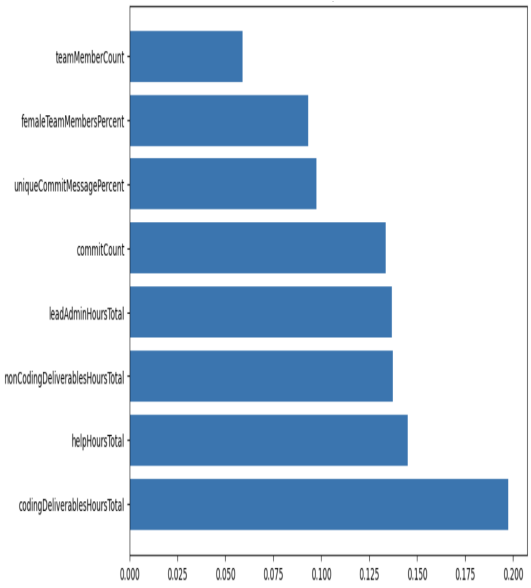


Figure 4.5: T3 Feature Importance Bar Plot

Variables	%
teamMemberCount	0.05
femaleTeamMembersPercent	0.10
codingDeliverablesHoursTotal	0.11
leadAdminHoursTotal	0.13
uniqueCommitMessagePercent	0.135
nonCodingDeliverablesHoursTotal	0.139
helpHoursTotal	0.14
commitCount	0.16

Table 4.16: A table with feature in percentage

In random forest analysis for T3 dataset, feature importance indicates the significance of each feature in predicting the target variable. Among the given features, **commitCount** (0.16) stands out as highly important, indicating its strong influence on the model's predictions. Other moderately important features include helpHoursTotal (0.14), **nonCodingDeliverablesHoursTotal** (0.139), and **uniqueCommitMessagePercent** (0.135). On the other hand, features like **leadAdminHoursTotal** (0.13), **codingDeliverablesHoursTotal** (0.11), **femaleTeamMembersPercent** (0.10), and **teamMemberCount** (0.05) have relatively less importance. Recognizing feature importance helps prioritize efforts in enhancing model accuracy and interpretability by focusing on the most influential features.

Feature importance of T4

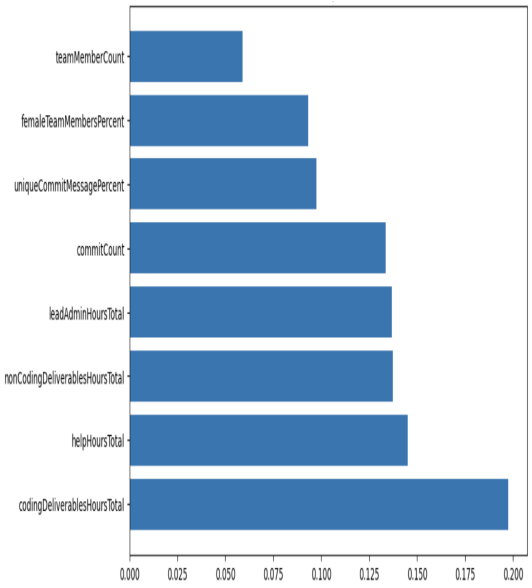


Figure 4.6: T4 Feature Importance Bar Plot

Variables	%
teamMemberCount	0.05
femaleTeamMembersPercent	0.093
uniqueCommitMessagePercent	0.097
commitCount	0.133
leadAdminHoursTotal	0.136
nonCodingDeliverablesHoursTotal	0.137
helpHoursTotal	0.145
codingDeliverablesHoursTotal	0.197

Table 4.17: A table with feature in percentage

For the T4 dataset, the provided features, **codingDeliverablesHoursTotal** (0.197) emerges as highly important, indicating its strong impact on the model's predictions. Other features with moderate importance include **helpHoursTotal** (0.145), **nonCodingDeliverablesHoursTotal** (0.137), **leadAdminHoursTotal** (0.136), and commitCount (0.133). On the other hand, **uniqueCommitMessagePercent** (0.097), **femaleTeamMembersPercent** (0.093), and **teamMemberCount** (0.05) hold relatively lower importance in the model. Understanding feature importance assists in prioritizing efforts to enhance model accuracy and interpretability by focusing on the most influential features.

Feature importance of T5

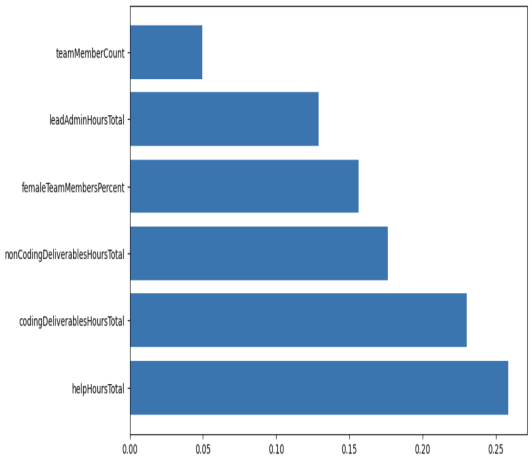


Figure 4.7: T4 Feature Importance Bar Plot

Variables	%
teamMemberCount	0.04
leadAdminHoursTotal	0.12
femaleTeamMembersPercent	0.15
nonCodingDeliverablesHoursTotal	0.17
codingDeliverablesHoursTotal	0.23
helpHoursTotal	0.25

Table 4.18: A table with feature in percentage

From the features in T5 dataset, **helpHoursTotal** (0.25) and **codingDeliverablesHoursTotal** (0.23) have high importance, while **femaleTeamMembersPercent** (0.15) holds moderate importance. **nonCodingDeliverablesHoursTotal** (0.17) and leadAdminHoursTotal (0.12) are moderately important as well. **teamMemberCount** (0.04) has the lowest importance. Understanding feature importance aids in improving accuracy and interpretability by focusing on influential features.

4.5.1 Sample Instances And LIME

Random Forest utilizes features from a set of decision trees to make decisions. To enhance interpretability and accuracy, LIME (Local Interpretable Model-Agnostic Explanations) is used. LIME approximates complex models with simpler ones for explaining predictions. By applying LIME to five datasets, it provides insights and assesses accuracy in decision-making. Random Forest features and LIME interpretation aid in informed decision-making across datasets.

T1

The model places more emphasis on features like the percentage of unique commit messages, the total hours spent on coding deliverables, and the amount of help provided. These features are likely the most influential factors in the model's decision-making process. Below is an analysis of instaces 6 and 8 from the T1 dataset with LIME results.

Table 4.19: T1 sample observations.

Instance number	uniqueCommitMessagePercent	codingDeliverablesHoursTotal	helpHoursTotal	nonCodingDeliverablesHoursTotal	productLetterGrade
6	1	9	3	0.16	1
8	0.9	28.5	5.1	0.14	1

To study the prediction process, focus on instances where these variables exhibit distinct characteristics. Based on observations, it is noted that high values of **uniqueCommitMessagePercent** and **codingDeliverablesHoursTotal** tend to indicate a high probability of success for a team.

Considering **uniqueCommitMessagePercent**. These variable measures the percentage of unique commit messages within a team's code repository. A high value of **uniqueCommitMessagePercent** suggests that team members are actively and consistently committing code changes with informative messages. This indicates good collaboration, effective communication, and an organized development process. Consequently, it is reasonable to assume that teams with a high **uniqueCommitMessagePercent** are more likely to succeed in their projects.

Secondly, **codingDeliverablesHoursTotal** represents the total number of hours dedicated to coding deliverables by a team. A higher value of **codingDeliverablesHoursTotal** implies that the team has invested a significant amount of time in actual coding activities. This shows a strong work enthusiasm, commitment, and a focus on producing positive results. Teams that allocate substantial time to coding deliverables are more likely to complete their tasks successfully and deliver high-quality outcomes.

By examining instances with high **uniqueCommitMessagePercent** and **codingDeliverablesHoursTotal**, we can delve deeper into the decision-making process of the model. By understanding the specific patterns and features that contribute to successful predictions, we can gain valuable insights into what the model deems as important for project success. By carefully selecting specific instances and focusing on the variables with the highest influence on prediction, such as **uniqueCommitMessagePercent** and **codingDeliverablesHoursTotal** , we can gain a better understanding of the model's decision-making process and identify key factors associated with project success.

Figure 4.8: T1 6 Local Interpretable Model-Agnostic Explanations Output

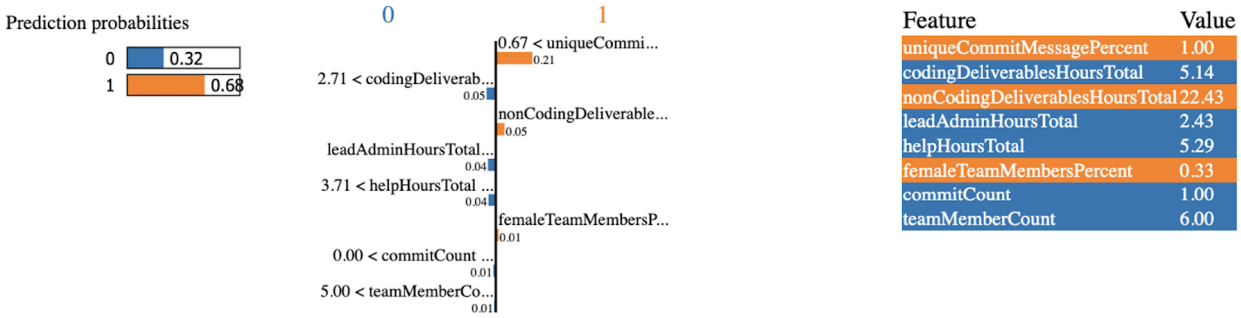
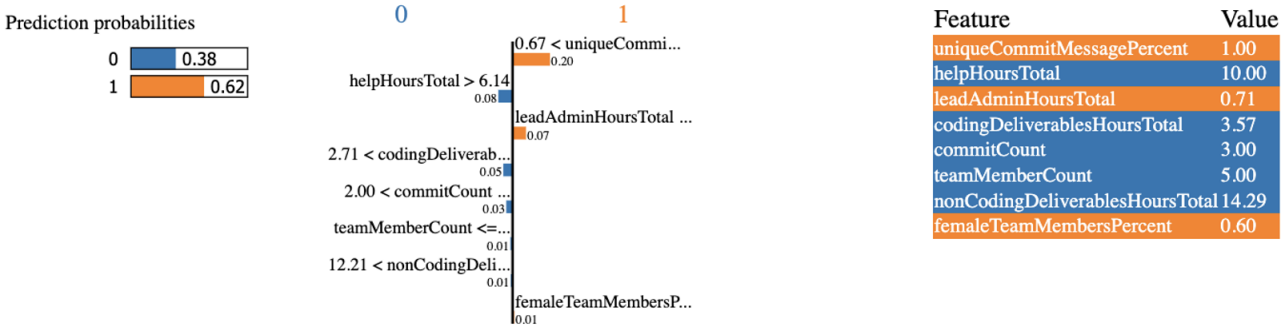


Figure 4.9: T1 8 Local Interpretable Model-Agnostic Explanations Output



- The LIME output for 6th instance provides predictions for the probabilities of two classes: 0 and 1. In this case, the predicted probability for class 0 is 32%, and the predicted probability for class 1 is 68%. These probabilities indicate the model's confidence in assigning the respective classes to the given input. Based on the variables, it determines if a team completes a milestone or not with percentages.

- The LIME output for 8th instance provides predictions for the probabilities of two classes: 0 and 1. In this case, the predicted probability for class 0 is 38%, and the predicted probability for class 1 is 62%. These probabilities indicate the model's confidence in assigning the respective classes to the given input. Based on the variables, it determines if a team completes a milestone or not with percentages.

T2 Sample Instances

The analysis of instances 4 and 12 in the T2 dataset.

Table 4.20: T2 sample observations.

Instance number	leadAdminHoursTotal	uniqueCommitMessagePercent	nonCodingDeliverablesHoursTotal	codingDeliverablesHoursTotal	productLetterGrade
4	3	0.9	72.5	19	1
12	2.5	0.2	34	5	1

In T2 dataset, the variables with the highest influence on the prediction are **uniqueCommitMessagePercent**, **codingDeliverablesHoursTotal**, **leadAdminHoursTotal**, and **nonCodingDeliverables**. Observations indicate that high values of **uniqueCommitMessagePercent** and **codingDeliverablesHoursTotal** are associated with a higher probability of team success. Teams with a high **uniqueCommitMessagePercent** demonstrate effective collaboration and communication, while those with a high **codingDeliverablesHoursTotal** display strong commitment and productivity in coding activities. By examining instances with these characteristics, we can gain insights into the model's decision-making process and identify key factors related to project success.

Figure 4.10: T2 4 Local Interpretable Model-Agnostic Explanations output

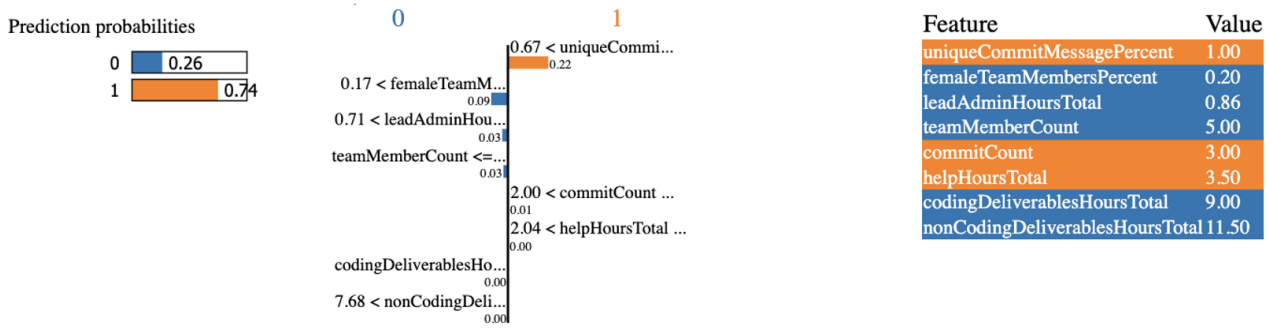
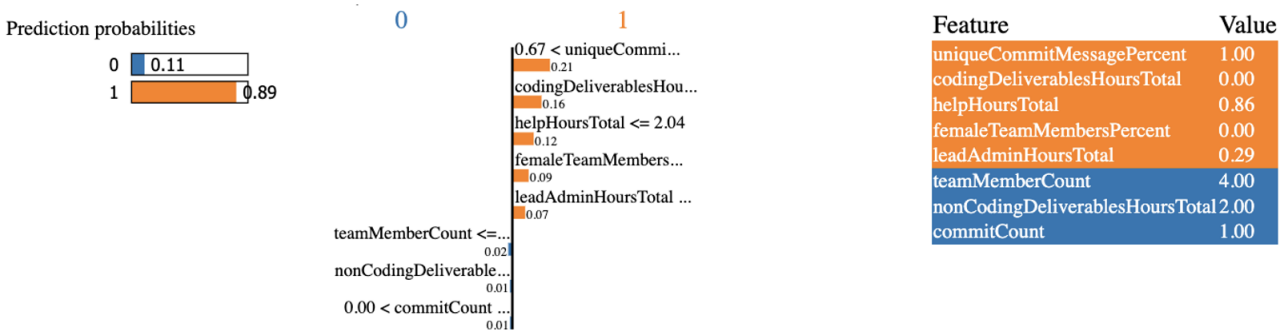


Figure 4.11: T2 12 Local Interpretable Model-Agnostic Explanations output



- The LIME output for 4th instance provides predictions for the probabilities of two classes: 0 and 1. In this case, the predicted probability for class 0 is 26%, and the predicted probability for class 1 is 74%. These probabilities indicate the model's confidence in assigning the respective classes to the given input. Based on the variables, it determines if a team completes a milestone or not with percentages.

- The LIME output for 12th instance provides predictions for the probabilities of two classes: 0 and 1. In this case, the predicted probability for class 0 is 11%, and the predicted probability for class 1 is 89%. These probabilities indicate the model's confidence in assigning the respective classes to the given input. Based on the variables, it determines if a team completes a milestone or not with percentages.

T3 Sample Instances

The analysis of instances 6 and 12 in the T3 dataset.

Table 4.21: T3 sample observations.

Instance number	codingDeliverablesHoursTotal	helpHoursTotal	nonCodingDeliverablesHoursTotal	uniqueCommitMessagePercent	productLetterGrade
6	212	39.0	26.4	0.6	0
12	7.8	11.8	10.4	0.6	0

The variables that have the highest impact on predictions in the T3 dataset are **uniqueCommitMessagePercent**, **codingDeliverablesHoursTotal**, **HelpHoursTotal**, and **nonCodingDeliverablesHoursTotal**. Analysis of the dataset suggests that higher values of **uniqueCommitMessagePercent** is linked to an increased likelihood of team success. Teams with a low **uniqueCommitMessagePercent** demonstrate low performance and communication, while those with a low **codingDeliverablesHoursTotal** show weak dedication and productivity in coding tasks. These may lead to failure in the reaching the milestone.

Figure 4.12: T3 6 Local Interpretable Model-Agnostic Explanations output

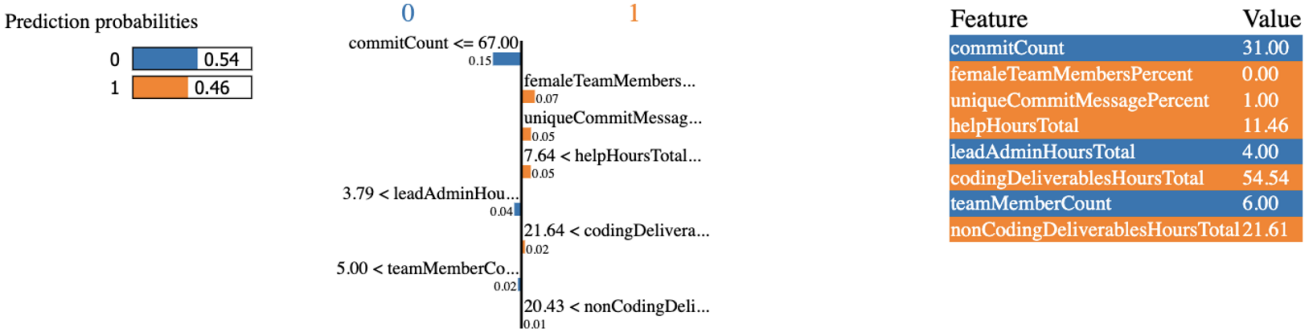
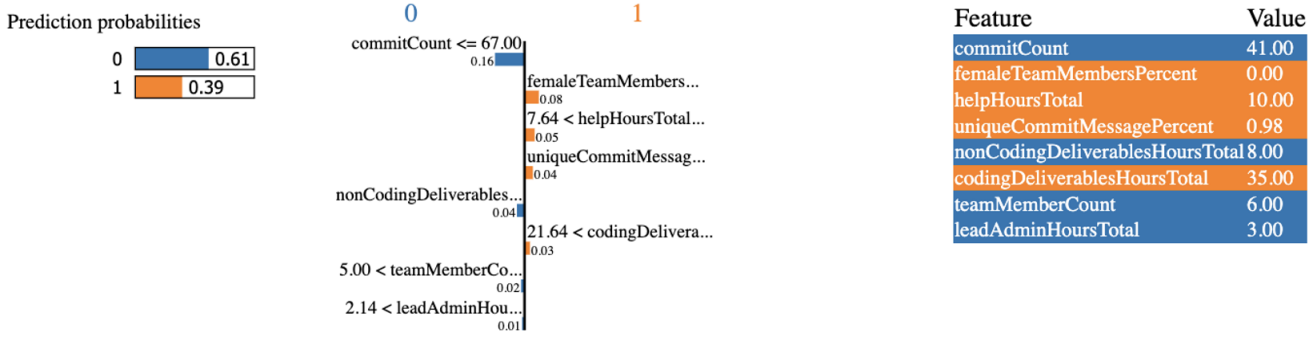


Figure 4.13: T3 12 Local Interpretable Model-Agnostic Explanations output



- The LIME output for 6th instance provides predictions for the probabilities of two classes: 0 and 1. In this case, the predicted probability for class 0 is 54%, and the predicted probability for class 1 is 46%. These probabilities indicate the model's confidence in assigning the respective classes to the given input. Based on the variables, it determines if a team completes a milestone or not with percentages.

- The LIME output for 12th instance provides predictions for the probabilities of two classes: 0 and 1. In this case, the predicted probability for class 0 is 11%, and the predicted probability for class 1 is 89%. These probabilities indicate the model's confidence in assigning the respective classes to the given input. Based on the variables, it determines if a team completes a milestone or not with percentages.

T4 Sample Instances

The analysis of instances 4 and 7 in the T4 dataset.

Table 4.22: T4 sample observation.

Instance number	helpHoursTotal	codingDeliverablesHoursTotal	nonCodingDeliverablesHoursTotal	leadAdminHoursTotal	productLetterGrade
4	61.4	46.1	25.2	2.8	1
7	72.5	21.6	13.1	4.7	0

In the T4 dataset, the variables codingDeliverablesHoursTotal, helpHoursTotal, nonCodingDeliverablesHoursTotal, uniqueCommitMessagePercent have the biggest effects on predictions. Higher helpHoursTotal and nonCodingDeliverablesHoursTotal values are associated with a higher probability of team success, according to an analysis of the dataset. on the other hand, low performance on both lead to failure in success as shown in the sample instances.

Figure 4.14: T4 4 Local Interpretable Model-Agnostic Explanations output.

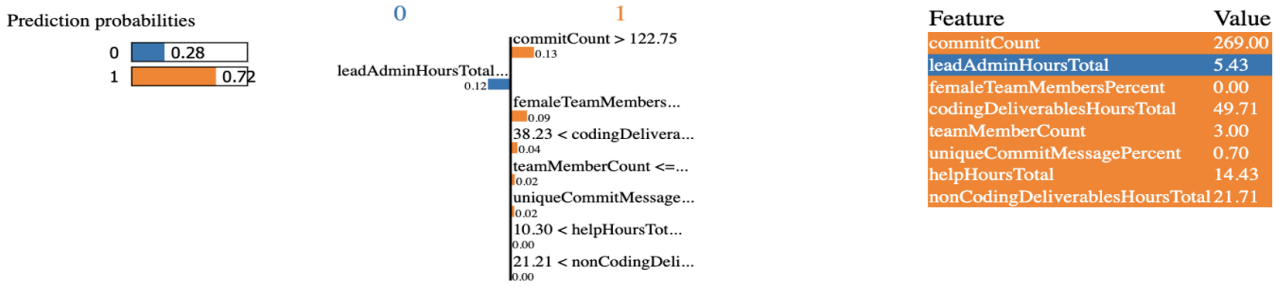
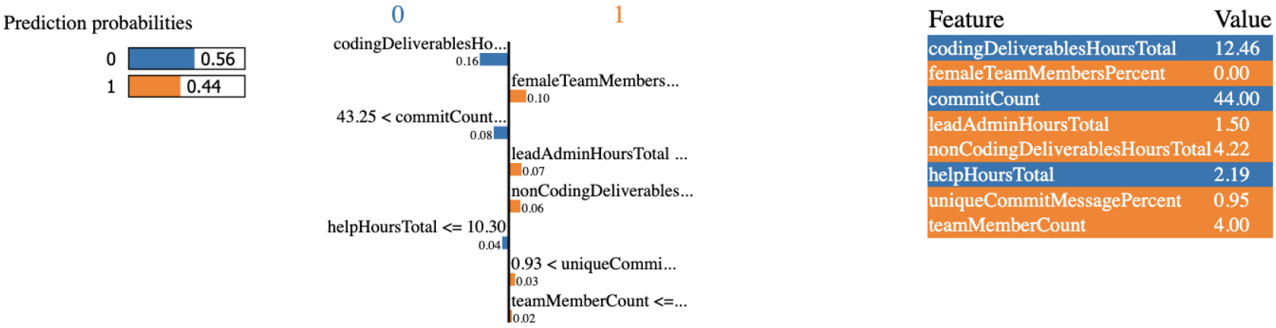


Figure 4.15: T4 7 Local Interpretable Model-Agnostic Explanations output.



- The LIME output for 4th instance provides predictions for the probabilities of two classes: 0 and 1. In this case, the predicted probability for class 0 is 32%, and the predicted probability for class 1 is 68%. These probabilities indicate the model's confidence in assigning the respective classes to the given input. Based on the variables, it determines if a team completes a milestone or not with percentages.

- The LIME output for 7th instance provides predictions for the probabilities of two classes: 0 and 1. In this case, the predicted probability for class 0 is 56%, and the predicted probability for class 1 is 44%. These probabilities indicate the model's confidence in assigning the respective classes to the given input. Based on the variables, it determines if a team completes a milestone or not with percentages.

T5 Sample Instances

The analysis of instances 5 and 7 in the T5 dataset.

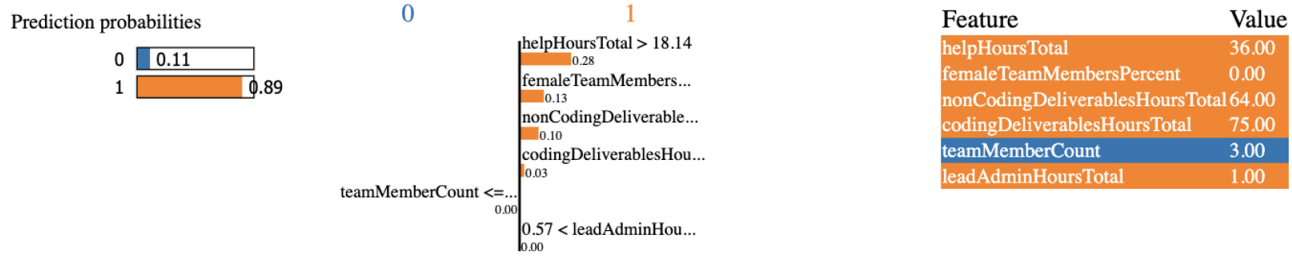
Table 4.23: T5 sample observations.

Instance number	helpHoursTotal	codingDeliverablesHoursTotal	nonCodingDeliverablesHoursTotal	femaleTeamMembersPercent	productLetterGrade
5	47	149	54.5	0	1
7	31.5	162.5	126.5	0	1

In T5 dataset, the variables with the highest influence on the prediction are **helpHoursTotal, codingDeliverablesHoursTotal helpHoursTotal, nonCodingDeliverablesTotal,** and **femaleTeamMembersPercent.** Observations indicate that high value of **helpHoursTotal** is associated with a higher probability of team success. Teams with a high **helpHoursTotal** tend to succed in achieving the milestone.

Figure 4.16: T5 5 Local Interpretable Model-Agnostic Explanations output.

Figure 4.17: T5 7 Local Interpretable Model-Agnostic Explanations output.



- The LIME output for 5th instance provides predictions for the probabilities of two classes: 0 and 1. In this case, the predicted probability for class 0 is 30%, and the predicted probability for class 1 is 70%. These probabilities indicate the model's confidence in assigning the respective classes to the given input. Based on the variables, it determines if a team completes a milestone or not with percentages.

- The LIME output for 7th instance provides predictions for the probabilities of two classes: 0 and 1. In this case, the predicted probability for class 0 is 11%, and the predicted probability for class 1 is 89%. These probabilities indicate the model's confidence in assigning the respective classes to the given input. Based on the variables, it determines if a team completes a milestone or not with percentages.

Chapter 5

Conclusion, Recommendations, and Future Work

5.1 Conclusion

The thesis presented an analysis of predicting team success toward milestone achievement. The research covered multiple stages, including data cleaning and processing, statistical tests for goodness of fit and normality, feature engineering and selection, and the application of machine learning algorithms.

The initial phase involved extensive data cleaning and processing to ensure dataset quality and reliability. Subsequently, various statistical tests were employed to assess the goodness of fit and normality of variables, providing a solid foundation for subsequent modeling phases. Feature engineering techniques were explored to improve prediction performance by creating additional meaningful features. This involved transforming existing variables, deriving new ones, and selecting the most relevant features for the prediction task.

Different machine learning methods, such as logistic regression, decision trees, special vector machine, and random forests, were used on the datasets. The SVM algorithm demonstrates effectiveness for specific datasets, but its accuracy may not be adequate for accurately and reliably predicting team success.Through comparative evaluation, random forest emerged as the

most effective algorithm for predicting team success toward milestone achievement, consistently demonstrating superior accuracy across the provided datasets. To optimize the chosen machine learning model, hyperparameter tuning was performed to fine-tune the model's performance. This optimization process ensured that the model was optimized for the best possible prediction accuracy. Furthermore, interpretability methods, particularly LIME, were utilized to gain insights into how the model predicted specific instances. This approach enhanced the understanding of the underlying factors driving team success.

The research's findings changed the initial hypothesis, reflecting the new insights gained through the analysis process. Also, the findings highlight the effectiveness of the random forest algorithm and offer valuable insights into the key variables contributing to team success in reaching milestones. After hyperparameter tuning, they significantly improved the prediction accuracy, increasing from an initial accuracy of 54% to 68%. Key variables such as uniqueCommitMessagePercent, codingDeliverablesHoursTotal, and helpHoursTotal were identified as significant contributors to the prediction accuracy. After careful analysis of the results, it was observed that certain variables demonstrated a strong contrivution with their higher frequency occurrences showing a higher probability of success. It is essential to understand the significance of these variables and take action on them when organising a team or group with the aim of succeeding.

5.2 Future Work

The following are recommendations for further work in relation to the analysis of predicting team success in achieving milestones:

- External validation on independent datasets from different domains or organizations.
- Longitudinal analysis to examine team success trends and predictors over time.
- Investigation of temporal dynamics and dependencies in team success prediction.
- Exploration of ensemble methods to improve predictive performance.

Several thorough suggestions for future study might be considered to improve the analysis of predicting team success in reaching milestones. Including more approaches and visualization, strategies can increase the model's explainability. This would provide a deeper understanding of how the model makes predictions and increase trust and transparency in the results. Additionally, developing a real-time prediction system would allow for continuous team performance monitoring, enabling firms to identify potential issues and take fast corrective action proactively. Providing current predictions based on the most recent data enables quick decision-making. Additionally, considering additional performance metrics beyond accuracy would provide a more comprehensive evaluation of the model's performance. Moreover, incorporating qualitative research methods like interviews or surveys would gather subjective insights from team members. This qualitative data would complement the quantitative analysis, providing a more holistic understanding of the factors influencing team success. Lastly, exploring causal inference techniques, such as propensity score matching or instrumental variable analysis, would allow for examining causal relationships between identified variables and team success. This analysis would help determine the direct impact of critical variables on success and uncover potential confounding factors that need to be considered. By pursuing these recommendations, the analysis can gain a deeper understanding of the underlying factors driving team success and improve the overall effectiveness of the predictive model.

5.3 APPENDIX

Table 5.1: Dataset T1 Summary statistics.

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
teamMemberCount	64.0	5.17	1.20	3.00	5.00	5.00	6.00	7.00
femaleTeamMembersPercent	64.0	0.18	0.17	0.00	0.00	0.17	0.25	0.83
teamMemberResponseCount	64.0	7.20	3.88	1.00	3.00	8.00	10.00	14.00
meetingHoursTotal	64.0	12.36	8.35	1.00	6.48	10.46	16.69	37.21
inPersonMeetingHoursTotal	64.0	9.34	6.33	1.00	4.67	8.57	12.32	26.36
nonCodingDeliverablesHoursTotal	64.0	13.43	8.66	1.00	7.77	11.43	18.65	45.00
codingDeliverablesHoursTotal	64.0	4.05	4.80	0.00	0.96	2.64	5.41	28.57
helpHoursTotal	64.0	4.52	3.42	0.25	1.91	3.64	6.63	15.50
leadAdminHoursResponseCount	64.0	1.38	0.72	0.00	1.00	1.00	2.00	3.00
leadAdminHoursResponseCount	64.0	1.38	0.72	0.00	1.00	1.00	2.00	3.00
leadAdminHoursTotal	56.0	1.30	1.00	0.29	0.71	0.96	1.45	4.57
commitCount	64.0	3.83	5.30	0.00	0.00	2.00	5.00	23.00
uniqueCommitCountMessageCount	64.0	3.09	4.83	0.00	0.00	1.50	3.00	22.00
uniqueCommitCountMessagePercent	64.0	0.54	0.44	0.00	0.00	0.67	1.00	1.00
commitMessageLengthTotal	64.0	96.09	128.76	0.00	0.00	47.50	130.75	488.00
issueCount	64.0	0.19	0.47	0.00	0.00	0.00	0.00	2.00
onTimeIssueCount	64.0	0.19	0.47	0.00	0.00	0.00	0.00	2.00

Table 5.2: Dataset T2 Summary statistics.

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
teamMemberCount	74.0	5.19	1.21	3.00	5.00	5.00	6.00	7.00
femaleTeamMembersPercent	74.0	0.17	0.17	0.00	0.00	0.17	0.20	0.83
teamMemberResponseCount	74.0	21.70	7.50	8.00	16.00	20.00	29.00	37.00
meetingHoursTotal	74.0	47.53	21.72	11.07	31.04	45.86	60.38	106.86
inPersonMeetingHoursTotal	74.0	36.49	18.78	8.14	20.84	34.66	46.72	90.11
nonCodingDeliverablesHoursTotal	74.0	41.45	18.62	10.43	25.61	41.25	53.44	92.21
codingDeliverablesHoursTotal	74.0	29.37	20.60	3.00	14.50	22.79	40.20	102.93
helpHoursTotal	74.0	17.25	8.77	1.14	11.34	16.25	21.36	44.50
leadAdminHoursResponseCount	74.0	4.04	1.34	0.00	3.25	4.00	5.00	7.00
leadAdminHoursTotal	71.0	4.67	3.58	0.43	2.54	3.64	5.57	18.14
commitCount	74.0	53.55	50.17	0.00	11.00	41.00	76.75	180.00
uniqueCommitCountMessageCount	74.0	36.26	35.79	0.00	7.00	28.50	56.50	133.00
uniqueCommitCountMessagePercent	74.0	0.62	0.30	0.00	0.48	0.67	0.85	1.00
commitMessageLengthTotal	74.0	1830	1932	0.00	193	1175	3095	6456
issueCount	74.0	1.19	1.30	0.00	0.00	1.00	2.00	7.00
onTimeIssueCount	74.0	0.92	1.00	0.00	0.00	1.00	1.00	5.00

Table 5.3: Dataset T3 Summary statistics.

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
teamMemberCount	74.0	5.19	1.21	3.00	5.00	5.00	6.00	7.00
femaleTeamMembersPercent	74.0	0.17	0.17	0.00	0.00	0.17	0.20	0.83
teamMemberResponseCount	74.0	16.88	7.66	2.00	10.00	20.00	22.75	36.00
meetingHoursTotal	74.0	41.25	25.88	7.15	20.89	37.11	52.29	122.57
inPersonMeetingHoursTotal	74.0	33.85	22.93	2.00	15.38	30.00	46.25	99.86
nonCodingDeliverablesHoursTotal	74.0	21.31	12.49	3.00	10.79	19.75	28.75	59.43
codingDeliverablesHoursTotal	74.0	59.14	41.92	1.00	22.79	55.13	84.33	207.50
helpHoursTotal	74.0	16.85	12.70	0.25	8.45	13.34	21.71	60.57
leadAdminHoursResponseCount	74.0	3.15	1.50	0.00	2.00	4.00	4.00	6.00
leadAdminHoursTotal	69.0	4.48	5.13	0.50	1.95	3.64	5.00	37.29
commitCount	74.0	123.26	92.91	0.00	65.00	105.50	164.00	463.00
uniqueCommitMessageCount	74.0	92.70	63.70	0.00	53.25	81.00	125.25	273.00
uniqueCommitMessagePercent	74.0	0.72	0.26	0.00	0.63	0.81	0.90	1.00
commitMessageLengthTotal	74.0	6624	4964	0.00	3143	5595	9211	21925
issueCount	74.0	1.57	1.45	0.00	1.00	1.00	2.00	7.00
onTimeIssueCount	74.0	1.28	1.22	0.00	0.00	1.00	2.00	5.00

Table 5.4: Dataset T4 Summary statistics.

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
teamMemberCount	63	5.21	1.12	3.00	5.00	5.00	6.00	7.00
femaleTeamMembersPercent	63	0.19	0.17	0.00	0.07	0.17	0.25	0.83
teamMemberResponseCount	63	17.63	7.64	6.00	13.50	15.00	22.50	41.00
meetingHoursTotal	63	42.40	24.63	3.64	23.29	36.29	55.29	107.14
inPersonMeetingHoursTotal	63	32.71	18.33	3.64	18.79	29.07	43.71	84.40
nonCodingDeliverablesHoursTotal	63	25.39	14.33	4.22	15.29	21.43	31.29	67.50
codingDeliverablesHoursTotal	63	76.75	45.00	6.43	38.61	70.29	104.29	172.57
helpHoursTotal	63	21.44	13.63	1.71	9.95	19.57	28.43	70.29
leadAdminHoursResponseCount	63	3.40	1.31	0.00	3.00	3.00	4.00	8.00
leadAdminHoursTotal	61	4.07	2.68	0.00	2.14	3.29	5.43	12.29
commitCount	63	84.76	65.36	0.00	43.50	74.00	125.50	269.00
uniqueCommitCountMessageCount	63	75.46	54.37	0.00	39.00	71.00	116.00	187.00
uniqueCommitCountMessagePercent	63	0.80	0.30	0.00	0.79	0.93	0.97	1.00
commitMessageLengthTotal	63	5022	4044	0.00	1716	4726	7417	15753
issueCount	63	0.70	0.73	0.00	0.00	1.00	1.00	3.00
onTimeIssueCount	63	0.51	0.72	0.00	0.00	0.00	1.00	3.00

Table 5.5: Dataset T5 Summary statistics.

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
teamMemberCount	74.0	5.19	1.21	3.00	5.00	5.00	6.00	7.00
femaleTeamMembersPercent	74.0	0.17	0.17	0.00	0.00	0.17	0.20	0.83
teamMemberResponseCount	74.0	11.89	9.80	1.00	6.00	9.00	13.00	38.00
meetingHoursTotal	74.0	27.28	28.83	0.86	6.36	14.64	32.89	105.50
inPersonMeetingHoursTotal	74.0	18.58	18.61	0.57	4.93	11.14	25.64	68.00
nonCodingDeliverablesHoursTotal	74.0	20.72	25.51	0.00	4.13	10.29	27.12	126.50
codingDeliverablesHoursTotal	74.0	49.83	50.99	1.14	13.16	29.14	73.07	196
helpHoursTotal	74.0	13.23	13.45	0.00	3.04	8.57	18.21	58.00
leadAdminHoursResponseCount	74.0	2.12	1.89	0.00	1.00	2.00	2.00	7.00
leadAdminHoursTotal	67.0	2.70	4.18	0.00	0.57	1.14	2.39	26.00
commitCount	74.0	11.70	18.22	0.00	0.00	2.00	18.50	84.00
uniqueCommitCountMessageCount	74.0	11.70	18.22	0.00	0.00	2.00	18.50	84.00
uniqueCommitCountMessagePercent	74.0	0.50	0.46	0.00	0.00	0.58	0.99	1.00
commitMessageLengthTotal	74.0	552.14	961.19	0.00	0.00	76.50	723.75	4530.00
issueCount	74.0	0.16	0.37	0.00	0.00	0.00	0.00	1.00
onTimeIssueCount	74.0	0.09	0.29	0.00	0.00	0.00	0.00	1.00

Figure 5.1: T1 histogram plots for distributions

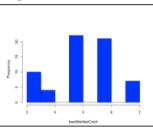
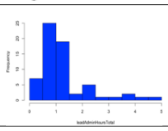
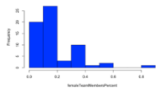
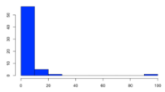
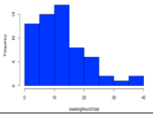
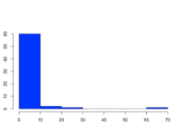
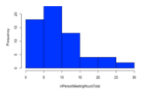
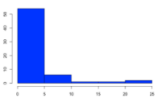
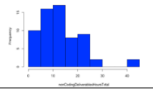
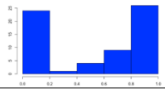
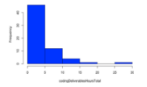
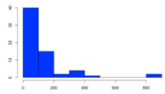
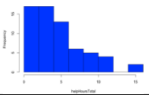
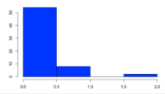
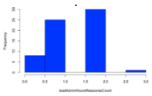
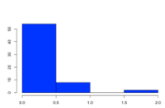
Histogram Plots	Descriptions	Histogram Plots	Descriptions
	**teamMemberCount** has most of the observations between 5 and 6 resulting to a statistical point of normal distributions.		The **leadAdminHoursTotal** Is having about half of the observation between 0-2, while the rest of are less above 2 hours.
	**FemaleTeamMembersPercent** shows a positively skewed distribution, where most of the points lay between 0 to 0.2. Also, there is a point that is considered to be an outlierthat is above 0.8.		The **commitCount** observations are mainly concentrated between 0 to 20. However, it can be considered a positively skewed distribution as shown by the plot. There are outliers between 90 -100.
	The **meetingHoursTotal** shows a positively skewed distribution, most points occur between 5 and 15.		The **uniqueCommitMessageCount** histogram is right-skewed, with the majority of observations concentrated between 0 and 5.
	**InPersonMeetingTotal** skews to the right with majority of the data between 5 and 10.		The **uniqueCommitMessagePercent** histogram is negatively skewed, with a high number of observations concentrated between 0 and 0.2.
	The **nonCodingDeliverablesTotal** distribution shows a rightly skewed distribution with observations that can be considered outliers**.**		The **uniqueCommitMessagePercent** histogram is negatively skewed, with a high number of observations concentrated between 0 and 0.2.
	The **codingDeliverablesTotal** histogram is right-skewed, with the majority of observations concentrated between 0 and 5. Also, some points are considered outliers from 25 to 30.		The **teamMemberResponseCount** plot shows more of a binomial distribution where most of the observations come from 8 to 12.
	The **helpHoursTotal** histogram is right-skewed, with the majority of observations concentrated between 0 and 5, and some points above 12 are considered outliers.		**InpersonmeetingTotal** skews to the right with majority of the data between 5 and 10.
	The **leadAdminHoursResponseCount** is relatively a normal distribution with more observations between 0.5 - 1, and 1.5 – 2.		The **leadAdminHoursResponseCount** is relatively a normal distribution with more observations between 0.5 - 1, and 1.5 – 2.

Figure 5.2: T1 pie charts for categorical variables

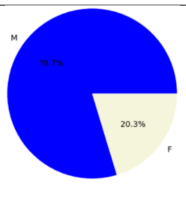
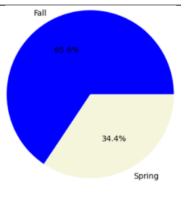
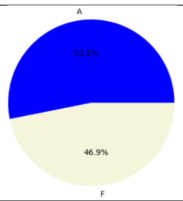
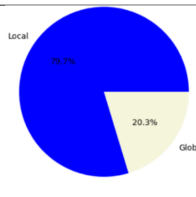
			
The **teamLeadGender** pie chart shows 79.7% of the team leaders are males while females make up the 20.3%.	According to the **semester** pie chart, 65.6% of semesters are fall, while 34.4% are spring.	The **productLetterGrade** pie chart shows that 53.1% of A success and 46.9% are F failures.	The **teamDistribution** pie graphic shows that Local teams make up 79.7% of all distributions and Global teams make up 20.3%.

Figure 5.3: T1 histogram plots for distributions

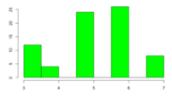
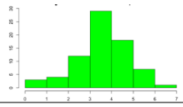
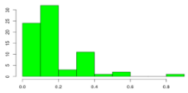
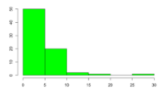
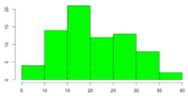
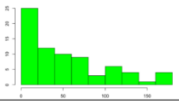
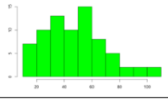
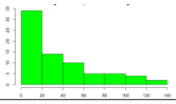
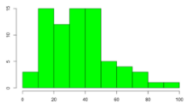
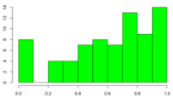
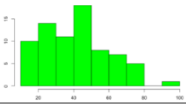
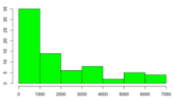
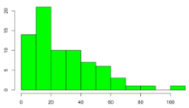
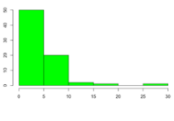
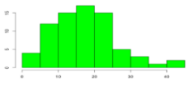
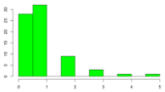
Histogram Plots	Description	Histogram Plots	Description
	**teamMemberCount** has most of the observations between 5 and 6 resulting to a statistical point of normal distributions.		**leadAdminHoursResponseCount** shown a clear bell curved shape making it a normal distribution
	**FemaleTeamMemberCount** shows a positively skewed distribution, where most of the points lay between 0 to 0.2. Also, there is a point that is an outlierthat is above 0.8.		**leadAdminHoursTotal** is a positively skewed distribution with an outlier between 25 and 30.
	**teamMemberResponseCount,** Most observations are in the middle, and this can be considered a normal distribution		**commitCount** has a gradual decrease in the in the observations making it right skewed
	**meetingHoursTotal** shows a normal distribution with smaller values from 80 and above		**UniqueCommitMessageCount** is a positively skewed distribution with most observations between 0 and 20.
	**inPersonMeetingHoursTotal** has the most observations between 10 to 50 leading a positively skewed distribution		**UniqueCommitMessagePercent** is a negatively skewed distribution with values considered outliers at 0.1
	**nonCodingDeliverablesHoursTotal** is a positively skewed distribution with a point considered outlier between 90 and 100		**commitMessageLengthTotal** is a positively skewed distribbution
	**codingDeliverablesHoursTotal** is a strong positively skewed distribution with an outlier above 100		**issueCount** has a positively skewed distribution
	**helpHoursTotal** show a normal distribution with most observations between 10 to 40		**onTimeIssueCount** is a positively skewed distribution with most observations between 0 and 1.

Figure 5.4: T1 pie charts for categorical variables

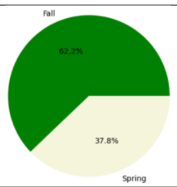
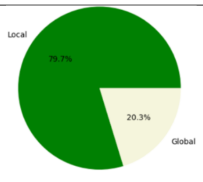
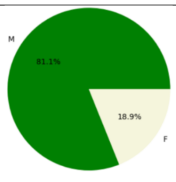
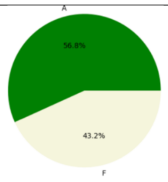
			
According to the **semester** pie chart, 62.2% of semesters are fall, while 37.8% are spring.	The **teamDistribution** pie graphic shows that Local teams make up 79.7% of all distributions and Global teams make up 20.3%.	The **teamLeadGender** pie chart shows 81.1% of the team leaders are males while females make up the 18.9%.	The **productLetterGrade** pie chart shows that 56.8% of A success and 43.2% are F failures.

Figure 5.5: T3 histograms for quantitative variables

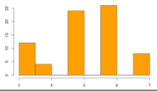
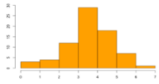
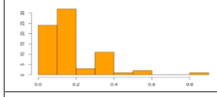
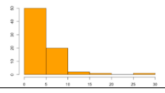
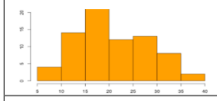
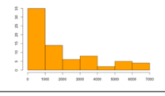
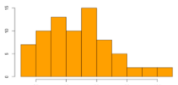
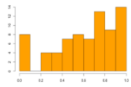
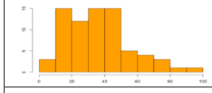
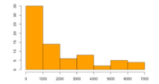
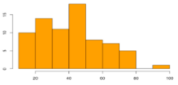
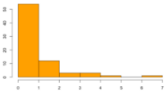
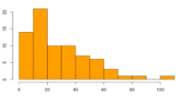
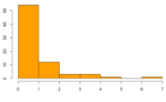
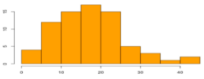
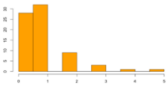
Histogram Plots	Description	Histogram Plots	Description
	**teamMemberCount** has most of the observations between 5 and 6 resulting to a statistical point of normal distributions.		**leadAdminHoursResponseCount** shown a clear bell curved shape making it a normal distribution.
	**FemaleTeamMemberCount** shows a positively skewed distribution, where most of the points lay between 0 to 0.2. Also, there is a point that is an outlierthat is above 0.8.		**leadAdminHoursTotal** is a positively skewed distribution with an outlier between 25 and 30.
	**teamMemberResponseCount** most observations are in the middle, and this can be considered a normal distribution.		**commitCount** has a gradual decrease in the in the observations making it right skewed.
	**meetingHoursTotal** shows a normal distribution with smaller values from 80 and above.		**UniqueCommitMessageCount** is is a positively skewed distribution with most observations between 0 and 20.
	**inPersonMeetingHoursTotal** has the most observations between 10 to 50 leading a positively skewed distribution.		**UniqueCommitMessagePercent** is a negatively skewed distribution with values considered outliers at 0.1
	**nonCodingDeliverablesHoursTotal** is a positively skewed distribution with a point considered outlier between 90 and 100.		**commitMessageLengthTotal** is a positively skewed distribution.
	**codingDeliverablesHoursTotal** is a strong positively skewed distribution with an outlier above 100.		**issueCount** has a positively skewed distribution.
	**helpHoursTotal** show a normal distribution with most observations between 10 to 40.		**onTimeIssueCount** is a positively skewed distribution with most observations between 0 and 1.

Figure 5.6: T3 pie charts for categorical variables

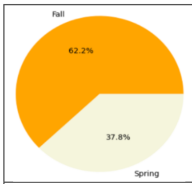
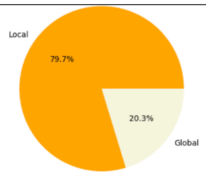
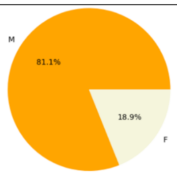
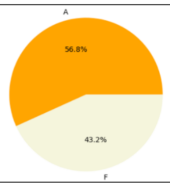
			
According to the **semester** pie chart, 62.2% of semesters are fall, while 37.8% are spring.	The **teamDistribution** pie graphic shows that Local teams make up 79.7% of all distributions and Global teams make up 20.3%.	The **teamLeadGender** pie chart shows 81.1% of the team leaders are males while females make up the 18.9%.	The **productLetterGrade** pie chart shows that 56.8% of A success and 43.2% are F failures.

Figure 5.7: T5 Histograms for quantitative variables

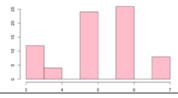
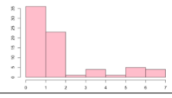
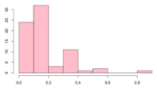
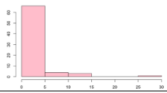
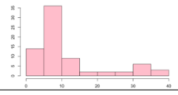
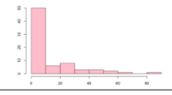
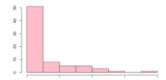
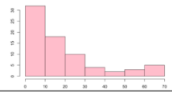
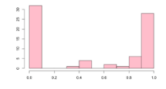
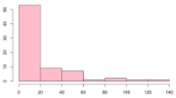
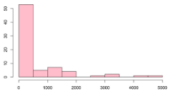
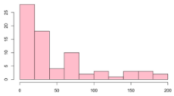
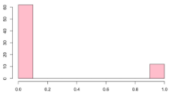
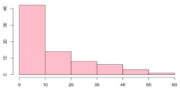
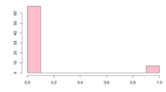
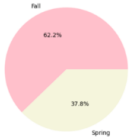
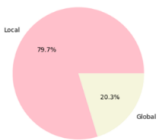
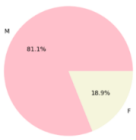
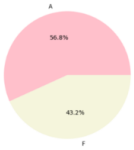
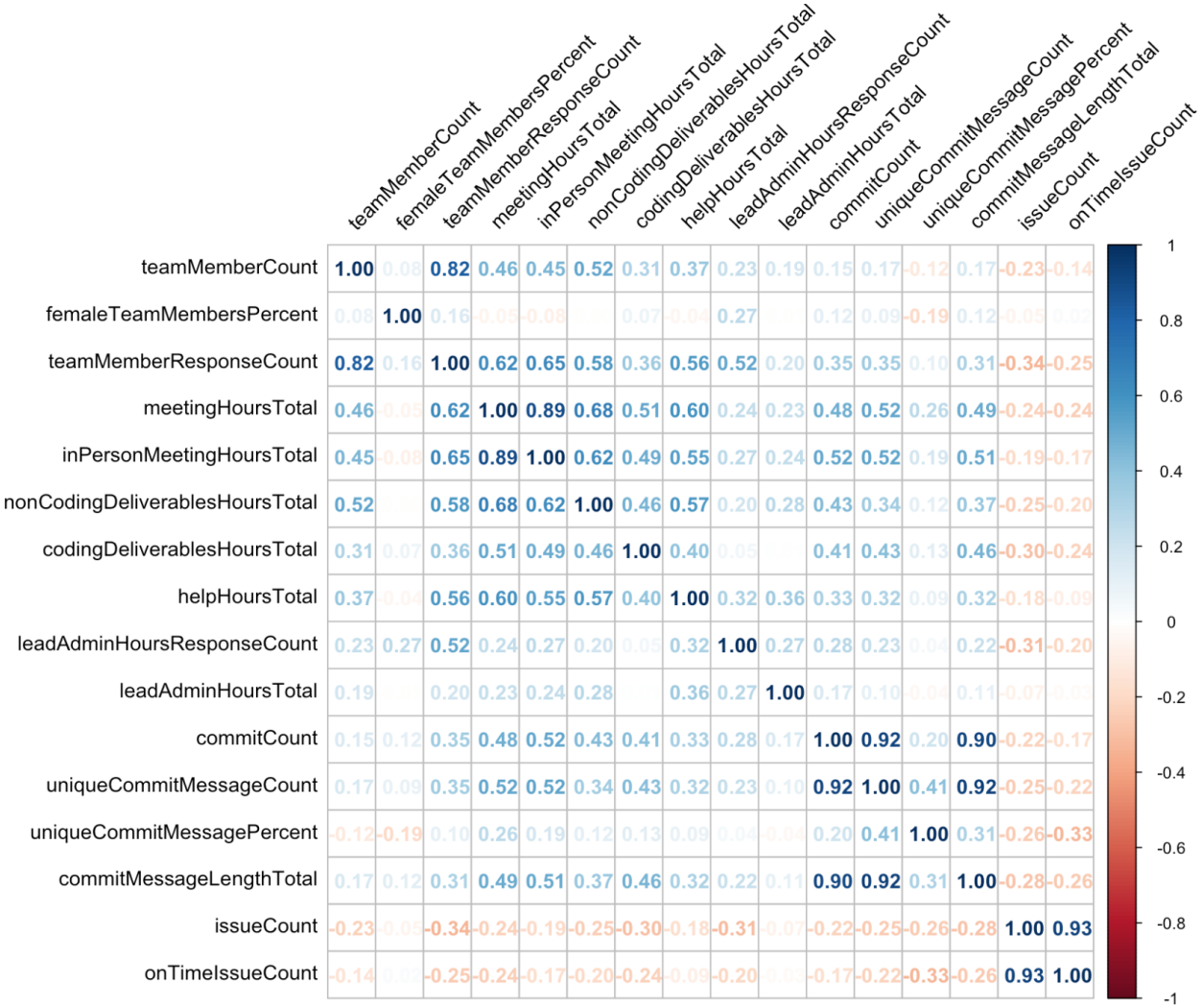
Histogram Plots	Description	Histogram Plots	Description
	**teamMemberCount** has most of the observations between 5 and 6 resulting to a statistical point of normal distributions.		**leadAdminHoursResponseCount** shows more of the observations are between 0 and 2.
	**FemaleTeamMemberCount** shows a positively skewed distribution, where most of the points lay between 0 to 0.2. Also, there is a point that is an outlierthat is above 0.8.		**leadAdminHoursTotal** is a positively skewed distribution with an outlier between 25 and 30.
	**teamMemberResponseCount** shows a considerable normal distribution.		**commitCount** is skewed to the right with some observations considered outliers above 80.
	**meetingHoursTotal** shows a positively skewed distribution with values from 90 and considered outliers.		**UniqueCommitMessageCount** is is a positively skewed distribution with most observations between 0 and 20.
	**inPersonMeetingHoursTotal** has the most observations between 0 to 20 leading a positively skewed distribution.		**UniqueCommitMessagePercent** is a negatively skewed distribution with values considered outliers at 0.1
	**nonCodingDeliverablesHoursTotal** is a positively skewed distribution.		**commitMessageLengthTotal** is a positively skewed distribution, qith most observations between 0 and 500.
	**codingDeliverablesHoursTotal** is a positively skewed distribution with most observations between 0 and 40.		**issueCount** has most observations between 0 to 0.1, and few at 0.9 to 1.
	**helpHoursTotal plot** shows a strong positively skewed distribution with most observations between 0 to 10.		**onTimeIssueCount** has most observations between 0 to 0.1, and few at 0.9 to 1.

Figure 5.8: T5 pie charts for categorical variables

			
The **semester** pie chart shows that 55.6% of semesters are autumn semesters and 44.4% are spring semesters.	The **teamDistribution** pie chart shows that Local teams make up 82.5% of all distributions and Global teams make up 17.5%.	The **teamLeadGender** pie chart shows 77.8% of the team leaders are males while females make up the 22.2%.	The **productLetterGrade** pie chart shows that 50.8% of A success and 49.2% are F failures.

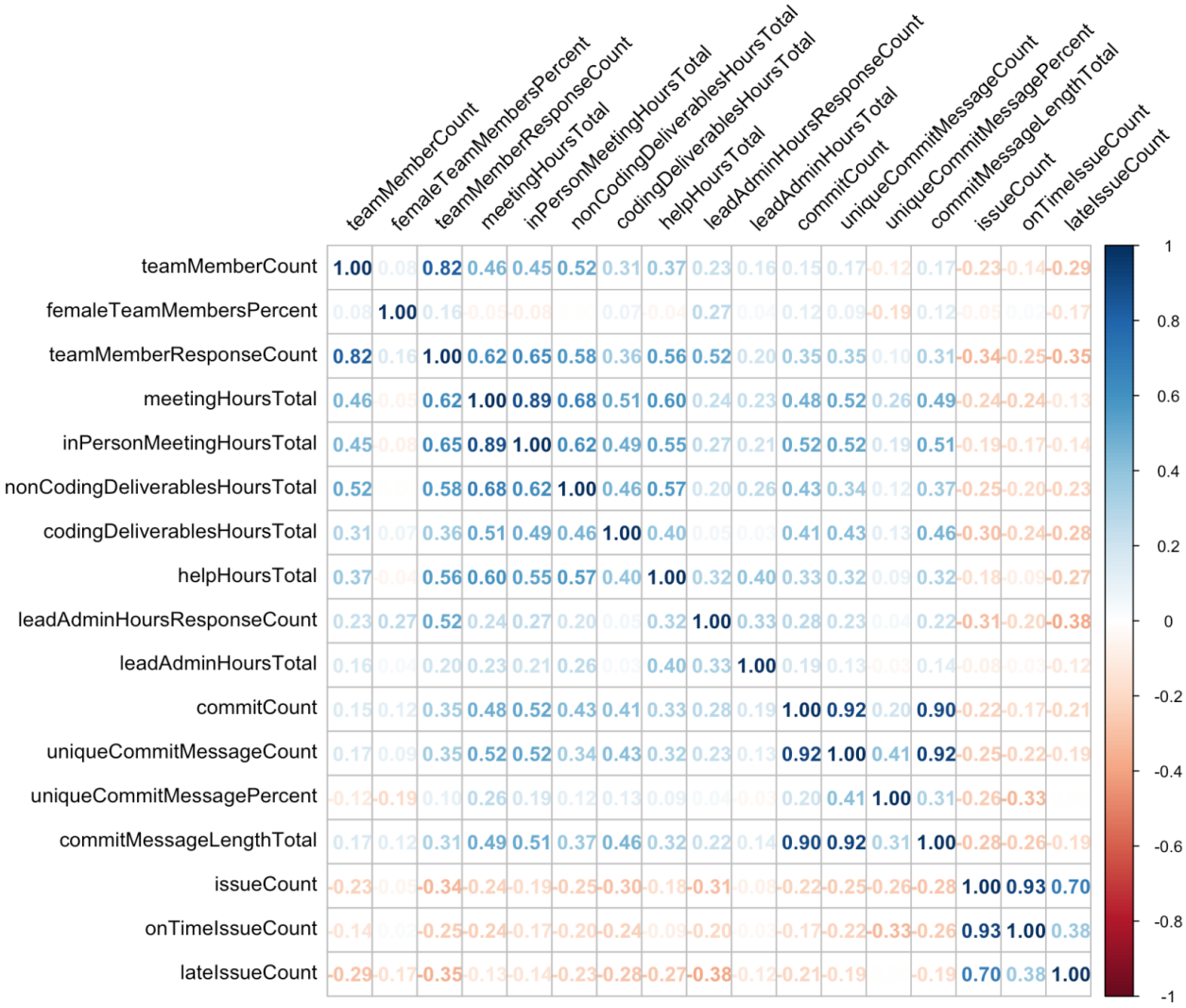
Correlation plot of dataset T1



Using a threshold of 0.8, the Pearson correlation output plot illustrates the correlations between the variables in a dataset. It aids in determining the significance and direction of correlations. A strong positive correlation is present when the correlation value between two variables exceeds the threshold. Below are the observed variables that are strongly correlated.

- *codingDeliverablesHoursTotal* and teamMemberResponseCount
- *leadAdminHoursResponseCount* and teamMemberResponseCount
- *helpHoursTotal* and meetingHoursTotal

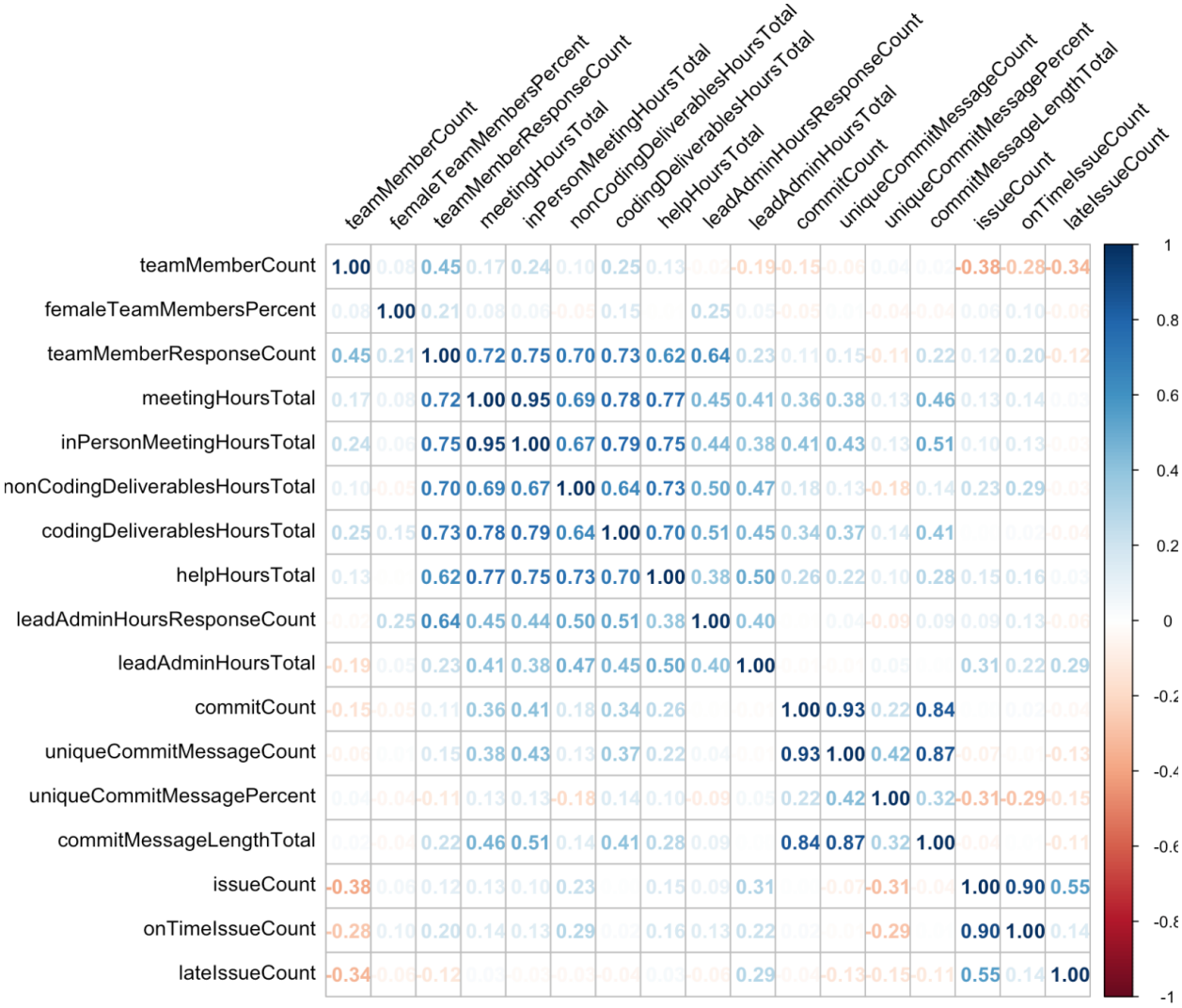
Correlation plot of dataset T2



Using a threshold of 0.8, the Pearson correlation output plot illustrates the correlations between the variables in a dataset. It aids in determining the significance and direction of correlations. A strong positive correlation is present when the correlation value between two variables exceeds the threshold. Below are some of the observed variables that are strongly correlated.

- *commitMessageLengthTotal* and leadAdminHoursTotal
- *commitMessageLengthTotal* and uniqueCommitMessagePercent
- *onTimeIssueCount* and commitMessageLengthTotal

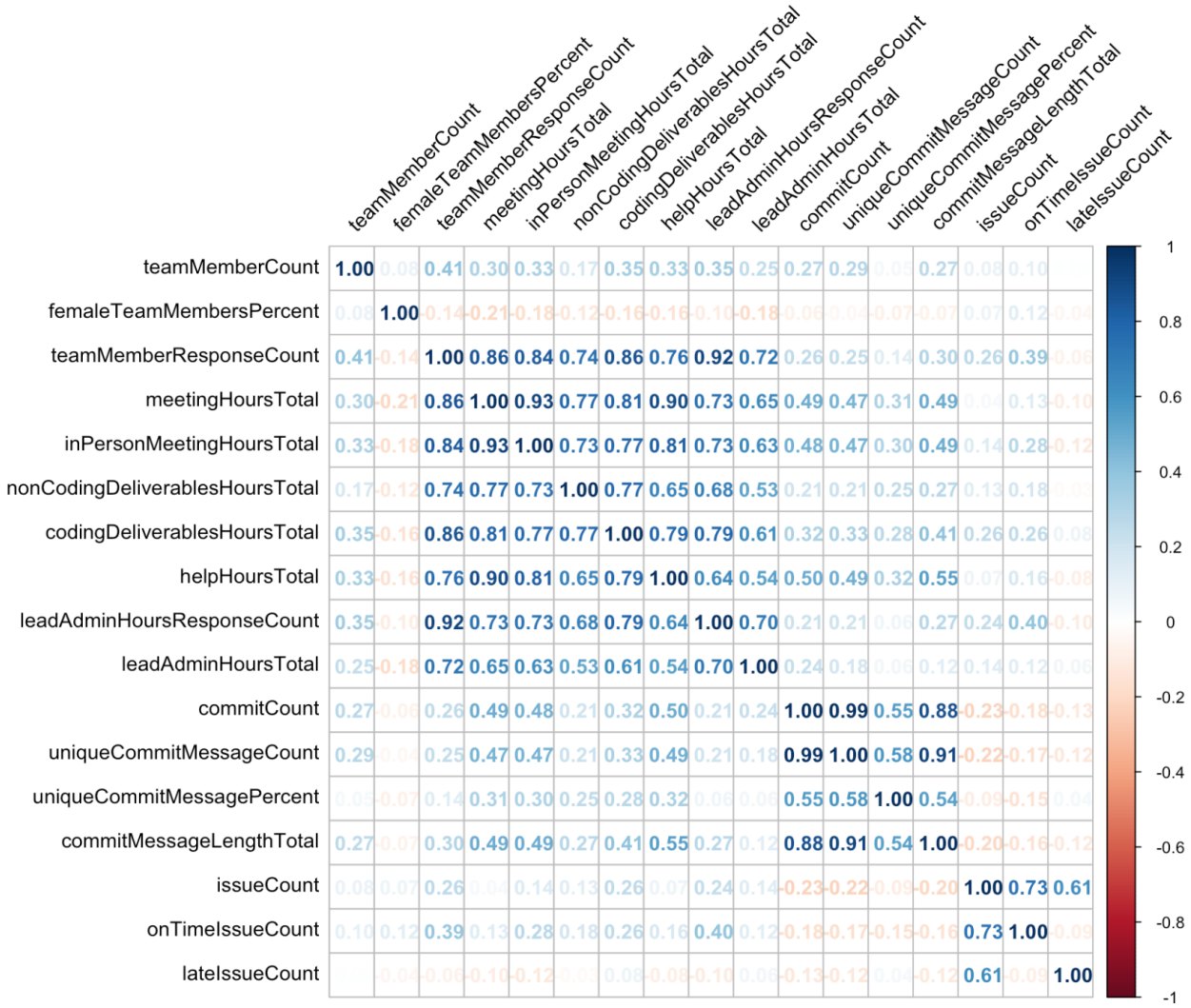
Correlation plot of dataset T3



Using a threshold of 0.8, the Pearson correlation output plot illustrates the correlations between the variables in a dataset. It aids in determining the significance and direction of correlations. A strong positive correlation is present when the correlation value between two variables exceeds the threshold. Below are some of the observed variables that are strongly correlated.

- *commitMessageLengthTotal* and countCommit
- *countCommit* and uniqueCommitMessagePercent

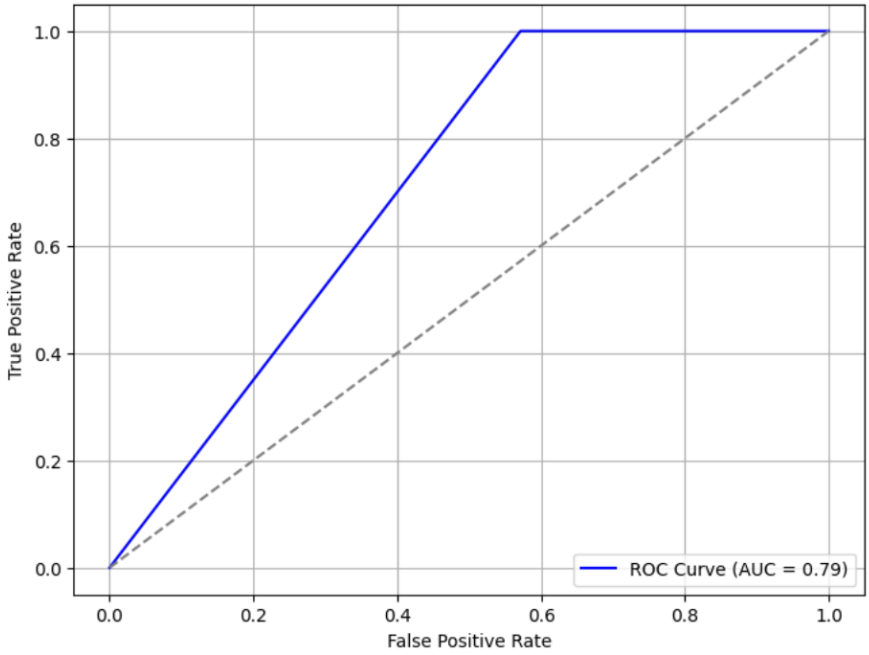
Correlation plot of dataset T5



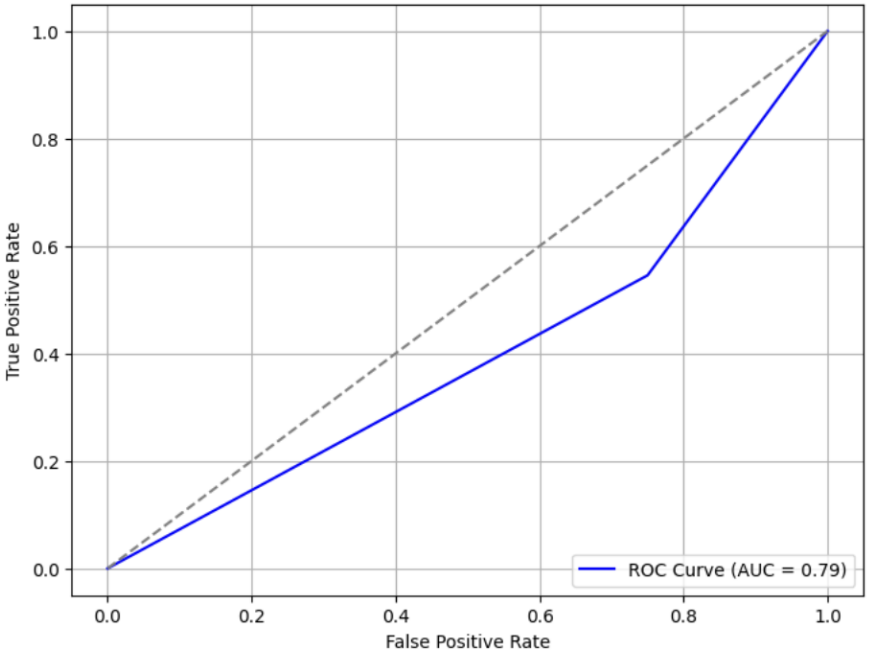
Using a threshold of 0.8, the Pearson correlation output plot illustrates the correlations between the variables in a dataset. It aids in determining the significance and direction of correlations. A strong positive correlation is present when the correlation value between two variables exceeds the threshold. Below are some of the observed variables that are strongly correlated.

- *commitMessageLengthTotal* and leadAdminHoursTotal
- *commitMessageLengthTotal* and uniqueCommitMessagePercent

Figure 5.9: AUC

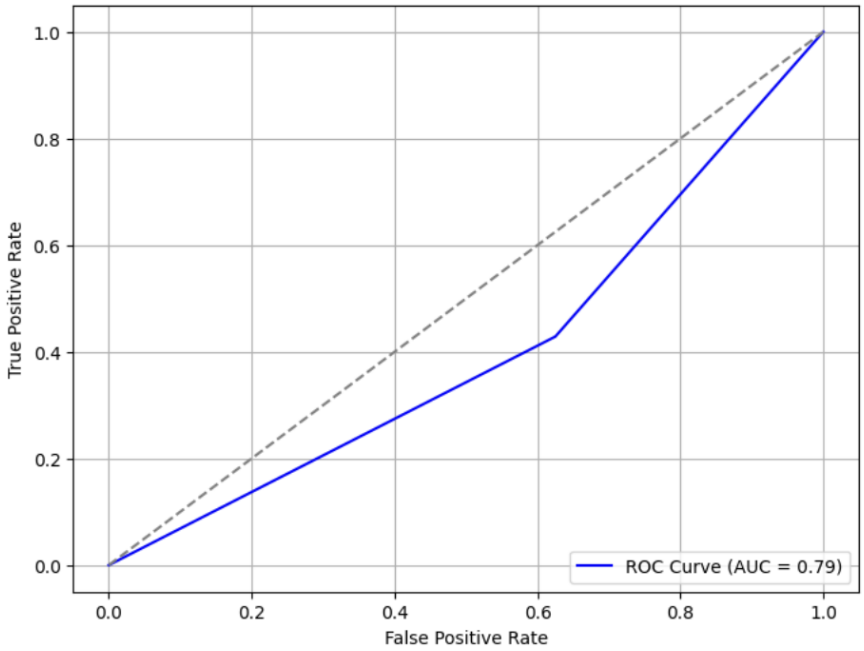


(a) AUC for T1.

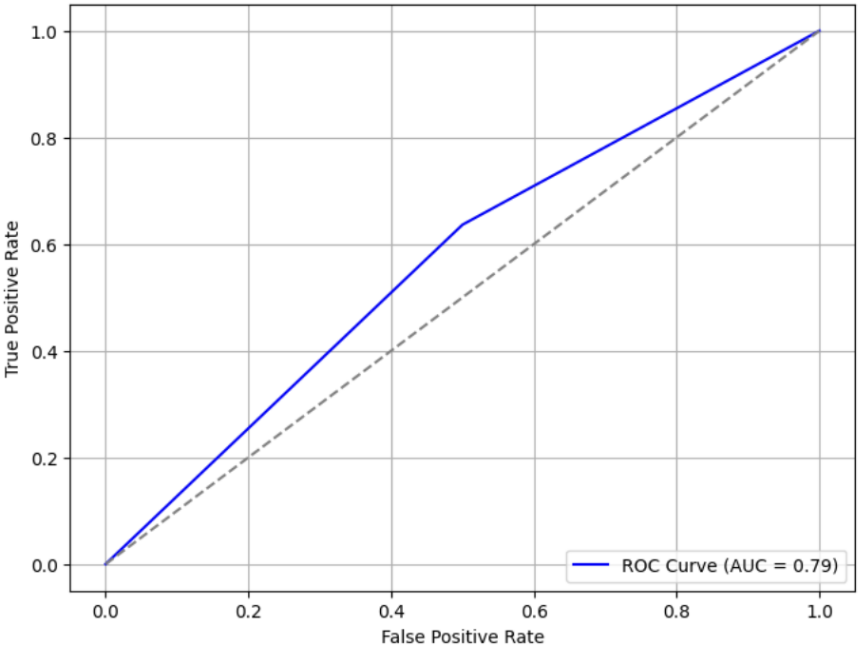


(b) AUC for T2.

Figure 5.10: AUC



(a) AUC for T3.



(b) AUC for T5.

Bibliography

Biswajit Bera. Empirical power comparison of goodness of fit tests for normality in the presence of outliers, 2013.

Daniel Berrar. Data science laboratory, 2013.

Emily Johnson Brian Conner. Descriptive statistics, 2017.

Marc Sosnick-Perez D. Petkovic. Using the random forest classifier to assess and predict student learning of software engineering teamwork frontiers in education, 2016.

A.-Shen W.-M. Weber R. [Famili, 1997] Famili and E. Simoudis. ata prepro- cessing and intelligent data analysis. intelligent data analysis, 1997.

Gammermann. Support vector machine learning algorithm and transduction, 2000.

Sungchul Yang Jaeho Son. A new approach to machine learning model development for prediction of concrete fatigue life under uniaxial compression, 2022.

Yoshua Bengio James Bergstra. Random search for hyper-parameter optimization, 2013.

Will Koehrsen. Hyperparameter tuning the random forest in python, 2018.

Minghui Hu. M.A. Ganaie. Ensemble deep learning: A review, 2022.

Emily Amor Balase Mayette Saculinggan. Empirical power comparison of goodness of fit tests for normality in the presence of outliers, 2013.

J. R. QUINLA. Learning decision tree classifiers, 1996.

Itamar Reis. Probabilistic random forest: A machine learning algorithm for noisy data sets, 2018.

Simon Dixon Saumitra Mishra, Bob L. Sturm. Local interpretable model-agnostic explanations for music content analysis, 2017.

Int. J. Remote Sens. Application of logistic regression model and its validation for landslide susceptibility mapping using gis and remote sensing data, 2004.

Lockhart J. Magazzeni D.v Villani, M. Random search for hyper-parameter optimization, 2022.

Virenrehal. Shapiro wilk test for normality, 2023.

Qi-Guang MIAO Ying CAO. Advance and prospects of adaboost algorithm, 2013.

Julián Chaparro-Peláez Ángel Hernández García. Predicting teamwork group assessment using log data-based learning analytics, 2018.