INT2X: Explanation for the Causes of a Prediction

TMP-23-142

Project Proposal Report

Britto T.A

B.Sc. (Hons) Degree in Information Technology

(Specialization in Data Science)

Department of Information Technology

Sri Lanka Institute of Information Technology

Sri Lanka

March 2023

INT2X: Explanation for the Causes of a Prediction

TMP-23-142

Project Proposal Report

B.Sc. (Hons) Degree in Information Technology

(Specialization in Data Science)

Department of Information Technology

Sri Lanka Institute of Information Technology

Sri Lanka

March 2023

DECLARATION

We declare that this is our own work, and this proposal does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or institute of higher learning, and to the best of our knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Name	Student ID	Signature
Britto T.A	IT20100698	Anjalika

The above candidate is carrying out research for the undergraduate Dissertation under my supervision.

Signature of the supervisor   Date

ABSTRACT

Explainable Artificial Intelligence(XAI) is an emerging area of research that aims to develop AI systems that can provide explanations for their decision-making processes. XAI can describe how AI came to a specific conclusion (such as classification or object detection)[1]. Explainability is essential in critical applications like healthcare, social media, law, the military, transportation, and financeThe General Data Protection Regulation (GDPR) of the European Union significantly influenced the growth of explainable AI (XAI). The GDPR, which was introduced in 2016, gives consumers the right to obtain "meaningful information about the logic involved" in automatic decision-making, also known as the "right to an explanation[2]."This study focuses develop an XAI solution for a text classification dataset using the Random Forest ensemble model. Many AI models have a "black-box" design that makes it challenging to understand how decisions are made. This research,main objective is to provide an XAI solution based on counterfactual analysis for text classification tasks. The proposed method specifically targets decision tree and random forest models for text classification tasks. Although these models perform well in various contexts, their black-box nature makes interpretation challenging. Throughout the research, the drawbacks of the existing methods and the weak points will be identified and addressed using several techniques.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
GDPR	General Data Protection Regulation
XAI	Explainable Artificial Intelligence
SHAP	Shaply Additive Explanations
LIME	Local Interpretable Model-Agnostic Explanations
AIX360	AI Explainability 360
ELI5	Explain Like I'm 5

1. INTRODUCTION

The interpretability or explainability of machine learning models has become a critical aspect of AI and ML systems. In recent years, the application of machine learning models in real-world scenarios has increased significantly, and many of these applications involve making important decisions that have a significant impact on people's lives. Therefore, it has become essential to ensure that the decisions made by these models are not only accurate but also explainable. Machine learning and artificial intelligence solutions are being used in various domains to automate workflows, make decisions, and improve overall efficiency. For instance, in the healthcare industry, machine learning algorithms are being used to predict and diagnose diseases, identify risk factors, and improve patient outcomes. Similarly, in the social media platform, machine learning is used for Ad targeting, Content moderation, and Sentiment Analysis.

1.1 Definition of Explainability/Interpretability

There are various definitions for explainability/interpretability in machine learning, and different authors and researchers may use slightly different definitions depending on the context and the specific techniques being used. The most popular and widely used definitions are "The degree to which a human can understand the cause of a decision or the reasoning behind a prediction made by a machine learning algorithm." [3]and “Interpretability is the degree to which a human can consistently predict the model's result”[4]

1.2. Need for Explainable Artificial Intelligence (XAI)

To understand the need for explainable artificial intelligence (XAI), it is crucial to understand the problems associated with machine learning (ML) models. As an example, Machine learning models are being increasingly used to predict the efficacy of potential drug compounds and to identify promising drug targets. However, if the decision-making process of the model is not transparent or interpretable, it may be difficult to understand why a particular drug compound was identified as promising or why a particular target was selected. Not only in drug development this level of reasoning is required in a variety of domains such as social media, legal, transportation, military, and finance.

Explainability can be applied to a variety of tasks in a variety of fields, including justifying, controlling, improving, and discovering.[5]

a. Explain to justify.

Some situations in AI/ML-based systems provide biased or discriminatory results. Therefore, the explanation of AI-based results is essential. The XAI systems provide the required information to justify the result when unexpected decisions are made.

b. Explain to control.

During the explainability process, users gain insights into the system's behavior. Greater visibility of undiscovered vulnerabilities and flaws is provided, which makes mistakes easier to spot and quickly fix.

c. Explain to improve.

A model that can be explained can be improved easily. This is because users comprehend the relationship between input variables and the resulting output, and how to make the output smarter.

d. Explain to discover

Explanation about the process is helpful to learn new facts, collect information, and gain knowledge. XAI models are useful to find new and hidden laws in various scenarios.

Counterfactual explanations offer valuable insights into an ML model's decision-making process by presenting alternative scenarios that would have led to a different outcome. These explanations are crucial for understanding the factors that contributed to a specific prediction and can assist users in identifying potential biases or shortcomings within the model. For example, consider a loan approval scenario. If an applicant is denied a loan by the ML model, a counterfactual explanation might reveal that the applicant would have been approved if their credit score were slightly higher or if they had a lower debt-to-income ratio. By presenting these alternative scenarios, counterfactual explanations enable users to examine the effects of various factors on the model's decisions and uncover any potential biases or discriminatory practices in the model's behavior. Furthermore, these explanations can serve as a guide for model improvements by emphasizing areas where adjustments could be made to promote fairness and reduce discrimination in the model's predictions.

1.3. Area of the research

Explainability techniques can be used to detect, what are the changes that need to be done to the model features to flip the prediction. Counterfactual explanations are a type of explainability method in artificial intelligence and machine learning, designed to help users understand the decision-making process of complex models.many researchers tried to address this many techniques have been developed and some of them show promising results. SHAP, LIME, and ELI5 are a few of the most popular frameworks developed recently.

In SHAP (Shapley Additive Explanations), the influence, relevance, or importance of each feature is ranked or measured in relation to the output using classic Shapley values. Any black-box classifier with two or more classes can be explained using LIME (Local Interpretable Model-agnostic Explanations). The classifier only needs to implement a function that accepts raw text or a NumPy array and outputs a probability for each class. Furthermore, the ELI5 (Explain Like I'm 5) tool assigns weights for input features and the ranking according to the weights.

The field of XAI has made significant progress in recent years, but there is still a need for more research and tools to improve the explainability of machine learning models. Our research focuses on developing a framework to improve the explainability of the ML model.

1.4. Component overview

In our research on developing a Random Forest solution for text classification, we emphasize the significance of incorporating counterfactual explanations to enhance model interpretability and fairness. The primary goal of this study is to create a novel counterfactual rule generation mechanism that can effectively explain the decisions of the Random Forest model. By incorporating counterfactual explanations, our solution will enable users to comprehend the minimum changes required in input features to alter the model's prediction. This understanding promotes transparency and allows users to identify potential biases or discriminatory patterns in the models' behavior.

2. LITERATURE REVIEW

2.1. Background

The rapid expansion of AI models and their applications has drawn the focus of researchers and communities to the field of Explainable Artificial Intelligence (XAI). The necessity and significance of XAI have emerged as people evaluate the trustworthiness and transparency of AI-driven decision-making processes. In recent years, numerous model-explainable methods have been proposed by researchers for the scientific community. However, since knowledge in this domain is still evolving, the majority of these techniques remain in the experimental stage.

As discussed in the introduction the concept of XAI has been focusing on black box models. In text classification, a Random Forest ensemble model can become a black box due to its complex structure and high-dimensional input features.

Random forest is a supervised learning technique, that is used for classification, and regression tasks. It is an ensemble method that consists of multiple decision trees, each built on different subsets of the given dataset. By averaging their individual predictions, the Random Forest model improves the overall predictive accuracy of the dataset. Instead of relying solely on a single decision tree, the Random Forest model considers the predictions from each tree and selects the final output based on the majority votes of those predictions, thereby ensuring a more robust and accurate decision-making process. Increasing the number of trees in a forest contributes to improved accuracy and helps avoid overfitting issues. Our research aims to provide an XAI solution for the text classification domain using Random Forests, enhancing interpretability and understanding of the model's decisions.

In text classification, Random Forest ensemble models can sometimes be perceived as black boxes, even though decision trees are generally considered interpretable and transparent. This perception arises due to the complexity introduced by the combination of multiple decision trees, making it difficult to trace the reasoning behind individual decisions.

When applied to text classification, the high-dimensional nature of the data and a large number of features (i.e., words) further complicate the understanding of Random Forest models. The final classification decision is based on the output of many decision trees. It makes it difficult to understand the overall decision-making process. As a result, determining the impact of specific features on the model's decision-making process becomes increasingly challenging.

This black-box nature of Random Forest models in text classification raises concerns about transparency and trustworthiness. As a result, there is an increasing demand for creating and applying interpretability methods that enable users to gain deeper insights into the decision-making mechanisms of these models, ensuring their fairness and reliability.

Explainable artificial intelligence (XAI) has advanced significantly in recent years, and researchers have categorized XAI methods into intrinsic or post-hoc methods[5]. Intrinsic methods aim to make machine learning models interpretable by design, often by simplifying the model architecture or incorporating domain-specific knowledge. Post-hoc methods, on the other hand, analyze the model after it has been trained to generate explanations for its behavior. Moreover, the explanation method used can be model-centric, tailored to a specific model architecture, or model-agnostic, applicable to different types of models[9]. Model-centric methods can provide more specific and detailed explanations for a particular model, while model-agnostic methods can provide more general insights into machine learning models in general.

A counterfactual explanation is a type of explanation used in the context of machine learning models to help users understand and interpret the model's predictions. It does so by providing an alternative scenario, where the model's prediction would have been different if certain input features were changed. In other words, it answers the question, "What would need to change in the input data for the model to make a different decision?"The ultimate goal of this research is to overcome model explainability issues in the random forest model by providing a better solution using XAI techniques. For the XAI solution development, use the Sentimental Analysis dataset(text classification data).

2.1 Literature Survey

XAI is dedicated to demystifying the black boxes by properly explaining the internal process of AI/ML models. It improves the trustworthiness, transparency, and responsibility of AI-based decision-making processes. Since the XAI research area is still in the starting era most of the techniques are in the research phase and researchers had built communities for do contributions to the XAI domain.

The author of [6] mentioned there are several explanation methods and strategies had proposed by researchers to make AI systems explainable. Among those techniques "Scoop Related Methods" and "Model related Methods" are commonly used for the explainable process. Scoop related methods are classified as global interpretability and local interpretability. Global interpretability facilitates the understanding of the whole mechanism of the model and the entire reasonings cause for the final output. Local interpretability focuses on explaining the reasons for a specific decision or a single prediction. Model-related methods are also classified into two categories as model specific interpretability (methods are limited to a specific model) and model-agnostic interpretability (methods are not tied to a specific ML model).

Several techniques have gained popularity within the research community, with new contributions being built upon these methods. LIME (Local Interpretable Model-agnostic Explanation) [7] and SHAP (SHapely Additive exPlanations) [8] serve as prime examples. Focusing on LIME, it offers faithful interpretation and explanation of a model's classification. LIME achieves local optimal explanations by calculating the significance of features through the generation of feature vector samples, which follow a normal distribution. After obtaining predictions from these samples, LIME assigns weights to each row to gauge their proximity to the original sample.

Subsequently, LIME employs a feature selection technique to pinpoint the most crucial features.

TreeSHAP: Explainable AI for Trees [10]

In this research, they have introduced a novel model-specific methodology that explains an approach to enhance the explainability and interpretability of tree-based models, such as decision trees and Random Forests. The authors address the growing need for improved interpretability and transparency in tree-based machine-learning models. Tree-based models, such as decision trees and Random Forests, have gained popularity due to their robust performance, but often suffer from limited explainability. To tackle this challenge, Lundberg and his colleagues introduce a model-agnostic method called SHAP (Shapley Additive Explanations) that provides local explanations for individual predictions made by tree-based models.

The paper discusses the importance of explainable AI (XAI), especially in the context of critical decision-making processes that utilize machine learning models. The authors demonstrate the application of SHAP values to decision trees, Random Forests, and gradient-boosted trees, offering a valuable contribution to the field of XAI. SHAP values help users comprehend the logic underlying the model's decisions by quantifying the contribution of each feature to the prediction for a particular instance.

In addition to local explanations, the paper extends the SHAP framework to enable a global understanding of tree-based models. By aggregating SHAP values across multiple instances, the authors present a method to identify the most important features and understand the overall structure of the model's decision-making process. Various visualization techniques are also discussed for interpreting the model behavior at a global level.

This paper by Lundberg et al. represents a significant advancement in the field of XAI, specifically in providing explanations and interpretability for tree-based models. Their work has the potential to improve transparency, trust, and fairness in machine learning applications that use decision trees, Random Forests, and other tree-based models.

Explaining Machine Learning Classifiers via Diverse Counterfactual Explanations [11]

The concept of Counterfactuals involves imagining hypothetical possibilities that could have occurred apart from the current situation. To comprehend the notion of Counterfactuals in this research, a real-life scenario that may arise in everyday life is presented and discussed below.

Consider a situation where a loan application is denied. Employing interpretable models, the organization may explain this decision, such as a "poor credit" history. However, from the perspective of the person who applied for the loan, this explanation does not help them determine their next steps. If the system could offer suggestions for improving their chances of securing a loan in the future, it would be more beneficial for all parties involved.

There may be instances where the most critical factors, such as gender or race, are insufficient to alter or change the prediction. In such cases, it is crucial to provide alternative, actionable guidelines. The paper offers the following counterfactual explanation for the loan example mentioned above.

"You would have received the loan if your income was higher by $10,000."

Upon receiving this type of explanation, the individual can identify which aspects need improvement to qualify for the loan in the future.

AI Explainability 360 (AIX360)[12]

The AI Explainability 360 (AIX360) toolkit is an open-source library developed by IBM Research. It provides a comprehensive suite of algorithms and resources to promote explainable AI (XAI) in machine learning models. By offering a variety of techniques, the toolkit caters to different users, ranging from data scientists to domain experts and end-users.

The AIX360 toolkit includes several explainability algorithms, each with its unique approach to generating explanations. These algorithms can be broadly categorized into local and global explanation techniques, as well as model-specific and model-agnostic methods. Local explanation techniques focus on providing insights into individual predictions, whereas global techniques aim to offer a broader understanding of the entire model.

The toolkit also provides a range of resources, including tutorials, example code, and detailed documentation to help users understand and apply these explainability techniques to their machine-learning models. By making these resources readily available, the AIX360 toolkit aims to facilitate the adoption of explainable AI practices and to encourage the development of transparent, accountable, and trustworthy AI systems.

Overall, the AIX360 toolkit is a valuable resource for researchers, practitioners, and stakeholders interested in implementing explainable AI in their machine-learning models. The toolkit's diverse set of techniques and resources makes it a versatile solution for addressing the growing need for transparency and accountability in AI systems.

GoogleXAI[13]

In "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI" (Arrieta et al., 2020), the authors offer an in-depth analysis of the XAI field, emphasizing the importance of interpretability, transparency, and accountability in AI systems. The article explores the evolution of XAI, its various methodologies, and taxonomies for classifying XAI methods. It also discusses the opportunities and benefits of XAI across numerous industries, as well as the challenges and limitations in its implementation. By shedding light on the current state of XAI and its potential influence on AI development and deployment, this paper serves as a valuable resource for researchers, practitioners, and policymakers, underscoring the need for continued research to achieve responsible and trustworthy AI systems.

3. RESEARCH GAP

The primary objective of this research is to enhance the interpretability of Decision Tree models, specifically in the context of text classification. A counterfactual rules generation mechanism can be employed to suggest modifications in input features that would result in changes to the model's predictions, thereby making the Decision Tree model more explainable. Existing tools, such as LIME[1], SHAP[5], XAI360[4], and Google XAI[10], offer mechanisms for explaining text classification models, but only XAI360[13] provides a counterfactual explanation for deciphering the internal behavior of black box models.

Moreover, most XAI tools present visualizations that depict model explainability but are often complex and comprehensible only to subject matter experts. It is crucial to develop user-friendly and easily understandable visualizations of the internal processes of black box models for the benefit of end users. By making minor alterations to the input features, counterfactual rule-based explanations can provide valuable insights into the model's decision-making process, resulting in improved transparency and trust in AI systems.

	TreeSHAP [10]	**Diverse Counterfactual Explanations[11]**	**LIME** [7]	**XAI360** [12]	**Google XAI[13]**	**SHAP [8]**	**Proposed Random Forest Explainable Method**
Model Specific Approach.	√	***X***	***X***	***X***	***X***	***X***	√
Provide a Counterfactual Rule Generation based Explainable Mechanism.	***X***	√	***X***	√	***X***	***X***	√
Text classification explainable task	√	√	√	√	√	√	√
Having a user-friendly visualization	***X***	***X***	***X***	***X***	***X***	***X***	√
Locally Explainable	√	√	√	√	√	√	√

Table 1:Comparison between existing methods and proposed method

4. RESEARCH PROBLEM

The research problem focuses on generating counterfactual rule-based explanations for Random Forest classifiers when they become black boxes in text classification tasks. Text data typically consists of a vast number of unique features or words, making it challenging to comprehend the contribution of each feature to the model's decision-making process. This complexity can lead to a lack of transparency and interpretability in the model's predictions, which can hinder user trust and limit the model's practical applicability.

In this context, counterfactual explanations can provide valuable insights by identifying alternative scenarios where the model's prediction would change, thus offering users a better understanding of the factors that influence the model's decisions. The primary goal of this research is to develop a method for generating counterfactual rule-based explanations tailored for Random Forest classifiers in text classification tasks, shedding light on their decision-making process and enhancing their explainability.

Addressing this research problem could lead to significant advancements in the field of explainable artificial intelligence (XAI), particularly for ensemble models like Random Forests, which are known for their performance but often lack transparency in high-dimensional data, such as text classification. Developing a counterfactual rule generation-based explanation method could help improve trust, fairness, and accountability in AI systems that rely on Random Forest classifiers for text classification tasks.

5. OBJECTIVES

5.1 Main Objective

Provide a novel post-hoc ,model-specific, local XAI solution to enhance the model interpretability of ensemble models focus on Random forest by developing a novel counterfactual rule generation mechanism related to the text classification domain.

5.2 Specific Objectives

To achieve the main objective study must achieve several sub objectives within the study period.

- Prepare the dataset and implement the Random forest classifier.
- Creating a counterfactual rule generate mechanism to explain the random forest model's behavior.
- Test the output with existing explainable methods.
- Do experiments to improve the XAI solution more.
- Do the visualization using the most appropriate Graphical User Interface (GUI) techniques.

5.3Work Breakdown Structure

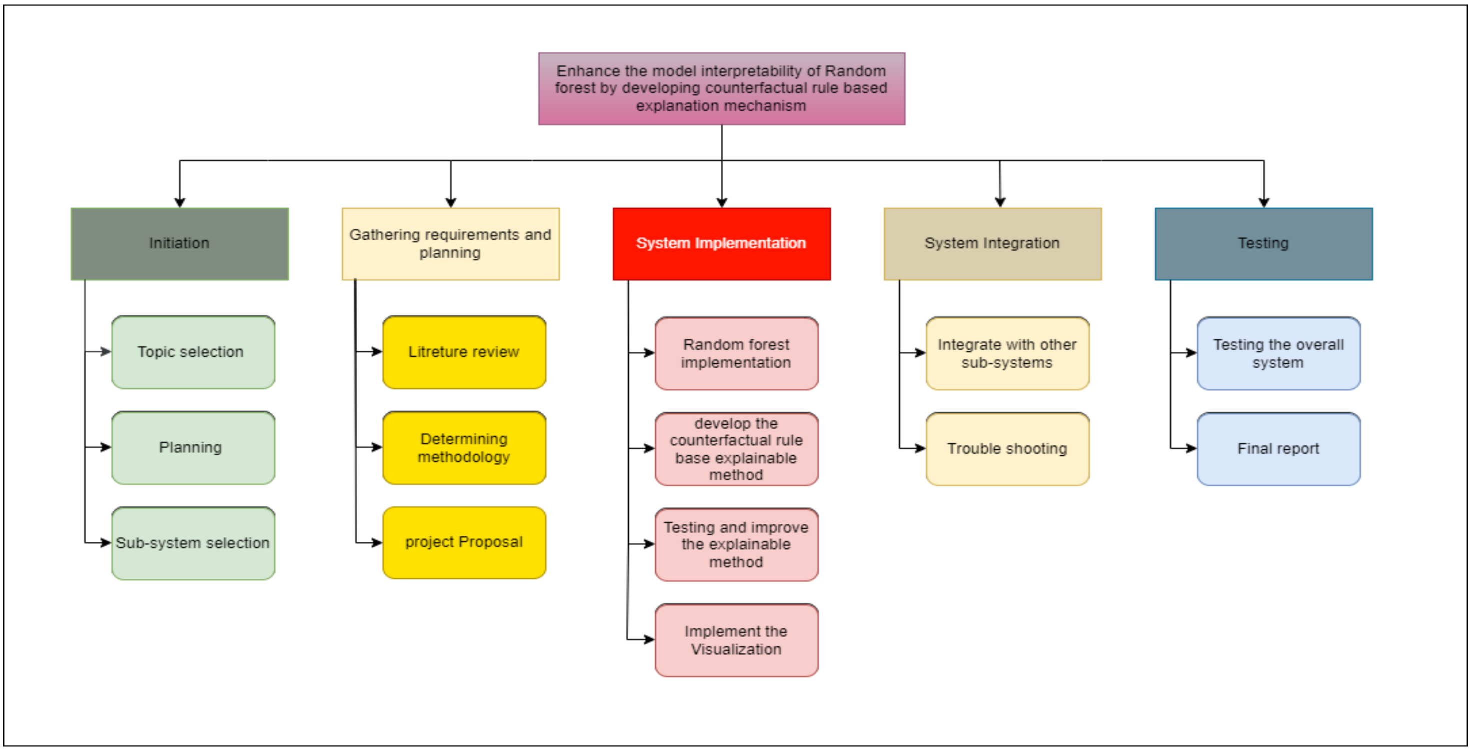

Enhance the model interpretability of Random forest by developing counterfactual rule based explanation mechanism
Initiation
Topic selection
Planning
Sub-system selection
Gathering requirements and planning
Litreture review
Determining methodology
project Proposal
System Implementation
Random forest implementation
develop the counterfactual rule base explainable method
Testing and improve the explainable method
Implement the Visualization
System Integration
Integrate with other sub-systems
Trouble shooting
Testing
Testing the overall system
Final report

Figure 1:Work Breakdown Structure

6. METHODOLOGY

There are few milestones that needs to be complete to develop the proposed system. These goals that needs to be achieved to complete this research is mentioned in this section.

6.1. System Architecture Diagram

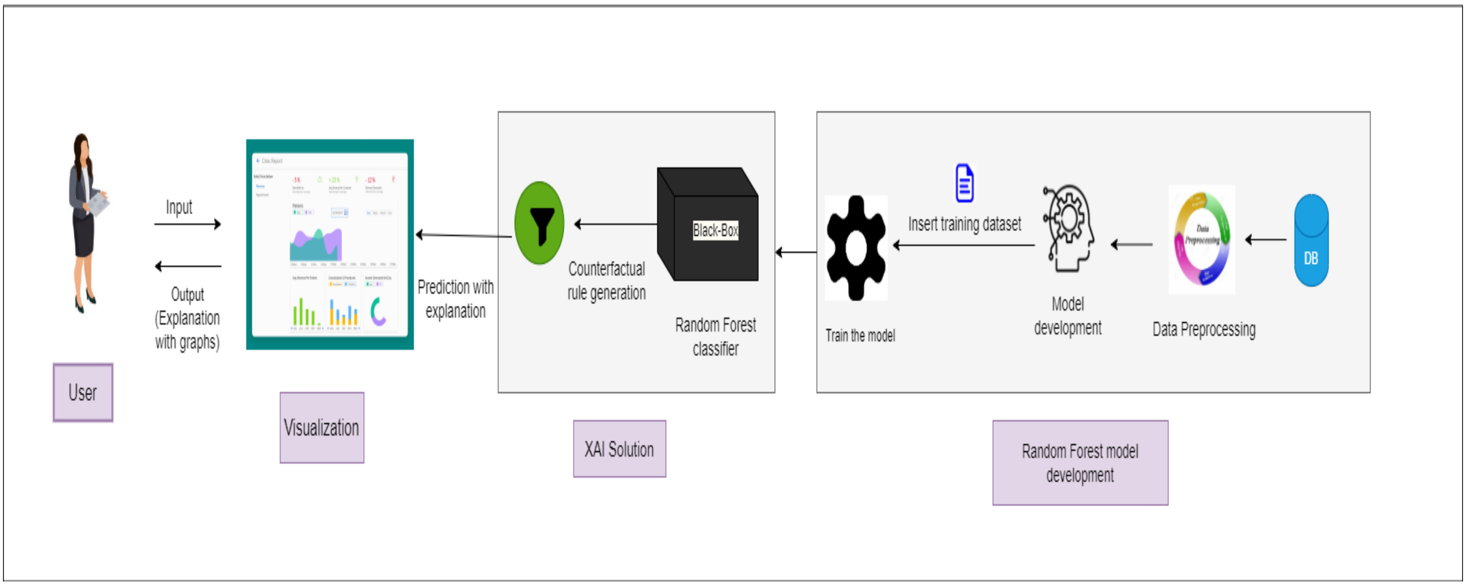


Figure 2:System Architecture Diagram

6.2 Tools and Technologies

Frontend:

- ReactJS
- Flask
- Boostrap

Backend:

- Python

Version Control:

- GitHub

Tools:

- VS Code
- Google Colab

7. PROJECT REQUIREMENTS

User Requirements

- User should have a knowledge of decision-making systems based on machine learning.
- Sometimes the researchers will be the users .
- Dataset should be pre-processed, and appropriate data engineering techniques should be applied.
- Instance that needs to be predicted should be provided by the user.

Functional Requirements

- Provide the counterfactual rules.
- System should be able to provide appropriate visualizations when needed.
- Model accuracies should be provided by the system.

Non-Functional Requirements

- Output should be understandable.
- Visualization should be user-friendly, accurate and interactive.

8. GANTT CHART

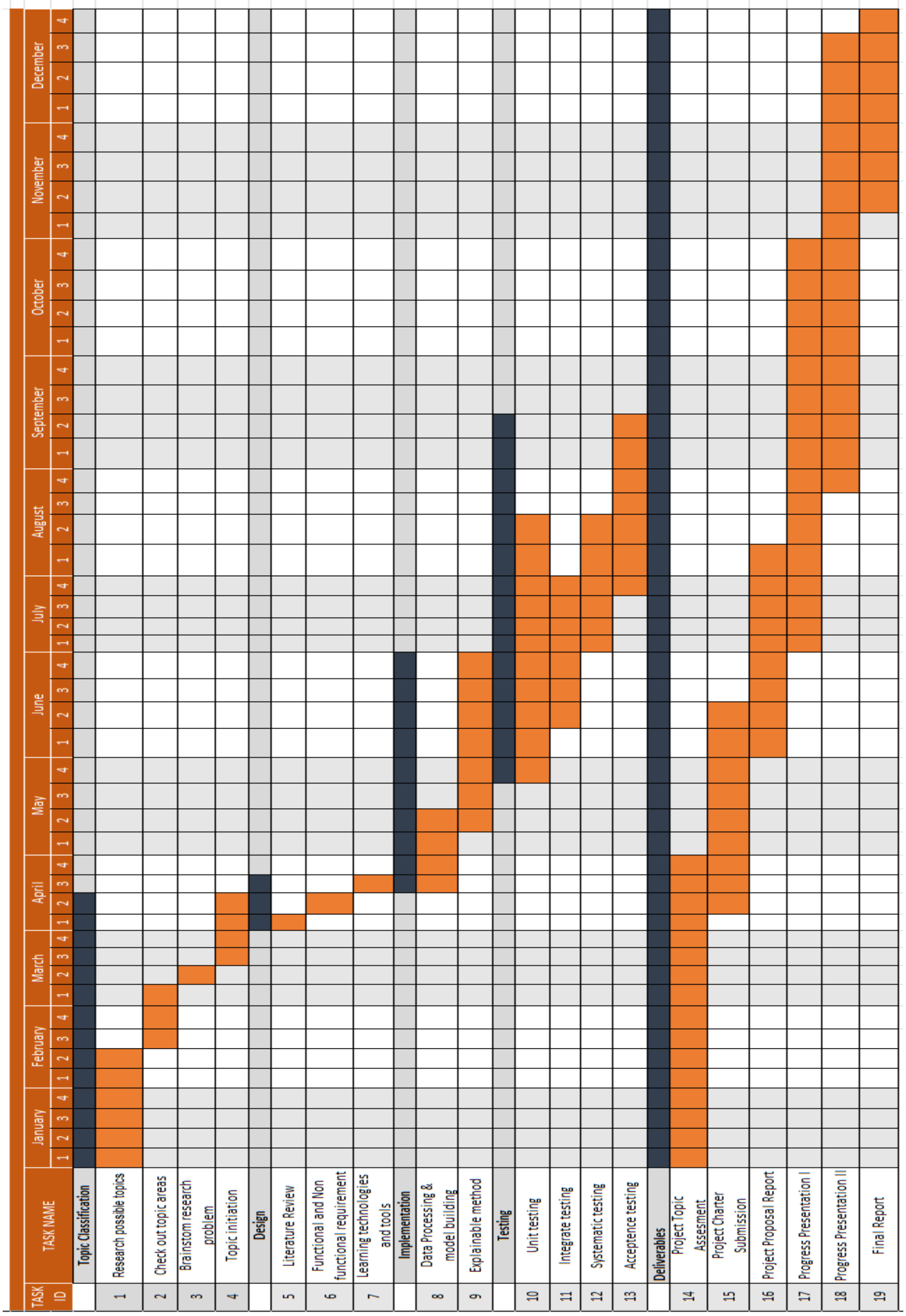


Figure 3:Tentative Gantt Chart

9. REFERENCES

[1] Gohel, P., Singh, P., & Mohanty, M. (2021). Explainable AI: current status and future directions. *arXiv preprint arXiv:2107.07045*.

[2] Mitrou, L. (2018). Data protection, artificial intelligence and cognitive services: is the general data protection regulation (GDPR)‘artificial intelligence-proof’?. Artificial Intelligence and Cognitive Services: Is the General Data Protection Regulation (GDPR)‘Artificial Intelligence-Proof.

[3] Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." arXiv Preprint arXiv:1706.07269. (2017).

[4] Kim, Been, Rajiv Khanna, and Oluwasanmi O. Koyejo. "Examples are not enough, learn to criticize! Criticism for interpretability." Advances in Neural Information Processing Systems (2016).

[5] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," in IEEE Access, vol. 6, pp. 52138-52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[6] S. M. Lundberg and S. I. Lee, ‘‘A unified approach to interpreting model predictions,’’ in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 4768–4777.

[7] Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/

[8] S. M. Lundberg and S. I. Lee, “A unified approach to interpreting model predictions,” Adv. Neural Inf. Process. Syst., vol. 2017-Decem, no. Section 2, pp. 4766–4775, 2017.

[9] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Model-agnostic interpretability of machine learning." ICML Workshop on Human Interpretability in Machine Learning. (2016).

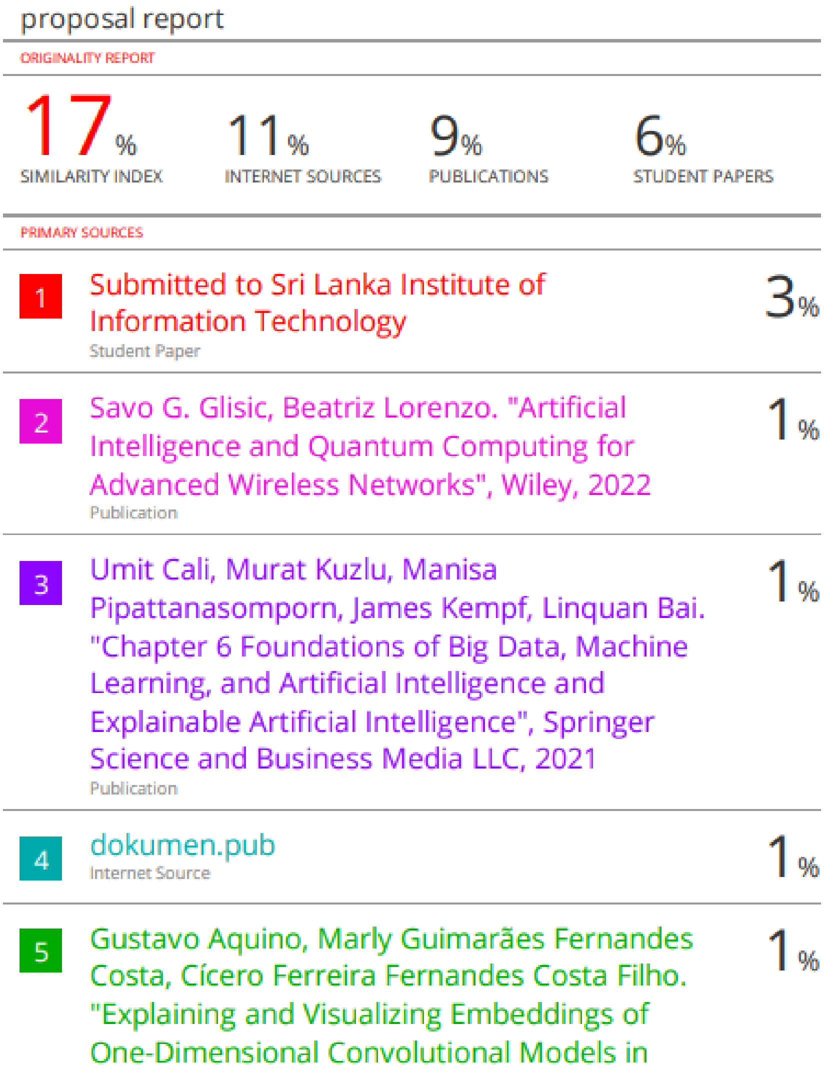
[10] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... & Lee, S. I. (2019). Explainable AI for trees: From local explanations to global understanding. *arXiv preprint arXiv:1905.04610*.

[11] R. K. Mothilal and C. Tan, “Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations.”, 2019.

[12] https://aix360.readthedocs.io/en/latest/

[13] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82-115

APPENDICES



proposal report

ORIGINALITY REPORT

17%	11%	9%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Sri Lanka Institute of Information Technology Student Paper	3%
2	Savo G. Glisic, Beatriz Lorenzo. "Artificial Intelligence and Quantum Computing for Advanced Wireless Networks", Wiley, 2022 Publication	1%
3	Umit Cali, Murat Kuzlu, Manisa Pipattanasomporn, James Kempf, Linquan Bai. "Chapter 6 Foundations of Big Data, Machine Learning, and Artificial Intelligence and Explainable Artificial Intelligence", Springer Science and Business Media LLC, 2021 Publication	1%
4	dokumen.pub Internet Source	1%
5	Gustavo Aquino, Marly Guimarães Fernandes Costa, Cícero Ferreira Fernandes Costa Filho. "Explaining and Visualizing Embeddings of One-Dimensional Convolutional Models in	1%