INT$_2$ X: Explanation For the Causes of a Prediction

TMP-23-142

Our Team

Mr. Prasanna Sumathipala
Supervisor

Mr.Jeewaka Perera
Co – supervisor

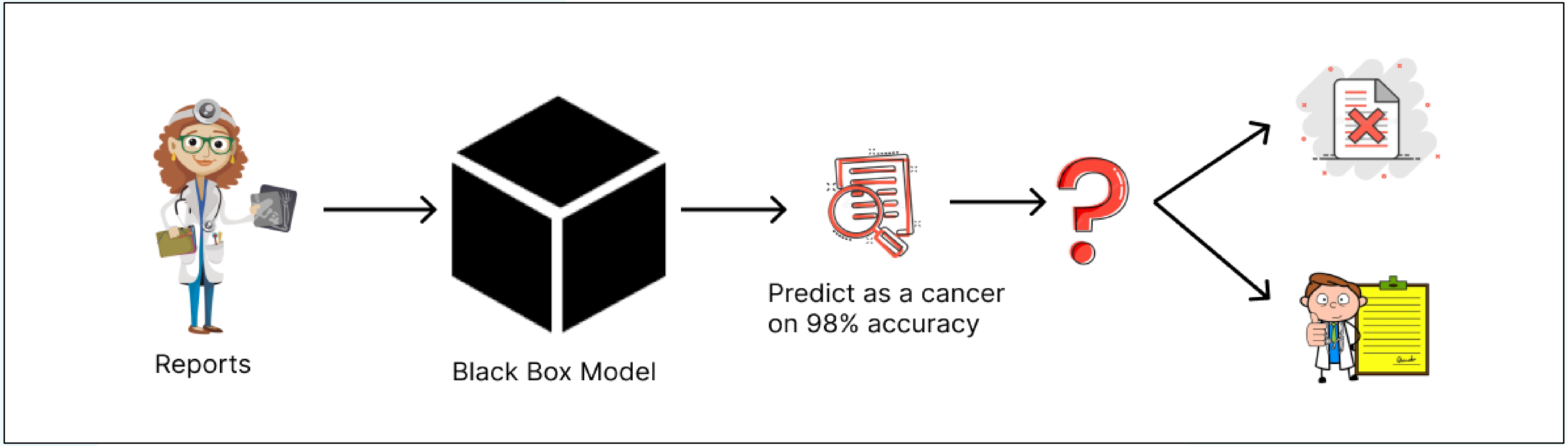
IT20097660
Warnasooriya S.D

IT18161298
Srinidee Methmal H.M

IT20100698
Britto T.A

IT20013950
Lakshani N.V.M

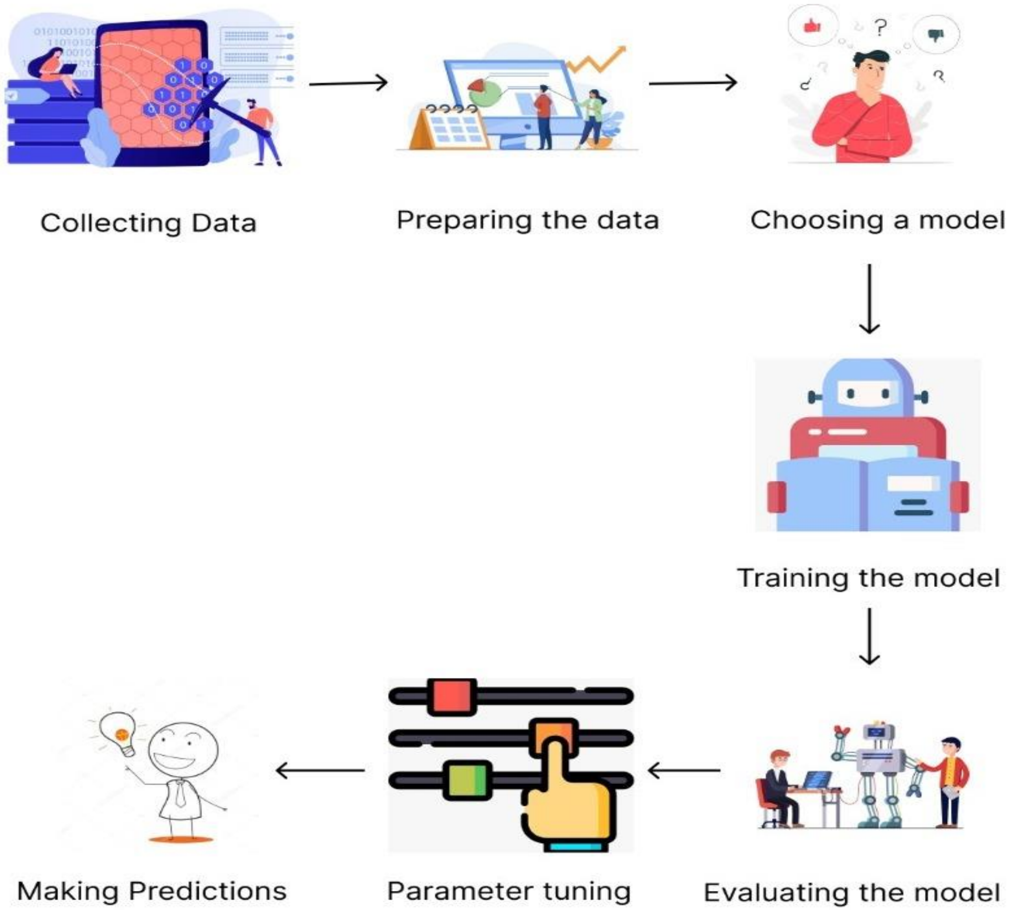
Predictions made by the ML models are 100% Accurate?



- The model predicts the patient has cancer with 98% accuracy, **So should the doctor confirm that the patient has cancer?**

What are Predictions?

- The outcomes or results of a model that uses input data to forecast future events or behaviors.

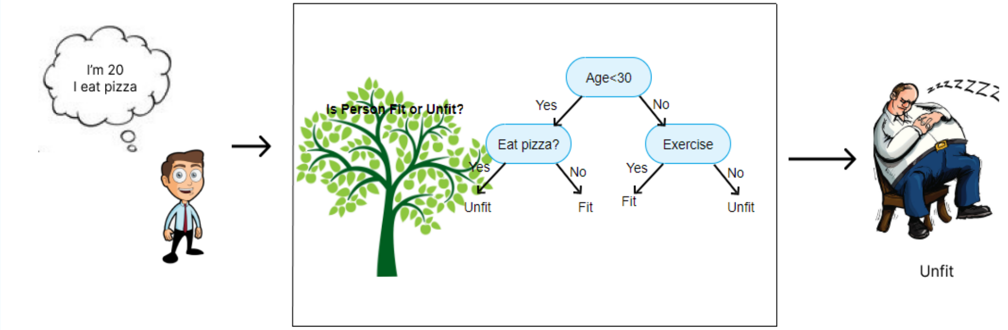


What are Predictive models ?

- Making predictions about future events or outcomes based on historical data.

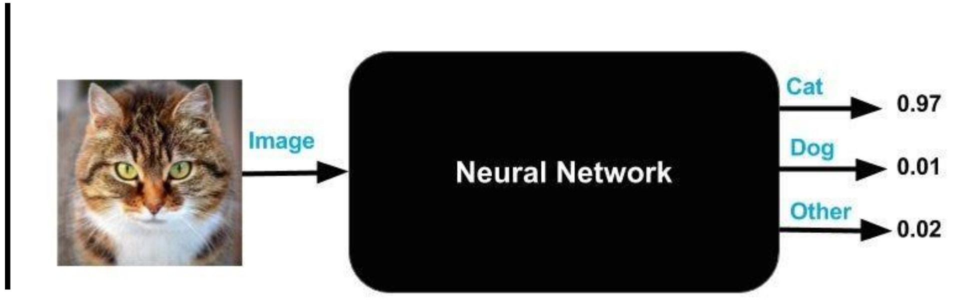
Predictive models

White Box



- Linear Regression
- Decision tree
- Naive bayes

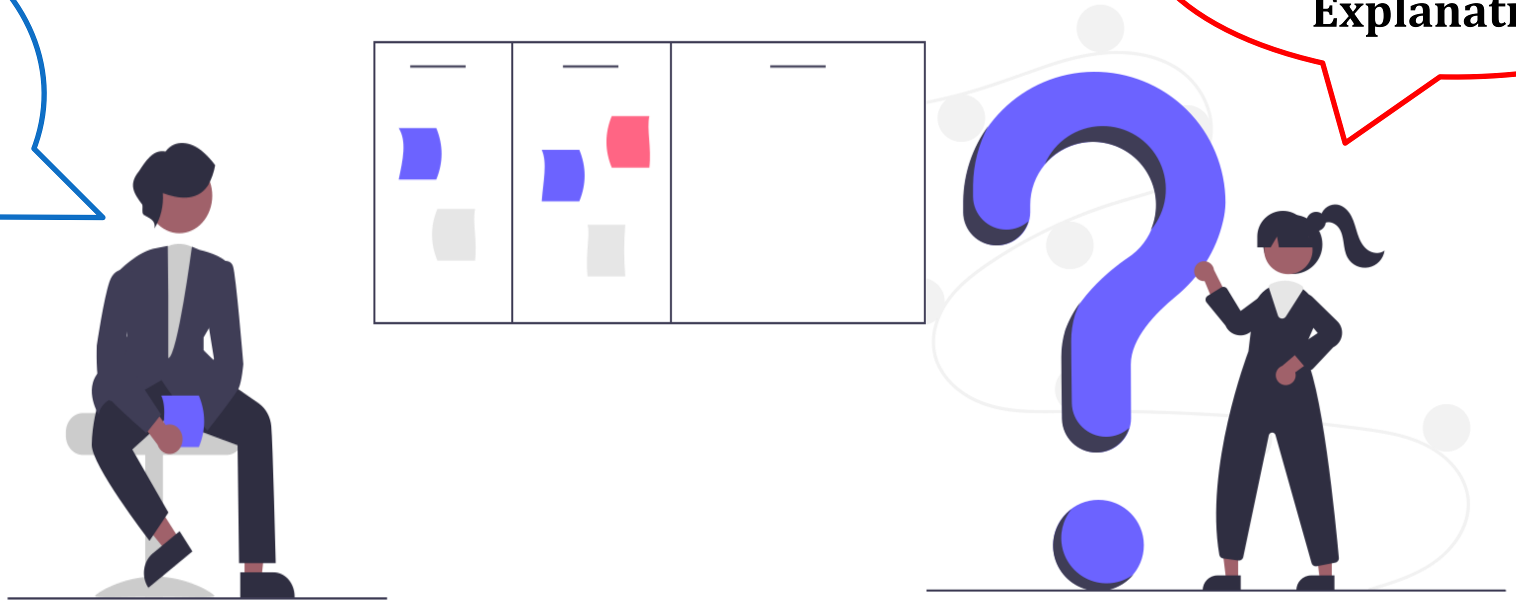
Black Box



- Random Forest,
- SVM
- K nearest neighbour
- Deep Neural Networks

- **Why** we need to **Explain Black Box** models?

Why certain decisions/ predictions have been made ?



Can we trust a **Decision** made by a machine **without Explanation**?

European Union's General Data Protection Regulation(GDPR)

- GDPR stipulates right to obtain *"meaningful information about the logic involved"* commonly interpreted as a "right to an explanation" for consumers affected by an automatic decisions.

On 21 January 2019, the French Data Protection Authority (*Commission Nationale Informatique et Liberté* – "CNIL") imposed a fine of € 50 million on Google for infringing the General Data Protection Regulation 2016/679 (the "GDPR")

Applications of XAI

- Transportation- Self driving cars
- Healthcare- Diagnose diseases
- Legal- Court cases
- Finance- Improve the services
- Military- Autonomous systems used in military operations
- Sentimental Analysis- Bias detection

This research mainly focuses to enhance the model interpretability of the Black-Box models related to a text classification task.

Where the model explainability useful in text classification ?

- **Interpretability**

  In legal or medical contexts, it may be necessary to explain how a particular classification decision was reached.

- **Debugging**

  When text classification models are not performing as expected, model explainability can help identify specific areas where the model having errors.

- **Bias Detection and Mitigation**

  Model explainability can help detect and mitigate bias

- **Improving Performance**

  By identifying specific features that are important for classification, help to the develop more effective models.

Explainability Methods

- **Local/Global explanation methods:**
  - Local- Explanations that are specific to a single instance or prediction made by the model.
  - Global-Provide insights into the overall behavior and performance of the model across the entire dataset.

- **Model-specific/Model-agnostic explanation methods:**
  - Model-specific- Designed for a particular machine learning model or algorithm
  - Model-agnostic- More general and can be applied to any machine learning model.

➢Post hoc/Intrinsic explanation methods:

- Post hoc-Explain a trained model's decisions after it has been trained.
- Intrinsic-Explain a model's decisions by analyzing its internal structure and parameters.

Overall Objectives

Main Objective:

Provide a novel post-hoc ,model-specific, local XAI solution to enhance the model interpretability of Black-Box models focus on Random Forest, Support Vector Machine, K Nearest Neighbor and Logistic Regression by developing a novel counterfactual rule generation mechanism related to the text classification domain.

Sub Objectives:

- To develop novel explainable method to enhance the model interpretability of function-based classification models focus on SVM .
- To develop novel explainable method to enhance the model interpretability of ensemble models focus on Random forest .
- To develop novel explainable method to enhance the model interpretability of distance-based classification models focus on KNN .
- To develop novel explainable method to enhance the model interpretability of regression-based classification models focus on Logistic regression .

What are the Counterfactual Explanations?

- **To flip the prediction, what are the changes that need be done to the model features.**

  Ex: Suppose a company has a machine learning model that predicts whether a customer is likely to churn

  A customer has churned, and the company wants to know why. The company uses counterfactual rule generation to provide an explanation for the decision.

  Counterfactual explanation: "If the customer had received a response to their support ticket within 24 hours, they would not have churned."

- **Diverse Counterfactual Explanations is a popular counterfactual framework.**

Existing XAI Tools

- **Shapley Additive explanations (SHAP)**
  - SHAP is a unified approach that has been developed based on **coalitional game theory**.
  - SHAP assigns important value for each and every feature according to their contribution for the prediction.
  - It maps the input features with the output results based on that Sharpley value.

- **Local interpretable model agnostic explanations (LIME)**
  - Provides local optimum explanations which compute the important features by generating normally distributed samples of the feature vector.
  - Then it assigns weights to each of the rows how close they are from original sample.
  - After it uses feature selection techniques like PCA(Principal Component Analysis) to get significant features.

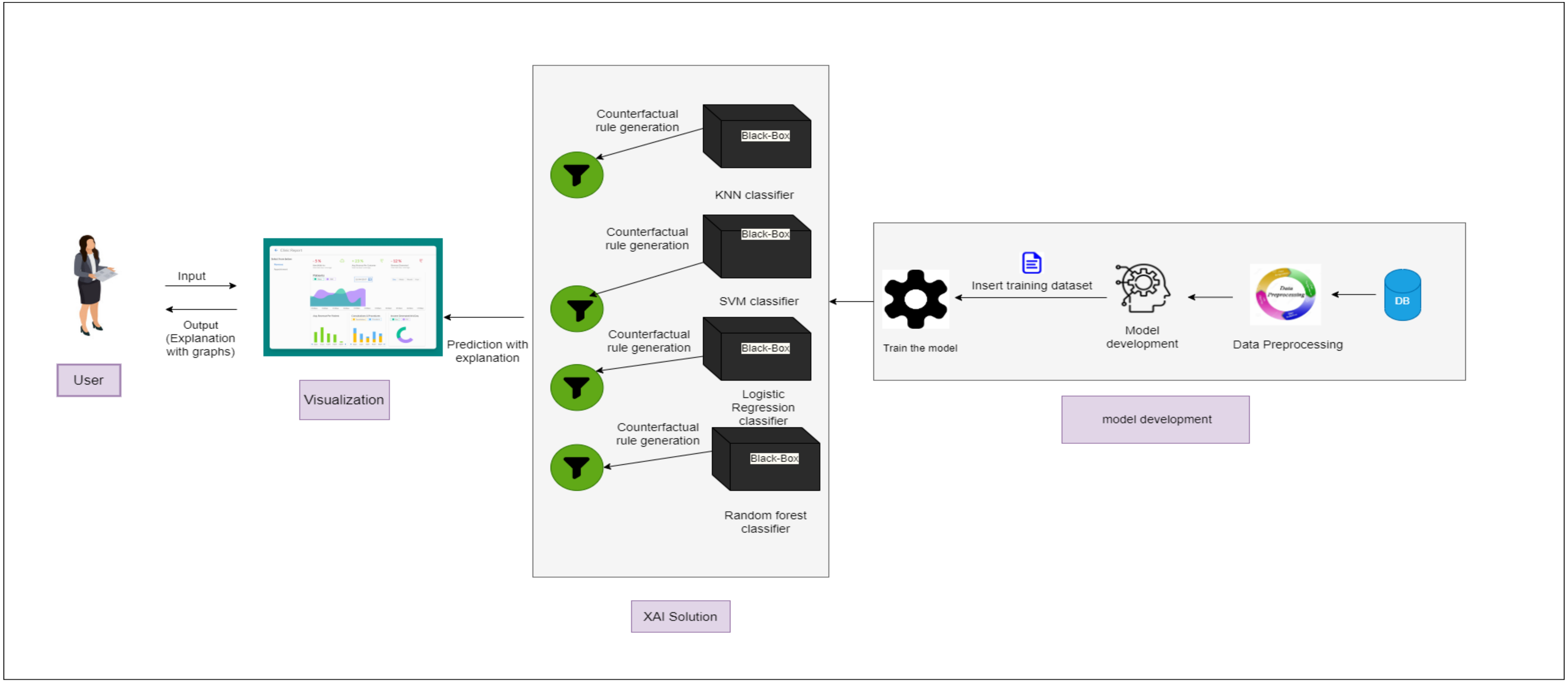
➢ XAI360

- XAI360 is model-agnostic tool which Provide both local and global explanations.
- It provides explanations for machine learning models using various techniques.
    - ❑ Partial dependence plots
    - ❑ Feature importance rankings
    - ❑ Decision trees
    - ❑ Counterfactual explanations

➢ Google XAI

- Google XAI is model-agnostic tool which Provide both local and global explanations.
- It provides explanations for machine learning models using various techniques
    - ❑ LIME
    - ❑ SHAP
    - ❑ Model cards
    - ❑ Integrated Gradients
    - ❑ TCAV (Testing with Concept Activation Vectors)

Overall System Diagram



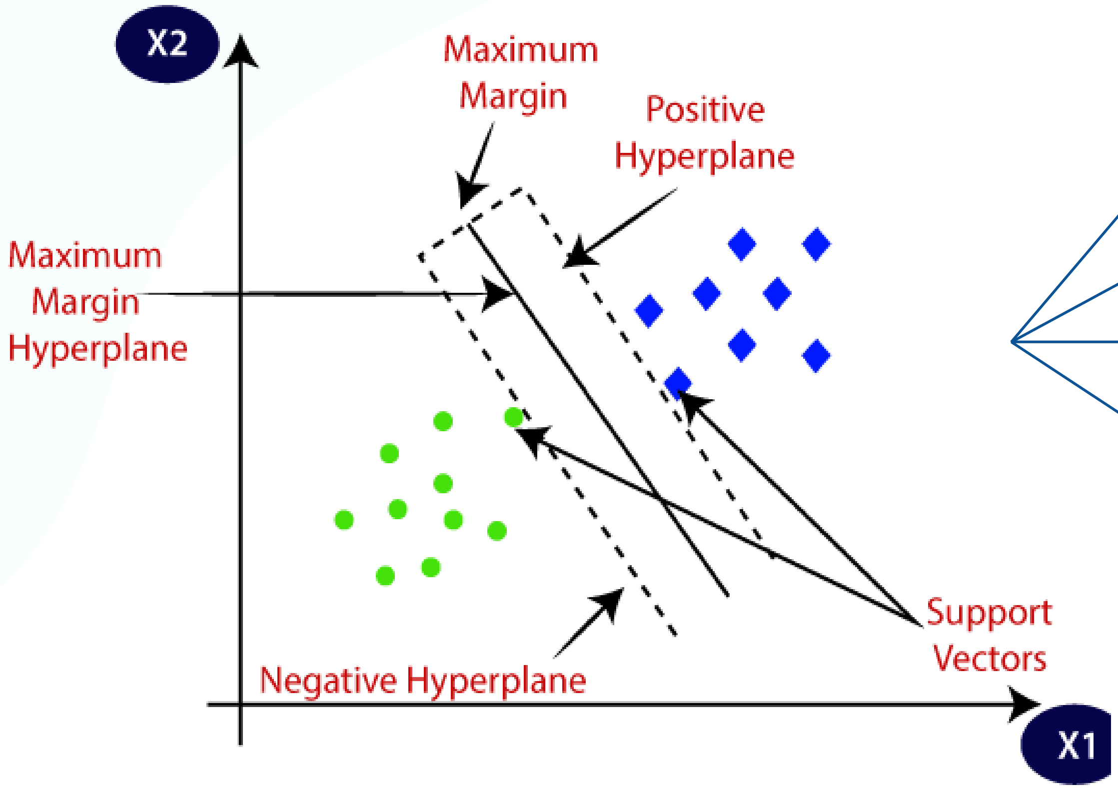
IT20097660 | WARNASOORIYA S.D

Specializing in Data Science

Support Vector machine (SVM).

Introduction

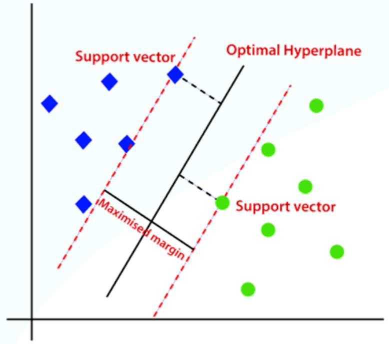
What is **Support Vector Machine (SVM)** ?



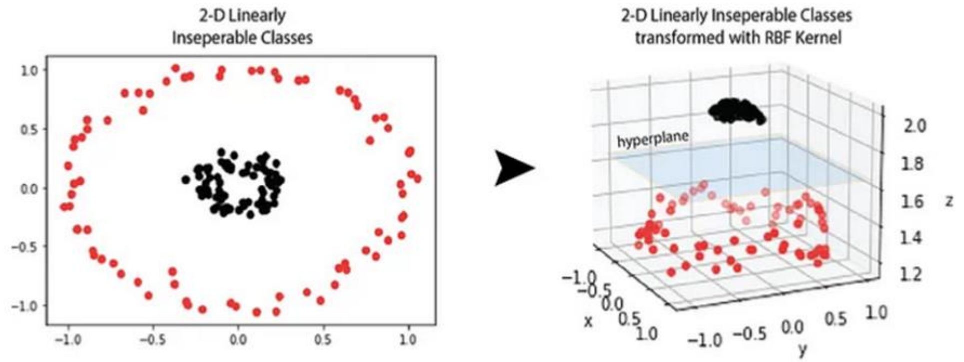
- Supervised Learning
- Mainly for classification task
- Goal is to create the best decision boundary with optimal hyperplane.
- Performs well with both linear and non-linear data using "Kernels".

Background

SVM Behavior with linearly separable and non-linearly separable data.



Linearly Separable Data

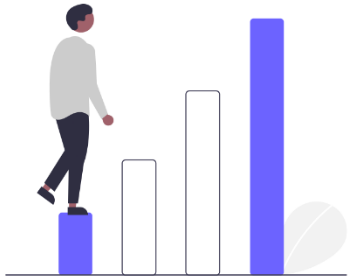


Non-Linearly Separable data

- SVM can use a straight line to separate the data into two classes.
- The model can be trained to find the optimal decision boundary that maximizes the margin between the two classes.
- Once the optimal hyperplane is found, the SVM can make predictions on new input data by simply evaluating which side of the hyperplane the input data falls on.

- SVM use a non-linear kernel function to map the input data into a higher-dimensional space where a linear decision boundary can be found.
- The choice of kernel function and its associated parameters can greatly impact the performance of the SVM

Why we need to explain SVM?



- The black box behavior of the SVM become more pronounced,

  When SVM is used to classify non-linearly separable data.

- When the input data is transformed into a higher-dimensional space using a non-linear kernel,
  - The decision boundary can become highly complex and difficult to visualize.
  - It may not be clear how the SVM is making its predictions.
  - Can be challenging to understand the role of each input feature in the decision-making process.

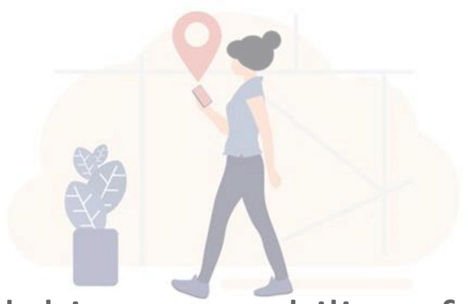
Research Gap

	Anchors [1]	LORE [2]	SHAP-FOIL [3]	XAI360[4]	GoogleXAI [5]	Diverse Counterfactual Explanations [6]	LIME [7]	SHAP [8]	Proposed SVM Explainable Method
Model Specific Approach.	X	X	√	X	X	X	X	X	√
Text classification explainable task	√	X	X	√	√	√	√	√	√
Provide a Counterfactual Rule Generation based Explainable Method.	X	X	X	√	X	√	X	X	√
Provide a user-friendly visualization	X	X	X	X	X	X	X	X	√

Research Problem

- How to get a counterfactual rule generation-based explanation for the Support Vector Machine classifier, when it handle non-linear separable data in text classification?

Objectives



Specific Objective

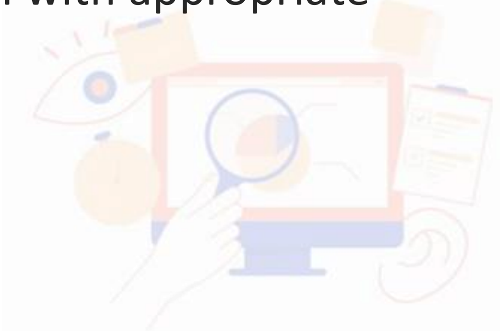
- Provide a novel post-hoc ,model-specific, local XAI solution to enhance the model interpretability of function-based classification models focus on SVM by developing a novel counterfactual rule generation mechanism related to the text classification domain.

Sub Objectives

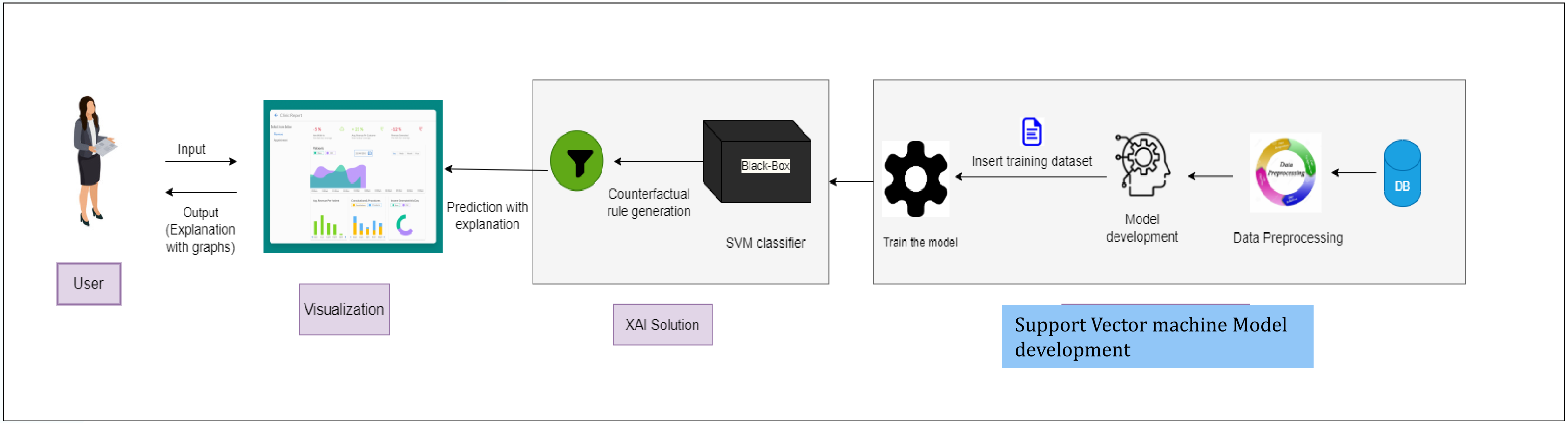
- Prepare the dataset and implement the SVM classifier.
- Develop the novel counterfactual rule generation mechanism related to the text classification task.
- Test the output with existing explainable methods.
- Do experiments to improve the XAI solution more.
- Do the visualization using the most appropriate Graphical User Interface (GUI) technique.

Methodology

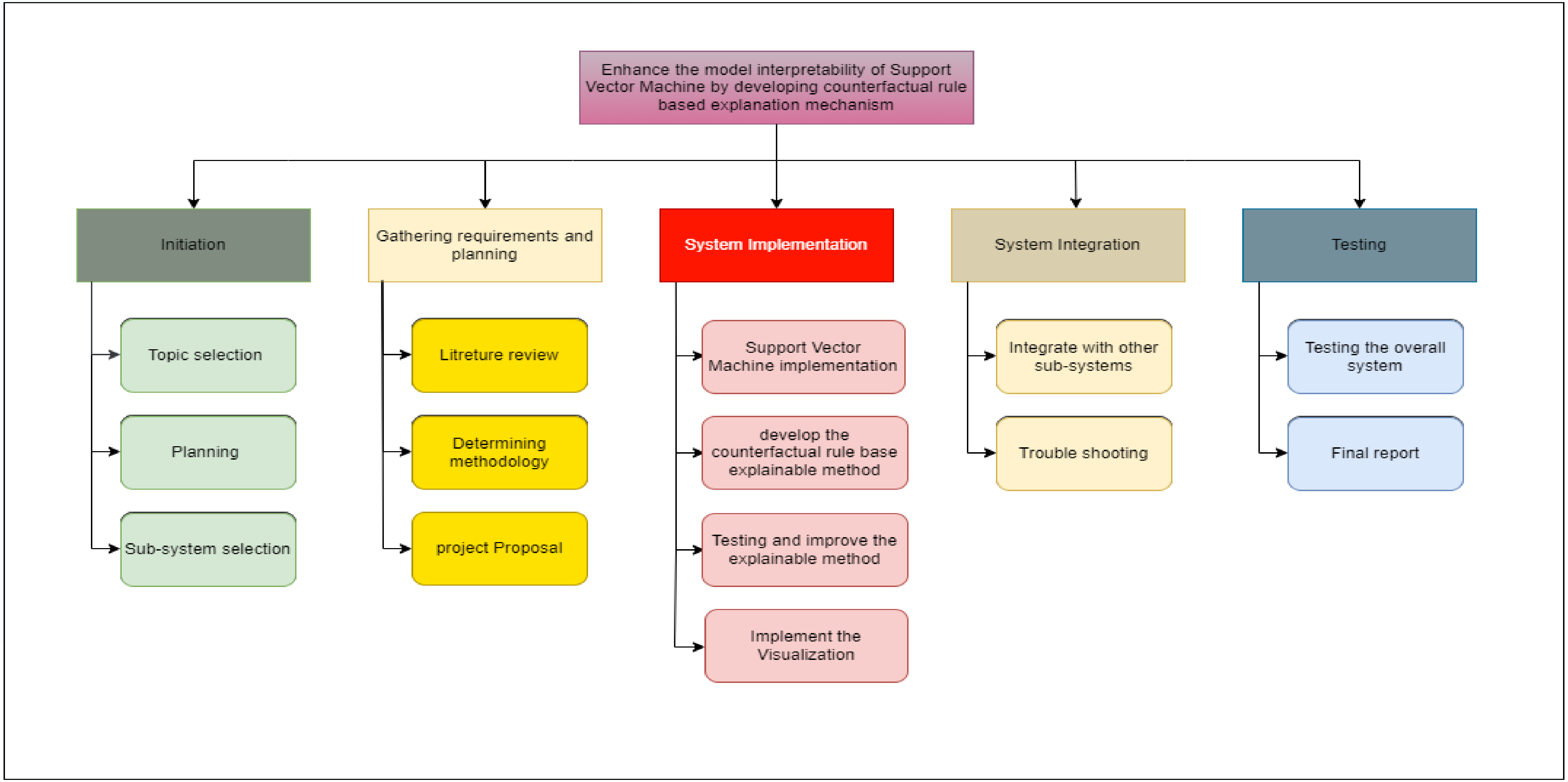
- The user has to provide the relevant training dataset and the instance that need to be predicted to the system through a GUI.
- The provided data will be applied to the support vector machine (SVM) model to get the prediction.
- Apply the novel counterfactual rule generation mechanism to the SVM and extract the counterfactual rules.
- Evaluate the novel explanation mechanism by testing with existing explainable tools and analysing expert's feedbacks
- The outcomes of the process (Counterfactual Rules) will be transferred to the GUI with appropriate visualizations to be more understandable for the users.



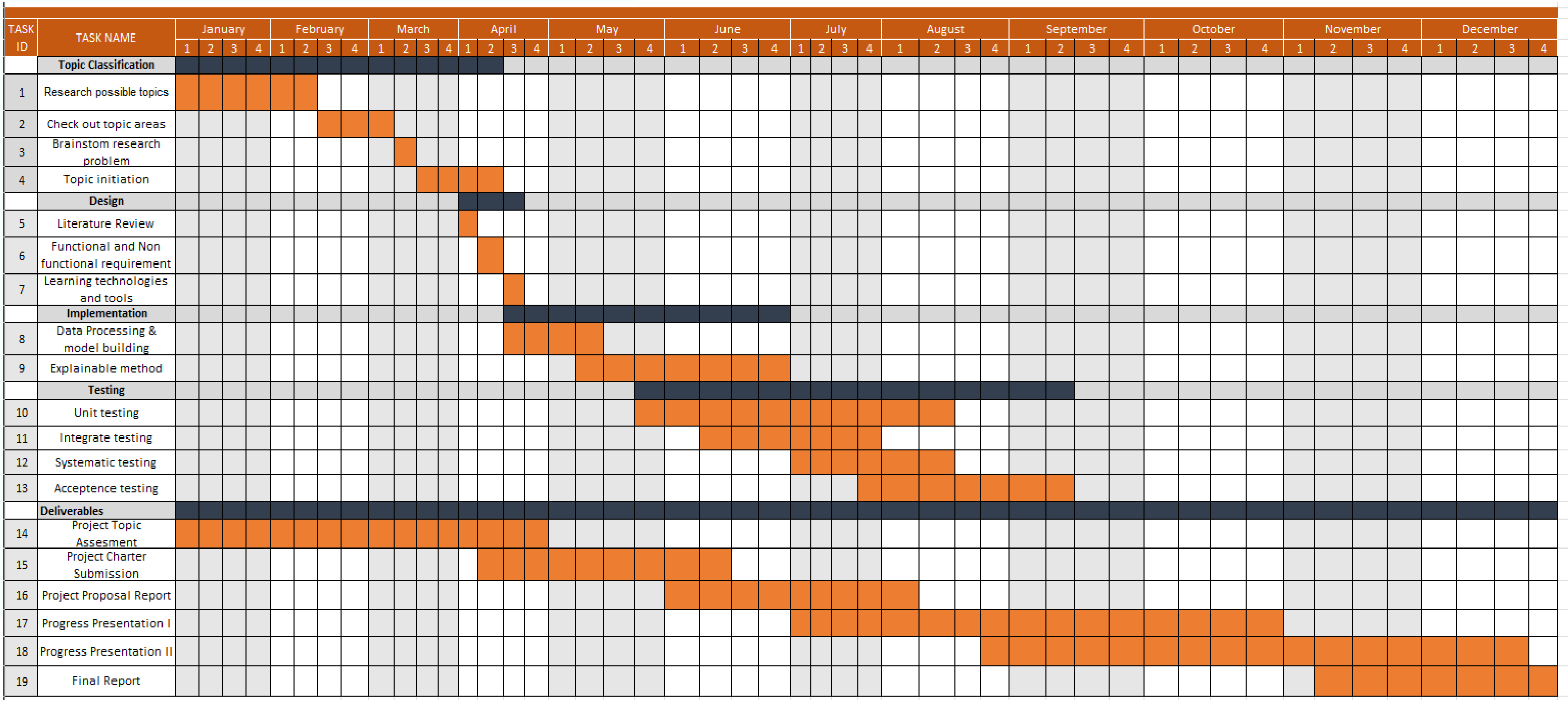
System Diagram



WBS


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

Gantt Chart


TASK ID
TASK NAME
January
February
March
April
May
June
July
August
September
October
November
December
1 2 3 4
Topic Classification
1 Research possible topics
2 Check out topic areas
3 Brainstom research problem
4 Topic initiation
Design
5 Literature Review
6 Functional and Non functional requirement
7 Learning technologies and tools
Implementation
8 Data Processing & model building
9 Explainable method
Testing
10 Unit testing
11 Integrate testing
12 Systematic testing
13 Acceptence testing
Deliverables
14 Project Topic Assesment
15 Project Charter Submission
16 Project Proposal Report
17 Progress Presentation I
18 Progress Presentation II
19 Final Report

References

[1] M. T. Ribeiro and C. Guestrin, “Anchors : High-Precision Model-Agnostic Explanations,” pp. 1527–1535.

[2] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti, “Local rule-based explanations of black box decision systems,” *arXiv*, no. May, 2018.

[3] F. Shakerin and G. Gupta, “White-box Induction from SVM Models: Explainable AI with Logic Programming,” *Theory Pract. Log. Program.*, vol. 20, no. 5, pp. 656–670, 2020, doi: 10.1017/S1471068420000356.

[4] A. Adadi and M. Berrada, “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI),” *IEEE Access*, 2018, doi: 10.1109/ACCESS.2018.2870052.

[5] 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

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

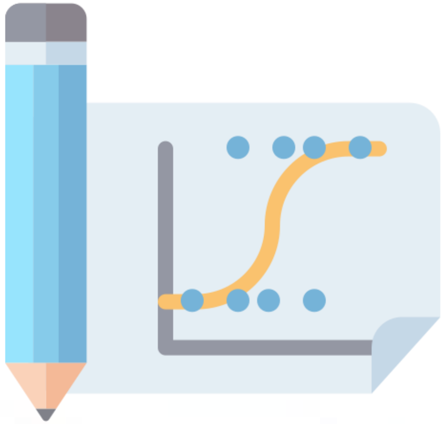
[7] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should i trust you?’ Explaining the predictions of any classifier,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, vol. 13-17-Augu, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[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.

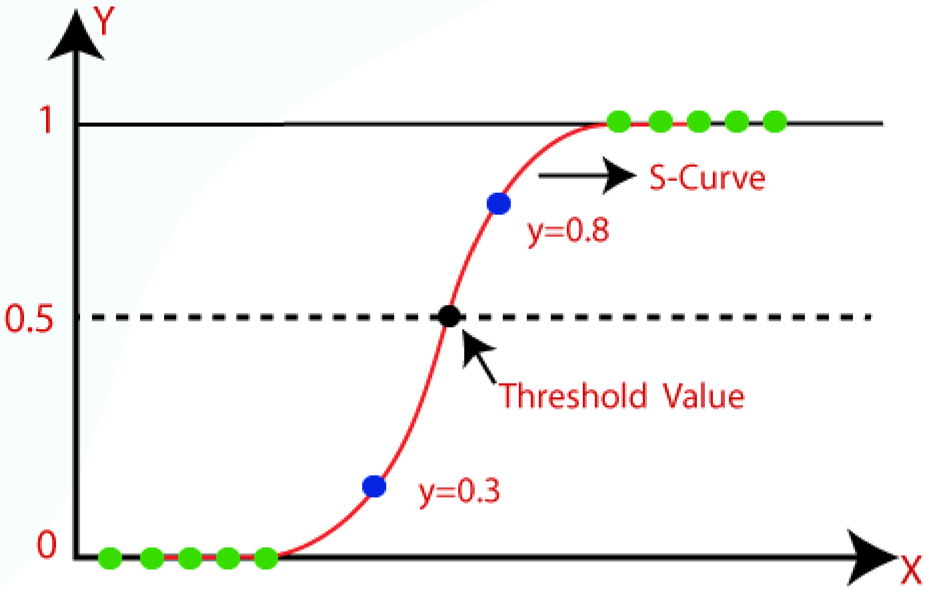
IT18161298 | SRINDEE METHMAL H.M

Specializing in Software Engineering

Logistic Regression



Introduction



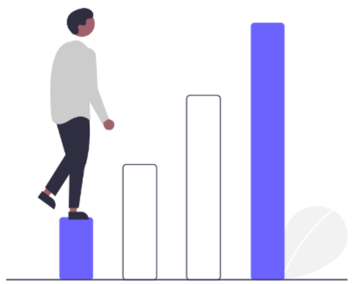
Supervised Learning

Used for predicting the categorical dependent variable using a given set of independent variables.

Used as a binary classification for text classification

Independent variable should not have multi-collinearity.

Introduction Cont.



- How Logistic Regression works?

  - Model the probability of an instance belonging to a particular class using a logistic function

  - For that it keeps minimizing the difference between predicted probabilities and true labels

  - This is done using a technique called maximum likelihood Estimation

Background

- Do we need **interpretability** for **Logistic Regression**?
  - **YES , of Course!**

- Why?
  - Logistic Regression Model can become black boxes,

    When it has difficulty in understanding how the algorithm arrives at a classification decision for an input text.

Background Cont.

What are the main stages of Logistic Regression for text classification?

- Fetching text data
- Preprocessing
- Text feature extraction and training the classifiers
- Evaluated using confusion matrix to show the accuracy rate for text classifiers

Research gap

	TCAV [1]	ALIBI [2]	Diverse Counterfactual Explanations [3]	XAI360[4]	GoogleXAI[5]	LIME [6]	SHAP [7]	Proposed LR Explainable Method
Model Specific Approach	√	√	√	X	X	X	X	√
Text classification explainable task	√	√	√	√	√	√	√	√
Provide a counterfactual Rule Generation based Explainable Method	X	X	√	√	√	X	X	√
Provide a user-friendly visualizations	√	√	X	X	X	√	√	√

Research Problem

- How to get a counterfactual rule generation-based explanation for the Logistic Regression classifier when it becomes black box in text classification?

Objectives

Specific Objective

- Providing model specific ,local ,post-hoc explanations using counterfactual mechanisms to improve the interpretability of the system.

Sub Objectives

- Implementing a mechanism to calculating the predicted probabilities based on current features and coefficients.
- Then manipulating and recalculate probabilities
- So, the difference between original and predicted probability after manipulating can be used to identify impact of each feature on model prediction.

Methodology

- The user has to provide the relevant training dataset and the instance that need to be predicted to the system through a GUI.

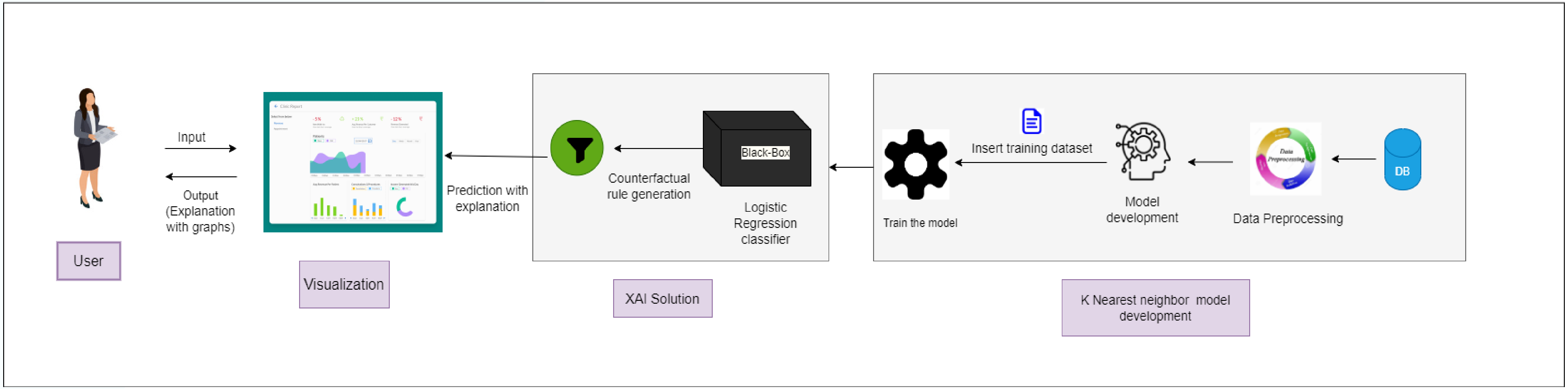
- The provided data will be applied to the logistic regression model to get the prediction.

- Apply the novel counterfactual rule generation mechanism to the logistic regression and extract the counterfactual rules

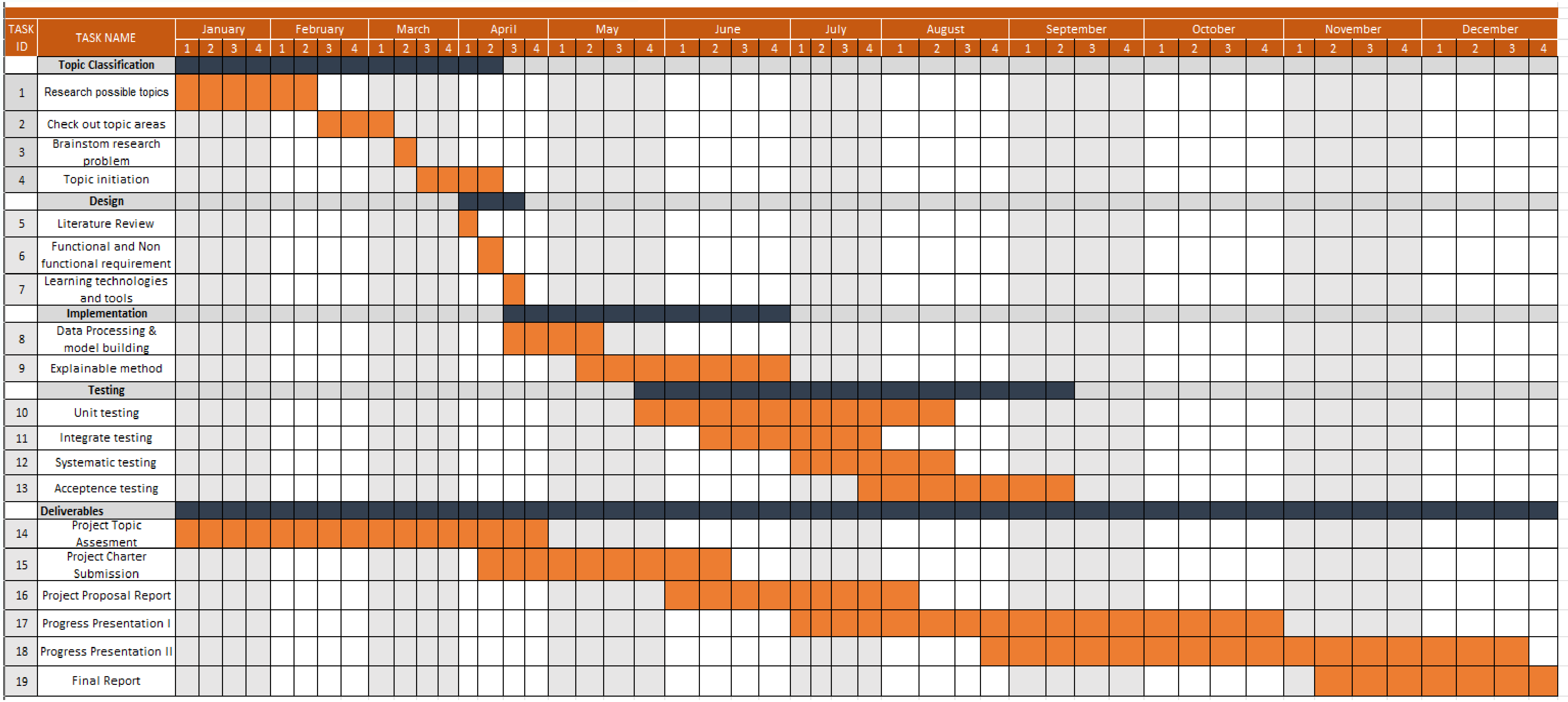
- Evaluate the novel explanation mechanism by testing with existing explainable tools and analysing expert’s feedbacks

- The outcomes of the process (Counterfactual Rules) will be transferred to the GUI with appropriate visualizations to be more understandable for the users.

System Diagram



Gantt Chart


TASK ID
TASK NAME
January
February
March
April
May
June
July
August
September
October
November
December
Topic Classification
1 Research possible topics
2 Check out topic areas
3 Brainstom research problem
4 Topic initiation
Design
5 Literature Review
6 Functional and Non functional requirement
7 Learning technologies and tools
Implementation
8 Data Processing & model building
9 Explainable method
Testing
10 Unit testing
11 Integrate testing
12 Systematic testing
13 Acceptence testing
Deliverables
14 Project Topic Assesment
15 Project Charter Submission
16 Project Proposal Report
17 Progress Presentation I
18 Progress Presentation II
19 Final Report

References

[1]Holzinger, A., Saranti, A., Molnar, C., Biecek, P., & Samek, W. (2022, April). Explainable AI methods-a brief overview. In xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers (pp. 13-38). Cham: Springer International Publishing.

[2]Mishra, P. (2021). Counterfactual Explanations for XAI Models. In Practical Explainable AI Using Python: Artificial Intelligence Model Explanations Using Python-based Libraries, Extensions, and Frameworks (pp. 265-278). Berkeley, CA: Apress.

[3]Dieber, J., & Kirrane, S. (2020). Why model why? Assessing the strengths and limitations of LIME. arXiv preprint arXiv:2012.00093.

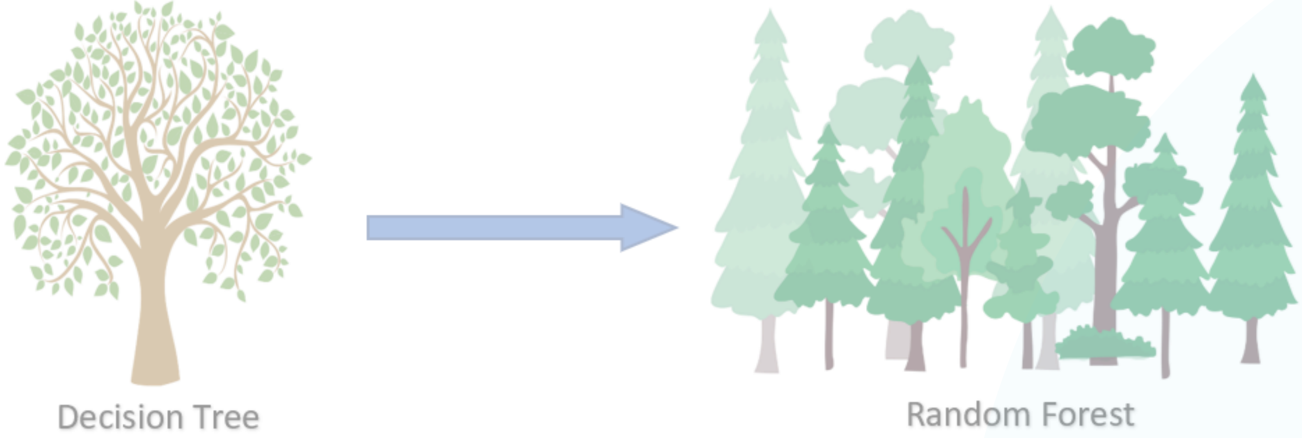
[4]Van den Broeck, G., Lykov, A., Schleich, M., & Suciu, D. (2022). On the tractability of SHAP explanations. Journal of Artificial Intelligence Research, 74, 851-886.

[5]Mothilal, R. K., Sharma, A., & Tan, C. (2020, January). Explaining machine learning classifiers through diverse counterfactual explanations. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 607-617).

IT20100698 | BRITTO T.A

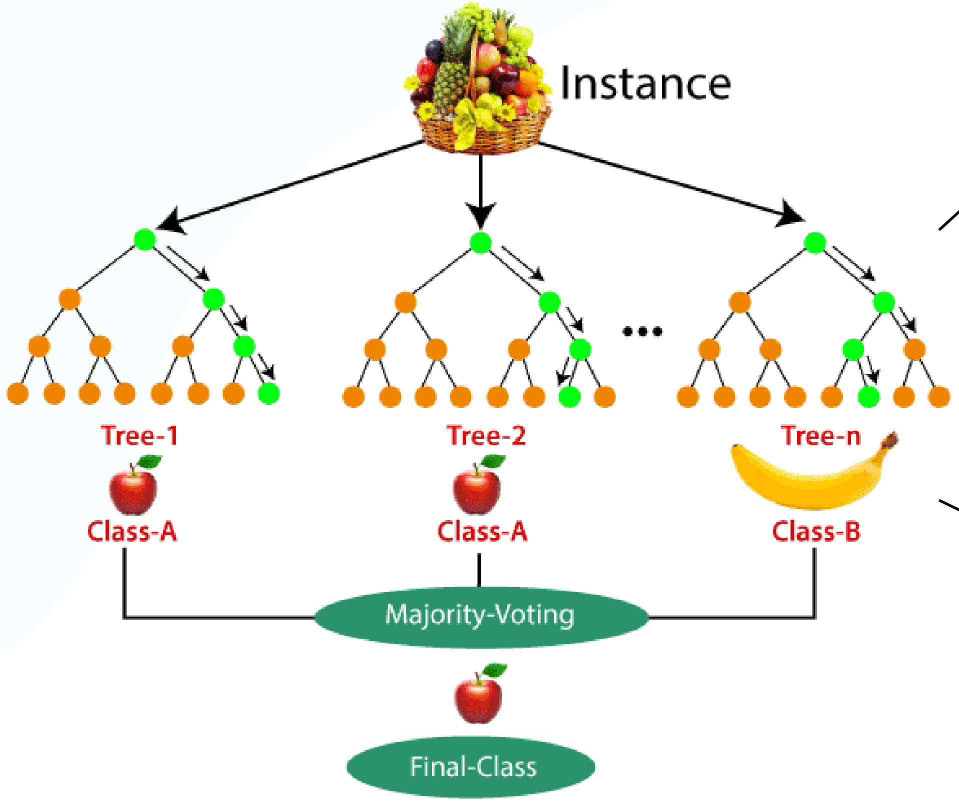
Specializing in Data Science

Random Forest



Introduction

What is Random forest?



Supervised Learning

Ensemble Learning Method

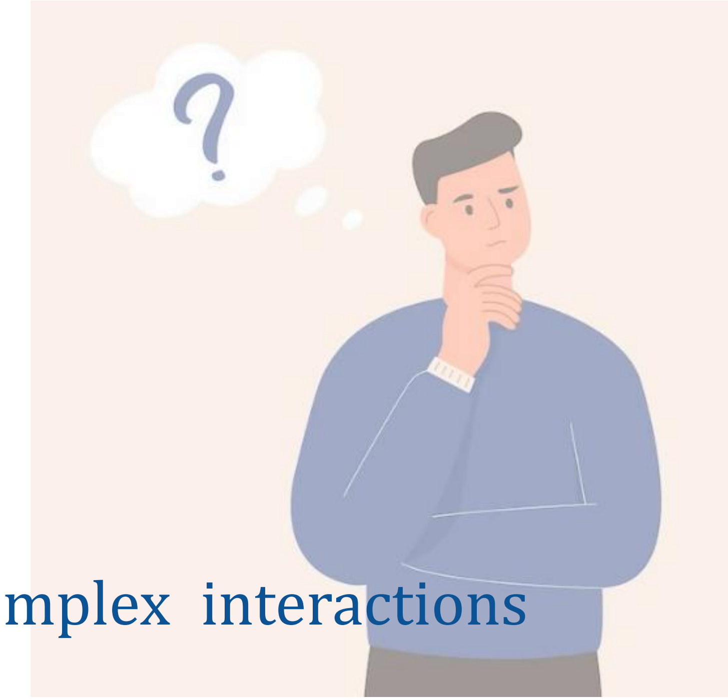
Contains a number of decision trees on various subsets

Why we use Random forest?

How Random forest work?

Background

Why we need to explain Random forest?



- Random forest models can become black boxes,
  when it has a large number of decision trees with complex interactions between the features.

- The final classification decision is based on the output of many decision trees. It make difficult to understand the overall decision-making process.

- The complexity of the dataset and the high number of input features also affect to the black box behavior.

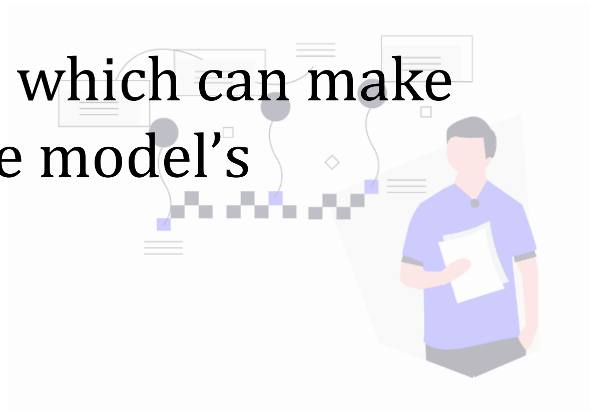
Research Gap

	TreeSHAP [1]	Diverse Counterfactual Explanations [2]	LIME [3]	XAI360[4]	Google XAI[5]	SHAP [6]	Proposed Random Forest Explainable Method
Model Specific Approach.	√	X	X	X	X	X	√
Provide a Counterfactual Rule Generation based Explainable Mechanism.	X	√	X	√	√	X	√
Text classification explainable task	√	√	√	√	√	√	√
Having a user-friendly visualization	X	X	X	X	X	X	√
Locally Explainable	√	√	√	√	√	√	√

Research Problem

- How to get a counterfactual rule generation-based explanation for the Random Forest classifier, when it becomes black box in text classification?

- Text data can have a large number of unique features or words, which can make it difficult to understand how each feature is contributing to the model's decision-making process.



Objectives

Specific 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.

Sub Objectives

- 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.

Methodology

- The user has to provide the relevant training dataset and the instance that need to be predicted to the system through a GUI.

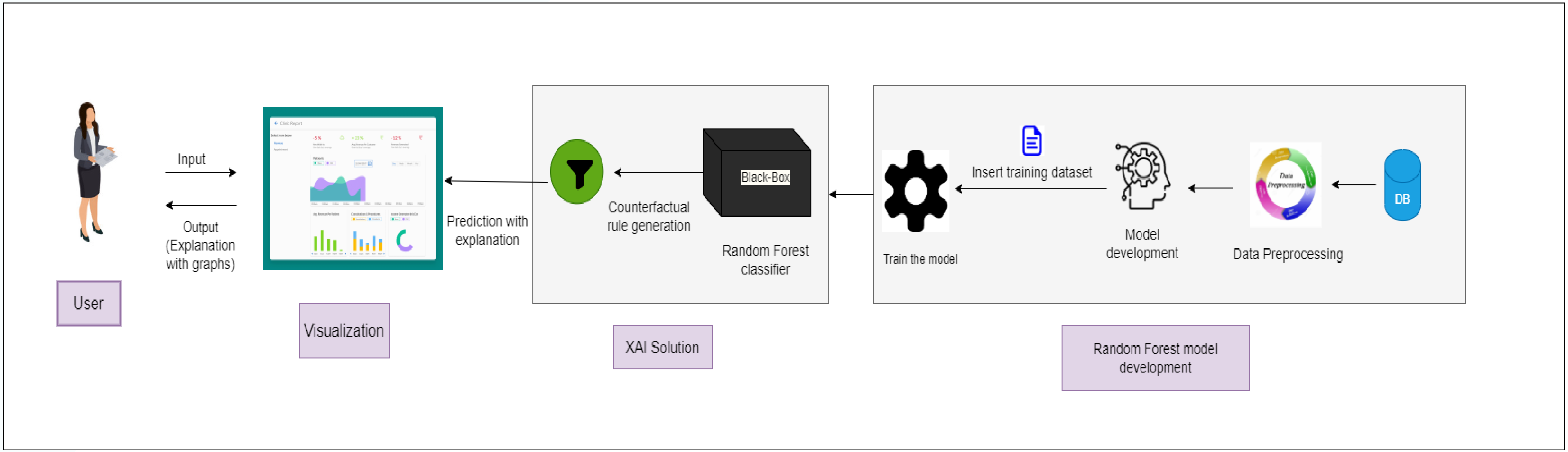
- The provided data will be applied to the Random Forest model to get the prediction.

- Apply the novel counterfactual rule generation mechanism to the Random Forest and extract the counterfactual rules.

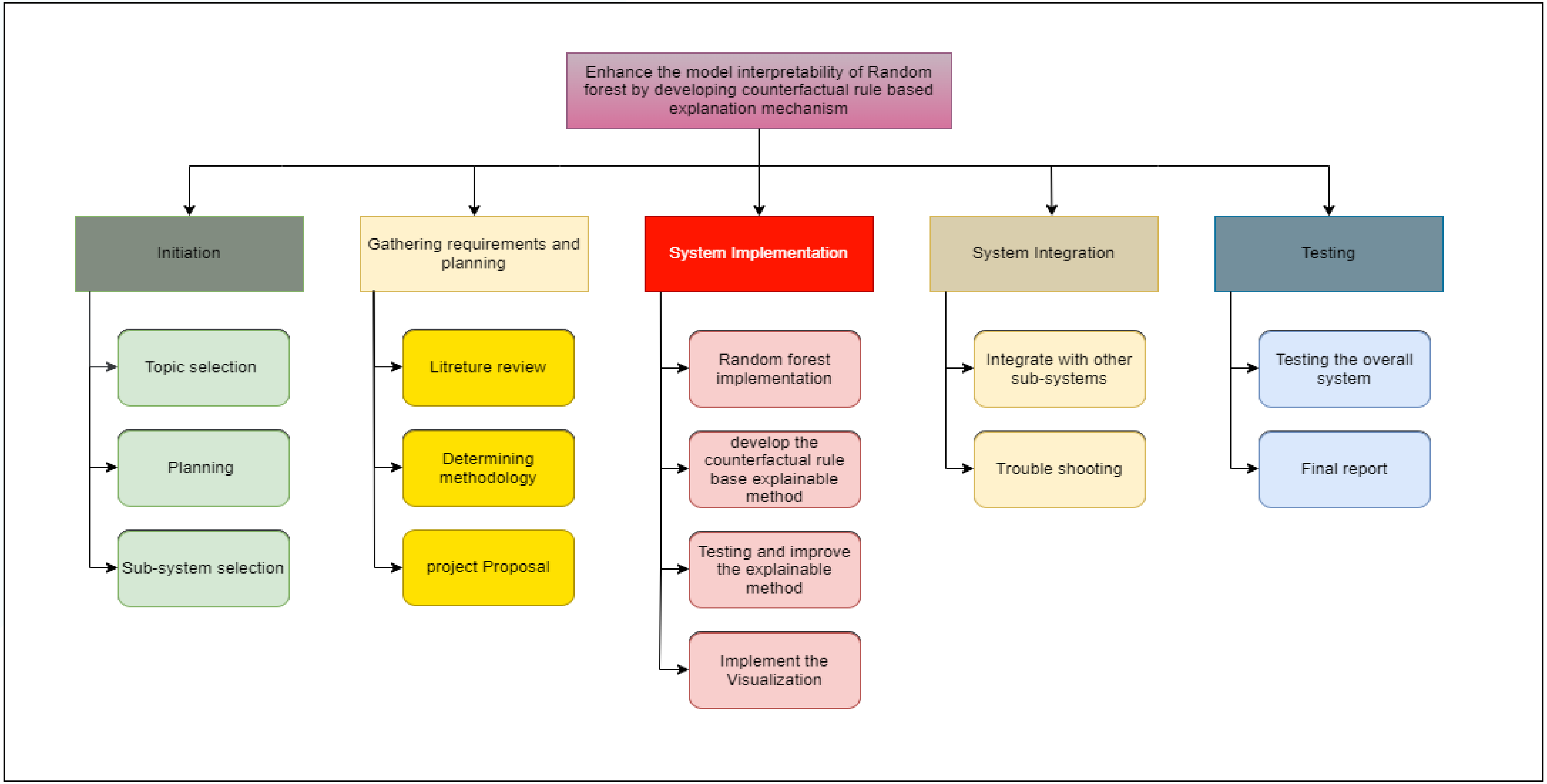
- Evaluate the novel explanation mechanism by testing with existing explainable tools and analysing expert’s feedbacks.

- The outcomes of the process (Counterfactual Rules) will be transferred to the GUI with appropriate visualizations to be more understandable for the users.

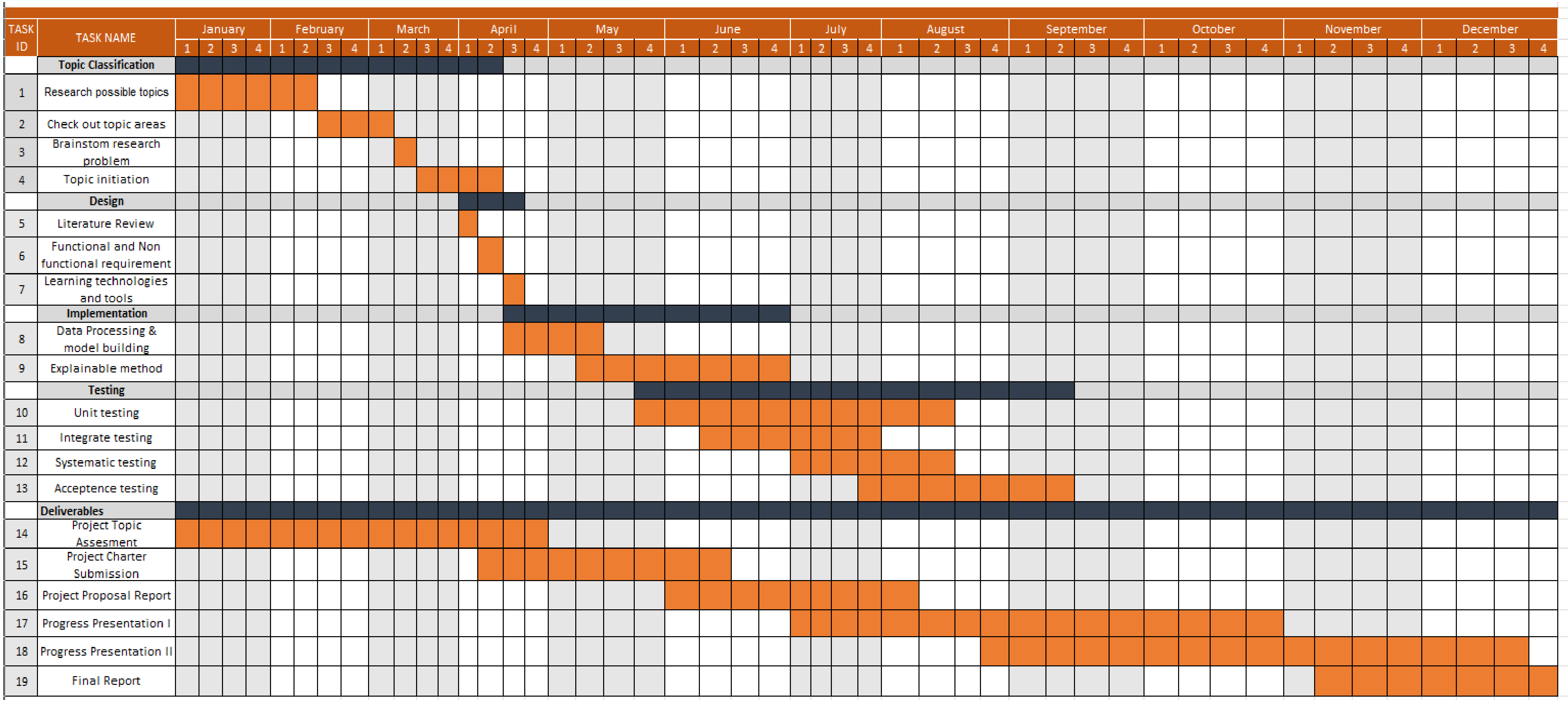
System Diagram



WBS


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

Gannt Chart


TASK ID
TASK NAME
January
February
March
April
May
June
July
August
September
October
November
December
Topic Classification
1 Research possible topics
2 Check out topic areas
3 Brainstom research problem
4 Topic initiation
Design
5 Literature Review
6 Functional and Non functional requirement
7 Learning technologies and tools
Implementation
8 Data Processing & model building
9 Explainable method
Testing
10 Unit testing
11 Integrate testing
12 Systematic testing
13 Acceptence testing
Deliverables
14 Project Topic Assesment
15 Project Charter Submission
16 Project Proposal Report
17 Progress Presentation I
18 Progress Presentation II
19 Final Report

References

[1] 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*.

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

[3] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should i trust you?’ Explaining the predictions of any classifier,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, vol. 13-17-Augu, pp. 1135–1144, doi: 10.1145/2939672.2939778.

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

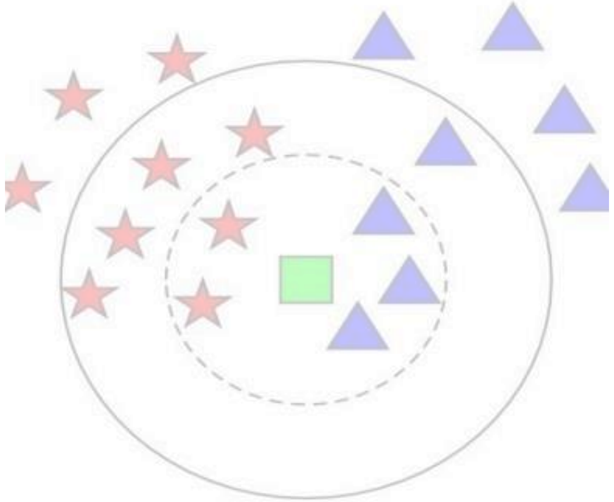
[5] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82-115

[6] 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.

IT20013950 | LAKSHANI N.V.M

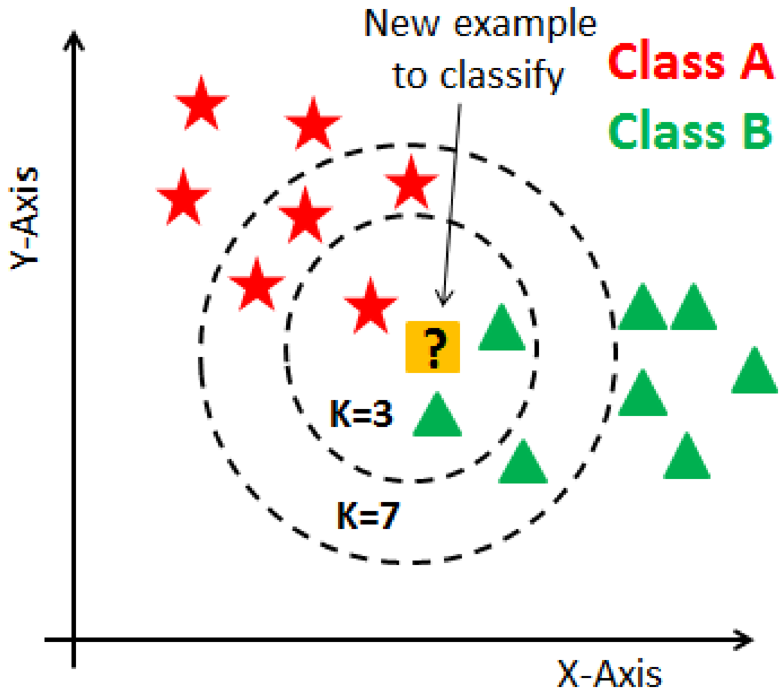
Specializing in Software Engineering

K-Nearest Neighbour



Introduction

- What is **K-Nearest Neighbor Algorithm** (k-NN) ?



Supervised Learning

Non-Parametric

Lazy Learning

Classification Problems

Regression Problems

Background

- k-NN is a distance-based algorithm.
- k-NN is a white-box model when there is only few attributes.
- The Black-Box situation of k-NN algorithm occurs when it has large number of dimensions which is referred to ‘Curse of Dimensionality’ problem in k-NN.
- How to overcome ‘Curse of Dimensionality’ ?
    1. Add more data to ensure that you have enough data density even as you add more dimensions.
    2. Concept of Dimensionality reduction

Research Gap

	LIME[1]	SHAP[2]	Diverse Counterfactual Explanations[3]	XAI360[4]	GoogleXAI[5]	Explaining and Improving Model Behavior with k Nearest Neighbor Representations[6]	Proposed KNN Explainable Method[7]
Model specific approach	X	X	X	X	X	X	√
Provide a Counterfactual Rule Generation based Explainable Method.	X	X	√	√	√	X	√
Locally Explainable	√	X	√	√	√	√	√
Text classification explainable task	√	√	√	√	√	√	√
Provide a user-friendly visualizations	X	X	X	X	X	X	√

Research Problem

- How to get a counterfactual rule generation-based explanation for the k-NN classifier, when it handle Curse of Dimensionality problem in text classification?

Objectives

Specific Objective

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

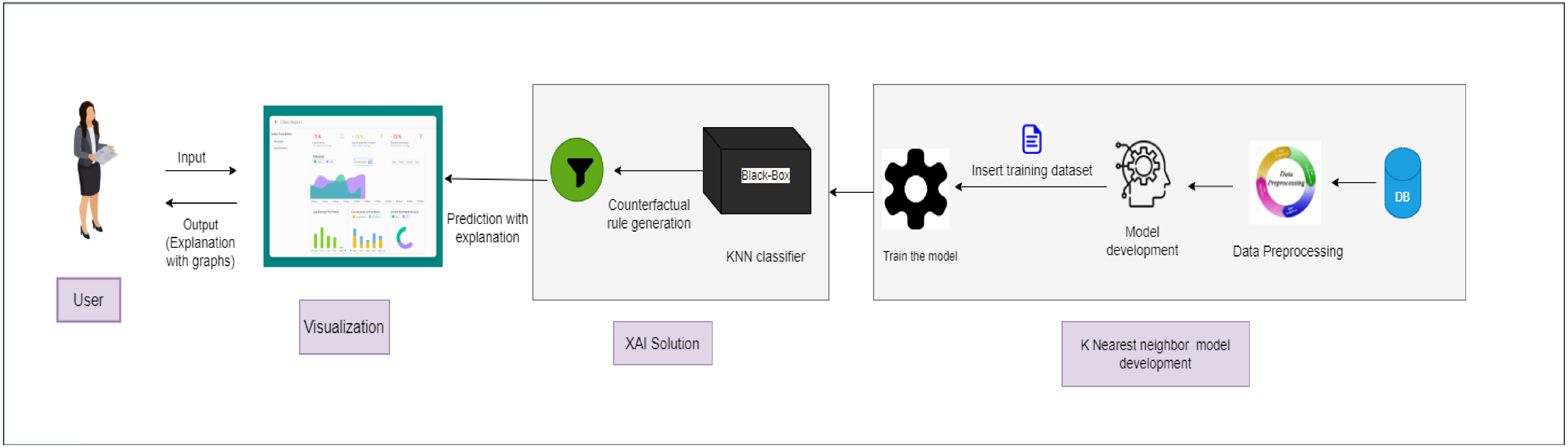
Sub Objectives

- Prepare the dataset and implement the k-NN text classifier.
- Develop the novel counterfactual rule generation mechanism related to the text classification task.
- Test the output with existing explainable methods.
- Do experiments to improve the XAI solution more.

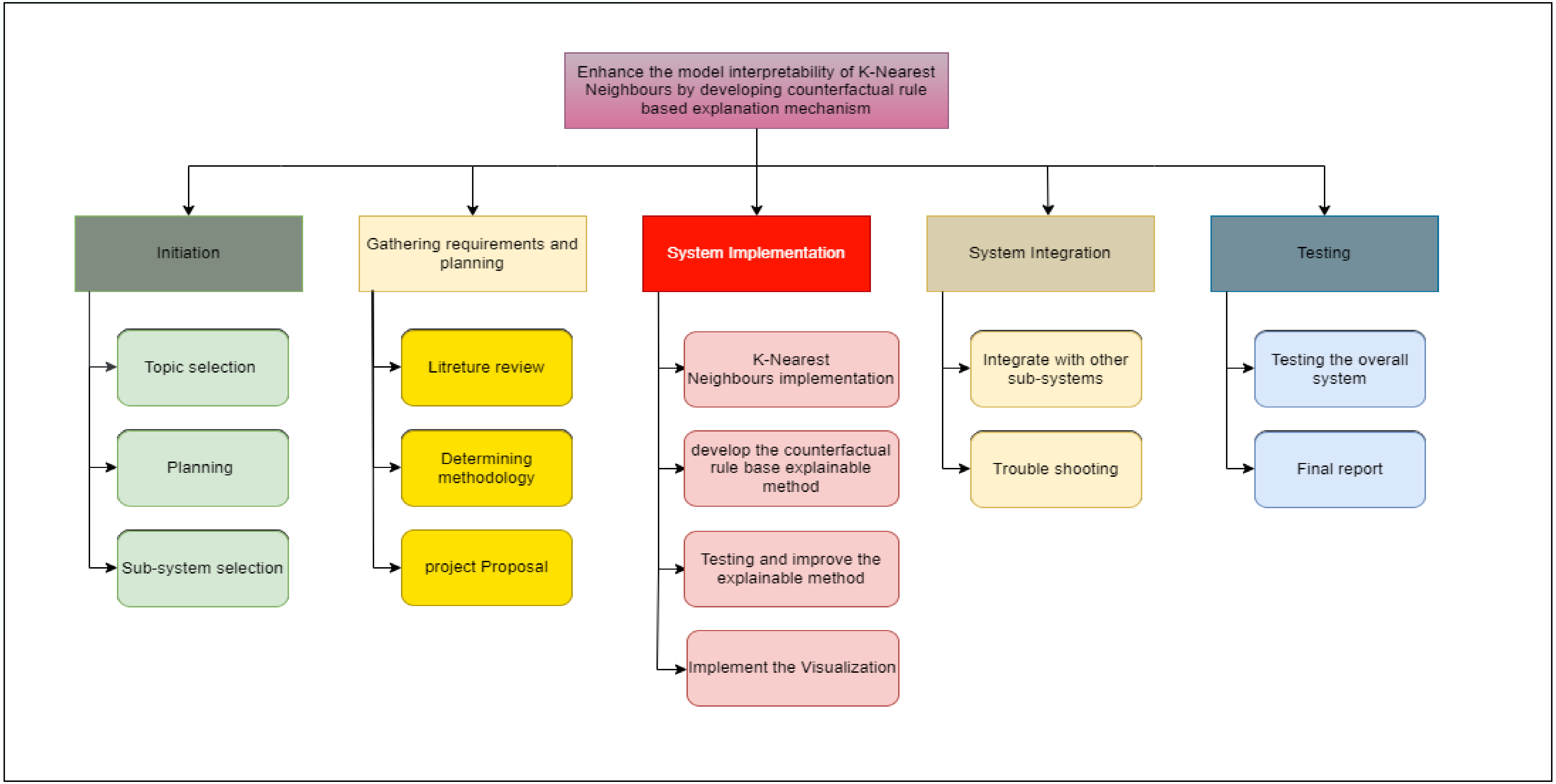
Methodology

- The user has to provide the relevant training dataset and the instance that need to be predicted to the system through a GUI.
- The provided data will be applied to the K Nearest Neighbour model to get the prediction.
- Apply the novel counterfactual rule generation mechanism to the K Nearest Neighbour and extract the counterfactual rules.
- Evaluate the novel explanation mechanism by testing with existing explainable tools and analysing expert’s feedbacks
- The outcomes of the process (Counterfactual Rules) will be transferred to the GUI with appropriate visualizations to be more understandable for the users.

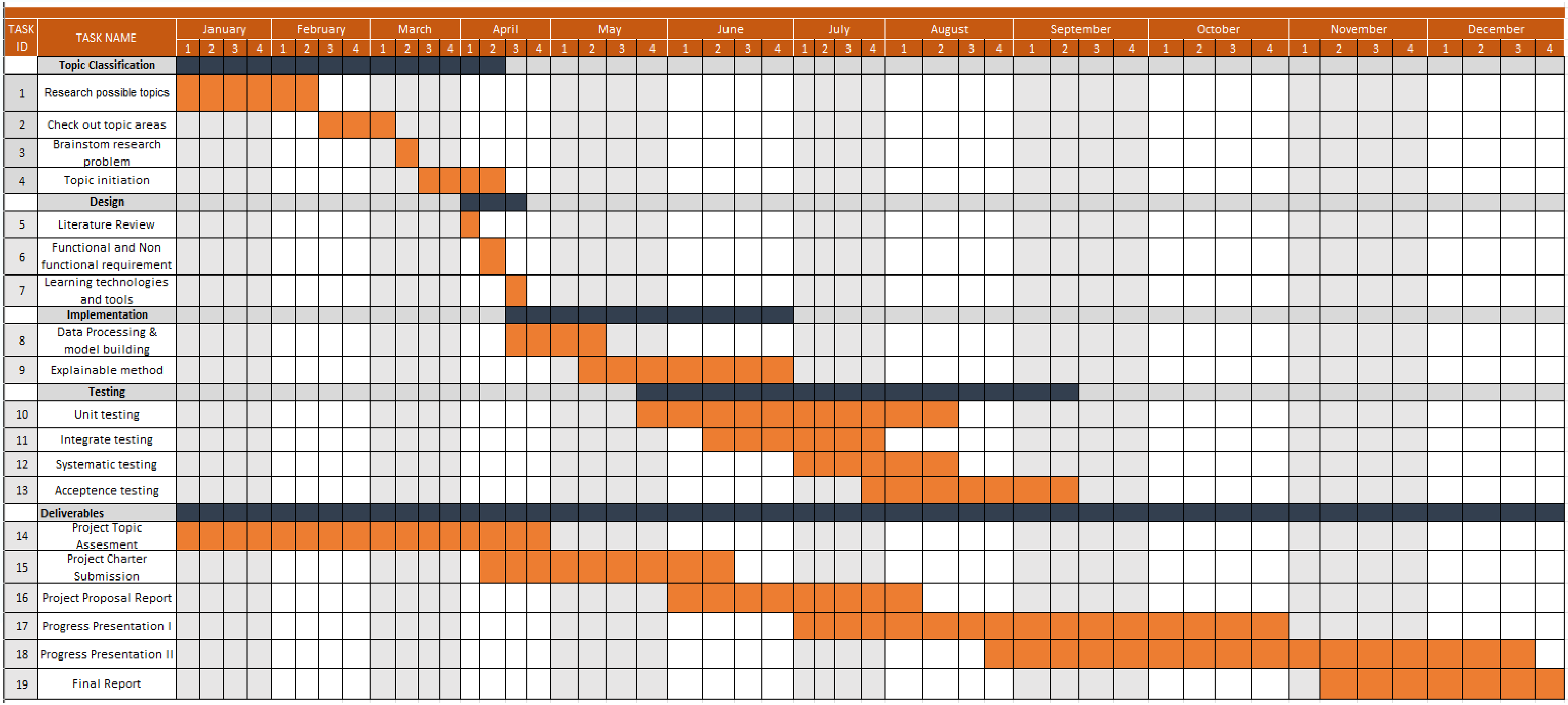
System Diagram



WBS


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

Gannt Chart


TASK ID
TASK NAME
January
February
March
April
May
June
July
August
September
October
November
December
1 2 3 4
Topic Classification
1 Research possible topics
2 Check out topic areas
3 Brainstom research problem
4 Topic initiation
Design
5 Literature Review
6 Functional and Non functional requirement
7 Learning technologies and tools
Implementation
8 Data Processing & model building
9 Explainable method
Testing
10 Unit testing
11 Integrate testing
12 Systematic testing
13 Acceptence testing
Deliverables
14 Project Topic Assesment
15 Project Charter Submission
16 Project Proposal Report
17 Progress Presentation I
18 Progress Presentation II
19 Final Report

References

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' Explaining the predictions of any classifier," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, vol. 13-17-Augu, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[2] 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

[3] R. K. Mothilal and C. Tan, "Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations.", 2019

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' Explaining the predictions of any classifier," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, vol. 13-17-Augu, pp. 1135–1144, doi: 10.1145/2939672.2939778.

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

[6] 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

[7] 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.

Tools and Technologies

- **Frontend:**
  - ReactJS
  - Flask
  - Boostrap
- **Backend:**
  - Python
- **Version Control:**
  - GitHub
- **Tools:**
  - VS Code
  - Google Colab

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 neede
- Model accuracies should be provided by the system.

➢ Non-Functional Requirements

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

Q & A

Thank You !