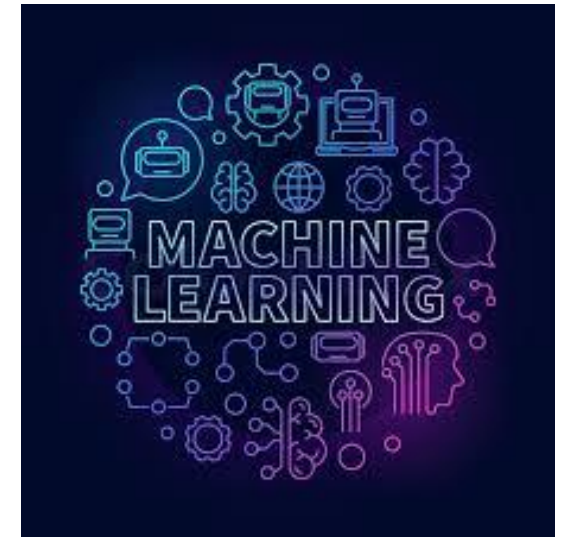
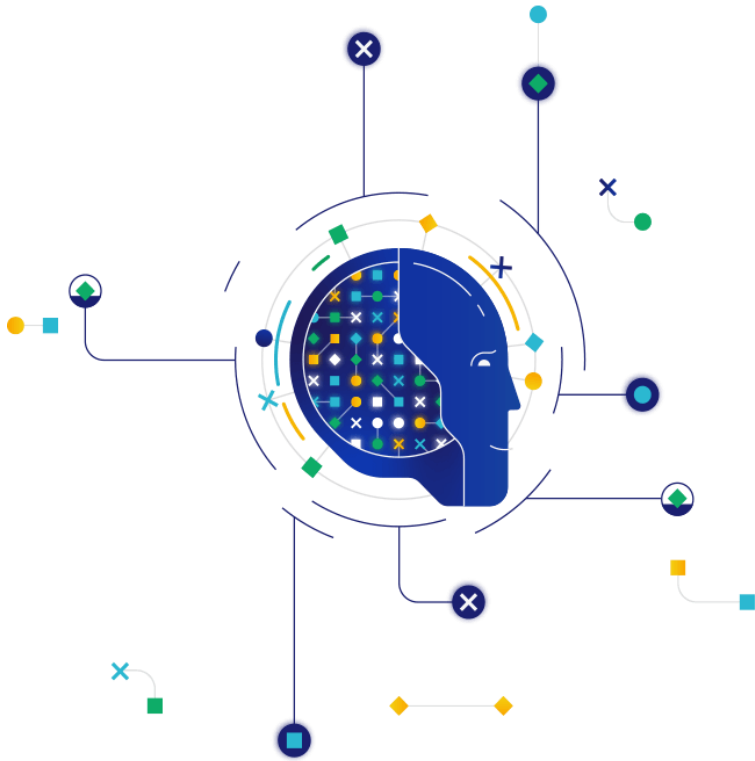


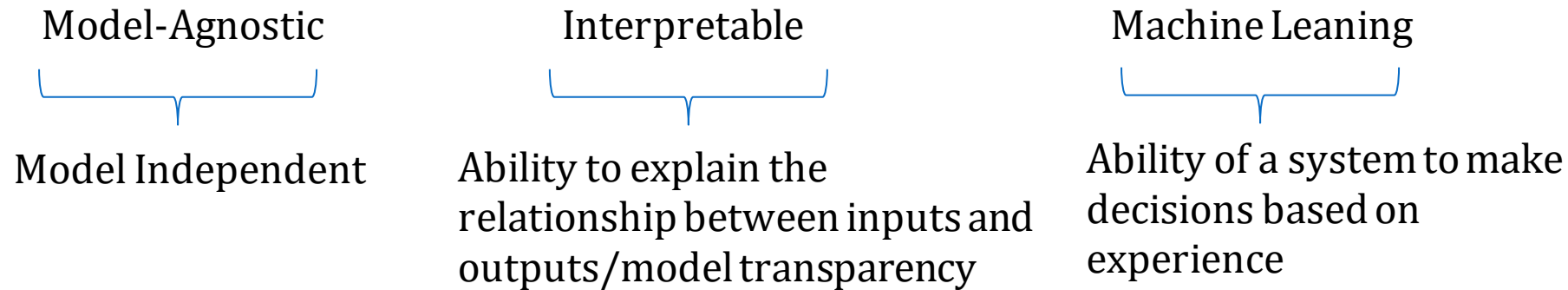
MAIML: Model-Agnostic Interpretable Machine Learning

2021-117



INTRODUCTION

- What is Model-Agnostic Interpretable Machine Learning?

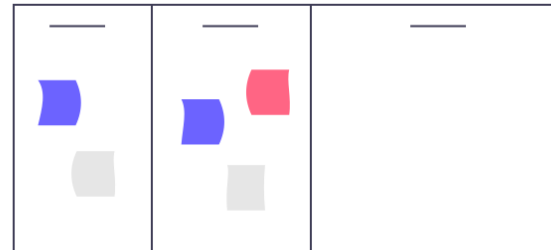
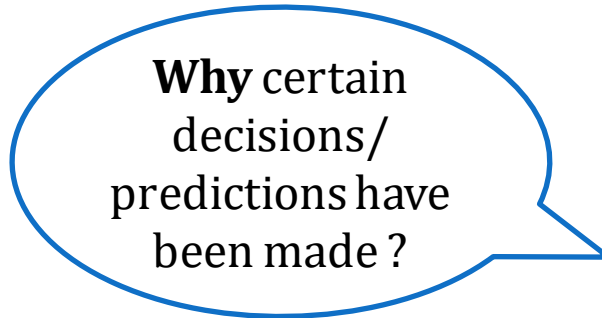


- In research communities, this area is also known as[6].
 - Explainable AI
 - XAI
 - Interpretable AI
 - Model Interpretability



INTRODUCTION

- What is **'Interpretable Machine Learning'** & **Why** we need it?
 - Providing explanations for algorithmic decisions
 - to improve the reliability of algorithmic decisions



BACKGROUND

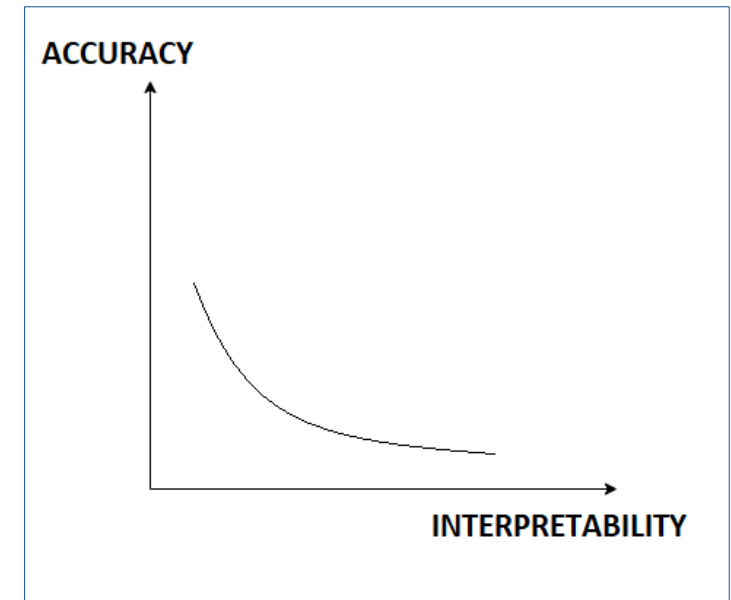
- There are two types of models can be identified in explainability
 - White-box
 - Black-box
- Black Box models perform well and produce better accuracies,
Ex: SVM, Deep Neural Nets, XGBoost, KNN etc
- Black-box models are high complex and the internal logic cannot understand by the users
- Concept of Explainable AI has been come into the picture to overcome that situation

BACKGROUND

- When the model complexity goes **up**, the accuracy also goes **up**
- When the complexity is **high** the interpretability is **low**

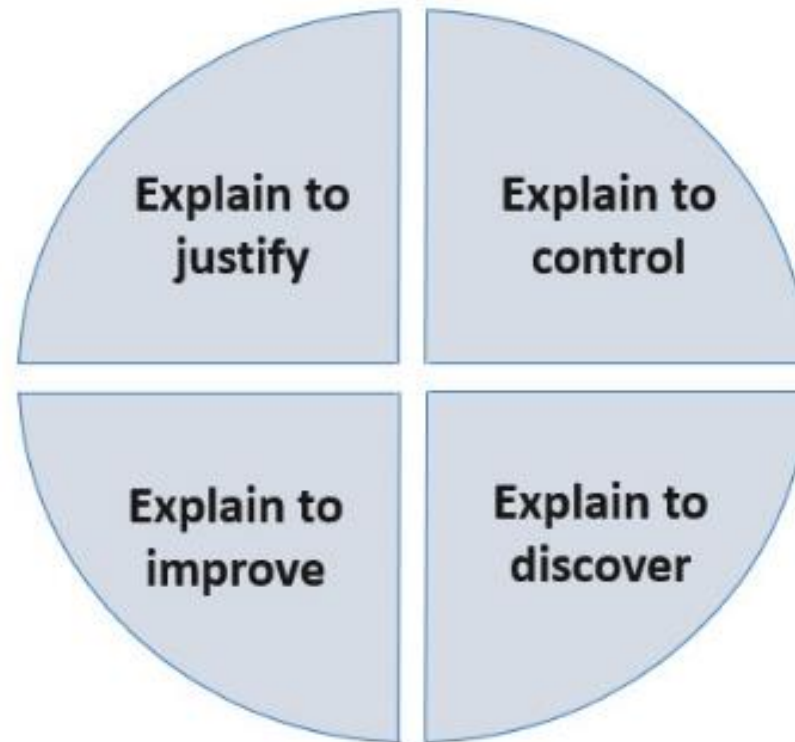
- That means,
 - At **high accuracy**,
the **model interpretability is low**

Accuracy vs Interpretability



BACKGROUND

- A. Adadi and M. Berrada shows four type of use cases for XAI in their survey (2018) [1]



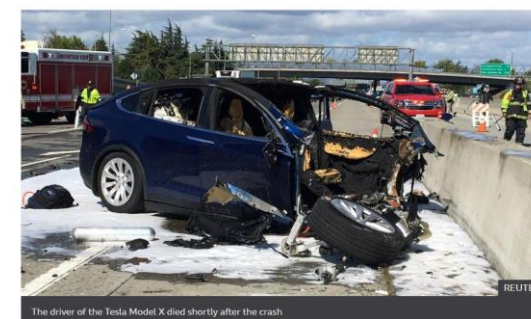
BACKGROUND

- Why it is that much important?

- Transportation
- Health
- Legal
- Finance
- Military

Tesla Autopilot crash driver 'was playing video game'

© 26 February 2020



The driver of the Tesla Model X died shortly after the crash

Self-driving Uber kills Arizona woman in first fatal crash involving pedestrian

Tempe police said car was in autonomous mode at the time of the crash and that the vehicle hit a woman who later died at a hospital



[bbc.com/news/technology-51645566](https://www.bbc.com/news/technology-51645566)

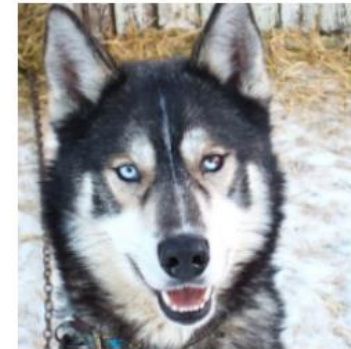
- According to the laws of many countries, machine generated decisions **should be able to provide justifications**

Ex: European Union's GDPR

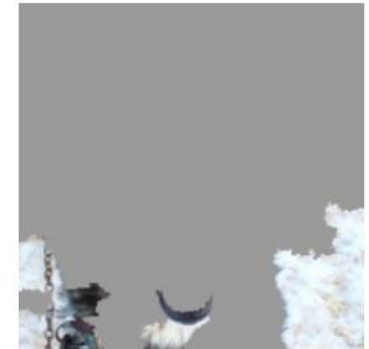
- The concept allows end users to identify the correctness of the results and the biasness

BACKGROUND

- Proposed methods,
 - **LIME**[8], **SHAP** [3], Local rule-based explanations, etc
- A popular scenario given by LIME,
 - A Husky has been classified as a wolf
 - Model considered only the snow background
 - Classifier has misclassified the image



(a) Husky classified as wolf



(b) Explanation



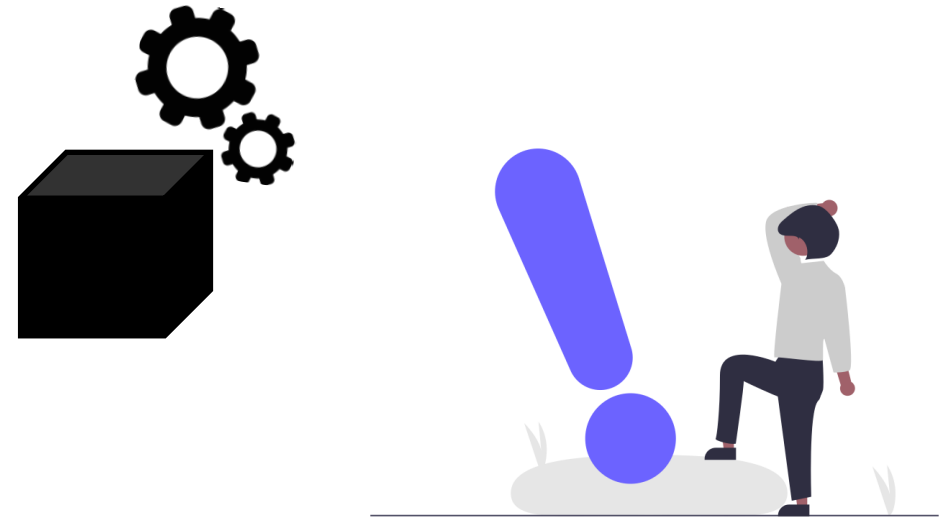
LIME



SHAP

RESEARCH PROBLEM

- Internal logic of the **Black-Box Models** cannot be understood by the users
- Existing methods are difficult to understand by the general users



Objectives

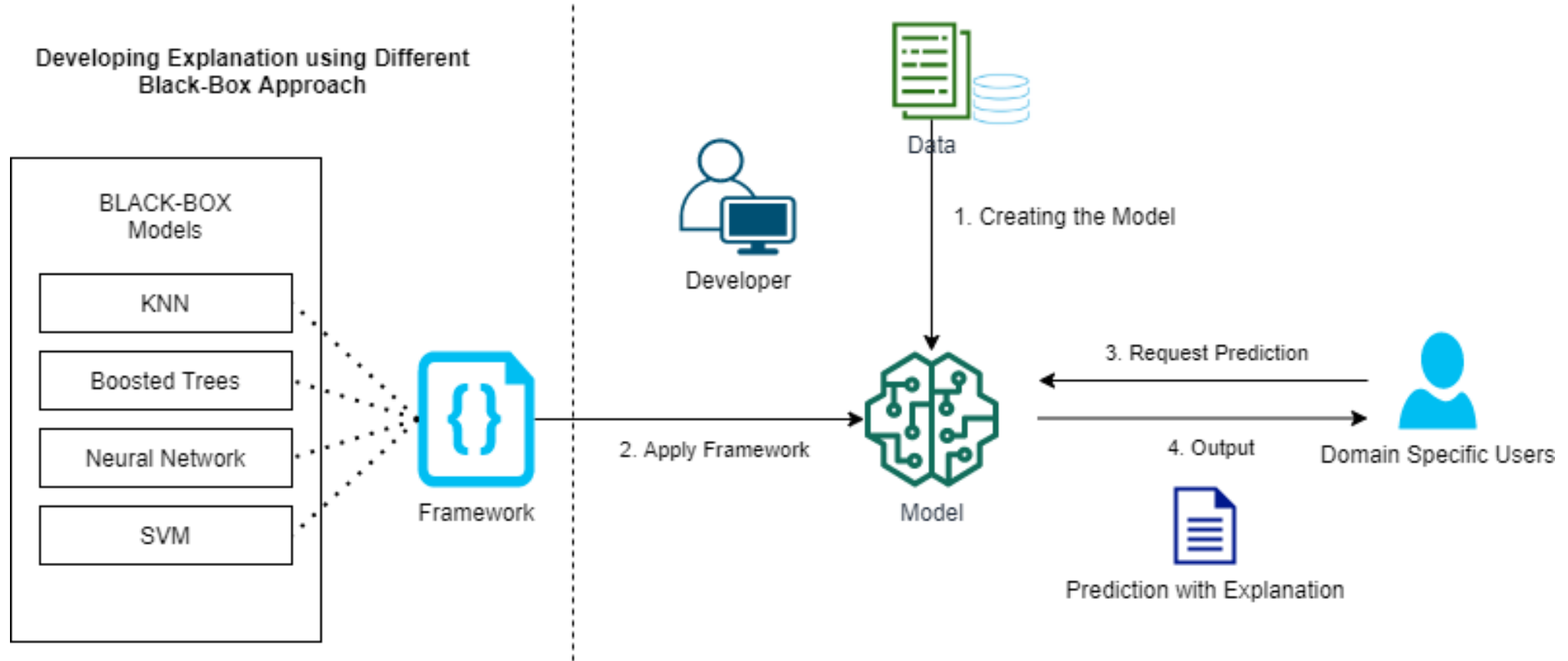
Specific Objective

- Implement a framework to enhance the interpretability of the black box models and provide more user-friendly explanations than existing methods

Sub Objectives

- Create a ML library to measure bias in neural networks using python
- Identify and mitigate the unintuitive feature attributions assign by TreeSHAP
- Enhance the model interpretability of SVM model and provide counterfactual explanations
- To Enhance the model interpretability of k-NN model overcoming 'Curse of Dimensionality' problem

SYSTEM DIAGRAM FOR OVERALL SYSTEM





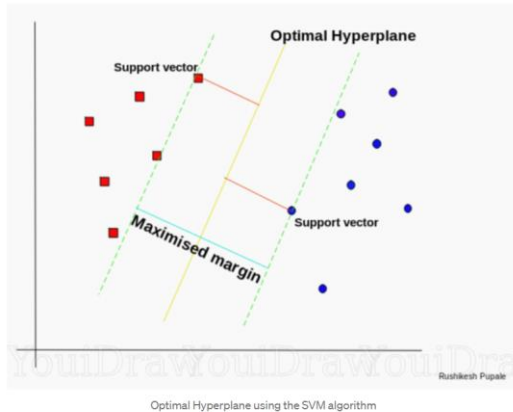
IT 18113600 | PERERA G.Y.N

Data Science

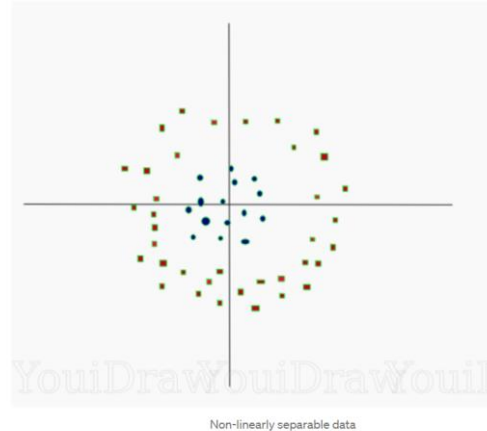
INTRODUCTION

BACKGROUND

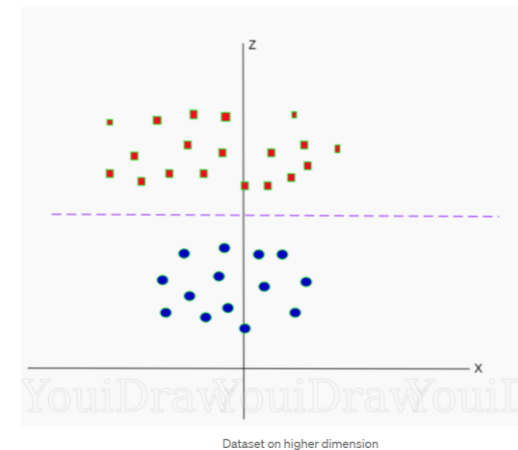
- What is Support Vector Machine(SVM)?
 - Support vector machine is a **supervised learning model** which works well in practical situations and provides better accuracies.
 - The model is mainly used for binary classification tasks
 - SVM performs well with both linear and non-linear data using “Kernels”.



Linear data



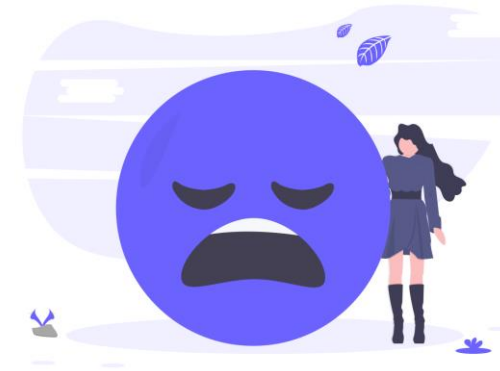
Non-linear data



- Real World usages of SVM: Bioinformatics, Face detections, Text classifications etc.

BACKGROUND cont..

- How is the interpretability of the model?
 - Unfortunately, it is a **Black-box** model[6].
- Do we need to care about it?
 - Model should have better interpretability, if the usage is very critical
 - It must explain the relationship between inputs and outputs



BACKGROUND cont..

- Are there any existing works done regarding the **explainability** of the model?
 - Local interpretable model agnostic explanations (LIME)[8]
 - Shapley Additive explanations (SHAP)[3]
 - Local Rule-based Explanation (LORE)[19, 20]
 - Anchors: High-Precision Model-Agnostic Explanations [21]
 - Model specific method by using SHAP-FOIL algorithm. SHAP-FOIL has the ability to learn the global behavior of SVM using logic programs [18]

BACKGROUND cont..

- What are the Counterfactual Explanations?
 - To flip the prediction, what are the changes that need be done to the model features.

Ex:

“you would have received the loan
if your income was higher by \$10, 000”

Feature: income

Counterfactual explanation: if income > \$10,000,
applicable for the loan



- Diverse Counterfactual Explanations (DiCE)[7] is a popular counterfactual framework

RESEARCH GAP



The existing local rule-based methods have not applied for **multivalued classification**



Support vector machine will be used as the model that need to be explained



In default SVM supports for binary classifications.



Counterfactual rules only consider the binary classes.

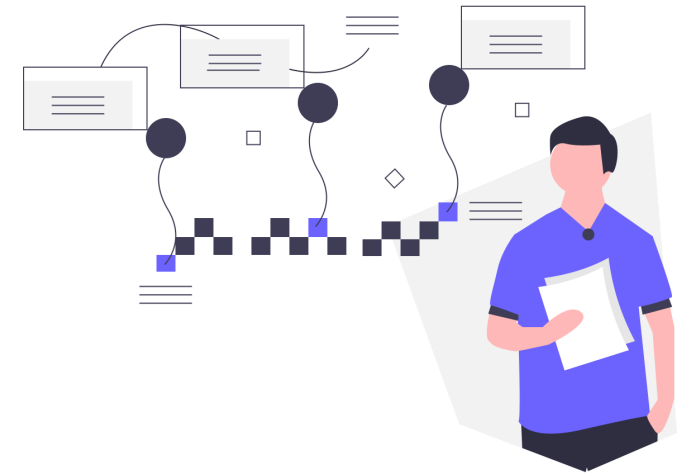
RESEARCH GAP cont..

	Existing Methods	Proposed Methods
Binary Classification	✓	✓
Multivalued Classification	X	✓
Counterfactual rule generation for binary predictors	✓	✓
Counterfactual rule generation for multivalued classifications	X	✓

- Existing methods do not have proper visualizations

RESEARCH PROBLEM

- Existing methods mainly focusing on **binary predictions**
- The approach has not been properly applied with multivalued classifications
- Counterfactual rule generations also have been done for binary predictions only
- Final outcomes are not properly visualized



OBJECTIVES

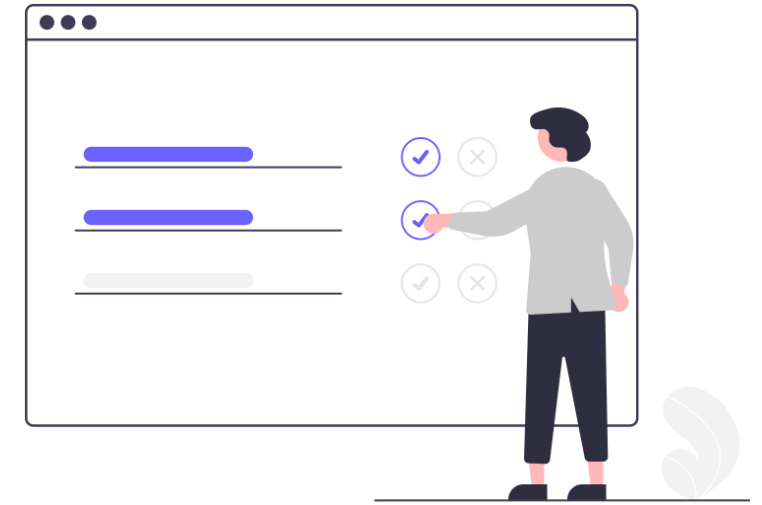
Specific Objective

- Enhance the model interpretability of SVM model and provide counterfactual explanations

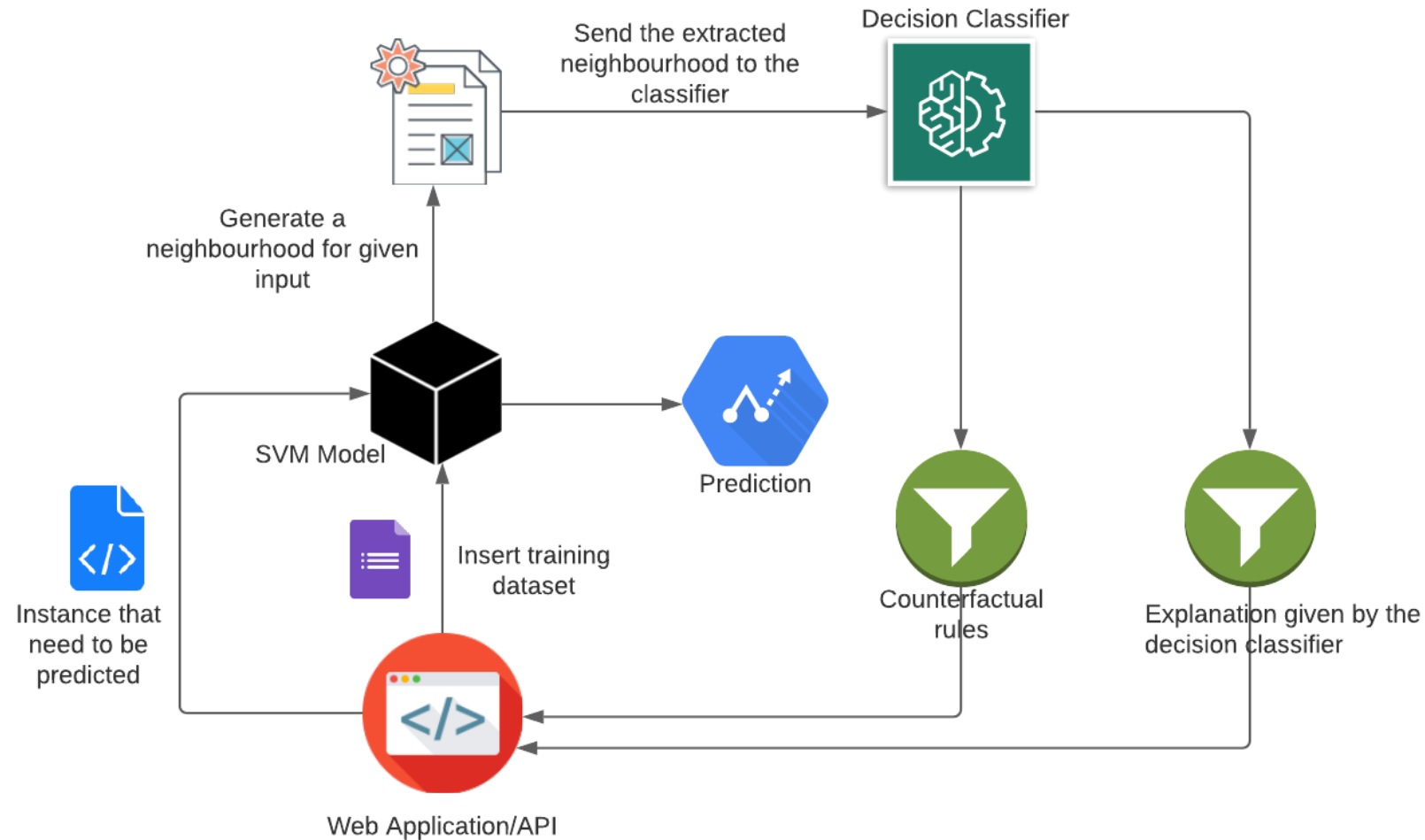
Sub Objectives

- Prepare the datasets
- Generate neighbourhood with training data
- Extract the explanation rule from decision tree classifier
- Extract counterfactual explanations from decision tree classifier

RESEARCH METHODOLOGY



SYSTEM OVERVIEW DIAGRAM



TECHNOLOGIES, TECHNIQUES AND ALGORITHMS

Technologies

- Development: Python
- VCS: Gitlab

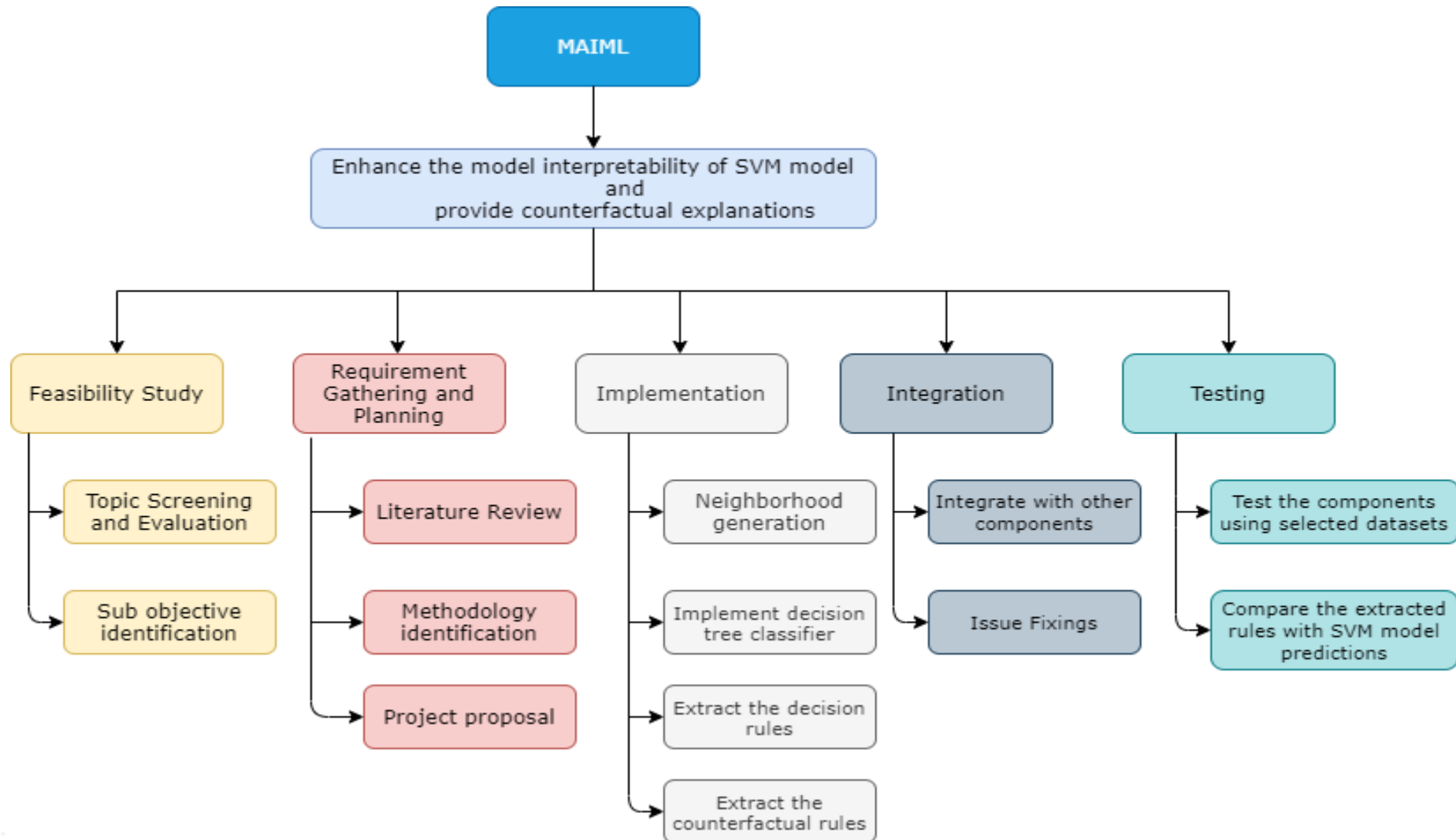
Techniques

- Genetic Algorithms, Distance functions for neighbourhood generations
- Vectorization techniques

Algorithms

- SVM
- Decision tree classifier

WORK BREAKDOWN STRUCTURE





IT 18134704 | J.K. KENNETH CHAMARA

Data Science

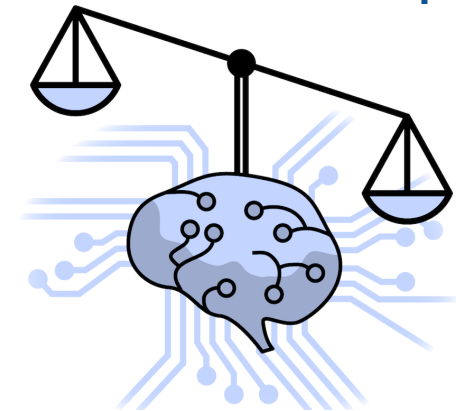
INTRODUCTION

BACKGROUND

- What is fairness In ML ?

“...discrimination is considered to be present if for two individuals that have the same characteristic relevant to the decision making and differ only in the sensitive attribute (e.g., gender/race) a model results in different decisions”

ToonCalders and Indrė Žliobaitė. 2013 [1].



BACKGROUND cont..



Why we need fairness in ML?

Improve reliability
Avoid ethical concerns
lead to more accurate decisions



Why ML models don't have fairness?

ML algorithms collect unwanted bias in training data.

BACKGROUND cont..

- How is this relevant to XAI ?
 - Implementing fair ML models is a main Goal in XAI.
- Explainability and Fairness
 - The Black box model Discrimination
 - Model Explainability
 - Fairness Metrics
 - Bias Mitigation Algorithms

BACKGROUND cont..

- **Measuring Fairness**

- Pre- Processing

- Learning Fair Representation (LFR) [2]

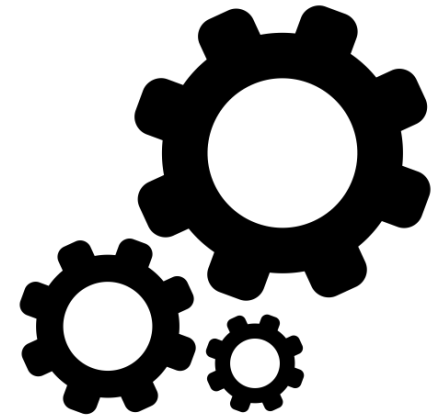
- Post- Processing

- Equality of Odd (EQO) [3]
 - Reject Option Based Classification (ORBC) [4]

- **Avoiding Bias**

- In – Processing [5]

- add a constraint or a regularization term

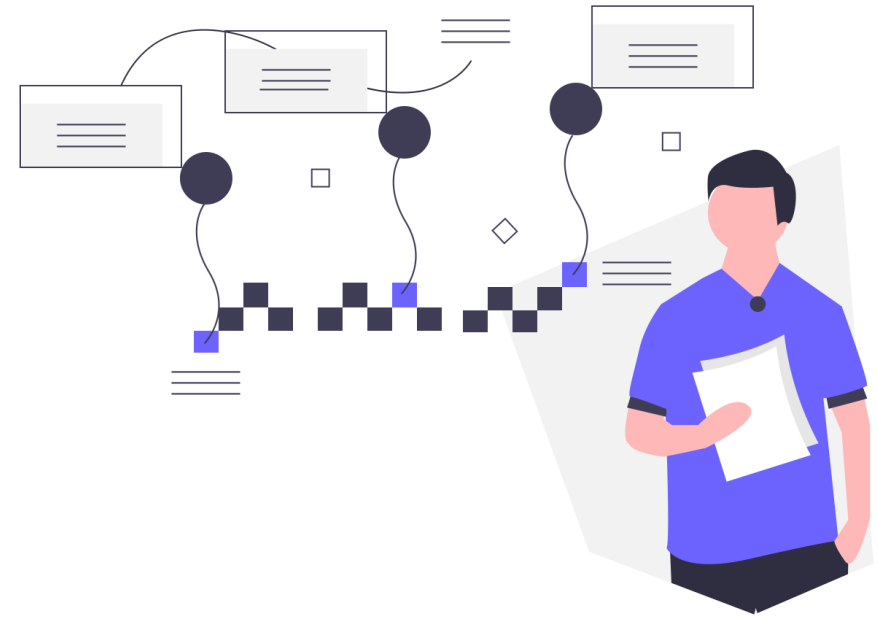


RESEARCH GAP

	Measuring Bias	Avoiding Bias
Pre-processing	Yes	Yes
In-processing	No	Yes
Post-processing	Yes	No

RESEARCH PROBLEM

- How to measure the **Bias** in a **Neural Network** in **Processing Time** ?



OBJECTIVES

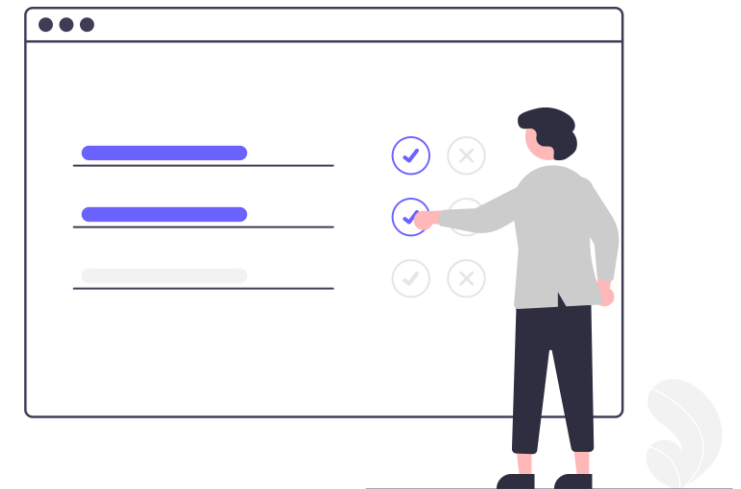
Specific Objective

- Create a ML library to measure bias in neural networks using python

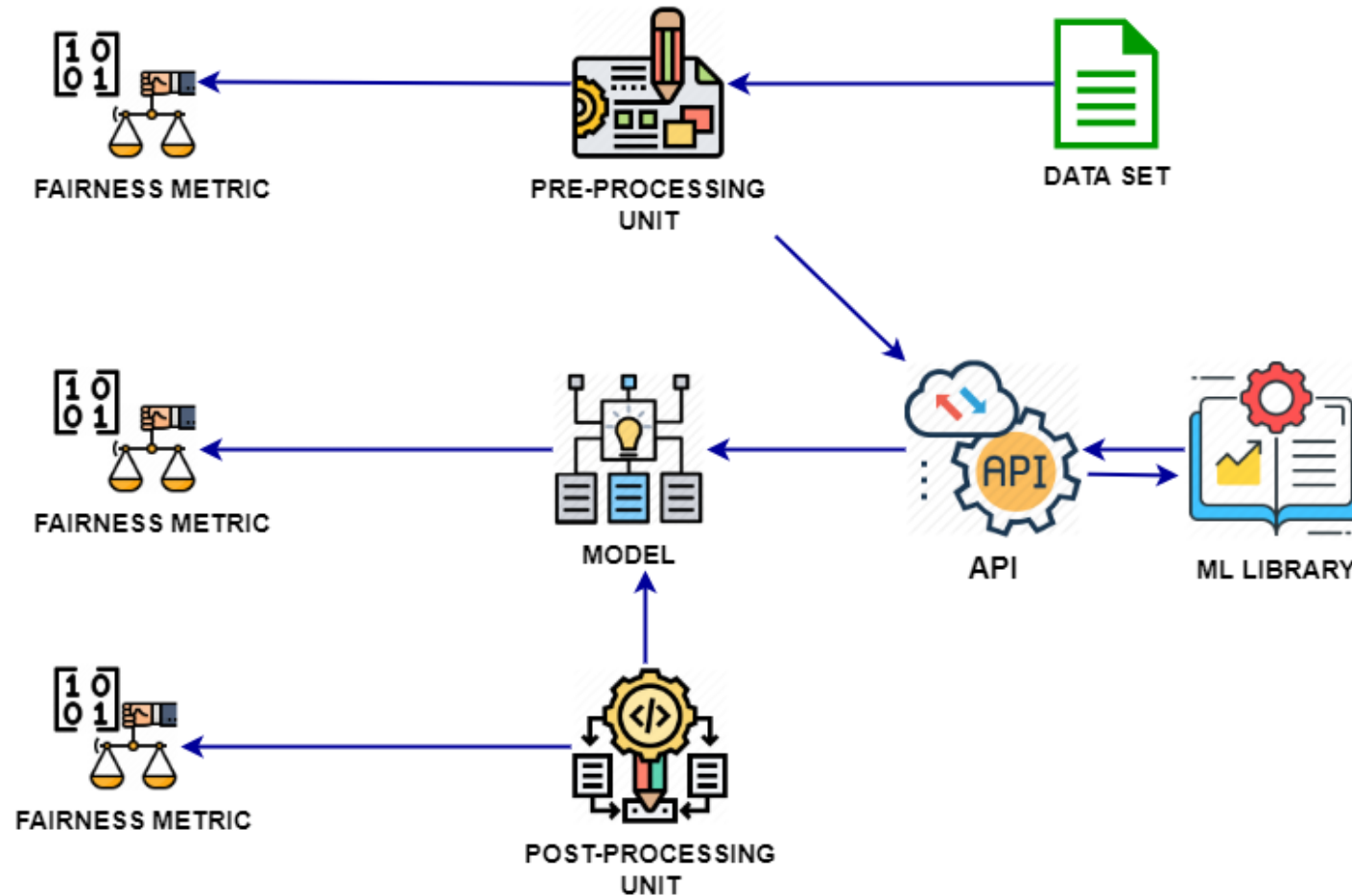
Sub Objectives

- Implementing and testing a neural **network** library
- creating a methodology to measure bias in processing time
- implement the methodology in the library

RESEARCH METHODOLOGY



SYSTEM OVERVIEW DIAGRAM



TECHNOLOGIES, TECHNIQUES & ALGORITHMS

Technologies

- Development : Python

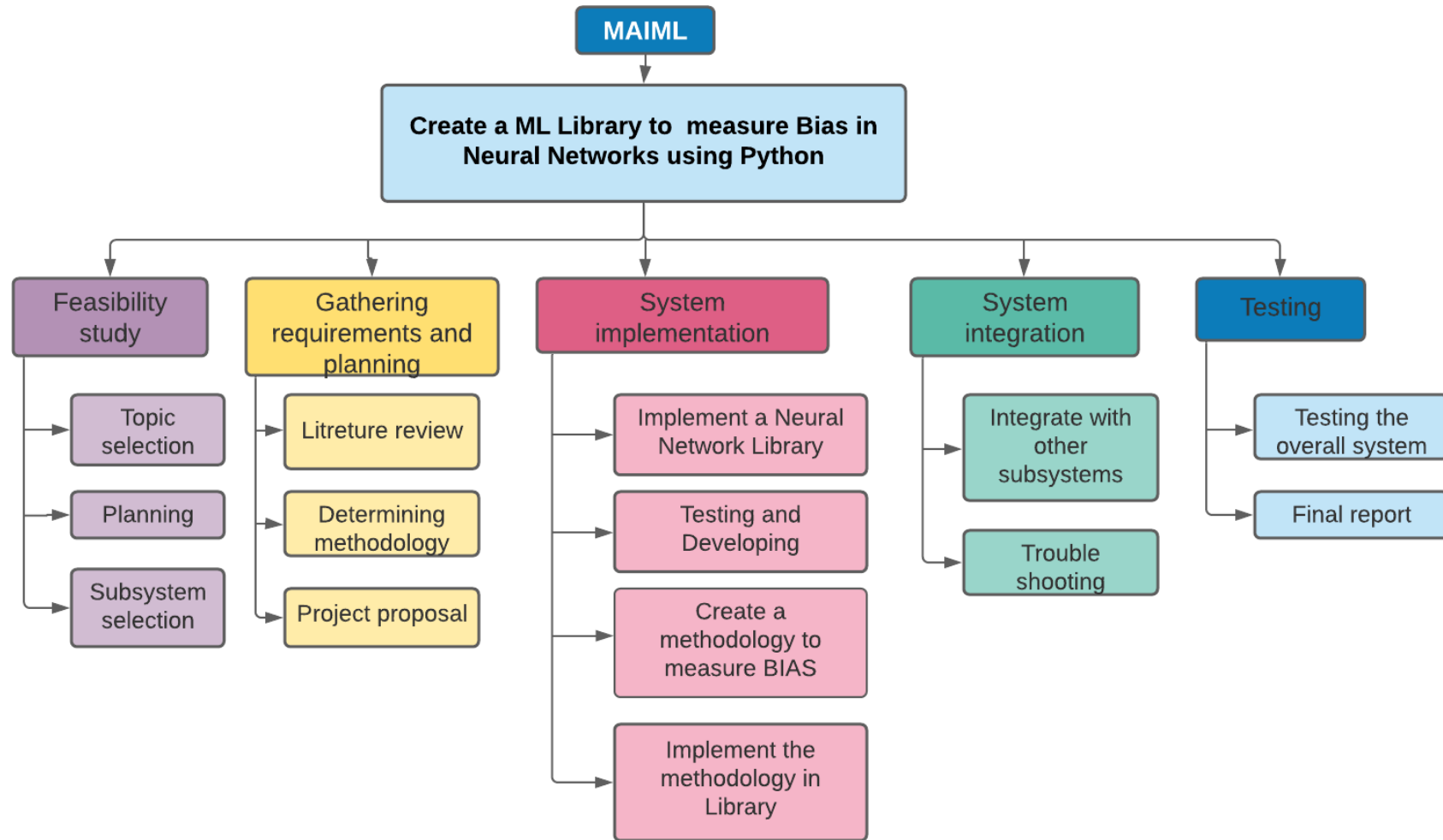
Techniques

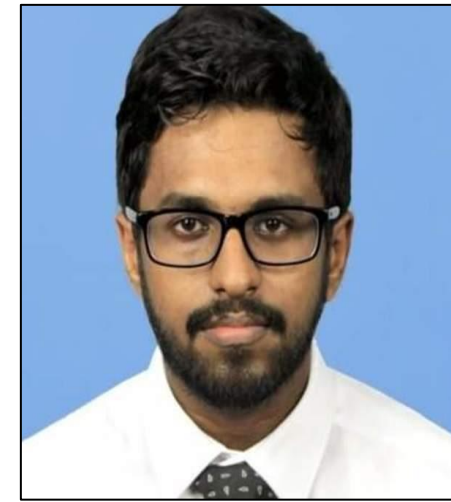
- Probability and Statistics
 - Inferential Statistics
 - Game theory
- Vectorization

Algorithms

- Deep neural network algorithm for binary classification

WORK BREAKDOWN STRUCTURE

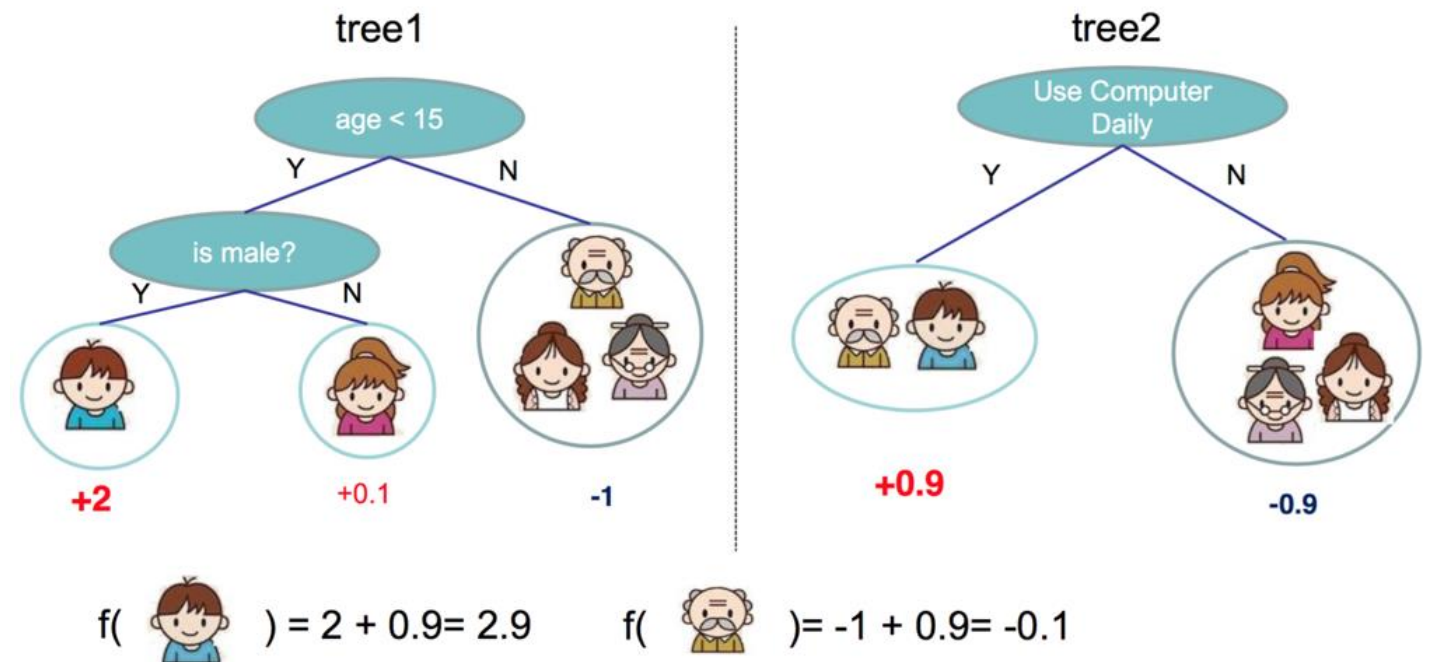




IT18059564 | ABEYAGUNASEKERA S.H.P

Data Science

INTRODUCTION



BACKGROUND



Single decision tree is interpretable and cannot handle complex decision-making processes.



Ensemble tree-based models are built up with many trees.



Several Ensemble tree algorithms

Random forest models produce final decision by aggregating the individual tree results.[7]
Xgboost use gradient based approach to arrive the decision.[8]



They differ from each other by the way they arrived the final decision.



Though the ensemble models increase the accuracy it hinders the interpretability.



Tree based ensemble models are the most used method in dealing with tabular data.

BACKGROUND cont.

In coalitional/cooperative game theory Shapley values used to distribute the payoff among participants fairly.



Why is it fair? [9]

Axiom 1 (symmetry): Interchangeable agents should receive the same payment.

Axiom 2 (Dummy player): Dummy players should receive nothing

Axiom 3 (Additivity): Ability to decompose the payment



Tree explainer uses Shapley values to explain the predictions given by tree ensemble models. [10]

BACKGROUND cont.

- How to calculate it?

$$\phi_i(v) = \sum_{S \subseteq N \setminus i} \underbrace{\frac{|S|!(N - |S| - 1)!}{N!}}_{\text{Weight}} \underbrace{(v(S \cup i) - v(S))}_{\text{Marginal contribution}}$$



- **Weight:** Number of arrangements which we able to calculate the given marginal contribution of a given instance divided by total possible arrangements.
- **Marginal Contribution:** Value added by that instance when it added with the other instances.

BACKGROUND cont.

- Let's think about a cooperative game where the yield of player X1 is $Val(X1)$ alone is = 1 and the player X2 which is $Val(X2) = 2$ if the combined yield is $Val(X1, X2) = 4$ then the distribution of yield for X1 according to the cooperative game theory we can be shown as.

		Value added by X1
{X1}	{X1, X2}	$Val(\{X1\}) = 1$
{X2}	{X2, X1}	$Val(\{X2, X1\}) - Val(\{X2\}) = 4 - 2 = 2$

- Here $N!$ no of combinations = 2
- Shapley value for first instance = $1/2 * 1 = 0.5$
- Shapley value for second instance = $1/2 * 2 = 1$
- Therefore, the payout for X1 is $1 + 0.5 = 1.5$ and X2 is $4 - 1.5 = 2.5$ (axiom 3)
- Are Shapley values same as ratios ? No
 - $4/3$ not as same as 1.5

RESEARCH GAP

Achievements \ Research papers	[9]	[10]	[15] (2020)	[16] (2020)	[17] (2020)
Calculation of Shapley values efficiently	✓	X	X	X	X
The interaction effect is the additional combined feature effect after accounting for the individual feature effects	X	✓	X	X	X
It is possible to create intentionally misleading interpretations with SHAP, which can hide biases	X	X	✓	X	X
Use tree-based model in health care sector	X	X	X	✓	X
"TreeSHAP can produce unintuitive feature attributions"	X	X	X	X	X

RESEARCH PROBLEM

- "TreeSHAP can produce unintuitive feature attributions".[11] (2019)
- "Features that have no influence on the prediction can get a TreeSHAP value different from zero."[11]
- Experimental evidence.
- Do customer ID really effect for the customer churning?
- <https://colab.research.google.com/drive/1b9CrnAzmo9SzKToOY0DKZAkQlteTArt0?usp=sharing>



OBJECTIVES

Specific Objective

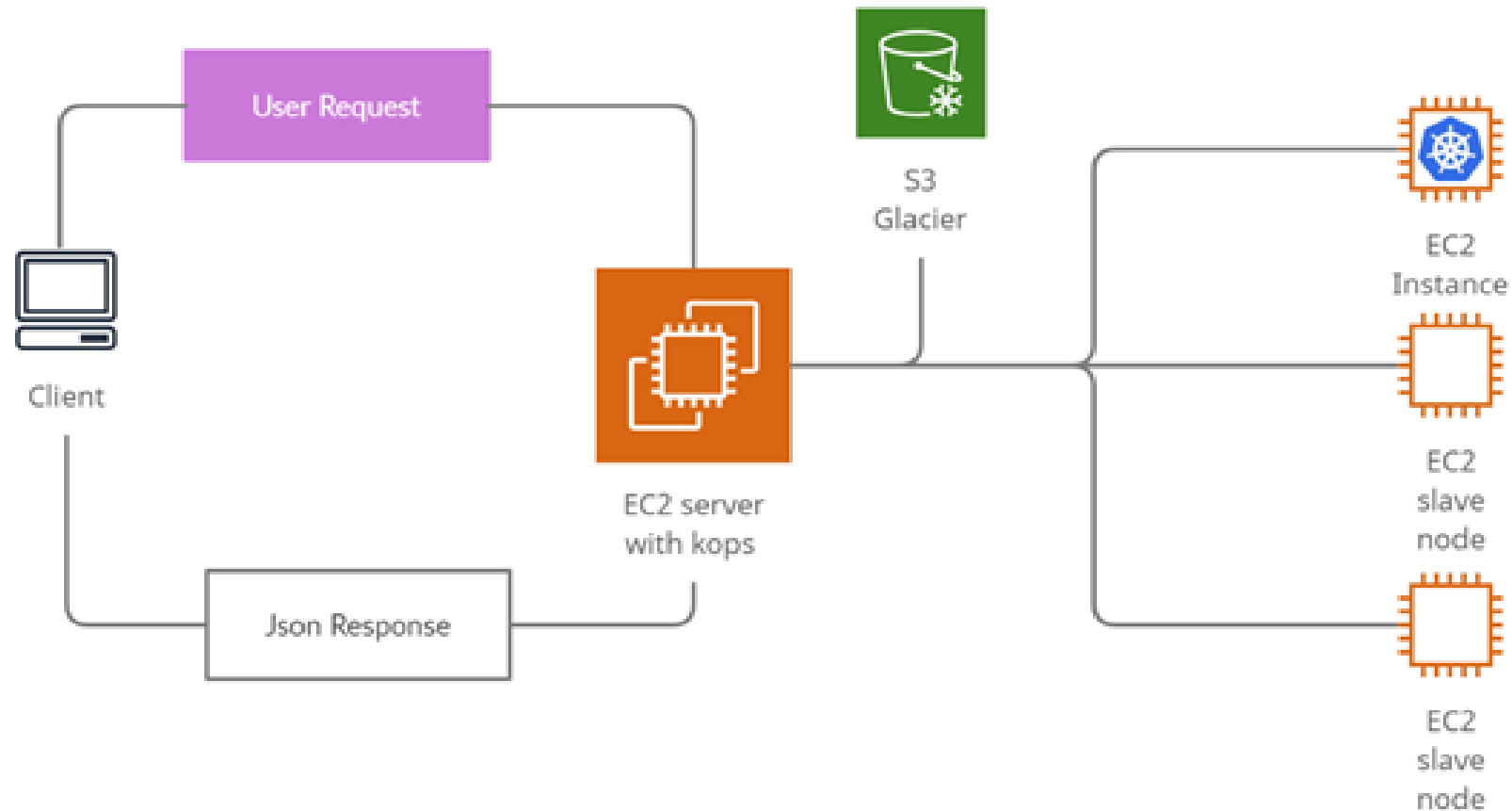
- Identify and mitigate the unintuitive feature attributions assign by TreeSHAP.

Sub Objectives

- Increase the interpretability of ensemble tree models by using multiple model agnostic interpretable methods.[12]
- A web-based solution with and algorithm to minimize or mitigate the limitation of TreeShap.
- Expect to use other model agnostic techniques like counterfactual explanations [13] and LIME [14] to increase the explainability of the decision.
- Unified approach of existing model agnostic methods to increase the explainability of the ensemble tree models.

RESEARCH METHODOLOGY

SYSTEM OVERVIEW DIAGRAM



TECHNOLOGY, TECHNIQUES & ALGORITHMS

Technology

- Development:
 - Python
- Web Technologies
 - Rest APIs using Flask
- Ensure Availability
 - Kubernetes
- Front End:
 - D3.js for graphs

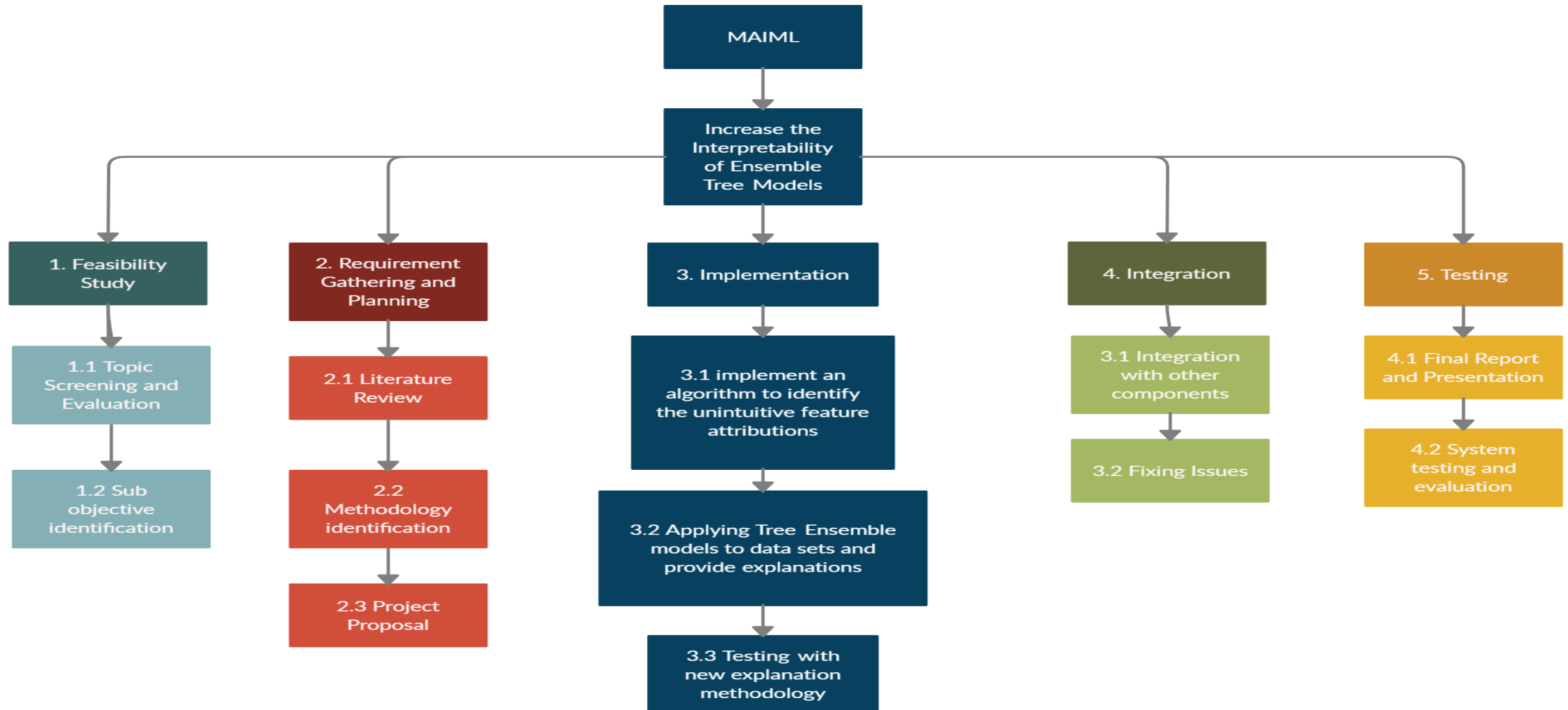
Techniques

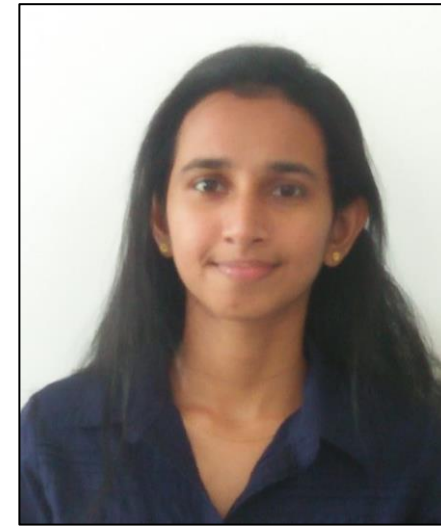
- Probability and Statistical techniques
- Game theoretical techniques
 - Shapley values-based CORE concept in collaborative game theory.

Algorithms

- Tree SHAP (Shapley Additive Explanations) algorithm.[10]

WORK BREAKDOWN STRUCTURE





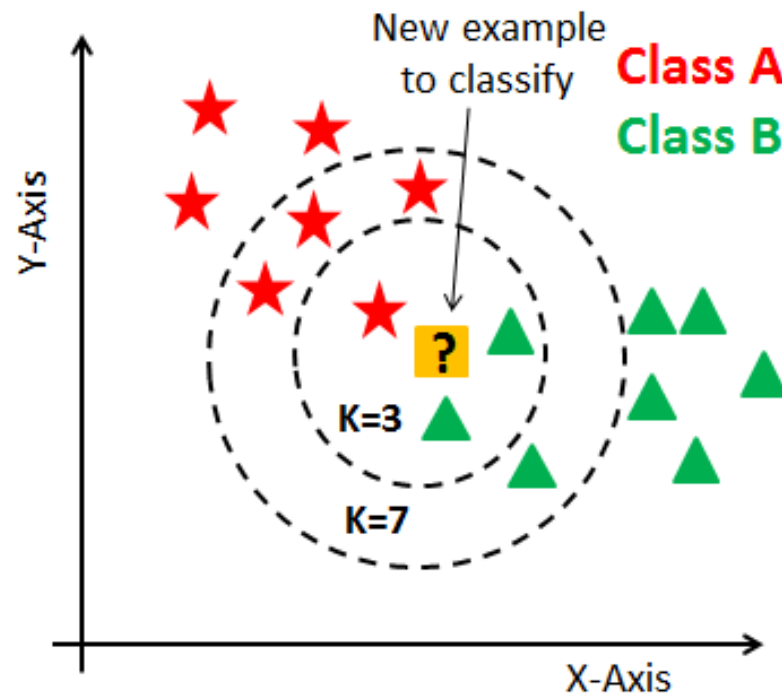
IT 18056976 | G.L. UDARI KAUSHALYA

Data Science

INTRODUCTION

BACKGROUND

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



Supervised Learning

Non-Parametric

Lazy Learning

Classification Problems

Regression Problems

BACKGROUND Cont.

How k-NN Works ?

- Most Common class

Two Tricky Parts ...

- K value
- Distance

BACKGROUND Cont.



Do we need **Interpretability** for **k-NN Algorithm** ?

YES, of Course!



Why ?

When it applies to **Sensitive** areas...

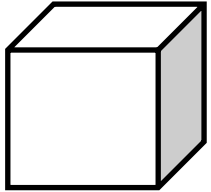
What happens if the decision becomes **wrong**?

How can we **trust** the decision ?





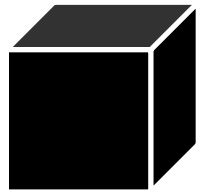
That is why we need this...

BACKGROUND Cont.



Is K-NN **Whitebox** ?

- Non-Parametric  No Interpretability on a Modular Level
- Interpretability depends on  Interpretability of Single Instances of Data Set
- If an Instance Consists of **Hundreds of Features [19] ???**



BACKGROUND Cont.

- What is **CURSE OF DIMENSIONALITY ?**
 - No of dimensions increases↑, the volume of the input space increases↑ at an exponential rate.
- In such instances, points that **appear to be similar** may be separated by **a large distance**.
- All points will be **far away from each other**

BACKGROUND Cont.

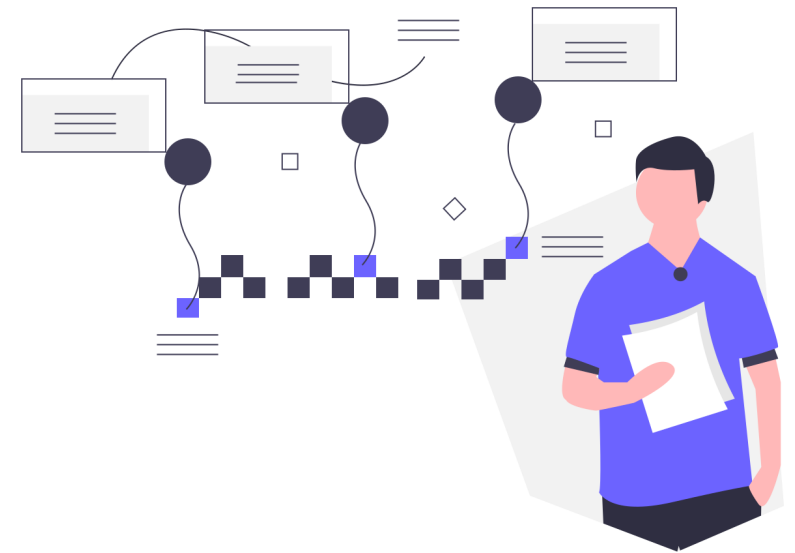
- This might feel unintuitive at first
 - Since this becomes **BLACK-BOX** when it has more dimensions
 - **Dimension Reduction** is applicable
- There are several ways of **Attribute Reduction** for Interpretable k-NN model...

RESEARCH GAP

PREVIOUS WORKS	SWKNN An Improved KNN algorithm Based on Kernel Methods and Attribute Reduction [20]	Classification of Functional Data Interpretable Ensemble Approaches [21]	The Effects of Dimensionality Curse in High Dimensional k-NN Search [22]	iDStar [23]	<i>K</i> important neighbors (KIN) [24]	Explaining and Improving Model Behavior with k Nearest Neighbor Representations [25]	Proposed Framework
Dimension Reduction	✓	✓	✓	✓	✓	✗	✓
Give an Explanation	✗	✗	✗	✗	✗	✓	✓
Effect Of K Value/ address K value defining	✓	✗	✓	✗	✗	✗	✓

RESEARCH PROBLEM

- How to get an **Explanation** for **k-Nearest Neighbors** model outcome, **overcoming** the '**Curse of Dimensionality**' ?



OBJECTIVES

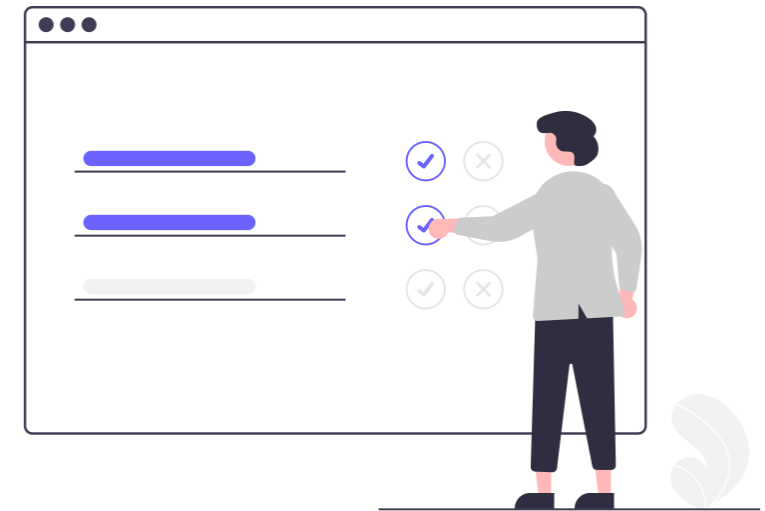
Specific Objective

- To Enhance the model interpretability of k-NN model overcoming 'Curse of Dimensionality' problem

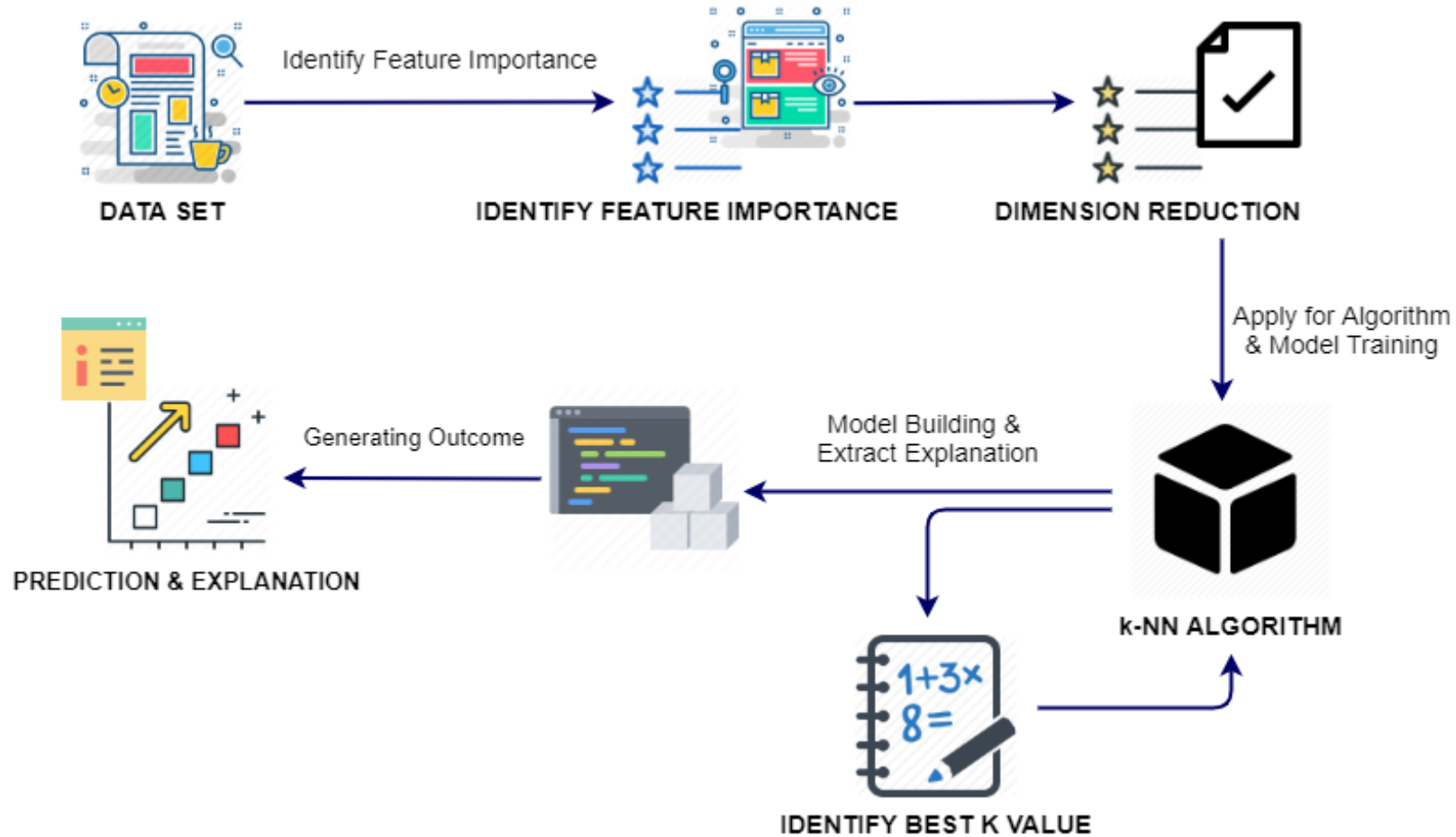
Sub Objectives

- Identify feature importance
- Identify and apply ensemble methodologies for **dimension reduction and get explanation**
- Dimension Reducing to get rid of **Curse of Dimensionality**
- **Extract Explanation** for k-NN using developed methodologies
- Apply for Datasets

RESEARCH METHODOLOGY



SYSTEM OVERVIEW DIAGRAM



TECHNOLOGIES, TECHNIQUES & ALGORITHMS

Technologies

- Development : Python

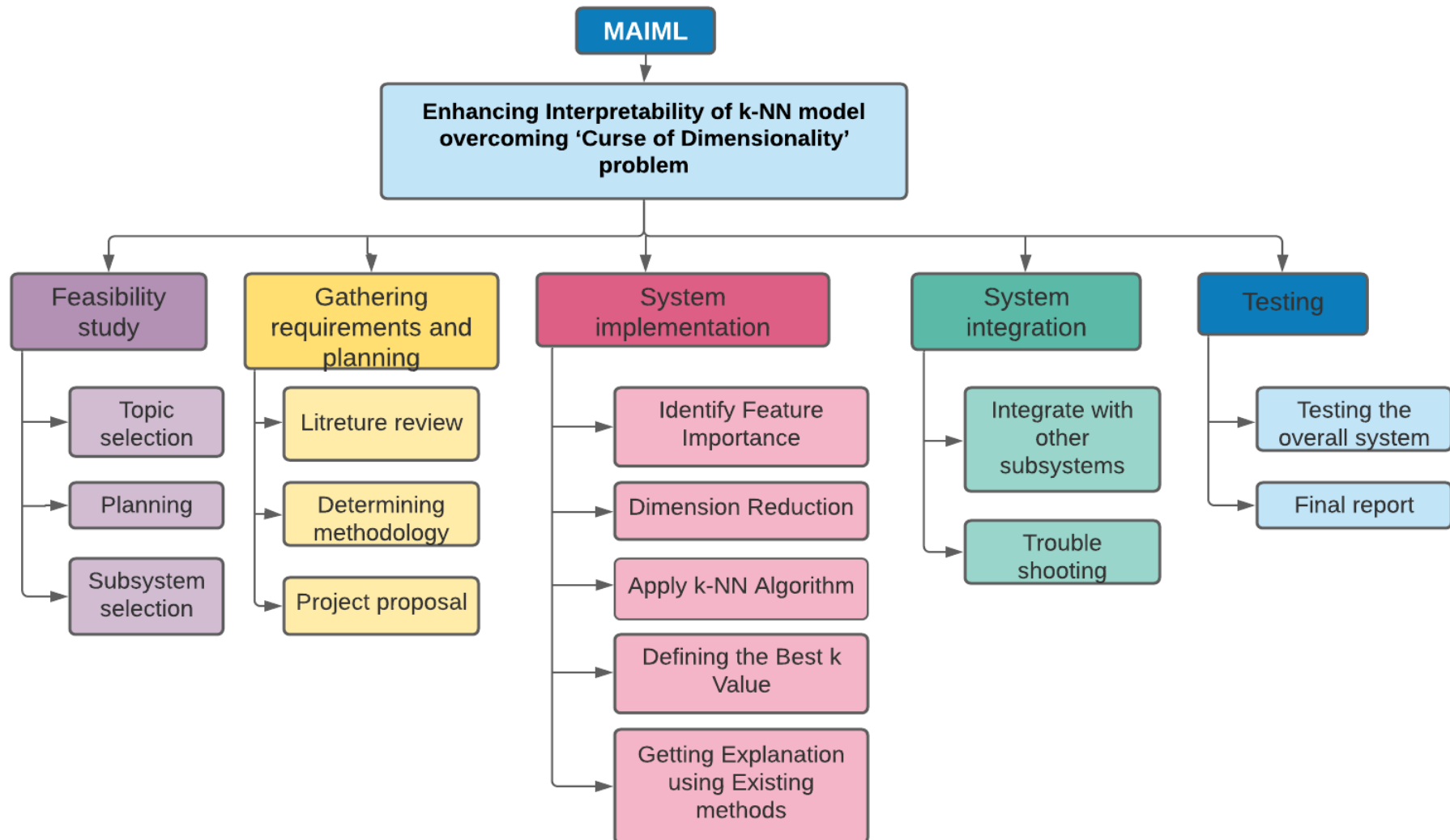
Techniques

- Dimension Reduction
- Sensitivity Analysis

Algorithms

- K-Nearest Neighbour basic and improved algorithms

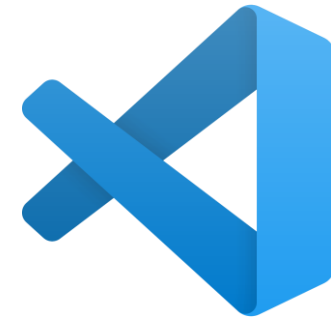
WORK BREAKDOWN STRUCTURE



SUPPORTIVE INFORMATION

SYSTEM, PERSONNEL & SOFTWARE SPECIFICATION REQUIREMENTS

- Software Specification Requirements
 - PyCharm
 - Visual Studio Code
 - GitLab



GitLab

COMMERCIALIZATION

- Modern Trend of AI in different domains
- Task Automation
- GDPR is not giving access
- Problem of Reliability, Transparency
- Explainable tools give an answer this Q...
- This may give a better impact for this...
- May get a huge demand



Budget (if any)

- Web Hosting
- Internet Costs

REFERENCES

- [1] ToonCaldersandIndr'eŽliobait'e.2013. Why Unbiased Computational Processes CanLeadtoDiscriminativeDecisionProcedures.SpringerBerlinHeidelberg,Berlin, Heidelberg,43–57. https://doi.org/10.1007/978-3-642-30487-3_3
- [2] R. Zemel, Y. Wu, K. Swersky, T. Pitassi and C. Dwork, "Learning fair representations", *International Conference on Machine Learning*, pp. 325-333, 2013.
- [3] M. Hardt, E. Price and N. Srebro, "Equality of Opportunity in Supervised Learning", *Advances in Neural Information Processing Systems*, pp. 3315-3323, 2016.
- [4] F. Kamiran, A. Karim and X. Zhang, "Decision theory for discrimination-aware classification", *IEEE 12th International Conference on Data Mining*, pp. 924-929, 2012.
- [5] M. Joseph, M. Kearns, J. Morgenstern, S. Neel and A. Roth, "Rawlsian Fairness for Machine Learning", *arXiv Prepr. arXiv1610.09559*, vol. 1, no. 2, pp. 1-26, 2016.
- [6] A. Stevens, P. Deruyck, Z. V. Veldhoven and J. Vanthienen, "Explainability and Fairness in Machine Learning: Improve Fair End-to-end Lending for Kiva," *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, Canberra, ACT, Australia, 2020, pp. 1241-1248, doi: 10.1109/SSCI47803.2020.9308371.
- [7] Leo Breiman "Random Forest" 2001
- [8] Tianqui Chen, Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". 2016
- [9] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in Neural Information Processing Systems*. 2017
- [10] Lundberg, Scott M., Gabriel G. Erion, and Su-In Lee. "Consistent individualized feature attribution for tree ensembles." *arXiv preprint arXiv:1802.03888* (2018)
- [11] Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. <https://christophm.github.io/interpretable-ml-book/>.

REFERENCES Cont.

- [12] 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.
- [13] Mothilal, Ramaravind K., Amit Sharma, and Chenhao Tan. "Explaining machine learning classifiers through diverse counterfactual explanations." *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020
- [14] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should I trust you?: Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM (2016)
- [15] Slack, Dylan, et al. "Fooling lime and shap: Adversarial attacks on post hoc explanation methods." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020
- [16] Athanasiou, Konstantina Sfrintzeri, Konstantia Zarkogianni, Anastasia C. Thanopoulou, Konstantina S. Nikita "An explainable XGBoost-based approach towards assessing the risk of cardiovascular disease in patients with Type 2 Diabetes Mellitus Maria" [arXiv:2009.06629v2](https://arxiv.org/abs/2009.06629v2). 2020
- [17] Nazanin Fouladgar, Kary Framling "XAI-P-T: A Brief Review of Explainable Artificial Intelligence from Practice to Theory". 2020
- [18] 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.
- [19] 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.
- [20] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, and F. Turini, "Factual and Counterfactual Explanations for Black Box Decision Making," *IEEE Intell. Syst.*, vol. 34, no. 6, pp. 14–23, 2019, doi: 10.1109/MIS.2019.2957223.
- [21] M. T. Ribeiro and C. Guestrin, "Anchors : High-Precision Model-Agnostic Explanations," pp. 1527–1535.

REFERENCES Cont.

- [22] Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019.
<https://christophm.github.io/interpretable-ml-book/>.
- [23] W. Xueli, J. Zhiyong and Y. Dahai, "An Improved KNN Algorithm Based on Kernel Methods and Attribute Reduction," 2015 Fifth International Conference on Instrumentation and Measurement, Computer, Communication and Control (IMCCC), Qinhuangdao, China, 2015, pp. 567-570, doi: 10.1109/IMCCC.2015.125.
- [24] Maierhofer, Thomas (2017). Classification of Functional Data. Interpretable Ensemble Approaches. *Master Thesis, Faculty of Mathematics, Computer Science and Statistics, Ludwig-Maximilians-Universität München*.
- [25] Kouroukidis, N., & Evangelidis, G. (2011). The Effects of Dimensionality Curse in High Dimensional kNN Search. *2011 15th Panhellenic Conference on Informatics*, 41-45.
- [26] Schuh, M.A., Wylie, T., & Angryk, R. (2014). Mitigating the Curse of Dimensionality for Exact kNN Retrieval. *FLAIRS Conference*.
- [27] Shahraki, H.R., Pourahmad, S., & Zare, N. (2017). K Important Neighbors: A Novel Approach to Binary Classification in High Dimensional Data. *BioMed Research International*, 2017.
- [28] Nazneen Fatema Rajani, Ben Krause, Wengpeng Yin, Tong Niu, Richard Socher & Caiming Xiong. (2020). Explaining and Improving Model Behavior with k Nearest Neighbor Representations.

THANK YOU!