Counterfactual Explanations for SVM, RF, LR and KNN

TMP-23-142

Our Team

IT20097660
Warnasooriya S.D

IT20100698
Britto T.A

IT18161298
Srinidee Methmal H.M

IT20013950
Lakshani N.V.M

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
- Logistic Regression

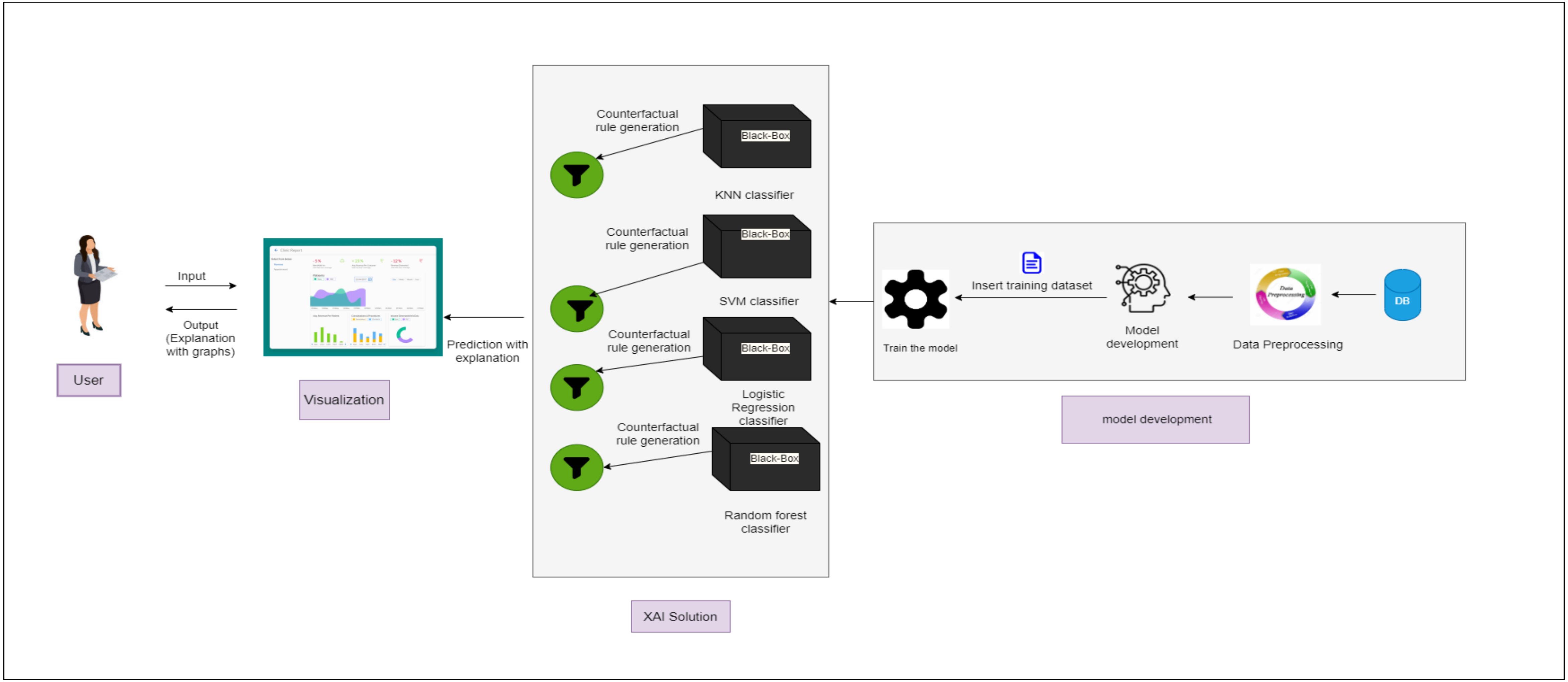
by developing a novel counterfactual rule generation mechanism related to the text classification domain.

Sub Objectives:

To develop novel explainable method to enhance the model explainability of :

- **function-based classification models focus on SVM .**
- **ensemble models focus on Random forest .**
- **distance-based classification models focus on KNN .**
- **regression-based classification models focus on Logistic regression .**

Overall System Diagram



IT20097660 | WARNASOORIYA S.D

Specializing in Data Science

Support Vector machine (SVM).

Research Gap

What are the existing methods that used for generating counterfactual rules related to Support Vector Machine?

SHAP[6]
- Not Primarily Designed for Counterfactuals
- Assumption of Independence
- Computational Complexity
- model-agnostic Explanation

LIME[7]
- Not Primarily Designed for Counterfactuals
- explanations can be unstable
- model-agnostic Explanation

Diverse Counterfactual Explanation(DICE)[6]
- Assumption of Feature Independence
- Computationally Intensive
- Model agnostic explanation

Nearest Instance Counterfactual Explanations[NICE][9]
- aims to find the smallest and most meaningful changes to an instance that would alter the model's prediction
- Finding the nearest instance and ensuring feasibility can be computationally intensive.
- Model agnostic explanation
- Limited to binary classification problems.

Research Gap

Anchors: High-Precision Model-Agnostic Explanations [1]

- Not provide a counterfactual explanation
- Model agnostic explanation

Local rule-based explanations of black box decision systems[LORE][2]

- Can not be used to text classification explainable task.
- Not provide a counterfactual explanation
- Model agnostic explanation.

SHAP-FOIL Algorithm[3]

- Can not be used to text classification explainable task
- Not provide a counterfactual explanation.

Proposed Method:

- **Text classification explainable task.**
- **Provide a neighborhood based Counterfactual Rule Generation Explainable Method.**
- **Model Specific Approach.**

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 explainability 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.

Counterfactual Explanation for Support Vector Machine using mirror point encountering with cosine-similarity comparison in kernel space.

Step 1: Generate contradictory prompts for the given prompt using a finetuned T5 model/custom WordFlippingGenerator.

Step 2: Vectorize all the prompts using the TFIDF vectorizer into the vector space .

Step 3: Project all the vectors into the SVM's kernel space .

Step 4: Find the mirror point $(\boldsymbol{C})$ of the given prompt's TFIDF vector on the hyperplane of the SVM.

Step 5: As the final step, algorithm find the closest point to mirror point $(\boldsymbol{C})$ and retrieve the most accurate contradictionary prompt as the output. . Most accurate contradictonary prompt return using the cosine-similarity between the mirror point and the contradictonary prompts.

Word Flipping Generator

[Randomly flips words with defined POS tags to their antonyms]

1. User should define the POS tags relevant to the words that must be flipped.

2. The user should invoke the functionality by specifying an original sentence and the number of variations they want

3. The algorithm will tokenize the words

4. The algorithm will generate a mask list corresponding to these tokens by referring to the POS tags previously defined by the user. A truth value in this mask will represent that the word must be flipped and false will mean otherwise

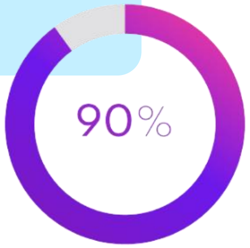
5. The algorithm will generate a set of lists of antonyms for the words to be flipped by referring to the mask above. These lists will have words ordered in the descending order of the probabilities of their occurrence

6. New sentences will be generated by merging the original words and antonyms appropriately throughout the implementation.

Completion and Future works

Completed Components

- ✓ Data preprocessed and built the SVM Model
- ✓ Find the novel methodology for generating counterfactual rule using cosine-similarity comparison.
- ✓ Complete the counterfactual solution related to SVM and generate counterfactual rule.
- ✓ Implemented the front-end user interface



Future Implementation

- Test the novel solution with existing tools and improve XAI solution
- Improve the user interface of the front-end.
- Integrate all the components and get the final output

Evidences for the Completion

Counterfactual Generator: WordFlipping

Predefined configuration

+ Code + Markdown

```python
from src.analyzers import SVMAnalyzer
analyzer = SVMAnalyzer(
    svm_path="./models/analysis-models/svm.pkl",
    vectorizer_path="./models/analysis-models/tfidf.pkl",
    cf_generator_config="./configs/models/wf-cf-generator.yaml"
)
review = "One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. They are right, as this is exactly what hap
search_space = 5
cf = analyzer(review, search_space)
explanation = analyzer.explanation()
print(explanation)
```

[2] Python

```
======== Analysis Report ========

Input text                     : One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. They are right, as this

Generated contradictory texts  :
        One of the other reviewers has mentioned that after watching just 1 Oz episode you 'll differ undercharge . They are right , as this differ exact
        One of the other reviewers abstain mentioned that after watching just 1 Oz episode you 'll be hooked . They differ right , as this differ exactly
        One of the other reviewers has mentioned that after watching just 1 Oz episode you 'll differ unhook . They differ right , as this differ exactly
        One of the other reviewers has mentioned that after watching just 1 Oz episode you 'll be hooked . They are right , as this is exactly what demate
        One of the other reviewers has mentioned that after watching just 1 Oz episode you 'll be unhook . They differ right , as this differ exactly what

Distances to the mirror point  :
        0.9996843344271062
        0.999649545631532
```

Analysis

Prompt

Variations
2

Test Cases

Name	Sampling Probability Decay Factor	Flipping Probability	Flipping Tags
Adjectives	0.2	1	JJ, JJR, JJS

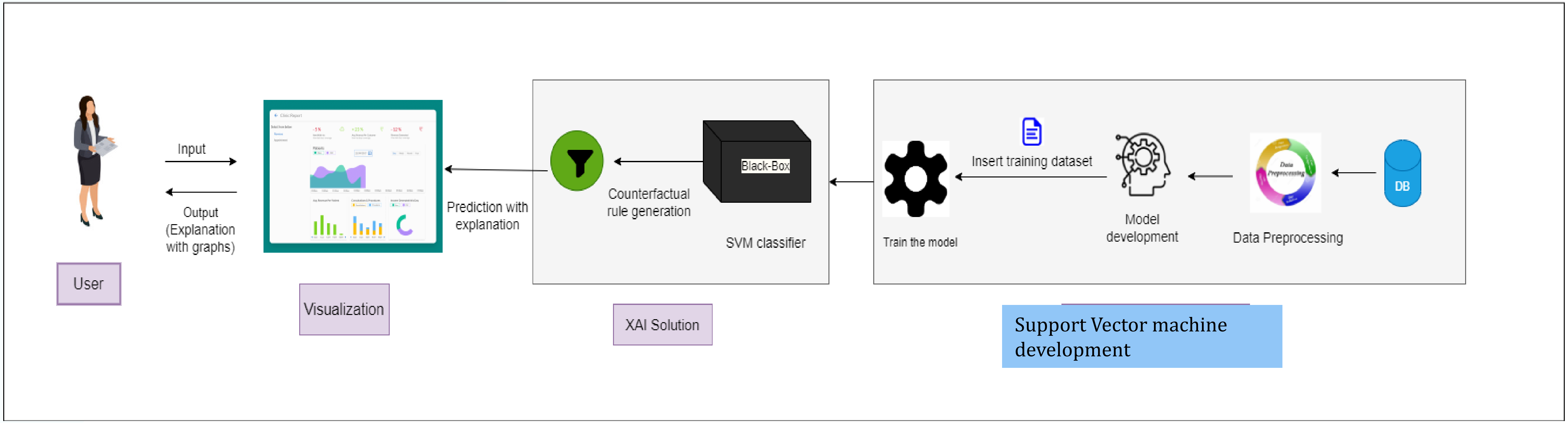
+

ANALYZE

Report

```
==== Configuration Adjectives (1) ====

======== Analysis Report ========

Input text                   : It was bad movie

Generated contradictory texts   :
        It was goodness movie
        It was unregretful movie

Distances to the mirror point   :
        0.99650689367582282
        0.9962301066637615

Closest contradictory text      : It was unregretful movie
```

System Diagram



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.

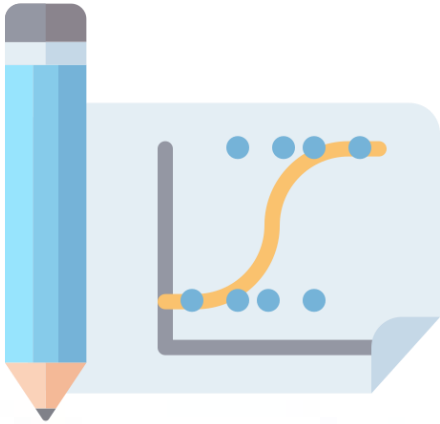
[8] S. M. Lundberg and S. I. Lee, “A unified approach to interpreting model predictions,” *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Section 2, pp. 4766–4775, 2017.

[9] Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, *31*, 841.

IT18161298 | SRINDEE METHMAL H.M

Specializing in Software Engineering

Logistic Regression



Research Gap

What are the existing methods that used for generating counterfactual rules related to Logistic Regression?

SHAP[4]
- Not Primarily Designed for Counterfactuals
- Assumption of Independence
- Computational Complexity
- model-agnostic Explanation

LIME[3]
- Not Primarily Designed for Counterfactuals
- explanations can be unstable
- model-agnostic Explanation

Diverse Counterfactual Explanation(DICE)[1]
- Assumption of Feature Independence
- Computationally Intensive
- Model agnostic explanation

Nearest Instance Counterfactual Explanations[NICE][2]
- aims to find the smallest and most meaningful changes to an instance that would alter the model's prediction
- Finding the nearest instance and ensuring feasibility can be computationally intensive.
- Model agnostic explanation
- Limited to binary classification problems.

Research Gap

RECOURSE

- Provide Counterfactuals.
- Model agnostic explanation.
- Rule-based approach.
- Optimization- based approach.
- Generative Adversarial Networks (GANs).
- Interactive Approaches

Proposed Method:

- **Text classification explainable task.**
- **Provide a weighted based Counterfactual Rule Generation Explainable Method.**
- **Model Specific Approach.**

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

Provide a novel post-hoc ,model-specific, local XAI solution to enhance the model explainability of

Binary classification-based models focus on LR

by developing a novel counterfactual rule generation mechanism related to the text classification domain.

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.

Antonym Replacement based Counterfactual Explanation for Logistic Regression Model.

Step 1: The input text is vectorized using TF-IDF vectorizer.

Step 2: Antonym selection and iterative replacement and removal.

Step 3: Then get the prediction score of the instance and classify it as positive or negative.

Step 4:Class change evaluation.

Step 5: Iterative replacement and removal process.

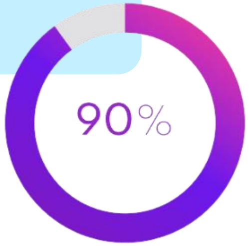
Step 6: Then get the feature importance of each word and sort the feature importance.

Step 7: Remove the most impactful features until we get a class change.

Completion and Future works

Completed Components

- ✓ Data preprocessed and built the Logistic Regression Model
- ✓ Find the novel methodology for generating counterfactual rule using random forest feature importance.
- ✓ Complete the counterfactual solution related to LR and generate counterfactual rule.
- ✓ Implemented the front-end user interface



Future Implementation

- Test the novel solution with existing tools and improve XAI solution
- Improve the user interface of the front-end.
- Integrate all the components and get the final output

Evidences for the Completion

```python
from src.analyzers import LRAnalyzer
%load_ext autoreload
%autoreload 2


explainer_lr = LRAnalyzer(
    "./models/analysis-models/lr.pkl",
    "./models/analysis-models/tfidf.pkl",
    threshold_classifier=0.49179999999978463,
    max_iter=50,
    time_maximum=120,
)

text = "One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. They are right, as this is exactly what happ
text = "hello"

explainer_lr(text, None)
```

✓ 4.5s Python

```
Start initialization...
initial sentence is ...
(1, 11612)
['..hello..']
score_predicted   [0.42906247]   initial_class   [0]
Initialization is complete.

 Elapsed time 3


 Iteration 1
```

Analysis

Prompt

Test Cases

Name	Classification Threshold	Maximum Iterations	Maximum Time
Case 1	0.491799999999785	50	120

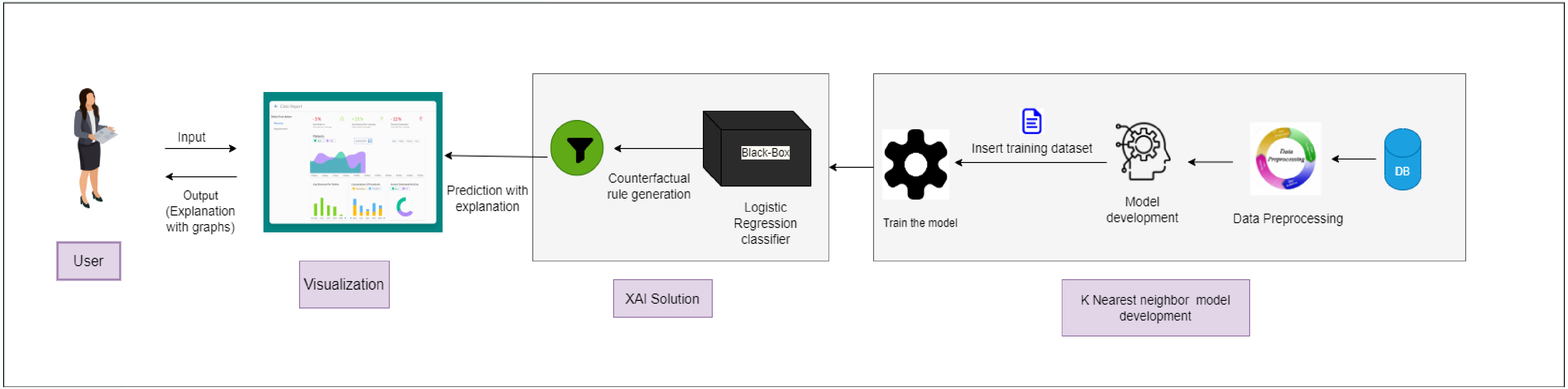
+

ANALYZE

Report


```
==== Configuration Case 1 (1) ====
{
    "input": {
        "text": "It was a bad movie",
        "score for positive": 2.782817425582194e-05,
        "initial class": 0
    },
    "output": {
        "Replacements": [
            {
                "feature": "bad",
                "replacement": "good"
            }
        ],
        "final_text": "It was a good movie",
        "final score for positive": 0.9110028599051992,
        "final class": 1
    },
    "process": {
        "final_exp": [
            [
                "bad",
                "good"
            ]
        ],
        "number active elements": 2,
        "number explanations found": 1,
        "size smallest explanation": 1,
        "time elapsed": 2.3372418880462646,
        "differences score": [
            0.08899714009480075
        ],
        "iterations": 1,
        "final_sentence": [
            "--bad--",
            "..movie..",
            "++good++",
            "  --> class 1 Score = ",
            0.9110028599051992
        ]
    }
}
```

System Diagram



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.

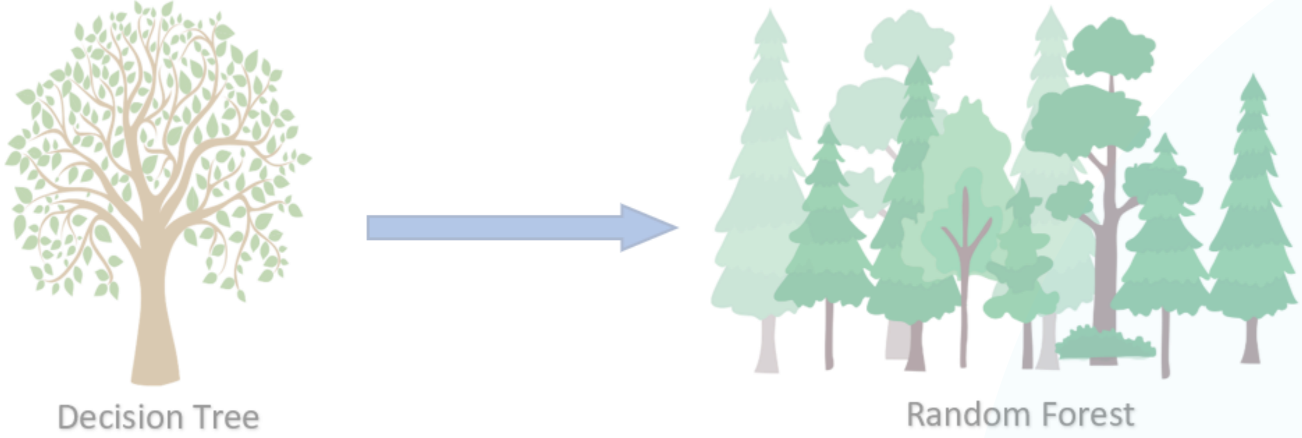
[5]Mothilal, R. K., Sharma, A., & Tan, C. (2020, January). Explaining machine learning classifiers through diverse counterfactual explanations. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 607-617).

[6] Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, *31*, 841.

IT20100698 | BRITTO T.A

Specializing in Data Science

Random Forest



Research Gap

What are the existing methods that used for generating counterfactual rules related to Random Forest?

SHAP[6]
- Not Primarily Designed for Counterfactuals
- Assumption of Independence
- Computational Complexity
- model-agnostic Explanation

LIME[7]
- Not Primarily Designed for Counterfactuals
- explanations can be unstable
- model-agnostic Explanation

Diverse Counterfactual Explanation(DICE)[6]
- Assumption of Feature Independence
- Computationally Intensive
- Model agnostic explanation

Nearest Instance Counterfactual Explanations[NICE][9]
- aims to find the smallest and most meaningful changes to an instance that would alter the model's prediction
- Finding the nearest instance and ensuring feasibility can be computationally intensive.
- Model agnostic explanation
- Limited to binary classification problems.

Research Gap

TreeSHAP
•Specially designed for tree-based models
•Model Specific Explanation

Proposed Method:

- **Text classification explainable task.**
- **Provide an instance-specific counterfactual explanation using feature importance in RF.**
- **Model Specific Approach.**

Research Problem

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

Objectives

Specific Objective

Provide a novel post-hoc ,model-specific, local XAI solution to enhance the model explainability of

Ensemble based classification 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.
- 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.

Instance-specific Counterfactual Explanation using Feature Importance in Random Forest Model.

Step 1:The given input text is preprocessed and makes an array.

Step 2: Extract feature importance from trained RF model and remove the features that are not related to the given instance.

Step 3:Get the prediction score of the instance and classify it as positive or negative.

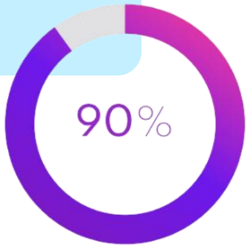
Step 4:Then get the feature importance of each word in the vector and sort the feature importance.

Step 5:Remove the most impact features iteratively until get a particular class change.

Completion and Future works

Completed Components

- ✓ Data preprocessed and built the Random Forest Model
- ✓ Find the novel methodology for generating counterfactual rule using random forest feature importance.
- ✓ Complete the counterfactual solution related to RF and generate counterfactual rule.
- ✓ Implemented the front-end user interface



Future Implementation

- Test the novel solution with existing tools and improve XAI solution
- Improve the user interface of the front-end.
- Integrate all the components and get the final output

Evidences for the Completion

Analyzer

```python
from src.analyzers import RFAnalyzer
%load_ext autoreload
%autoreload 2


explainer_rf = RFAnalyzer(
    "./models/analysis-models/rf.pkl",
    "./models/analysis-models/tfidf.pkl",
    threshold_classifier=0.49339999999983775,
    max_iter=50,
    time_maximum=300,
)

text = "Watching that film was a complete waste of time. The plot was dull from start to finish, and the performances were bad. I
#text = "I thought that Mukhsin has been wonderfully written. Its not just about entertainment. There's tonnes of subtle message
#ds.x_train[72, :]
#text = "it was a bad and dull movie but the end was amazing"
explainer_rf(text, None)
```

[3] ✓ 5.1s Python

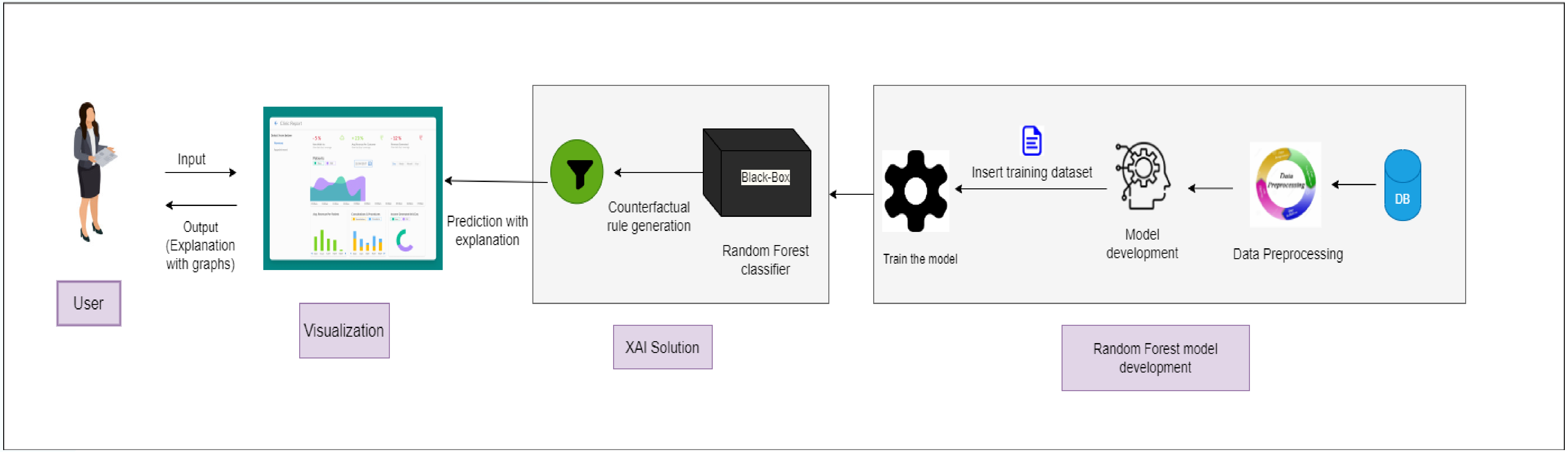
```
The autoreload extension is already loaded. To reload it, use:
  %reload_ext autoreload
Start initialization...
initial sentence is ...
(1, 11612)
['..anyone..', '..bad..', '..cant..', '..complete..', '..dull..', '..film..', '..finish..', '..performance..', '..plot..', '..recomm
score_predicted   [0.17755176]   initial_class   [0]
initial_score  0.17755175676474977
665   0.0006410131485748198 1 [0.17588509]
665   0.0006410131485748198
```

Analysis
Prompt
Test Cases
Name
Classification Threshold
Maximum Iterations
Maximum Time
case 1
0.493399999999838
50
120
+
ANALYZE
Report
==== Configuration case 1 (1) ====
{
"input": {
"text": "Watching that film was a complete waste of time. The plot was dull from start to finish, and the performances were bad. I
"score for positive": 0.17755175676474977,
"initial class": 0

http://localhost:3001


```
        {
            "start": 0.0010786052234948646
        },
        {
            "time": 0.0021253787888005603
        },
        {
            "waste": -0.02392262547906587
        },
        {
            "watch": 0.0018750162723304388
        }
    ],
    "output": {
        "Removed_words": [
            "bad",
            "waste",
            "plot"
        ],
        "final_text": "Watching that film was a complete of time The was dull from start to finish and the performances were --- I can t r
        "final score for positive": 0.6661218489176939,
        "final class": 1
    }
}
```

System Diagram



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] Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, *31*, 841.

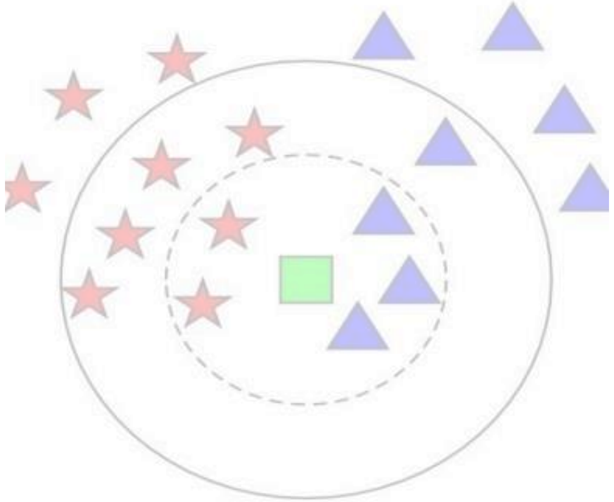
[5] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82-115

[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



Research Gap

What are the existing methods that used for generating counterfactual rules related to Random Forest?

SHAP[6]
- Not Primarily Designed for Counterfactuals
- Assumption of Independence
- Computational Complexity
- model-agnostic Explanation

LIME[7]
- Not Primarily Designed for Counterfactuals
- explanations can be unstable
- model-agnostic Explanation

Diverse Counterfactual Explanation(DICE)[6]
- Assumption of Feature Independence
- Computationally Intensive
- Model agnostic explanation

Nearest Instance Counterfactual Explanations[NICE][9]
- aims to find the smallest and most meaningful changes to an instance that would alter the model's prediction
- Finding the nearest instance and ensuring feasibility can be computationally intensive.
- Model agnostic explanation
- Limited to binary classification problems.

Research Gap

Explaining and Improving Model Behavior with k Nearest Neighbor Representations

•Locally explainable
•Model Agnostic Explanation

Proposed Method:

- **Text classification explainable task.**
- **Provide a distance based Counterfactual Rule Generation Explainable Method.**
- **Model Specific Approach.**

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 explainability of

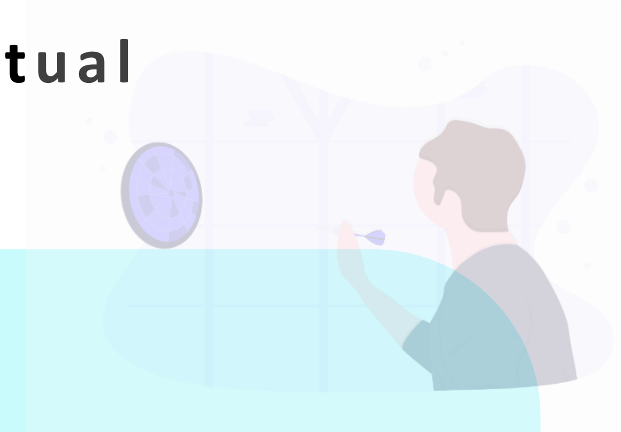
distance based classification models focus on KNN

by developing a novel counterfactual rule generation mechanism related to the text classification domain.

Sub Objectives

- Prepare the dataset and implement the KNN 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.

Feature Density Comparison based Counterfactual Explanation for K-Nearest Neighbor (KNN)



Step 1: Generate Counterfactuals using word flipping generator.

Step 2: Vectorize the counterfactuals and the original review.

Step 3:Get neighbour statistics (compare feature densities between the original review and counterfactuals).

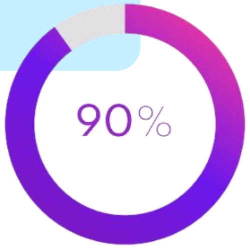
Step 4: Get probability of getting positive or negative classification for the counterfactuals.

Step 5: Select the best counterfactual using neighbour statistics.

Completion and Future works

Completed Components

- ✓ Data preprocessed and built the KNN Model
- ✓ Find the novel methodology for generating counterfactual rule using feature density comparison.
- ✓ Complete the counterfactual solution related to KNN and generate counterfactual rule.
- ✓ Implemented the front-end user interface



Future Implementation

- Test the novel solution with existing tools and improve XAI solution
- Improve the user interface of the front-end.
- Integrate all the components and get the final output

Evidences for the Completion

Predefined configuration

```python
from src.analyzers.knn import KNNAnalyzer
analyzer = KNNAnalyzer(
    knn_path="./models/analysis-models/knn.pkl",
    vectorizer_path="./models/analysis-models/tfidf.pkl",
    cf_generator_config="./configs/models/wf-cf-generator.yaml"
)
text="One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. They are right, as this is exactly what happene
analyzer(text, 2)
print(analyzer.explanation())
```

[1] Python

```
======== Analysis Report ========

Input text                  : One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. They are right, as this is
Input neighbor counts       : {'negative': 34, 'positive': 56}
Input class probabilities   : {'negative': 0.37777777777777777, 'positive': 0.6222222222222222}
Input class densities       : {'negative': 25.998423572007233, 'positive': 44.53242593645316}
Input review class          : positive

Contradictory texts:
        One of the other reviewers has mentioned that after watching just 1 Oz episode you 'll be unhook . They are right , as this is exactly what demate
        One of the other reviewers lack mentioned that after watching just 1 Oz episode you 'll be hooked . They are right , as this differ exactly what c
Contradictory neighbor counts:
        {'negative': 31, 'positive': 59}
        {'negative': 32, 'positive': 58}
Contradictory class probabilities:
        {'negative': 0.34444444444444444, 'positive': 0.6555555555555556}
        {'negative': 0.35555555555555557, 'positive': 0.6444444444444445}
Contradictory class densities:
        {'negative': 23.531772458395043, 'positive': 45.93646841277097}
```

Analysis

Prompt

Variations
2

Test Cases

Name	Sampling Probability Decay Factor	Flipping Probability	Flipping Tags
Adjectives	0.2	1	JJ, JJR, JJS

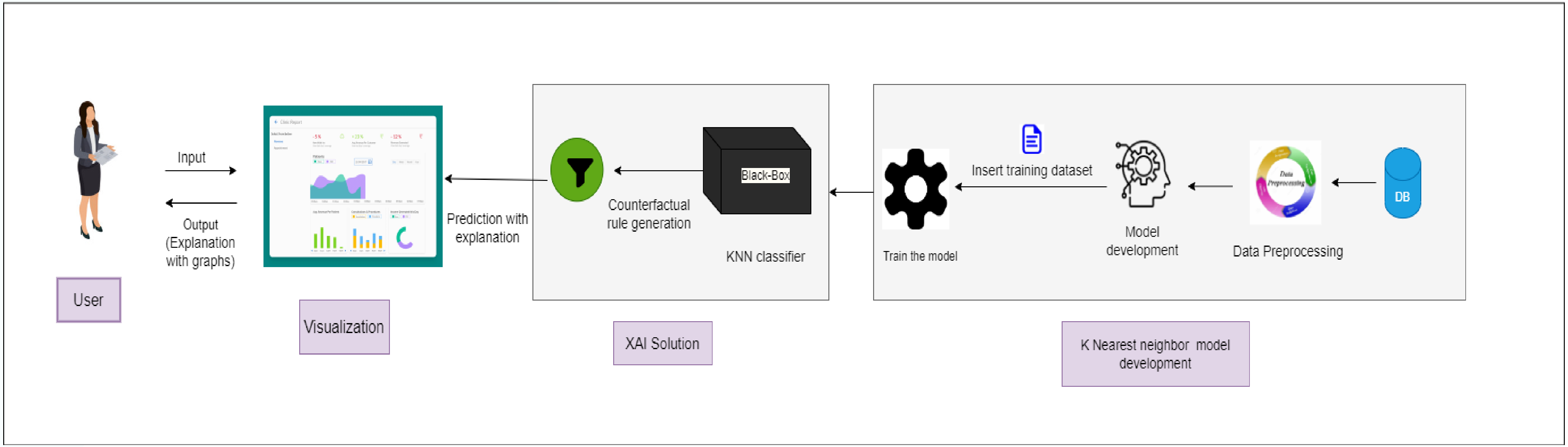
+

ANALYZE

Report

```
==== Configuration Adjectives (1) ====

======== Analysis Report ========

Input text                 : It was a good movie
Input neighbor counts      : {'negative': 43, 'positive': 47}
Input class probabilities  : {'negative': 0.4777777777777778, 'positive': 0.5222222222222223}
Input class densities      : {'negative': 37.52496132816629, 'positive': 41.20450637690619}
Input review class         : positive

Contradictory texts:
        It was a ill movie
        It was a evil movie
Contradictory neighbor counts:
        {'negative': 57, 'positive': 33}
        {'negative': 49, 'positive': 41}
Contradictory class probabilities:
        {'negative': 0.6333333333333333, 'positive': 0.36666666666666664}
        {'negative': 0.5444444444444444, 'positive': 0.45555555555555555}
Contradictory class densities:
        {'negative': 45.22733089368357, 'positive': 26.20004968036461}
        {'negative': 39.94622994853936, 'positive': 33.34828842485549}
Closest counterfactual ID: 1
```

System Diagram



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.

[8] Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, *31*, 841.

Tools and Technologies

- **Frontend:**
  - NestJs
  - Mantine UI

- **Backend:**
  - Python

- **Version Control:**
  - GitLab

- **Tools:**
  - VS Code
  - Google Colab

Q & A

Thank You !