Proceedings of the 10th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'24)
Barcelona, Spain - August 19 - 21, 2024
Paper No. CIST 158
DOI: 10.11159/cist24.158

Enhancing Model Explainability with CTGAN-LIME: A Novel Approach for Interpretable Machine Learning

Bodrunnessa Badhon[1], Ripon K. Chakrabortty[1], Sreenatha G. Anavatti[2]
[1]School of Systems & Computing, UNSW Canberra at ADFA
ACT-2610, Australia.
First Author Email: b.badhon@unsw.edu.au; Second Author Email: r.chakrabortty@unsw.edu.au
[2] School of Engineering & Technology, UNSW Canberra at ADFA
ACT-2610, Australia.
Third Author Email: s.anavatti@unsw.edu.au

Abstract – Machine learning (ML) has become integral in numerous industries, offering unparalleled data analysis, pattern recognition, and predictive modelling advantages. However, the opacity of ML models, often referred to as "black boxes," poses significant challenges in understanding their decision-making processes. Explainable Artificial Intelligence (XAI) techniques aim to address this challenge by providing transparency into ML models' inner workings, enhancing human comprehension and trust. This study proposes a novel approach, CTGAN-LIME, combining Conditional Tabular Generative Adversarial Networks (CTGAN) with the LIME (Local Interpretable Model-Agnostic Explanations) framework to enhance model explainability. CTGAN-LIME addresses LIME's limitations by structuring neighbourhood sample generation and considering class balance, thereby improving the reliability and stability of explanations. Empirical evaluations across diverse datasets demonstrate CTGAN-LIME's superiority in local fidelity, stability, and local concordance over traditional LIME, underscoring its effectiveness in enhancing trustworthiness across various black-box models.

***Keywords*:** Explainable Artificial Intelligence (XAI), Machine Learning, LIME, Black-box Model.

1. Introduction

Artificial intelligence (AI), particularly machine learning (ML), has become pivotal across various applications, revolutionising industries ranging from healthcare and finance to marketing and transportation. Its ability to analyse vast amounts of data, recognise patterns, and make predictions has led to significant advancements and efficiencies [1]. However, amidst its widespread adoption, a significant challenge arises: opacity. The intricate algorithms underlying AI systems often operate as "black boxes," rendering their decision-making processes mysterious to human understanding. This opacity poses profound implications, raising concerns about accountability, bias, and the ethical implications of AI-driven decision-making [2]. As AI continues to permeate our daily lives, addressing the opacity problem is critical for ensuring transparency, fairness, and trust in these transformative technologies.

One promising research area, Explainable Artificial Intelligence (XAI), aims to address this issue by providing transparency into the inner workings of ML models, enabling humans to comprehend and trust their outputs [3]. XAI techniques can illuminate the reasoning behind ML predictions, mitigating concerns related to understandability. There are many types of XAI methods, including model-agnostic- applied to any kind of black-box model and model-specific- applied to specific models, as well as post-hoc- applied after the black-box model has been trained and intricate model- incorporating explainability during model training [4]. Among XAI approaches, post-hoc model-agnostic methods are popular for their flexibility in explaining any ML model after training without altering the original structure [2]. A notable example is Local Interpretable Model-Agnostic Explanations (LIME) [5], which has gained significant attention due to its exceptional versatility, effectiveness, and widespread acceptance within the scientific community. It generates explanations by simulating data points around a specific instance through random perturbation and fitting a weighted sparse linear model over the predicted responses from these points.

While LIME proves effective in addressing tabular classification problems, it encounters limitations concerning stability [4]. Firstly, LIME randomly selects nearby data points, leading to selection variability across different runs.

This randomness results in different explanations being provided for similar cases, thereby diminishing the reliability and stability of its explanations. Secondly, LIME's generation of neighbourhood samples overlooks the importance of class balance, potentially introducing bias or skewness into its explanations and thereby impacting its overall performance. To overcome these limitations, we propose a novel model called CTGAN-LIME (Conditional Tabular Generative Adversarial Networks based on Local Interpretable Model-Agnostic Explanations) to enhance LIME's capabilities. Our contribution can be summarised as follows:

1. Firstly, CTGAN-LIME addresses the issue of random selection of nearby data points by employing a more structured approach through Conditional Tabular Generative Adversarial Networks (CTGAN). CTGAN, known for its ability to generate synthetic data points while preserving high-dimensional characteristics, ensures greater consistency in LIME's explanations across different runs. This structured generation of synthetic data enhances the reliability and stability of interpretations, thus contributing to the robustness of CTGAN-LIME.
2. Secondly, CTGAN-LIME considers the importance of class balance during the generation of neighbourhood samples using CTGAN by utilising conditional vector. By considering class balance, our model aims to reduce the potential for bias or skewness in explanations, thus improving LIME's overall fairness and performance.

To achieve our goals, this study begins by selecting a specific data point for explanation. Instead of utilising traditional random perturbations, a CTGAN generates synthetic neighbourhood samples around the chosen datapoint, thereby capturing the data distribution and ensuring class balance. Subsequently, predictions are generated by the trained black-box model for each neighbourhood instance, encompassing both real and synthetic samples. Finally, the local ridge regression model is fitted to capture explanations for the specific data point effectively.

The remaining sections of this manuscript are structured as follows: In Section 2, we conduct a comprehensive review of existing works in the field of local explanation methods. Section 3 introduces CTGAN-LIME, elucidating the details of the proposed method. The CTGAN-LIME is then rigorously evaluated in Section 4, and we subsequently conclude our work and outline future directions in Section 5.

2. Related Work

Our review focuses on locally interpretable post-hoc models, which aim to elucidate the relationship between input variables and the target variable within specific regions. Through an extensive examination of existing literature, we have identified key competitors and outlined their methodologies.

LIME [5] is a widely utilised post-hoc technique that constructs local surrogate models to mimic opaque models' behaviour. It achieves interpretability by perturbing input data to observe changes in model predictions, providing explanations for individual instances. However, LIME struggles with stability and capturing complex non-linear relationships [4]. ALIME (Autoencoder-based Local Interpretable Model-Agnostic Explanations) [6] utilises autoencoders to delineate local decision boundaries and offer interpretable explanations for predictions. Despite improving local fidelity, understanding the meaning of latent representations learned by autoencoders can be challenging, and the performance of autoencoders can be highly sensitive to hyperparameters [3]. DLIME (Deterministic Local Interpretable Model-Agnostic Explanations) [7] is tailored for computer-aided diagnosis systems, aiming to enhance transparency in predictions. It integrates Hierarchical Clustering to address instability issues yet struggles with high-dimensional data due to the curse of dimensionality [2]. G-LIME (Global priors-based LIME) [4] advances the interpretation of Deep Neural Network models by incorporating Bayesian linear regression with sparsity and informative global priors, improving consistency compared to LIME. However, the effectiveness of G-LIME relies on the quality of global priors utilised [8].

In the post-hoc XAI landscape, diverse methodologies exist, each with strengths and limitations. While capable of interpreting machine learning models, these approaches encounter challenges with stability and local fidelity within high-dimensional and facing data overfitting, particularly, when trained on limited or noisy data. Addressing these challenges requires tailored XAI approaches to develop more robust methodologies.

3. Proposed Methodology

3.1. LIME (Local Interpretable Model-Agnostic Explanations)

Before explaining LIME-CTGAN, a brief overview of the LIME [5] framework is provided in this section. The process of generating explanations with LIME involves several key steps. First, a specific data point of interest is selected for explanation. Then, LIME generates synthetic data samples around this data point by randomly perturbing the input features. These synthetic samples and the original data point are used to train a local surrogate model, such as linear regression. Once the surrogate model is trained, it provides insights into how the black-box model behaves in the local neighbourhood of the selected data point. The coefficients of the surrogate model indicate the importance of each feature in making predictions for that specific instance.

As mentioned earlier, it's crucial to acknowledge that LIME comes with challenges and limitations [4]. One notable limitation is the instability caused by the randomness in generating nearby data points and the potential for class imbalance in the generated neighbourhood samples. This randomness and imbalance can lead to variations in explanations provided by LIME, affecting the reliability and stability of its interpretations.

3.2. CTGAN-LIME

This section presents our proposed post-hoc XAI model, CTGAN-LIME, as depicted in Figure 1. CTGAN-LIME utilises CTGAN [9] to generate diverse and balanced neighbourhoods instead of relying on random perturbations, thus aiming to improve LIME's performance.

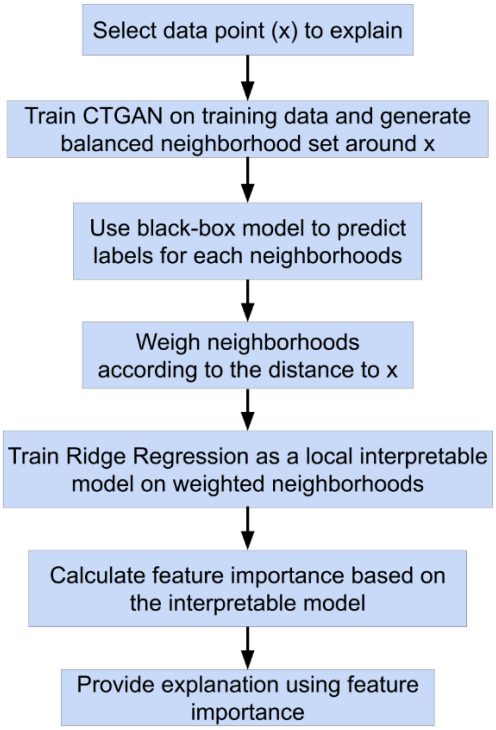


Figure 1: Flowchart of Proposed CTGAN-LIME Method

CTGAN extends the traditional GAN architecture to generate tabular data. Unlike GAN, CTGAN excels in capturing and reproducing dependencies and relationships between different features in structured datasets, making it proficient in handling mixed data types. A key innovation of CTGAN is its conditional mechanism, which allows for data generation while adhering to specific conditions or attributes within the dataset. This involves integrating a conditional generator and a training-by-sampling method, where the conditional vector guides the generator to learn the real data conditional distribution. For instance, in a dataset with categorical attributes like 'Colour,' CTGAN can enforce equal representation of all colour categories during data generation, promoting a more equitable distribution.

Algorithm 1 outlines the procedure for CTGAN-LIME. It commences by selecting the input data point (x) for explanation. Next, neighbourhood samples are generated using a pre-trained CTGAN model, capturing the local area surrounding the input data. These samples are retained for subsequent analysis. Following this, the black-box model predicts labels for the neighbourhood samples, and weights are computed based on their proximity to the input point. The algorithm proceeds to train the interpretable Ridge regression model using the neighbourhood samples, their predicted labels, and the calculated weights. Finally, employing the trained Ridge regression model, the algorithm provides detailed analyses, including coefficients, thereby furnishing explanations for the predictions made by the black-box model concerning the input data point.

Algorithm 1: CTGAN-LIME

1. Initialize:
   - For each instance *i*:
2. Generate Neighborhoods:
   - Utilize CTGAN_Model to generate neighborhood samples around the input *x*. Let these samples be denoted as $\boldsymbol{c_i'}$.
3. Prediction:
   - Apply the Black-box_Model to predict the outcomes for each generated sample $\boldsymbol{c_i'}$. Denote these predictions as $\boldsymbol{f(c_i')}$
4. Compute Similarity Weights:
   - Calculate the similarity weights for the generated samples with respect to the input *x*. Denote these weights as $\boldsymbol{w_x(c_i')}$
5. Store Results:
   - Save the tuples ($\boldsymbol{c_i', f(c_i'), w_x(c_i')}$) for each generated sample $\boldsymbol{c_i'}$.
6. Train Ridge Regression:
   - Use the generated samples $\boldsymbol{c_i'}$as features, their predictions $\boldsymbol{f(c_i')}$ as labels, and the similarity weights $\boldsymbol{w_x(c_i')}$ to train a Ridge Regression model.
7. Generate Explanation:
   - Calculate the feature importance from the trained Ridge Regression model to explain the prediction of the Black-box_Model at the input *x*.

Output:

- Explanation: A set of feature importances derived from Ridge Regression that elucidates the prediction of the Black-box_Model for the input *x*.

4. Result Analysis

In this section, we empirically evaluate the performance of CTGAN-LIME across three diverse datasets sourced from the UCI repository. These datasets span various domains and are utilised to assess the effectiveness of our proposed model. The descriptions of each dataset are delineated as follows:

.

I. Heart Disease dataset [10]: A widely utilised dataset comprising 1026 patient observations (instances and 13 distinct features utilised for predicting the presence or absence of heart disease.

II. Red wine quality dataset [11]: It consists of 1600 instances and 11 features and is employed to predict the quality of wines based on their physicochemical properties.

III. Census income dataset [12]: This dataset aims to predict whether an individual's income exceeds $50,000 per year based on census data. It consists of 48,842 instances, encompassing 10 features for each instance.

Figure 2 illustrates how our model provides explainability by approximating the behaviour of a black-box machine learning model at the individual data point level. We employed a feedforward neural network (NN) as our black-box model, comprising a single hidden layer with 40 neurons and utilising the rectified linear unit (ReLU) activation function with L2 regularisation to prevent overfitting. The input layer corresponds to the number of features in the dataset, while the output layer consists of two neurons representing the binary class output. The NN was trained using backpropagation with binary cross-entropy loss. The datasets were split into 70% for training and 30% for testing, achieving accuracies of 96%, 91%, and 93% for the Heart Disease, Red Wine Quality, and Census income datasets, respectively. In Figure 2, features with higher coefficients are considered more influential, as they have a greater impact on the predicted class likelihood. When a coefficient is positive and shown in green (e.g., age, trestbps, sex etc from Figure 2 (a)), it means that increasing the value of that feature makes it more likely for the predicted class to be true, while negative coefficients (e.g., restecg, chol, ca etc from Figure 2(a)) suggest the opposite. By examining these coefficients, users can understand why the black-box model made a particular prediction for the chosen data point and which features contributed most significantly to that prediction.

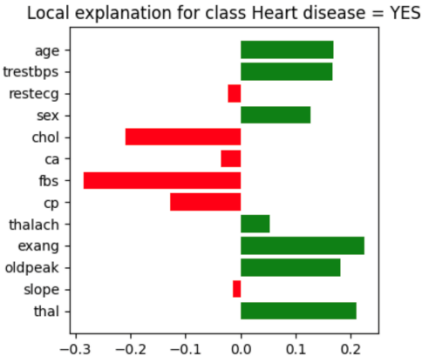


Figure 2(a): Local Explanation for Heart Disease Dataset

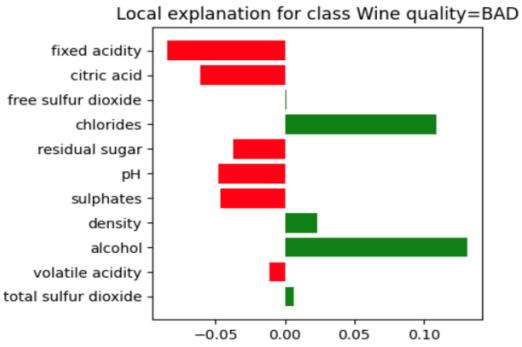


Figure 2(b) Local Explanation for Red Wine Quality Dataset

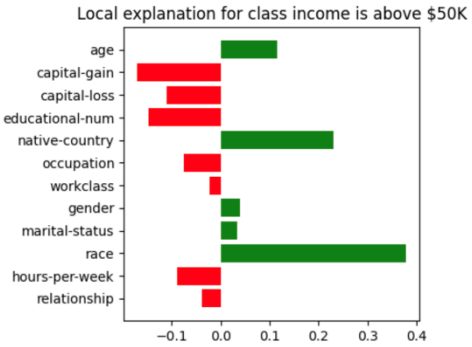


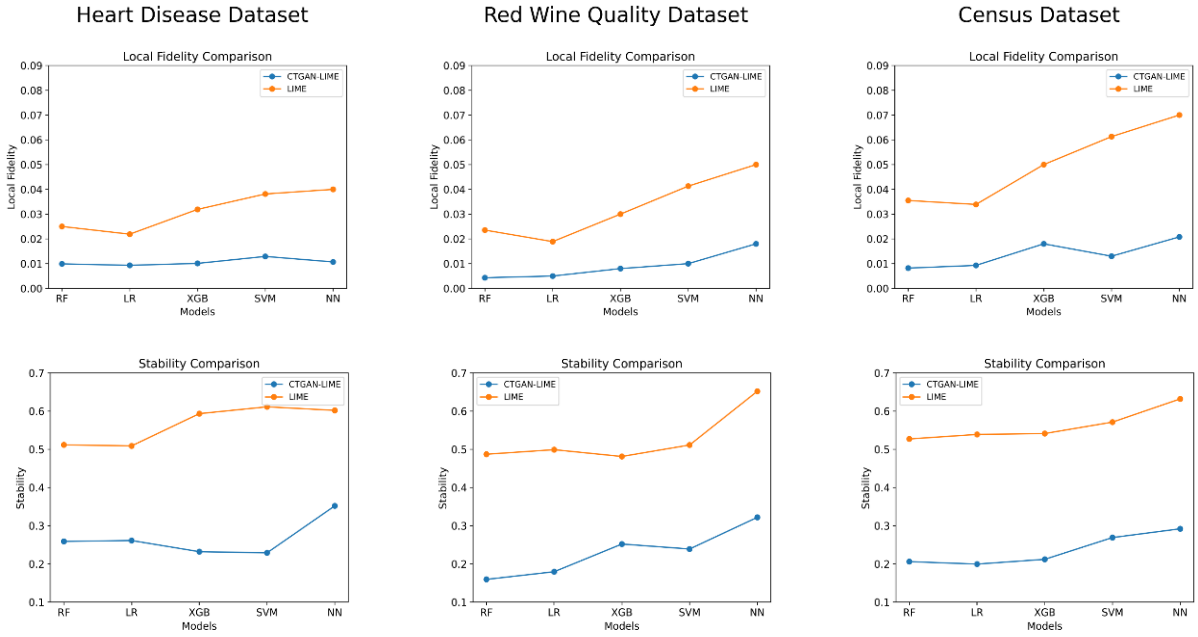
Figure 2(c): Local Explanation for Census Income Dataset

4.1. Performance Evaluation

In this section, we conduct a comparative analysis between CTGAN-LIME (proposed) and the original LIME method across various black-box models, including Random Forest (RF), Logistic Regression (LR), XGBoost (XGB), Support Vector Machine (SVM), and Neural Network (NN). To assess the performance of our proposed approach, we consider three metrics, as described in Table 1. In Figure 3, the first row illustrates the local fidelity across the three datasets. Our proposed CTGAN-LIME consistently outperforms LIME in terms of local fidelity across different models, demonstrating its superior capability to represent the underlying black-box models faithfully. Regarding stability (second row), CTGAN-LIME exhibits higher stability compared to LIME, showcasing its robustness, particularly with increasing neighbourhood size. Lastly, for local concordance (third row), CTGAN-LIME demonstrates better alignment between the local surrogate model and the original black-box model, emphasising its reliability in providing consistent explanations across various instances. This comparative analysis underscores the effectiveness of our proposed CTGAN-LIME model in enhancing trustworthiness across diverse datasets and its applicability in various black-box models.

Table 1: Performance Metrics for Model Explanation Evaluation

Metric Name	Descriptions	Equation
Local fidelity	Local fidelity is determined by a performance metric, typically Mean Squared Error (MSE), to compare the outputs of the black-box model (f_i) and the surrogate model (f_i'), revealing how faithfully the surrogate model replicates the black-box model within a local region.	$Local\ Fidelity\ (MSE) = \frac{1}{n}\sum_{i=1}^{n}(f_i' - f_i)$
Stability	Stability *(S)* is a metric that quantifies the robustness of an explanation method when applied to the same instance multiple times. It measures the degree to which the explanation remains stable when repeatedly applied to identical instances.	$S(f, e, N) = E[\|\|e(x, f)' - e(x, f)\|\|^2]$
Local concordance	Local concordance *(L)* assesses how well a surrogate model, g, approximates the black-box model, f, for a specific instance x while adhering to a conciseness constraint. It quantifies the agreement between $f(x)$ and $g(x)$ for this single instance, measured through the hinge loss function.	$L = \max(0, 1 - \lvert f(x) - g(x)\rvert)$



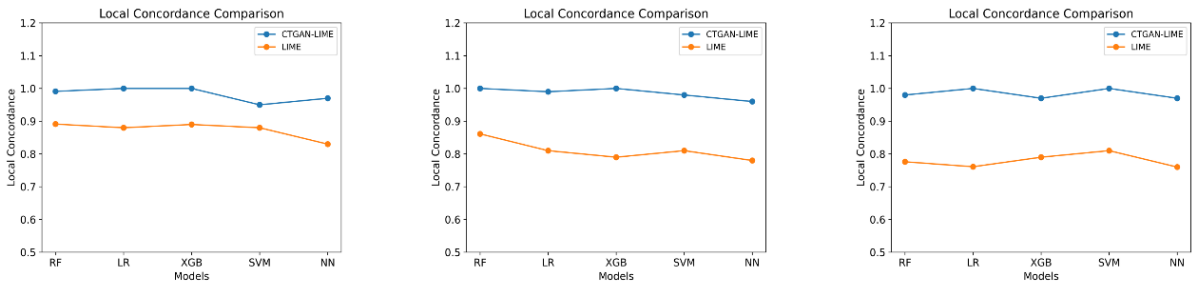


Figure 3: Comparison Between CTGAN-LIME and LIME in terms of Local Fidelity, Stability and Local Concordance for Heart Disease, Red Wine Quality and Census Income Dataset

5. Conclusion

In this study, we introduced CTGAN-LIME as a novel approach to enhance the explainability of black-box machine learning (ML) models. By leveraging CTGAN to generate more structured and balanced neighbourhood samples, CTGAN-LIME overcomes the limitations of LIME, particularly regarding the randomness in generating neighbourhood samples, leading to varied explanations for the same instance across different runs. Our empirical assessments across different datasets and black-box models underscored the superior performance of CTGAN-LIME in faithfully representing underlying models and providing consistent explanations. These findings demonstrate the effectiveness and applicability of CTGAN-LIME in addressing the opacity problem in ML, thereby fostering trust and transparency in ML-driven decision-making processes across various domains. Future directions for this work could involve investigating methods to improve the scalability of CTGAN-LIME for real-world applications, thereby advancing the field of Explainable Artificial Intelligence (XAI).

References

[1] Z. Jan, F. Ahamed, W. Mayer, N. Patel, G. Grossmann, M. Stumptner, and A. Kuusk, "Artificial intelligence for industry 4.0: Systematic review of applications, challenges, and opportunities," Expert Systems with Applications, vol. 216, no. 216, p. 119456, Apr. 2023, doi: https://doi.org/10.1016/j.eswa.2022.119456.

[2] W. Saeed and C. Omlin, "Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities," Knowledge-Based Systems, vol. 263, p. 110273, Jan. 2023, doi: https://doi.org/10.1016/j.knosys.2023.110273.

[3]. I. Ahmed, G. Jeon, and F. Piccialli, "From Artificial Intelligence to eXplainable Artificial Intelligence in Industry 4.0: A survey on What, How, and Where," IEEE Transactions on Industrial Informatics, vol. 18, no. 8, pp. 1–1, 2022, doi: https://doi.org/10.1109/tii.2022.3146552.

[4] X. Li, H. Xiong, X. Li, X. Zhang, J. Liu, H. Jiang, Z. Chen, and D. Dou, "G-LIME: Statistical learning for local interpretations of deep neural networks using global priors," Artificial Intelligence, vol. 314, pp. 103823–103823, Jan. 2023, doi: https://doi.org/10.1016/j.artint.2022.103823

[5] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?,'" Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, 2016, doi: https://doi.org/10.1145/2939672.2939778.

[6] S. M. Shankaranarayana and D. Runje, "ALIME: Autoencoder Based Approach for Local Interpretability," Intelligent Data Engineering and Automated Learning – IDEAL 2019, pp. 454–463, 2019, doi: https://doi.org/10.1007/978-3-030-33607-3_49.

[7] M. R. Zafar and N. Khan, "Deterministic Local Interpretable Model-Agnostic Explanations for Stable Explainability," Machine Learning and Knowledge Extraction, vol. 3, no. 3, pp. 525–541, Jun. 2021, doi: https://doi.org/10.3390/make3030027.

[8] X. Kong, S. Liu, and L. Zhu, “Toward Human-centered XAI in Practice: A survey,” Machine Intelligence Research, Jan. 2024, doi: https://doi.org/10.1007/s11633-022-1407-3.
[9] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, “Modeling Tabular data using Conditional GAN,” Neural Information Processing Systems, 2019.
[10] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, “Heart Disease,” UCI Machine Learning Repository, 2012.
[11] P. Cortez , A. Cerdeira, F. Almeida, T. Matos, and J. Reis, “Wine Quality,” UCI Machine Learning Repository, 2009.
[12] Kohavi,Ron. (1996). Census Income. UCI Machine Learning Repository. https://doi.org/10.24432/C5GP7S.