Analysis of UC Irvine Heart Disease Dataset Using Supervised Machine Learning Algorithms

Abstract—**The paper investigates the performance of several machine learning algorithms, including Logistic Regression, KNN, and Support Vector Machine, applied to the UC Irvine Heart Disease Dataset. The study emphasises the significance of data preprocessing and Visualisation techniques in gaining insights into the dataset's characteristics. Additionally, it addresses the class imbalance issue by adopting the SMOTE method. Noteworthy findings include achieving an accuracy of around 88% in binary classification tasks and approximately 60% in multiclass classification tasks post-data oversampling.**

Index Terms—**logistic regression, KNN, SVM, SMOTE, class imbalance, Normalisation, dataset Precision**

I. Introduction

Cardiovascular Diseases (CVDs), popularly known as heart diseases, are a group of diseases that affect the heart's functions, leading to significant health issues in a person irrespective of age. CVDs are one of the major causes of premature death in the world. This is one of the main motivations behind choosing this topic for the work. The recent integration of artificial intelligence with medical research has proven that machine learning and deep learning can be used for disease prediction and risk assessment in a patient. Although the paper does not focus much on the medical aspects, it is a good start for an enthusiast interested in these fields. The paper's main focus will be analysing using supervised machine learning algorithms on medical data.

The paper delves into the efficiency of different machine learning algorithms on the UC Irvine Heart Disease dataset. This paper focuses on three common supervised machine learning algorithms: logistic regression [1], K nearest neighbour(KNN) [2], and support vector machine(SVM) [3]. The paper also explores different data preprocessing techniques.

Using a correlation matrix, the paper focuses on understanding how each feature correlates. The focus is on understanding how age is a major contributing factor to whether a person in the US has a heart disease. Through these experiments, we also try to understand how class imbalance impacts multiclass classification. Generally, a dataset with a high-class imbalance can generate a biased model that will more accurately predict unseen data corresponding to the class with more data than the one with fewer data. To address this issue, the paper explores the SMOTE method. A discussion on the impact of introducing this method on the dataset is provided in this paper.

The work also focuses on exploring different metrics to determine the effectiveness of the models. Traditionally, researchers use accuracy as a method to find the efficiency of a model. In this work, other metrics like precision and confusion metrics are used.
The experiments are divided into two classification problems:

- Binary classification
- Multiclass classification

The structure of the paper is as follows:

1) Literature Review
2) Dataset
3) Methodologies and Experimental Setups
4) Observation and Inference
5) Conclusion

II. Literature Review

The main motivation for this paper was based on the first usage of this dataset, which was to develop a new discriminant functional model to estimate the probabilities of angiographic coronary diseases [4]. The model was then used to predict the prevalence of the disease in three patient groups. The paper focussed on comparing the predictability with a Bayesian model CADENZA. Both the models overpredicted the disease prevalence, but the new discriminant model seemed moderate with its predictions.
Another relevant work proposes an algorithm that measures the impact of significant features contributing to heart disease. The scores of significant features were computed using the weighted associative rule mining [5]. The study successfully observed rules and features that help diagnose heart disease.

A work that addresses the issue of class imbalance and feature selection proposes a combination of classification algorithms and feature selection techniques [6]. The work concludes that using Random Forest algorithms while sampling the data using the ADASYN algorithm yielded the best prediction.

A work focusing on myocardial infractions (heart attacks) is a larger literature review that inspects different techniques within machine learning and deep learning to understand how different approaches can predict heart disease, specifically myocardial infractions. This work aims to aid future works in this field that involve ML and DL techniques [7].

III. UC Irvine Heart Disease Dataset

The UC Irvine Heart Disease Dataset [8] consists of 76 features, but only fourteen of these variables are used

TABLE I
UC IRVINE DATASET

No.	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	that	num
1	63	1	1	145	233	1	2	150	0	2.3	3	0.0	6.0	0
2	67	1	4	160	286	0	2	108	1	1.5	2	3.0	3.0	2
3	67	1	4	120	229	0	2	129	1	2.6	2	2.0	7.0	1
4	37	1	3	130	250	0	0	187	0	3.5	3	0.0	3.0	0
5	41	0	2	130	204	0	2	172	0	1.4	1	0.0	3.0	0

for primary scientific research. Some of these features are age, sex, fasting blood sugar(fbs), chest pain(cp), serum cholesterol(chol) etc. The dataset comprises 303 patients from Cleveland, Hungary, Switzerland and VA Long Beach. The focus of the work is on the Cleveland database. The target has five classes, mainly 0, corresponding to no heart disease and class values 1,2,3,4, corresponding to various stages of heart disease. The dataset consists of numerical and categorical data; we will use the processed data. The table III is a small dataset sample. the dataset was acquired from Kaggle [9]

For data preprocessing, the following methods were used:

- *Linear Interpolation*: this method was used to fill the missing values in the dataset's 'ca' and 'thal' features. We could have dropped these data rows, but since the dataset is small, we cannot afford to lose any data; hence, we opted for linear interpolation to fill in the missing values.
- *Normalisation*: since variations in the ranges of different dataset features exist, we have introduced Normalisation to bring the field values within the interval [0,1].
- *One-Hot Encoding [10]*: This method converts the target classes to 0 or 1, where 0 indicates the person has no heart disease and one indicates the person has a heart disease. This is to prepare the data for binary classification.

The dataset is highly imbalanced with a bias towards class 0, i.e., which is visualised from the Fig.1 heart disease

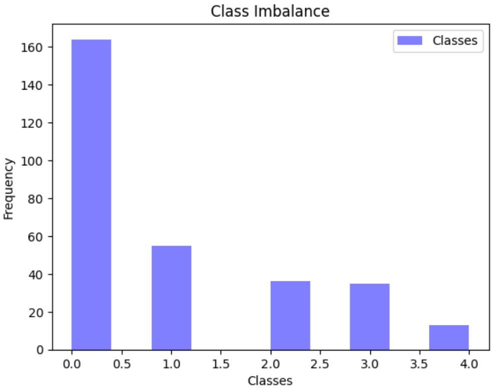

Fig. 1. Visualisation of class imbalance

The class imbalance issue was addressed by oversampling the data using the SMOTE [11] method. The class distribution after the oversampling is visualised in Fig.2

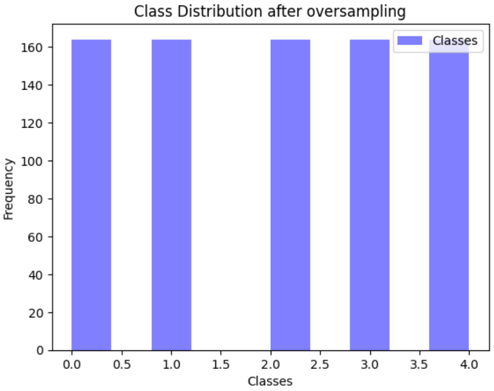

Fig. 2. Class Distribution After introducing Oversampled

To further gain insight into the data, a correlation matrix [12]. The correlation matrix is visualised in the Fig.3. Some of the points we can note are as follows:

- Age and trestbps (resting blood pressure): A positive correlation of 0.28 suggests that resting blood pressure also tends to increase as age increases.
- Cp (chest pain type) and thalach: There is a positive correlation, suggesting that different types of chest pain could be associated with higher maximum heart rates.
- Age and thalach (maximum heart rate achieved): A negative correlation indicates that the maximum heart rate achieved tends to decrease as age increases.

Although these insights are significant, we cannot imply them as causation, but we can use them to direct further how we approach the problem.

The work also considers the relation between age and having heart disease, although this is not implied from the correlation matrix but rather by visualising the available data in Fig.4. It is very evident from the graph that age contributes to understanding whether a person has a heart disease or not. It can be seen that there is a very high chance that a person above the age of fifty has a heart disease. The original dataset has been used for visualising Fig.3 and Fig.4 and not the dataset after introducing oversampling.

IV. METHODOLOGIES AND EXPERIMENTAL SETUPS

A. *Machine Learning Models*

For experimentation purposes, we have chosen three popular supervised learning methods:

- *Logistic Regression*: This method is easier to implement and faster to analyse when compared to other models.

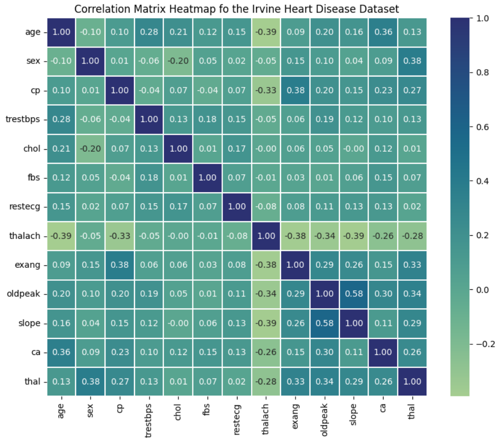

Fig. 3. The Correlation Matrix

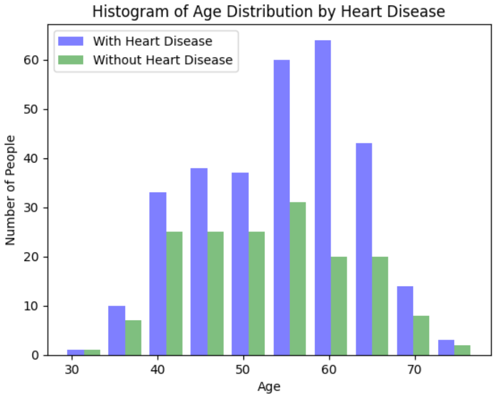

Fig. 4. Visualising the relation Between Age and Having a Heart Disease.

Also, most traditional classification problems that used logistic regression classifiers performed well in predictions.

- *KNN*: In KNN, the model does not rely on the underlying distribution of the data but rather focuses on its intuitive algorithm to classify classes based on naturally close neighbour data
- *SVM*: In the case of SVM, it can be used for both small and large datasets and is robust to overfitting.

Although there are many other classification models, our work will be restricted to these three models.

B. Methodologies

The dataset is split into train and test using the train_test_split() function from the Python [13] library scikit_learn [14]. The dataset is split at 0.2, i.e., 80% of the data will be used for training, and 20% of the data for testing, and the random seed is 42. this is the general data split for both binary and multiclass classifications.

For binary classification, the training strategy is straightforward: then the data is fed to the models, and the predictions are calculated on the test data after training. In binary classification, the class imbalance does not affect the efficiency of these models in their predictions. In a way, the One-Hot Encoding shadows the class imbalance issues as the total number of classes has been reduced to two now. In the case of Multiclass classification, we opt for two training strategies:

1) Training the dataset will class imbalance
2) Training the dataset after introducing oversampling to eliminate class imbalance.

The reason for opting for such a strategy is to see how class imbalance affects the predicting capability of a model.

V. Observation and Inference

To comprehend the predicting capability of the models, we will be using the following evaluation metrics:

- Accuracy
- Precision [15]
- Confusion Matrix [16]

All these evaluation metrics were implemented using the built-in function of scikit-learn.

A. Binary Classification

TableIII summarises the prediction results for the test data on binary classifications.

TABLE II
Accuracy and Precision on the Test and Train Data for Binary Classification

Classifier	Train Accuracy	Test Accuracy	Precision
Logistic regression	83.47%	88.52%	88.54%
KNN	86.77%	78.68%	83.23%
SVM	86.77%	78.68%	83.23%

As per expectations, all three classifiers could predict the test with almost good accuracy. The average accuracy of the three models is 88%, and the precision value is also around the same range. From this, we can infer that our traditional machine-learning models are exceptional for binary classification problems. This could be mainly due to the less complexity of the data. As there are only two prediction classes, the model complexity significantly decreases, making it easier for models to learn features. The below Fig.5 visualises the confusion matrix for the logistic regression. Refer to the Jupyter Notebook for the confusion matrix for the other model.

B. Multiclass Classification

For multiclass classification, the class imbalance issue has been a major hurdle in getting proper predictions. All three models could predict the test values with an accuracy of 55% on the test data. The values are summarised in the following Table.III After oversampling the dataset and training the classifiers again, we observed a slight improvement in the prediction capability of the KNN and SVM classifiers. Still, there was no improvement for the logistic regression classifier. While the KNN and SVM classifiers had an average accuracy of 67%, the logistic regression classifier had only 55% accuracy; the same applies to the precision values. The observations are summarised in the Table.IV below

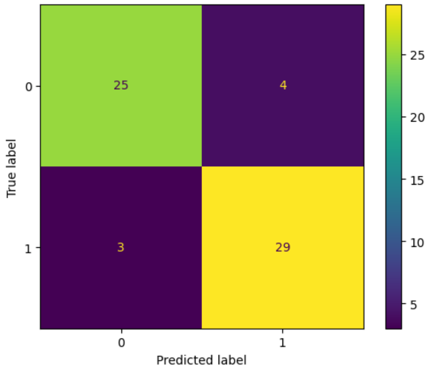

Fig. 5. Confusion Matrix for Logistic Regression

TABLE III
ACCURACY AND PRECISION ON THE TEST AND TRAIN DATA FOR MULTICLASS CLASSIFICATION

Classifier	Train Accuracy	Test Accuracy	Precision
Logistic regression	65.7%	52.4%	46.1%
KNN	68.5%	55.7%	46.5%
SVM	61.5%	60.6%	49.1%

Also adding the Visualisation of the confusion matrix for KNN classifier pre and post-oversampling in Fig.6 and Fig.7.

VI. CONCLUSION

In conclusion, the traditional classifying methods were able to predict with high accuracy for binary classification problems. Still, for multiclass classification, the classifiers were unable to perform well due to the class imbalance. Even after oversampling the data, the logistic regression classifier did not have much improvement compared to the other classifiers, and this may imply that given it is a simple design, the classifier is very much dependent on the size of the dataset on which it was trained to give better predictions. SVM and KNN classifier behaviour were in the way expected as these classifiers are intuitive and rely less on the underlying distribution of data.

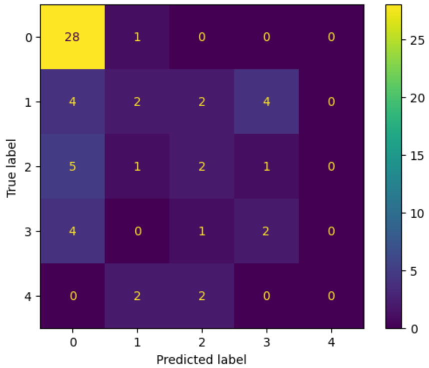

Fig. 6. Confusion Matrix of KNN Classifier Before Oversampling the data.

TABLE IV
ACCURACY AND PRECISION ON THE TEST AND TRAIN DATA FOR MULTICLASS CLASSIFICATION POST OVERSAMPLING THE DATA

Classifier	Train Accuracy	Test Accuracy	Precision
Logistic regression	56.5%	55.4%	54.7%
KNN	78.6%	69.5%	70.9%
SVM	72.1%	64.0%	64.4%

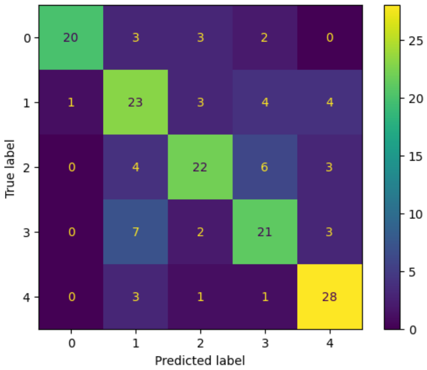

Fig. 7. Confusion Matrix of KNN Classifier After Oversampling the data.

The reason for not using undersampling as a method to eliminate class imbalance was because the dataset size is relatively small, and training classifiers on it can lead to underfitting. In future works, we can experiment with multi-model classifiers and different class imbalance methods.

REFERENCES

[1] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 20, no. 2, pp. 187–220, 1958.

[2] T. Cover and P. Hart, "Some methods of speeding up the computation of nearest neighbor," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[3] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[4] R. C. Detrano, A. Jánosi, W. Steinbrunn, M. E. Pfisterer, J.-J. Schmid, S. Sandhu, K. Guppy, S. Lee, and V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease." *The American journal of cardiology*, vol. 64 5, pp. 304–10, 1989. [Online]. Available: https://api.semanticscholar.org/CorpusID:23545303

[5] A. Yazdani, K. D. Varathan, Y. K. Chiam, A. W. Malik, and W. A. Wan Ahmad, "A novel approach for heart disease prediction using strength scores with significant predictors," *BMC medical informatics and decision making*, vol. 21, no. 1, p. 194, 2021.

[6] L. A. Sevastyanov and E. Y. Shchetinin, "On methods for improving the accuracy of multi-class classification on imbalanced data." *ITTMM*, vol. 20, pp. 70–82, 2020.

[7] M. Bhushan, A. Pandit, and A. Garg, "Machine learning and deep learning techniques for the analysis of heart disease: a systematic literature review, open challenges and future directions," *Artificial Intelligence Review*, vol. 56, no. 12, pp. 14 035–14 086, 2023.

[8] S. W. P. M. Janosi, Andras and R. Detrano, "Heart Disease," 1988, DOI: https://doi.org/10.24432/C52P4X.

[9] Kaggle. (2009) Kaggle: Your home for data science. [Online]. Available: https://www.kaggle.com/

[10] W. S. Robinson, "Ecological correlations and the behavior of individuals," *American Sociological Review*, vol. 15, no. 3, pp. 351–357, 1950.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[12] R. A. Fisher, “Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population,” *Biometrika*, vol. 10, no. 4, pp. 507–521, 1915.

[13] *Python*, Python Software Foundation, 1991. [Online]. Available: https://www.python.org/

[14] Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E., “Scikit-learn: Machine learning in python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: http://jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

[15] M. E. Maron and J. L. Kuhns, “On relevance, probabilistic indexing and information retrieval,” *Journal of the ACM*, vol. 7, no. 3, pp. 216–244, 1960.

[16] D. J. Spiegelhalter and R. P. Knill-Jones, “Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology,” *Journal of the Royal Statistical Society: Series A (General)*, vol. 147, no. 1, pp. 35–77, 1984.