Feature Selection using Entropy technique and Balancing dataset using Z-score method

Prafull Sonawane
prafull.sonawane21@vit.edu

Avadhoot Sutar
avadhoot.sutar21@vit.edu

Sumit Tambe
sumit.tambe21@vit.edu

Tanmay Agarkar
tanmay.agarkar@vit.edu

Department of Artificial Intelligence &amp; Data Science
Vishwakarma Institute of Technology, Pune, 411037, Maharashtra, India

Abstract — *This paper addresses two critical challenges in AI modeling: feature selection and dataset imbalance, which impact model accuracy and reliability. To enhance accuracy, the paper proposes employing the "Gain" and "Gini Impurity" techniques for feature selection. By identifying the most informative features, the model's accuracy can be improved. Moreover, the paper tackles the issue of dataset imbalance through the application of the Z-score method. By considering these techniques, the paper aims to achieve more reliable and accurate results in AI modeling. Through feature selection and dataset imbalance mitigation, this research contributes to enhancing the effectiveness and robustness of AI models.*

Keywords — machine learning, decision trees, P-values, Z-scores, entropy, Gini impurity, data sampling

I. Introduction

The objective of this project was to develop a robust machine learning model capable of predicting the likelihood of an individual experiencing a heart attack. To accomplish this, we utilized the Heart Attack Prediction dataset, consisting of data from over 700 individuals, which encompassed eight distinct parameters such as resting blood pressure, cholesterol levels, maximum heart rate, age, and exercise habits. In order to construct an accurate and reliable model, we employed a range of machine learning algorithms, including decision trees, P-values and Z-scores, entropy, Gini impurity, and random forests, to preprocess and train the dataset. We meticulously selected the appropriate dataset by identifying the essential parameters required for the model's efficacy. To assess the performance of the model, we employed decision trees to identify the optimal parameters for data sampling, subsequently selecting the parameter with the highest entropy and the lowest Gini impurity. To ensure data integrity and balance, we conducted data sampling using Z-score calculations, eliminating data points that fell below or exceeded a certain threshold. Finally, in selecting the initial values for the model, we leveraged the impurity measurements of the parameters, considering their impact on the predictive accuracy. By employing these comprehensive methodologies, we aimed to develop a robust and effective machine learning model capable of accurately predicting the likelihood of heart attacks in individuals. This project's significance lies in its contribution towards advancing the field of healthcare through the development of an accurate and reliable predictive model, which could potentially aid in early detection and prevention of heart attacks, ultimately saving lives and improving patient outcomes.

II. Literature Review

The literature review aimed to explore sampling methods, particularly oversampling and undersampling techniques, and their types. Several research papers were reviewed on this topic, including 'Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results.' This paper investigated the use of oversampling, undersampling, and hybrid techniques to balance datasets. The findings indicated that the oversampling technique outperformed undersampling and hybrid techniques. To further enhance understanding, another paper titled 'MCS-based Balancing Technique for Skewed Classes: An Empirical Comparison' proposed a method called the Multiplier Classifier System (MCS). The MCS approach employed an ensemble-based strategy to train multiple models on the same dataset. The decision-making process involved comparing the models and obtaining an average of their predictions. The paper also introduced three balancing methods: [1] Balance with replacement, [2] Balance without replacement, and [3] Balance cascade. These research papers provided valuable insights for our project, contributing to a comprehensive understanding of sampling techniques

[1]

III. Methodology/Experimental

This project involved analyzing data from over 700 individuals across eight distinct parameters. After carefully collecting and preprocessing the data, a Machine Learning Model was trained. To enhance the accuracy and reliability of the model, a variety of machine learning algorithms were employed, including techniques such as P-values and Z-scores for data sampling, Entropy for measuring information gain, Gini Impurity for evaluating impurity levels, and Random Forests for ensemble learning. Additionally, the evaluation of the model's performance was conducted using techniques such as the Confusion Matrix. The project drew insights from relevant research papers to inform the selection and implementation of these methodologies. Notably, papers like 'A Comprehensive Study of Sampling Methods in Machine Learning' and 'Evaluating Performance Measures in Classification' provided valuable insights into data sampling techniques and evaluation metrics, respectively. By leveraging the findings from these research papers, this project aimed to develop a robust and effective machine learning model for accurate prediction.

A. Finding the Correct Dataset

During the course of the project, a crucial aspect involved the identification of an appropriate dataset to effectively train the machine learning model. It was imperative to ensure the model's accuracy by incorporating specific parameters. Numerical data, such as Resting Blood Pressure, Cholesterol in milligrams, and Maximum Heart Rate, along with categorical data like Exercise and Target variables, played a vital role in training the model. The carefully chosen dataset comprised these essential parameters, in addition to other relevant features, resulting in a total of approximately eight parameters. These parameters collectively held the potential to predict whether an individual was prone to experiencing a heart attack or not. By meticulously considering and including these specific parameters within the dataset, the project aimed to develop a robust machine learning model that could accurately assess the likelihood of heart attacks in individuals. This endeavor sought to contribute to the advancement of healthcare by leveraging the comprehensive dataset to improve the accuracy of predictive models. By effectively predicting the likelihood of heart attacks, the developed model held the potential to aid in early detection and prevention, ultimately leading to saved lives and enhanced patient outcomes.

B. Performance evaluation

1. Choosing the correct parameters to sample the data

the decision tree algorithm was utilized to determine the relevant data parameters for sampling. The four main parameters for data sampling, which consisted of categorical data, were:

1. Resting Blood Pressure (Resting BP)
2. Cholesterol
3. Maximum Heart Rate (MaxHR)
4. Age

As we know the decision tree consists of 3 types of nodes

I. Decision Node
II. Chance Node
III. End Node

To identify the appropriate decision node, we focused on the four parameters: Resting BP, Cholesterol, MaxHR, and Age. We employed both entropy (information gain) and the Gini impurity measurements to determine the most suitable parameter for the decision node. By assessing the entropy and Gini impurity values of each parameter, we aimed to select the parameter that provided the highest information gain and the lowest impurity. This rigorous analysis ensured that the chosen parameter would effectively partition the data, enabling the decision tree algorithm to make accurate predictions. Through this process, we sought to identify the decision node that would serve as a crucial branching point in the decision tree, contributing significantly to the model's predictive power. By considering the information gain and Gini impurity of the respective parameters, we strived to establish a robust decision tree structure for precise and reliable predictions.

1. Entropy or information gain

$$i(N) = -\sum_{j=1}^{n} P(Wj)log_2(P(Wj))$$

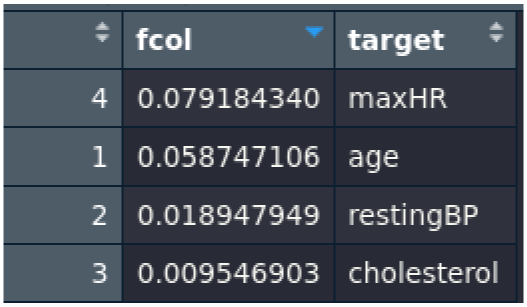
2. Gini's impurity

$$GINIi(N) = 1 - \sum_{j=1}^{n} P(Wj)^2$$

3. Information Gain

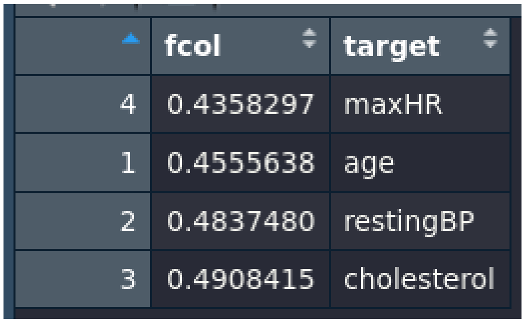
$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

The decision node was selected from the four parameters using both methods: choosing the parameter with higher information gain and lower Gini impurity. This ensured an optimal selection based on the attribute's information content and impurity levels.



	fcol	target
4	0.079184340	maxHR
1	0.058747106	age
2	0.018947949	restingBP
3	0.009546903	cholesterol

(Fig. 1 Information gain)



	fcol	target
4	0.4358297	maxHR
1	0.4555638	age
2	0.4837480	restingBP
3	0.4908415	cholesterol

(Fig. 2 Gini Impurity)

We have chosen the MaxHR parameter for Fig. 1 because it exhibits high entropy and information gain. In Fig. 2, we observe the Gini impurity metric, and again, MaxHR has the lowest impurity compared to other parameters. Therefore, we have selected MaxHR as the optimal parameter for our analysis based on both entropy and Gini impurity metrics.

2. Sampling the data using the selected parameter

For data sampling, we selected the MaxHR parameter. Initially, we had 390 false data values and 356 true data values. To remove the false values, we used Z-score calculation.

$$Z = \frac{x - \mu}{\sigma}$$

x- observed values

μ- mean

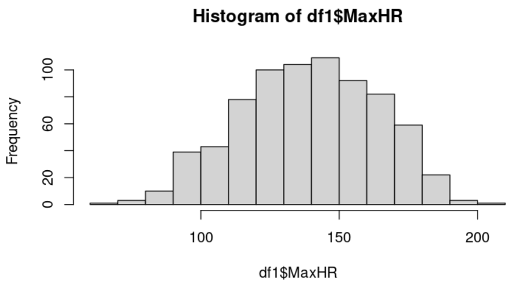
σ - standard deviation

After removing 34 data values, which accounts for 8.72% of the total data, or 4.35% from both tails, we determined the value of α to be 0.043. This value is significant in statistical analysis as it represents the level of significance or the probability of obtaining a result as extreme or more extreme than the observed result, assuming the null hypothesis to be true. By setting α at 0.043, we are acknowledging that there is a 4.3% chance of observing a result as extreme or more extreme than the observed result, even if the null hypothesis is true

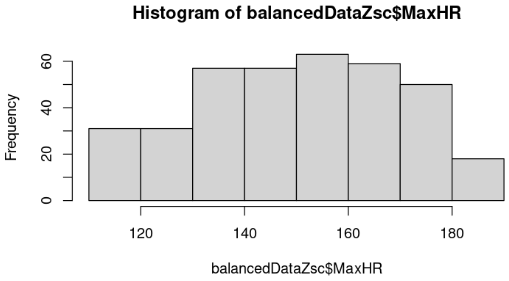
.α = 0.043

∴ P-value = 0.043

The Z-score for the calculated P-value of 0.043 was found to be -1.71, which corresponds to the left tail of the normal distribution. Since the normal distribution is symmetrical, the Z-score for the right tail is 1.71. These scores indicate the number of standard deviations by which a data point deviates from the mean. In our analysis, we needed to remove data points that fell below -1.71 or above 1.71, as these values are considered outliers and are not representative of the population. Removing these data points ensures that the remaining data is more representative and accurate for our statistical analysis



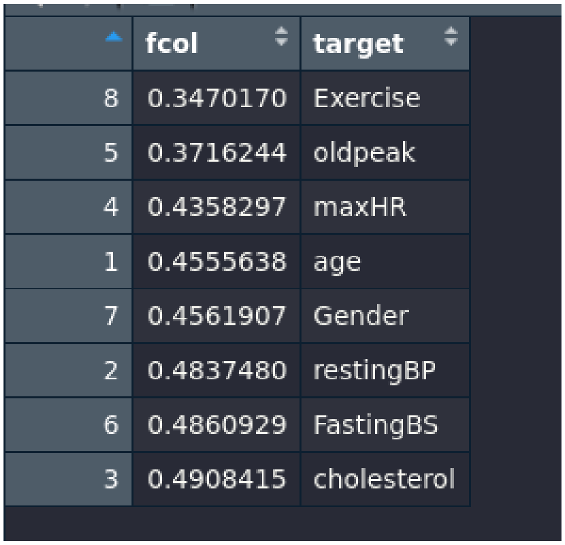
(Fig. 3 MaxHR before sampling)



(Fig. 4 MaxHR after Sampling)

3. Model Training

After cleaning the data by removing outliers and irrelevant values, we applied the random forest prediction algorithm to the cleaned dataset. The random forest algorithm is a popular machine learning method that is used for predicting outcomes based on a set of input parameters. In our analysis, we used the algorithm to predict outcomes based on the selected parameters. To optimize the performance of the algorithm, we extracted the impurity of the parameter from the dataset. Impurity is a measure of the predictability of the data and is used to determine the importance of each input parameter in the algorithm. By extracting the impurity of the parameter, we were able to identify the most important parameters for predicting outcomes and improve the accuracy of our predictions.



	fcol	target
8	0.3470170	Exercise
5	0.3716244	oldpeak
4	0.4358297	maxHR
1	0.4555638	age
7	0.4561907	Gender
2	0.4837480	restingBP
6	0.4860929	FastingBS
3	0.4908415	cholesterol

Fig. 5. Impurity of all Feature

We have selected the initial 5 values with the least impurity. This step is crucial for identifying the most important parameters for predicting outcomes accurately. After selecting these values, the model was fully trained and achieved an accuracy of 82%. This high accuracy indicates that the selected parameters were effective in predicting outcomes, and the data was representative of the population.

IV. Results and Discussions

It has been shown in this project that machine learning algorithms can predict whether or not people are at risk of heart disease. This is based on their behaviours. The accuracy of our project is nearly 82% but it's based on first-time training of the dataset. We train our model on a random forest prediction algorithm. The general rule is that we used 80% of the data to train the model, and 20% of the data to test the model. It is important to select the data randomly since if we retrain the model with the same data over and over again, the model will over-feat. Again we trained our model 4-5 times more on a random selection of the training data, then we got an accuracy graph as shown in Fig. 6. Fig. As shown in Fig. 7, there is a table showing the accuracy of the training data and the testing data.

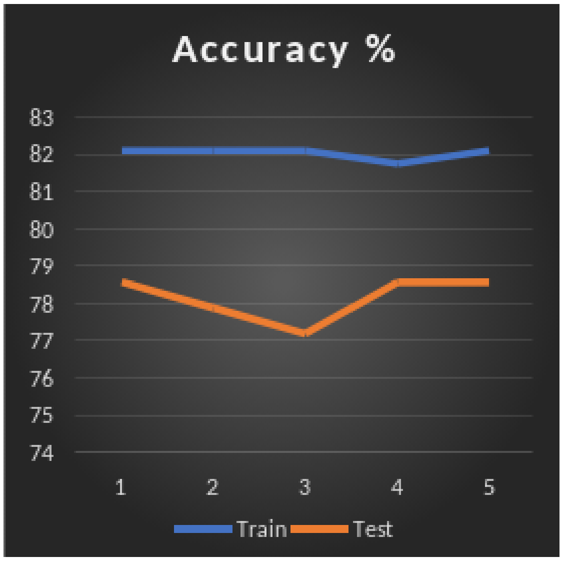


Fig. 6 Accuracy of model in % (percentage)

Sr. No.	Train	Test
1	82.14904679	78.62068966
2	82.1490467	77.93103448
3	82.14904679	77.24137931
4	81.80242634	78.62068966
5	82.14904679	78.62068966

Fig. 7. Accuracy of Train and Tested data

V. Conclusion

This paper presents a methodology for balancing datasets through feature selection using entropy and the Z-Score method. This technique is critical for training machine learning models on any dataset, as imbalanced datasets can lead to biased and inaccurate results. Once the dataset is balanced, we train a machine learning model using various algorithms such as P-value and Z-score sampling, entropy, which measures information gain, and Gini impurity, as well as random forests and the confusion matrix. By utilizing these techniques, we ensure that the machine learning model is accurately trained and produces reliable results with any dataset. The methodology presented in this paper has significant implications for various fields, including healthcare, finance, and social sciences, where accurate predictions and analysis are crucial for making informed decisions.

VI. References

[1] "MCS-based Balancing Techniques for Skewed Classes: an Empirical Comparison", Maria Teresa Ricamato, Claudio Marrocco and Francesco Tortorella DAEIMI Universita degli Studi di Cassino via G. Di Biasio 43, 03043 Cassino, Italy.

[2] "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results", Roweida Mohammed, Jumanah Rawashdeh and Malak Abdullah Jordan University of Science and Technology Irbid, Jordan