A Comparative Analysis of Data Mining Algorithms on CARDIOVASCULAR DATA (December 2023)

Jamyang Gelek Gurung, *Student, Kathmandu University*
Ngawang Choegyap Gurung, *Student, Kathmandu University*

***Abstract*—Data mining is a crucial process for extracting meaningful insights from vast datasets. This study conducts a comparative analysis of various data mining algorithms applied to cardiovascular data. The focus is on binary classification tasks employing algorithms such as Logistic Regression, RandomForest Classifier, Gaussian Naive Bayes Classifier, KNeighbors Classifier, and Support Vector Classifier (SVC).**

The investigation begins with a comprehensive exploration of the dataset, encompassing data preprocessing steps such as handling missing values, removing duplicates, and addressing outliers using the Interquartile Range (IQR) method. Attribute construction and discretization techniques are also applied to enhance the representational power of the dataset.The subsequent data exploration phase employs statistical summaries, box plots, distribution plots, correlation matrices, and count plots to gain insights into the dataset's characteristics.

Classification algorithms are then introduced, and their performances are rigorously evaluated using cross-validation techniques. Further analysis involves hyperparameter tuning to optimize the Logistic Regression model, enhancing its effectiveness in predicting cardiovascular outcomes. Additionally, a Neural Network model is introduced, demonstrating the growing influence of deep learning in predictive analytics.

The comparative analysis of classification algorithms provides valuable insights into their effectiveness for practitioners seeking effective models for binary classification tasks in the context of cardiovascular health analysis.

***Index Terms*—Binary Classification, Algorithms Comparison, Data Exploration, Data Cleaning, Cardiovascular Disease**

I. INTRODUCTION

DAta mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically [1].

The iterative procedures of Exploratory Data Analysis (EDA) and Data Preprocessing constitute essential steps in this process. Data preprocessing involves the methodical refinement of raw data, encompassing activities such as cleaning, transforming, reducing, integrating, and organizing the data into a structured and analytically useful format. Concurrently, in data exploration, we gain insights into the dataset through both graphical and non-graphical methods, unveiling relationships and discerning noteworthy patterns.

Data mining encompasses a range of tasks with key tasks including Classification, Prediction, Clustering, and Outlier Detection. These tasks collectively contribute to the overarching goal of data mining. The specific task chosen depends on the nature of the data and the objectives of the analysis. In the context of cardiovascular disease analysis, the objective involves binary classification. Specifically, we compare and analyze diverse classification algorithms to determine their effectiveness in accurately predicting outcomes related to cardiovascular health.

II. KNOWING YOUR DATA

The data-set, ”cardiovascular-disease-dataset” was acquired from Kaggle[2]. Notably, the data-set's singular format, presented as a CSV file, eliminates the necessity for data integration.

The data-set is structured with shape (70000, 13), indicating a total of 70,000 instances, each characterized by 13 attributes. Among these attributes, 7 are categorical, while 5 are numerical as the 'id' attribute is excluded from the data-set.

TABLE I
NUMERICAL AND CATEGORICAL ATTRIBUTES

Numerical Attributes	Categorical Attributes
age	gender
height	cholesterol
weight	gluc
ap_hi	smoke
ap_lo	alco
	active
	cardio

III. DATA PREPROCESSING

Data preprocessing, a crucial step is imperative for enhancing the quality of the data by addressing various issues associated with raw data such as inconsistencies, errors, and missing values, thereby laying the foundation for accurate and meaningful analyses. It also facilitates the identification and handling of outliers which are crucial to prevent them from unduly influencing the outcome of analyses and statistical models.

Various preprocessing steps, encompassing data cleaning activities such as removing redundancies and handling outliers, alongside data transformation tasks that involve attribute construction and discretization are carried out.

A. Missing and Duplicated Values

Upon analysis of the data, we verified the absence of missing values in all columns, ensuring the dataset's completeness and reliability. Additionally, we identified and removed 24 duplicated rows to maintain data integrity.

B. Handling Outliers

Outliers are data points that deviate significantly from the rest of the dataset, potentially influencing the results of analyses or models. With the help of data visualization tools like box plot, we identified the outliers which we were further identified with (IQR) Inter Quartile Range. The identified outliers were addressed by clipping method by setting predefined lower and upper bounds for specific numerical columns. Using pandas' 'clip' function, outlier values beyond these bounds are replaced. This approach preserves the overall data distribution, crucial for maintaining dataset integrity [3].

C. Attribute Construction

Attribute construction is performed in data preprocessing to enhance the representational power of the dataset by creating new features or attributes, derived from existing ones, to better capture complex relationships and patterns within the data.

The Body Mass Index (BMI) attribute is constructed by dividing the weight in kilograms by the square of the height in meters:

$$\mathrm{BMI} = \frac{\text{weight}}{\left(\frac{\text{height}}{100}\right)^2}$$

TABLE II
BMI

weight	height	bmi
62.0	168.0	22.0
85.0	156.0	34.9
64.0	165.0	23.5

D. Discretization

Discretization is carried out in data preprocessing to transform continuous numerical data into discrete intervals or categories, simplifying the data, reducing noise, and improving the interpretability of analyses and models that perform better on categorical or ordinal data.[1]

The age is categorized into Young($0\,\text{to}\,30$ years), Middle-aged ($31\,\text{to}\,45$) years, Senior ($46\,\text{to}\,60$ years) and Old: $61\,\text{to}\,100$) years. For this the age is first converted from days to years.

Whereas the blood pressure is categorized into three groups: Hypotension, Normal and Hypertension

IV. Data Exploration

Exploration can be graphical or non-graphical in nature and done with categorical or numerical attributes or both. It is done to reveal relationships, distributions, and patterns within the data.

A. Statistics Summary

Statistics Summary is a descriptive statistics that summarizes statistical information to get the gist of the information. It is used to find central tendency, statistical dispersion and the shape of the distribution.

TABLE III
Statistical Summary of Numeric Columns after Outliers Handling

	age	height	weight	ap_hi	ap_lo	bmi
count	69976.0	69976.0	69976.0	69976.0	69976.0	69976.0
mean	53.3	164.4	73.9	126.7	81.7	27.4
std	6.8	7.8	13.4	16.4	9.0	4.9
min	33.0	142.5	39.5	90.0	65.0	14.4
25%	48.0	159.0	65.0	120.0	80.0	23.9
50%	54.0	165.0	72.0	120.0	80.0	26.4
75%	58.0	170.0	82.0	140.0	90.0	30.2
max	65.0	186.5	107.5	170.0	105.0	39.7

B. Box Plot

Box plot or box and whiskers plot provides the visualization of the summary statistics.

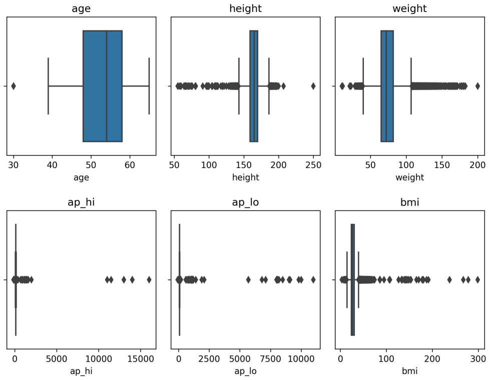

Fig. 1. Box-plot before outlier handling

From the box plot, the outliers such as negative values as well as very high values in columns "ap_hi" and "aph_lo" were identified. There also exists outliers in "height" and "weight" attributes as result outlier is also found in "bmi" attribute.

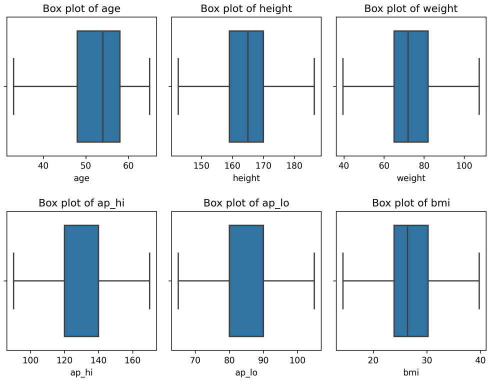

Fig. 2. Box-plot after outlier handling

In the figure 2, we can see that there is no outlier as it has been handled with clipping.

C. Distribution Plot

In distribution plot, histograms and kernel density estimator(kde) plot are done find the distribution of data such as whether the data are uniformly distributed or skewed on left or right side [4].

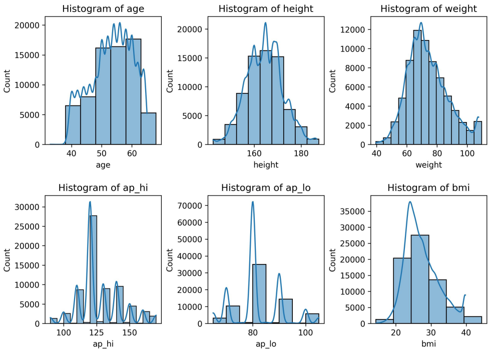

Fig. 3. Distributions of numeric attributes

In the figure 3 we can see the distribution of various numeric attributes and can see that they follow distribution similar to normal distribution.

D. Correlation Matrix

Pairwise correlations are assembled into the matrix to find statistical relationships between the attributes of the dataset.

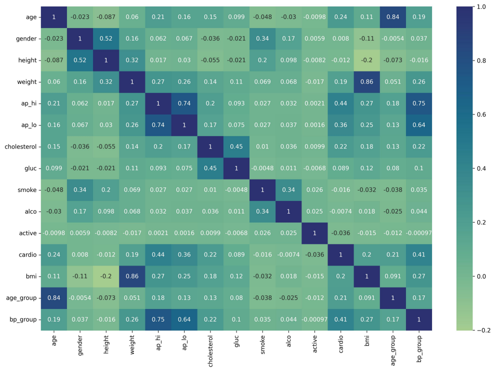

Fig. 4. Heat-map of Correlation Matrix

From the figure 4 we can see that the target attribute 'cardio' is strongly positively related to the "ap_hi" which is systolic pressure. Another interesting relation is the positive correlation of glucose and cholesterol level.

E. Count Plot

Count-plot explicitly counts the number of occurrences of the categories of a categorical attribute.

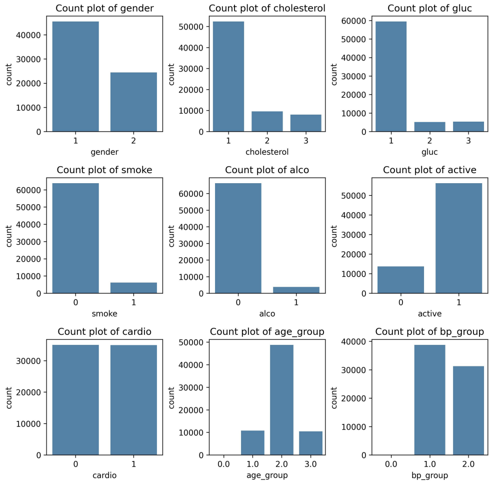

Fig. 5. Count-plot for all the categorical attributes

From the figure 5 we can see that class label 'cardio' has balanced ratio of yes and no cardio-vasular disease.

V. Classification Algorithms

Classification in data mining entails assigning predefined labels to instances based on their attributes. Utilizing a labeled training data-set, algorithms discern patterns and relationships between features and the target class. Classification finds applications in domains such as spam filtering, medical diagnosis, credit scoring, and image recognition, establishing its indispensability for decision-making across diverse fields [5].

In the classification task, 'cardio' is a binary attribute taken as target class (y) and the reaming attributes are taken as input features (X). The classification algorithms that are employed include:-

1. Logistic Regression

Logistic Regression is a statistical technique designed for binary classification problems, predicting the probability of an instance belonging to one of two classes. Its popularity in professional applications stems from its interpretability, computational efficiency, and simplicity. By utilizing the sigmoid function, logistic regression provides probability estimates and is particularly well-suited for scenarios where the outcome is binary.

2. RandomForest Classifier

RandomForest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees.

3. Gaussian Naive Bayes Classifier

The Gaussian Naive Bayes classifier is a probabilistic model based on Bayes' theorem. It assumes that features are conditionally independent given the class label and follows a Normal distribution.

4. KNeighbors Classifier

The KNeighbors classifier is an instance-based, distance-based learning algorithm where the function is approximated locally. It classifies a data point based on the majority class among its k-nearest neighbors.

5. Support Vector Classifier (SVC)

The Support Vector Classifier is a classification algorithm that seeks to find a hyperplane that best separates classes. It works well in high-dimensional spaces and is effective for both linear and non-linear boundaries through the use of kernels.

VI. Model Evaluation: Cross Validation

Cross-validation divides the data-set into training and testing sets, training the model on the former, and validating it on the latter, repeating this process multiple times. Cross-validation helps in comparing models and choosing the most suitable one for a specific predictive task, offering a powerful tool with lower bias compared to alternative methods. Various cross-validation techniques exist, including Hold-out, K-folds, Leave-one-out, and Stratified K-folds, each with its application depending on the nature of the data [6].

Model performance is rigorously assessed through metrics like accuracy (AC), precision, recall, f-1 score. Mitigating risks of overfitting and underfitting, along with hyperparameter tuning, ensures optimal model efficacy.

A. Train Test Split

In Train-test split or hold out cross validation, the data-set is divided into training and testing set. In our specific instance, the test size is designated as 0.2, signifying an 80% allocation for training data and a 20% allocation for testing data.

B. K-fold Cross Validation

k-Fold cross-validation mitigates the drawbacks of the hold-out method by dividing the dataset into k folds, training the model on k-1 folds, and validating it on the remaining fold in each iteration, allowing for multiple evaluations across the entire dataset and providing a more robust assessment of model performance.

TABLE IV
K-fold Cross Validation Score of Classifiers Models

Model	AC	Precision	Recall	F1Score	ROC-AUC
Logistic	0.7255	0.7272	0.7255	0.7249	0.7253
Random Forest	0.7115	0.7116	0.7115	0.7114	0.7114
Gaussian NB	0.7226	0.7243	0.7226	0.7220	0.7224
KNN	0.6896	0.6899	0.6896	0.6894	0.6895
SVC	0.7223	0.7297	0.7223	0.7198	0.7219

C. Hyperparameter tuning

Due to resource constraints, we initially employed cross-validation to identify the best model using default hyperparameters. Subsequently, hyperparameter tuning was exclusively performed on the selected optimal model, which was found to be logistic regression.

The grid search of hyperparameters, using a 5-fold cross-validation, utilizing a Logistic Regression model achieved a best score of 0.726. The optimal hyperparameters for this result were found to be a regularization strength (C) of 10 and a penalty term of ’l1’.

VII. Neural Network

Neural networks, or artificial neural networks (ANNs), form a subset of machine learning central to deep learning. Inspired by the human brain, they consist of layers of nodes—input, hidden, and output. Each node, or artificial neuron, has connections with associated weights and thresholds. Activation occurs when a node’s output exceeds the threshold, facilitating data transmission to the next layer. Neural networks improve accuracy over time through training data [7].

Our neural network is a sequential model with three layers: 32-unit input layer (ReLU), 16-unit hidden layer (ReLU), and a single-unit output layer (sigmoid) for binary classification. Compiled with Adam optimizer, binary crossentropy loss, and accuracy metric, it incorporates early stopping (10 epochs patience). Trained for 100 epochs on data (batch size 64), it achieves a test set accuracy of 0.729.

VIII. Conclusion

In conclusion, our comparative analysis of data mining algorithms on cardiovascular data reveals distinct strengths among Logistic Regression, RandomForest, Naive Bayes, KNeighbors, and Support Vector classifiers. Hyperparameter tuning enhances the predictive power of Logistic Regression, and the introduction of a Neural Network highlights the evolving landscape of predictive analytics. Each algorithm contributes valuable insights, providing practitioners with diverse tools for effective binary classification tasks in cardiovascular health analysis.

References

[1] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques.* 3rd ed., Elsevier, 2012.

[2] “Cardiovascular Disease dataset,” Kaggle. 20-Jan.-2019. [Online]. Available: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset. [Accessed: 5-Dec.-2023].

[3] S. Chaudhary, “Why “1.5” in IQR Method of Outlier Detection?”, Towards Data Science. 28-Sept.-2019. [Online]. Available: https://towardsdatascience.com/why-1-5-in-iqr-method-of-outlier-detection-5d07fdc82097. [Accessed: 18-Dec.-2023].

[4] “An introduction to seaborn — seaborn 0.12.2 documentation,” [Online]. Available: https://seaborn.pydata.org/tutorial/introduction.html. [Accessed: 25-Sept.-2023].

[5] Ian H. Witten, Eibe Frank, Mark A. Hall, *Data Mining Practical Machine Learning Tools and Techniques:* 3rd ed. Elsevier, 2012.

[6] V. Lyashenko, “Cross-Validation in Machine Learning: How to Do It Right,” 21-July-2022. [Online]. Available: https://neptune.ai/blog/cross-validation-in-machine-learning-how-to-do-it-right. [Accessed: 18-Dec.-2023].

[7] “What are Neural Networks? — IBM,” [Online]. Available: https://www.ibm.com/topics/neural-networks. [Accessed: 17-Dec.-2023].