Griffin Bjerke
University of Massachusetts Lowell
griffin_bjerke@student.uml.edu

Matt Serdukoff
University of Massachusetts Lowell
matthew_serdukoff@student.uml.edu

Predicting Cardiovascular Disease Using Machine Learning

***Abstract*— This research explores innovative approaches for cardiovascular disease detection using machine learning techniques. Four distinct notebooks are presented, each showcasing a unique methodology. The first notebook employs a Decision Tree algorithm, demonstrating its efficiency in classifying cardiovascular disease based on patient attributes. The second notebook introduces Linear Regression, emphasizing its predictive capabilities in the context of cardiovascular health. The third notebook delves into Logistic Regression, showcasing its accuracy and F1 score as evaluation metrics. The fourth notebook explores a Naive Bayes approach, both Gaussian and categorical, revealing promising results in disease prediction. The research collectively aims to contribute diverse insights to the evolving landscape of cardiovascular disease detection methodologies.**

Introduction

Cardiovascular diseases pose a significant global health threat, necessitating advanced and accessible diagnostic tools. In response, this paper investigates novel methodologies for cardiovascular disease detection, leveraging machine learning algorithms. Building on the success of existing screening methods, such as Decision Trees, Linear Regression, Logistic Regression, and Naive Bayes, our research aims to enhance accuracy, affordability, and accessibility in disease prediction. By combining innovative techniques with traditional approaches, we seek to contribute to the ongoing efforts in preventing cardiovascular mortality.

In this exploration, the Decision Tree notebook employs a tree-based model to discern patterns in patient data, offering transparency in decision-making. Moving beyond, the Linear Regression notebook introduces predictive modeling, emphasizing its potential for precise cardiovascular risk estimation. The Logistic Regression notebook extends the analysis, emphasizing the importance of accuracy and F1 score in disease classification. Additionally, the Naive Bayes notebook explores both Gaussian and categorical approaches, highlighting their effectiveness in probabilistic disease prediction.

Through these diverse methodologies, our research aims to provide a comprehensive understanding of their strengths and limitations. Furthermore, we explore the integration of low-cost technologies, such as mmWave radar sensor arrays, coupled with deep learning, to propose an alternative for accurate and affordable cardiovascular screening. This interdisciplinary approach seeks to bridge gaps in healthcare accessibility, particularly in resource-constrained settings.

In the subsequent sections, we delve into related works, presenting a contextual overview of advancements in medical imaging, predictive modeling, and cost-effective screening solutions. By contextualizing our research within the existing landscape, we aim to contribute meaningfully to the ongoing discourse on cardiovascular disease detection.

- *Related works*

Innovations in Machine Learning for Disease Prediction:

The application of machine learning algorithms has revolutionized disease prediction and risk assessment. Decision Tree algorithms, as demonstrated in this research, showcase the ability to discern intricate patterns within patient datasets, aiding in effective disease classification. Moreover, Linear Regression models contribute to precise risk estimation, offering valuable insights into cardiovascular health. Logistic Regression techniques, emphasizing accuracy and F1 score, further contribute to the evolving landscape of disease prediction.

Exploration of Bayesian Approaches:

Bayesian methodologies, such as Naive Bayes, have gained attention for their probabilistic approach to disease prediction. In the context of cardiovascular health, both Gaussian and categorical implementations offer promising results. These approaches leverage the conditional probabilities of observed features, enabling efficient and interpretable predictions. The research community is actively exploring Bayesian techniques to enhance the robustness of disease prediction models.

- *Description:*

1. Preprocessing:

- The preprocessing phase plays a crucial role in ensuring the quality and reliability of the dataset used for cardiovascular disease detection.

CardioDT:

- Jupyter notebook file for the Decision Tree Model

CardioLinReg:

- Jupyter notebook file for the Linear Regression Model

CardioLOGREG:

- Jupyter notebook file for the Logistic Regression Model

CardioNB:

- Jupyter notebook file for the Naive Bayes Model

2. Database Details:

- The cardiovascular disease dataset used in this research serves as the foundation for model development and evaluation. The dataset comprises diverse features, including patient demographics, medical history, and diagnostic indicators. Each entry is associated with a binary target variable indicating the presence or absence of cardiovascular disease. The dataset is split into training and testing sets to facilitate the model development and evaluation process.

3. Algorithm Details:

Decision Tree Classifier

- Implemented from scratch in CardioDT, the decision tree classifier is used for its ability to uncover decision boundaries based on select feature values.

Linear Regression:

- Implemented from scratch in the notebook CardioLinReg, this algorithm focuses on establishing relationships between input features and the associated target variable.

Logistic Regression:

- Implemented from scratch in the notebook CardioLOGREG, this algorithm is used to focus on binary classification. This is used to emphasize the importance of accurate predictions and F1 score.

Naive Bayes:

- Implemented from scratch in the notebook CardioNB, used as a classifier applied to the test data so evaluation metrics such as the confusion matrix and F1 score can be computed.

4. Training and Testing Techniques:

- The dataset is split into training and testing sets to assess model performance accurately. The training set is used to teach the algorithms the underlying patterns in the data, enabling them to make predictions. The testing set, distinct from the training set, serves as an independent dataset to evaluate the model's generalization performance. Techniques such as cross-validation may be employed during training to enhance model robustness. Model evaluation metrics, including accuracy, F1 score, and confusion matrices, provide a comprehensive understanding of the algorithms' predictive capabilities.

Experimental Evaluation

1. Library Details:

- The experimental evaluation leverages popular data science and machine learning libraries to implement and assess the performance of the cardiovascular disease detection models. Key libraries include:

NumPy and Pandas:

- Used for data manipulation, handling, and exploration. NumPy provides efficient numerical operations, while Pandas facilitates data manipulation and preprocessing.

Matplotlib and Seaborn:

- Employed for data visualization and result representation. Matplotlib provides versatile plotting capabilities, and Seaborn enhances the aesthetics of visualizations.

Scikit-learn:

- A comprehensive machine learning library used for implementing and evaluating various algorithms. Specific modules, such as DecisionTreeClassifier and LinearRegression, are utilized for model development, and evaluation metrics from sklearn.metrics aid in assessing model performance.

2. Available Codes and References:

- The implementation builds upon foundational codes and references, ensuring reliability and adherence to best practices:

Decision Tree Classifier:

- This model splits the data into branches based on the specific features such as resting heart rate or fasting blood sugar. Each branch represents an outcome of a test on a feature.

Linear Regression:

- Linear Regression models the relationship between a dependent variable and independent variable(s) by means of fitting a linear equation to fit the data. This equation is used to predict the likelihood of the desired outcome.

Logistic Regression:

- Logistic Regression takes a binary approach to the statistical analysis of data, meaning that 0/1 is equal to yes/no. Using a logistic function, it models a binary dependent variable.

Naive Bayes:

- Naive Bayes uses Bayes' theorem with the 'naive' assumption for independence between every pair of features being analyzed.It is primarily used for classification.

3. Results Details:

- The experimental results provide insights into the efficacy of each model in cardiovascular disease detection. Key result details include:
- Naive Bayes
  - Accuracy: 95.5%
  - f1 score: 0.96
- Linear Regression
  - Accuracy: 85%
  - f1 score: 0.85

- Logistic Regression
    - Accuracy: 91.5%
    - f1 score: 0.93
- Decision Tree
    - Accuracy: 95.5%
    - f1 score: 0.96

Decision Tree Classifier:

- The accuracy of the Decision Tree Classifier is evaluated on the test set, showcasing the model's ability to discern patterns in the dataset. The decision tree structure, along with feature importance, provides interpretability.

Linear Regression and Logistic Regression:

- Results include coefficients, intercepts, and evaluation metrics such as accuracy and F1 score. These metrics quantify the models' predictive performance and reveal their effectiveness in capturing underlying relationships.

Naive Bayes:

- Gaussian Naive Bayes results encompass confusion matrices, accuracy, and F1 score. These metrics reflect the probabilistic nature of the Naive Bayes algorithm and its ability to handle continuous and categorical features.

Limitations and Future Works:

This research was not without its limitations, as we did not implement a neural network, which would have likely yielded the highest accuracy for cardiovascular disease classification. Additionally, we desired to implement two different Naive Bayes models, one of which did not work.

Conclusion:

This research explored different approaches for cardiovascular disease detection using machine learning techniques. Four distinct methodologies were used and each demonstrated strength in its task. The decision tree algorithm showcased its brawn by classifying cardiovascular disease based on various patient attributes. Linear Regression introduced predictive modeling and precise risk estimation while Logistic Regression took the binary approach to classification. Naive Bayes also showed promising accuracy in predicting heart disease. Our work lies within the broader context of advancements in machine learning for disease prediction, highlighting the importance of the aforementioned models. In conclusion, this research contributes to the many ongoing efforts to combat disease by implementing new technologies and innovative solutions which will drive more accurate, affordable, and accessible methods for detecting cardiovascular disease in humans.

References

K. G. Dinesh, K. Arumugaraj, K. D. Santhosh and V. Mareeswari, "Prediction of Cardiovascular Disease Using Machine Learning Algorithms," *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, Coimbatore, India, 2018, pp. 1-7, doi: 10.1109/ICCTCT.2018.8550857.

N. S. Rajliwall, R. Davey and G. Chetty, "Machine Learning Based Models for Cardiovascular Risk Prediction," 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), Sydney, NSW, Australia, 2018, pp. 142-148, doi: 10.1109/iCMLDE.2018.00034.