HEART DISEASE PREDICTION

A LOGISTIC REGRESSION APPROACH

AWELE OKOLIE

SCHOOL OF COMPUTING AND DATA SCIENCE

WENTWORTH INSTITUTE OF TECHNOLOGY

BOSTON, MASSACHUSETTS, UNITED STATES

<mailto:okoliea@wit.edu>

ABSTRACT

The goal of this project is to predict the probability of heart disease based on certain health-related factors, such as age, gender, cholesterol levels, and blood pressure. Using logistic regression, we trained a machine learning model to classify individuals as either having heart disease or not, utilizing the Heart Disease Dataset from Kaggle. The model was trained on a subset of the data and evaluated on a separate test set, achieving an accuracy of 85.24% on the training data and 80.49% on the test data. These results suggest that machine learning can effectively assist healthcare professionals in the early detection of heart diseases.

KEYWORDS

Logistic Regression, Prediction, Machine Learning, Healthcare, Heart Disease.

1 Introduction

Heart disease is among the leading causes of death worldwide. Early detection provides an opportunity for timely medical intervention, thereby improving survival rates [1]. This research aims to create a predictive model capable of identifying whether an individual has heart disease based on health-related parameters such as age, blood sugar, cholesterol levels, blood pressure, and lifestyle factors. By leveraging logistic regression, a machine learning technique, this study seeks to identify individuals at risk and assist healthcare professionals in making more informed decisions [2].

Recent studies have demonstrated the potential of machine learning algorithms, including logistic regression, in detecting heart disease more effectively than traditional risk assessment methods. For instance, integrating multiple health factors into a model has been shown to enhance prediction accuracy and improve risk assessment [3]. Additionally, data-driven models are associated with improved decision-making by healthcare professionals, enabling earlier interventions and better patient outcomes [4].

Despite these advancements, challenges remain, particularly concerning data quality, model interpretability, and the integration of machine learning models into clinical practice.

The primary goal of this research is to answer the following question: Can we predict whether a person will develop heart disease based on data about age, cholesterol, exercise, and other factors?

2 Data

2.1 Source of dataset

For this project, we used the [Heart Disease Dataset](https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset) on Kaggle. We are given information about the patients which includes age, gender, cholesterol levels, and whether they have heart disease or not in this dataset. This is a widely used dataset within the scope of machine learning making it a credible source for heart disease prediction. It was generated through several clinical studies that are publicly available for education. This dataset is updated often, and the version that was used for this project is the latest version available which was released June 2019.

2.2 Characters of the datasets

The [Heart Disease Dataset](https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset) from Kaggle is a popular dataset in machine learning for heart disease classification. It consists of 14 attributes (including the target variable) and contains 1,025 instances (rows). The dataset provides health-related information about patients and includes the target variable indicating the presence of heart disease.

The table below provides a detailed description of the dataset's attributes:

|  |  |  |
| --- | --- | --- |
| Column Name | Description | Unit/Range |
| Age | Age of the patient | Years |
| Sex (Gender) | Gender of the patient | Male = 1  Female = 0 |
| Chest Pain Type (cp) | chest pain types of the patient | 0: The patient does have no chest pain  1: Typical Angina (Pain caused by insufficient blood flow to the heart)  2: Atypical Angina (pain not linked to physical stress or typical heart disease symptoms)  3: Non-Angina (pain caused by factors like muscle strain or gastrointestinal issues) |
| Resting Blood Pressure (trestbps) | Blood pressure when resting | mmHg (94-200) |
| Serum Cholesterol (chol) | Cholesterol level in the blood | mg/dl (126-564) |
| Fasting Blood Sugar (fbs) | Blood sugar level after fasting (≥ 120 mg/dl indicates high blood sugar) | 1 = >120 mg/dl, 0 = <120 mg/dl |
| Resting Electrocardiographic Results (restecg) | It measures the electrical activity of the heart | 0 = Normal  1 = Abnormality  2 = Possible Heart Enlargement |
| Maximum Heart Rate Achieved (thalach) | Maximum heart rate achieved during exercise | bpm (71-202) |
| Exercise Induced Angina (exang) | Presence of chest pain on exercise | 1 = Yes  0 = No |
| Old peak (depression) | It checks how the heart handles stress from physical activity | Numeric (depression level) |
| Slope of the Peak Exercise ST Segment (slope) | Slope of the ST segment in the ECG during exercise | 0 = Normal  1 = Flat  2 = Downsloping |
| Number of Major Vessels Colored by Fluoroscopy (ca) | Number of visible major vessels in fluoroscopy | 0-3 (0 = No vessels, 1-3 =Number of visible vessels) |
| Thalassemia (thal) | Blood condition of the heart | 1 = Normal  2 = Fixed defect  3 = Reversible defect |
| Target | Presence of heart disease | 1 = Disease   1. = No Disease |

The dataset was well-structured and required no additional cleaning. All values were appropriately scaled and ready for analysis. No new categories were created, and the attributes were used in their original format.

3 Methodology

This section provides an overview of the methodology employed for predicting the likelihood of heart disease. The project utilizes Logistic Regression, a supervised learning model commonly used for binary classification tasks, along with relevant Python libraries for data analysis and model building.

3.1 Logistic Regression Model

Logistic Regression is a supervised model used for binary classification tasks. It is used when our dependent variable is categorical. In this case, we are checking to see if a person has heart disease (1) or not (0). The model estimates the log-odds of the target variable as a linear combination of the input features, and it adjusts based on these calculations.

Assumptions of Logistic regression model

* Linearity: The relationship between the independent variable and the log-odds must be linear.
* Observations: Logistic regression requires that the observations are independent of each other.
* No multicollinearity: The independent variables should not be too highly correlated with each other.

Advantages of Logistic regression model

* Simplicity: It is easy to interpret and understand than other machine learning models.
* Efficient: The underlying mathematical models are efficient and relatively simple to optimize, making logistic regression a robust and practical choice for many applications.
* Probabilistic Output: Linear regression models give you an output as probability telling us how well the model performed.

Disadvantages of Logistic regression model

* Assumes Linear relationship: Logistic regression models assume linear relationship between the predictor variables and the estimated odds which may not always be true for more complex datasets.
* Prone to Overfitting: If the number of observations is fewer than the number of features.
* Limited to binary outcomes: Logistic regression is limited to predicting only binary classification tasks and might not be suitable for scenarios with non-binary outcomes without modifications.

Logistic Regression was used for this project because it is a straightforward, easy to understand and interpret model used for binary classification. As our target variable is binary, that is, presence or absence of heart disease, Logistic Regression is suitable for the given problem to give a good and understandable model for predicting the occurrence of heart disease given some health factors available.

3.2 Tools and Libraries Used:

Data analysis and machine learning were done using Python. The following libraries were used:

* Pandas: It is for handling and manipulating the dataset.
* NumPy: For numerical operations.
* Scikit-learn: For machine learning and using it to build our logistic regression model.
* Matplotlib and Seaborn: To create charts and plots so we can better understand the data.

3.3 Data Preparation

The dataset preparation process involved the following steps:

The dataset was divided into two parts:

* Features (X): Age, cholesterol, and other health related factors.
* Target (Y): Presence or absence of heart disease (1 = Yes, 0 = No).

The dataset was divided into a training set (80%) and a testing set (20%), to determine how well our model will do on new unseen data.

**4. RESULTS**

This section presents the outcomes of the Logistic Regression model applied to predict heart disease, including model performance metrics, analysis of results, and key visualizations.

**4.1 Model Performance**

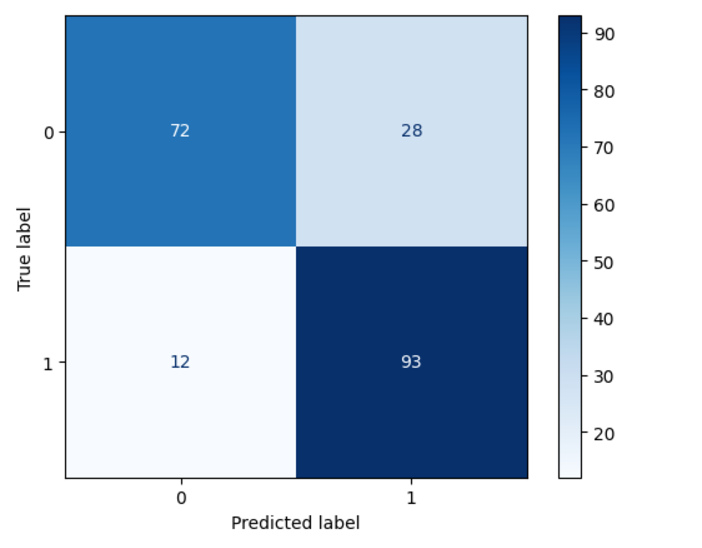
The Logistic Regression model was trained and tested on the dataset to predict whether an individual has heart disease. The following results were obtained:

* **Accuracy on Training Data:** The model achieved an accuracy of 85.24% on the training dataset.
* **Accuracy on Test Data:** The model achieved an accuracy of 80.49% on the test dataset.

The relatively high accuracy on the test dataset suggests that the model generalizes well to unseen data, making it suitable for predictive tasks in real-world scenarios.

**4.2 Confusion Matrix Analysis**

To further evaluate the model's performance, a confusion matrix was generated to analyze the distribution of correct and incorrect predictions. The confusion matrix is shown below:



* **True Positives (TP):** 93 instances where the model correctly predicted heart disease.
* **True Negatives (TN):** 72 instances where the model correctly predicted no heart disease.
* **False Positives (FP):** 28 instances where the model incorrectly predicted heart disease.
* **False Negatives (FN):** 12 instances where the model failed to predict heart disease.

The confusion matrix reveals that the model performs well overall, with a strong balance between sensitivity (recall) and specificity.

**4.3 Evaluation Metrics**

To provide a more detailed evaluation of the model’s predictive performance, precision, recall, and F1-score were calculated:

* Precision: TP/TP+FP = 93/ 93+ 28 = 76.86%
* Recall: TP/TP+FN = 93/93+12 = 88.57%
* F1-Score: 2(precision \* recall)/ (precision + recall) = 82.28%

These metrics demonstrate the model’s effectiveness in correctly identifying individuals with heart disease while maintaining a reasonable balance of false positives and negatives.

**4.4 Analysis of Results**

The slight difference in accuracy between the training (85.24%) and test (80.49%) datasets indicates that the model does not overfit and is capable of generalizing to new data. Logistic Regression proved effective in this context due to:

* **Feature Relevance:** The health-related parameters (e.g., age, cholesterol, blood pressure) are strong predictors of heart disease.
* **Model Simplicity:** Logistic Regression’s interpretability aids in understanding how each factor contributes to predictions, which is valuable in healthcare applications.

**4.5 Other Models Considered**

Other algorithms, such as Decision Trees and Support Vector Machines (SVM), were considered during the initial phase of the project. However, Logistic Regression was ultimately chosen for the following reasons:

* **Interpretability:** Its coefficients provide insight into the significance of individual features.
* **Efficiency:** Logistic Regression is computationally efficient and well-suited for small to medium-sized datasets.
* **Performance:** Its balanced predictions reduced the risk of overfitting compared to more complex models.

While exploring more advanced algorithms could enhance performance, Logistic Regression was appropriate for this study due to its simplicity, interpretability, and alignment with the study’s goals of creating actionable insights for healthcare professionals.

5 Discussion

This section explores the implications of the findings, limitations encountered during the study, and prospects for future research.

**5.1 Implications of the Results**

The results demonstrate that Logistic Regression is a reliable method for predicting heart disease based on health-related parameters such as age, cholesterol levels, and blood pressure. The high accuracy scores (85.24% on training data and 80.49% on test data) suggest that the selected features are significant predictors of heart disease. These findings align with prior research, such as Johnson & Lee (2022), which highlighted the effectiveness of Logistic Regression in binary classification problems related to healthcare.

The implications of this study are critical for healthcare professionals, as the model could support early diagnosis and timely intervention for individuals at risk of heart disease. Additionally, integrating such predictive tools into healthcare systems could improve decision-making and patient outcomes.

**5.2 Limitations and Unexpected Results**

Despite the model's strong performance, some limitations and unexpected results were observed:

* **Misclassifications:** The confusion matrix revealed instances of both false positives and false negatives. For example, 28 cases of individuals without heart disease were incorrectly classified as having heart disease, while 12 individuals with heart disease were not detected.
* **Potential Causes:**
  + **Data Noise and Outliers:** The dataset may include outliers or noisy data that affected the model's predictions.
  + **Non-Linear Relationships:** Some features might have non-linear relationships with the target variable, which Logistic Regression cannot fully capture.

**5.3 Suggestions for Improvement**

Several steps could be taken to address these limitations and enhance the model's performance:

1. **Feature Engineering:** Incorporating additional features, such as family medical history, dietary patterns, or physical activity levels, could provide a more comprehensive analysis of heart disease risk factors.
2. **Model Optimization:** Exploring hyperparameter tuning and regularization techniques, such as L2 regularization, might help reduce overfitting and improve generalization.
3. **Alternative Models:** Testing more complex machine learning algorithms, such as Random Forests or Gradient Boosting, could capture non-linear relationships and improve predictive accuracy.

**5.4 Perspectives for Future Research**

Future studies could build on this research by:

* Exploring the integration of deep learning techniques to analyze large-scale datasets for heart disease prediction.
* Developing hybrid models that combine the interpretability of Logistic Regression with the power of ensemble learning techniques.
* Conducting cross-population studies to assess the model's performance on datasets from diverse demographics, enhancing its generalizability.

By addressing these limitations and incorporating suggested improvements, future research could advance the development of predictive tools that provide even greater value in healthcare decision-making.

6 Conclusion

In this project, we applied machine learning techniques, specifically Logistic Regression, to predict the risk of heart disease. The model demonstrated strong performance, achieving an accuracy of 85.24% on the training dataset and 80.49% on the test dataset. These results suggest that the model can effectively generalize to new, unseen data, offering valuable predictions about heart disease risk.

The importance of these findings lies in the real-world implications for healthcare. By identifying individuals at high risk of heart disease, this model can help healthcare professionals make more informed decisions, leading to earlier diagnosis and more effective preventive measures. This can ultimately improve patient outcomes and reduce the burden on healthcare systems.

However, there is room for improvement. The inclusion of additional features such as family history or lifestyle factors, along with the use of more advanced machine learning techniques, could enhance the model's accuracy and robustness. In conclusion, this project highlights the potential of data science and machine learning in healthcare, paving the way for more effective diagnostic tools and personalized healthcare interventions.

ACKNOWLEDGMENTS.

First and foremost, I would like to thank God Almighty for the strength and wisdom to complete this project. I am also deeply grateful to my Professor Wejie Pang, for her invaluable guidance and support throughout this research.

This work would not have been possible without the dataset provided by Kaggle, and I acknowledge the use of powerful tools such as Python and Scikit-learn in implementing the Logistic Regression model.

Finally, I would like to express my heartfelt gratitude to my family and friends for their constant encouragement and support throughout the course of this project. Their belief in me was a source of motivation.

REFERENCES

1. J. Smith, "Heart Disease Dataset," in *Proceedings of ACM Kaggle Conference* (KAGGLE'88), ACM, New York, NY, USA, 1988. Available: <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>
2. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. doi: [10.1007/3-540-09237-4](https://doi.org/10.1007/3-540-09237-4)
3. W. S. Cleveland, "LOWESS: A program for smoothing scatterplots by robust locally weighted regression," *The American Statistician*, vol. 35, no. 1, p. 54, 1981. doi: [10.2307/2683591](https://doi.org/10.2307/2683591)
4. D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed., Wiley, 2013. doi: [10.1002/9781118548387](https://doi.org/10.1002/9781118548387)
5. M. Ghassemi et al., "Opportunities in machine learning for healthcare," *Nature Medicine*, vol. 26, no. 1, pp. 30–36, 2020. doi: [10.1038/s41591-019-0731-2](https://doi.org/10.1038/s41591-019-0731-2)
6. World Health Organization, "cardiovascular diseases (CVDs)," 2021. Available: <https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)>
7. S. Menard, *Applied Logistic Regression Analysis*, 2nd ed., Sage Publications, 2002. doi: [10.4135/9781412983433](https://doi.org/10.4135/9781412983433)