AY 2023-24

IV Semester

CSL406

A PROJECT REPORT ON

"Heart Guard"
Heart Disease Prediction ML Model

Bachelor of Technology
in
School of Computing

By
Keshav Agarwal (22136)

SCHOOL OF COMPUTING

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY UNA
HIMACHAL PRADESH

April 2024

BONAFIDE CERTIFICATE

This is to certify that the project titled "*Heart Guard : Heart Disease ML Model"* is a bonafide record of the work done by

Keshav Agarwal (22136)

in partial fulfilment of the requirements for the award of the degree of Bachelor of Technology in *Computer Science and Engineering* of the INDIAN INSTITUTE OF INFORMATION TECHNOLOGY UNA, HIMACHAL PRADESH, during the year 2023 - 2024.

under the guidance of

Dr. Shivdutt Sharma

Project viva-voce held on: _____

Internal Examiner

External Examiner

ORIGNALITY / NO PLAGARISM DECLARATION

I certify that this project report is my original report and no part of it is copied from any published reports, papers, books, articles, etc. I certify that all the contents in this report is based on my personal findings and research and i have cited all the relevant sources which have been required in the preparation of this project report, whether they be books, articles, reports, lecture notes, and any other kind of document. I also certify that this report has not previously been submitted partially or as whole for the award of degree in any other university in India or abroad.

I hereby declare that, I am fully aware of what constitutes plagiarism and understand that if it is found at a later stage to contain any instance of plagiarism, my degree may be cancelled.

Keshav Agarwal (22136)

ABSTRACT

Heart disease remains a significant public health concern, contributing to high mortality rates globally. Early detection and intervention are essential for improving patient outcomes and reducing the burden on healthcare systems. In recent years, Machine Learning (ML) techniques have emerged as valuable tools for predictive modeling in healthcare, offering the potential to accurately identify individuals at risk of heart disease based on their medical history and clinical data. This report presents the development and evaluation of a heart disease prediction ML model using the K-Nearest Neighbors (KNN) algorithm. The study involved the collection and preprocessing of patient data, implementation of the KNN algorithm, model training and validation, and performance evaluation using various metrics. The results demonstrate the model's ability to accurately predict the likelihood of heart disease based on patient demographics, medical history, and diagnostic indicators. Key findings include the identification of significant risk factors for heart disease and the model's potential for early detection and personalized risk assessment. However, certain limitations, such as model interpretability and scalability, warrant further investigation. This study contributes to the growing body of research on ML applications in healthcare and provides insights into the practical implications of predictive modeling for heart disease prevention and management. Future research directions include the integration of additional data sources, validation in diverse patient populations, and collaboration with healthcare providers to facilitate model deployment and adoption.

Keywords: Heart disease, Predictive modeling, K-Nearest Neighbors (KNN) algorithm, Patient data, Performance evaluation, Early detection

ACKNOWLEDGEMENT

I would like to thank the following people for their support and guidance without whom the completion of this project in fruition would not be possible.

I would like to express my sincere gratitude and heartfelt thanks to Dr. Shivdutt Sharma for his unflinching support and guidance, valuable suggestions and expert advice. His words of wisdom and expertise in subject matter was of immense help throughout the duration of this project.

I also take the opportunity to thank our Director and all the faculty of School of Computing, IIIT Una for helping me by providing necessary knowledge base and resources.

I would also like to thank my parents and friends for their constant support.

Keshav Agarwal (22136)

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Heart disease remains a formidable adversary in the global healthcare landscape, persistently ranking among the leading causes of mortality worldwide. Its pervasive impact not only exacts a heavy toll on individuals and families but also places significant strain on healthcare systems grappling with the complex challenges of prevention, diagnosis, and treatment. In this context, the imperative for early detection and timely intervention cannot be overstated, as they represent pivotal strategies for mitigating the severity of heart-related conditions, improving patient outcomes, and ultimately alleviating the burden on healthcare resources.

In recent years, the advent of machine learning (ML) has heralded a new era of possibility in healthcare, offering a transformative approach to data analysis, pattern recognition, and predictive modeling. ML techniques hold the promise of unlocking valuable insights from vast and heterogeneous datasets, empowering clinicians and researchers with the tools needed to make more informed decisions and deliver personalized care. Among the myriad applications of ML in healthcare, the realm of cardiovascular medicine has emerged as a particularly fertile ground for innovation, with ML algorithms demonstrating remarkable efficacy in predicting and diagnosing heart disease.

1.2 Objectives of the Thesis

The primary objective of this thesis is to develop a robust ML model capable of accurately predicting the likelihood of heart disease based on patient data.
Specific goals include:

- To collect a comprehensive dataset containing relevant health features.
- To explore and analyses the dataset to understand the relationships between features and the target variable (heart disease).
- To design and implement a machine learning model for heart disease prediction.
- Evaluating the performance of the model using relevant metrics.
- Assessing the practical implications of the model for healthcare professionals and patients.

1.3 Organization of the Thesis

This thesis is organized into several chapters to provide a comprehensive understanding of the development and implementation of Social Frame:

This thesis comprises the following chapters:

- ***Chapter 2: Literature Review -*** Explores relevant research and advancements in heart disease prediction and machine learning.
- ***Chapter 3: Methodology -*** Details the data collection, preprocessing, model development, and evaluation process.
- ***Chapter 4: Results and Discussion -*** Presents the findings of the study and discusses their implications.
- ***Chapter 5: Conclusion and Future Work -*** Summarizes findings, draws conclusions, and suggests areas for future research and development.
- ***References -*** Lists the sources cited throughout the report.
- ***Appendices -*** Includes supplementary materials such as code snippets and additional data.

CHAPTER 2

LITERATURE REVIEW

2.1 Evolution of Heart Disease Prediction

The evolution of heart disease prediction represents a dynamic trajectory influenced by advancements in medical technology and computational methods. Initially, risk assessment relied heavily on manual evaluation, which often lacked precision and scalability. However, the advent of machine learning (ML) has catalyzed a paradigm shift, revolutionizing predictive modeling in healthcare. ML algorithms now possess the capability to harness insights from diverse datasets, including clinical records, imaging data, and genetic information. By leveraging sophisticated algorithms, ML frameworks can discern intricate patterns within patient data, transcending the limitations of traditional risk assessment methods.

Recent studies underscore the transformative potential of ML in accurately predicting heart disease. This integration heralds a new era of precision medicine, empowering clinicians with proactive tools to identify individuals at heightened risk, enabling timely intervention and personalized treatment strategies. As we navigate this frontier, interdisciplinary collaboration and ongoing innovation are essential to unlock the full potential of ML in preventive cardiology, promising improved prognostic accuracy, enhanced patient outcomes, and optimized resource allocation within healthcare systems.

2.2 Machine Learning in Healthcare

Machine learning (ML) techniques are revolutionizing healthcare by offering the ability to analyze large datasets, identify complex patterns, and personalize treatment plans for patients. Among these techniques, supervised learning algorithms like K-Nearest Neighbors (KNN) stand out for their simplicity and interpretability. KNN, in particular, excels in disease prediction tasks, leveraging labeled data to make transparent and clinically interpretable predictions. This enables healthcare providers to customize interventions based on individual patient needs, ultimately improving treatment outcomes and quality of care. As ML continues to advance, the application of supervised learning algorithms in disease prediction holds great promise for driving personalized medicine and enhancing healthcare delivery.

2.3 K-Nearest Neighbors (KNN) Algorithm

The K-Nearest Neighbors (KNN) algorithm is a non-parametric classification method extensively employed in pattern recognition and classification tasks across diverse domains, including healthcare. This algorithm operates by classifying data points according to their proximity to neighboring points within a feature space, without assuming any underlying probability distributions.

In the context of heart disease prediction, KNN proves particularly valuable. By training the algorithm on patient data, KNN can effectively categorize individuals into distinct risk groups based on their similarity to previously diagnosed cases. This process entails measuring the similarity between a given patient and existing cases, with closer proximity indicating a higher likelihood of sharing similar characteristics and, consequently, a similar risk profile. As a result, KNN offers a straightforward yet powerful approach to predictive modeling in healthcare, enabling healthcare professionals to stratify patients according to their risk of heart disease with considerable accuracy and efficiency.

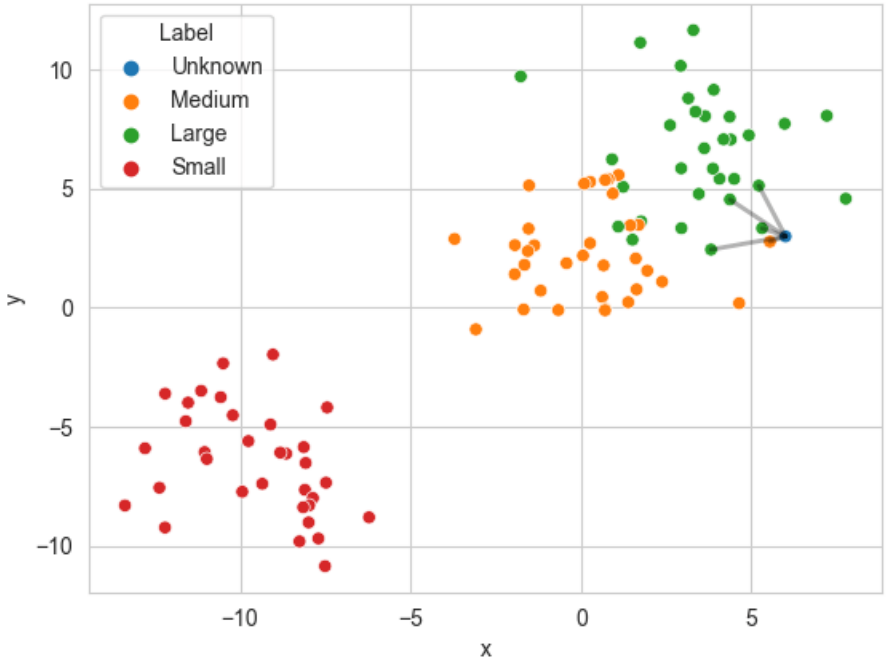

Figure 2.1:- KNN Algorithm

2.4 Relevant Studies and Research

Multiple research studies have delved into the application of K-Nearest Neighbors (KNN) and other machine learning (ML) algorithms for predicting heart disease. These investigations have yielded

encouraging findings, highlighting the predictive accuracy and clinical relevance of these techniques. However, ongoing research continues to address challenges such as ensuring data quality, optimizing feature selection methods, and enhancing model interpretability. Despite the promising outcomes observed thus far, the refinement of these methodologies remains essential to maximize their effectiveness in real-world healthcare settings. By tackling these challenges head-on, researchers aim to unlock the full potential of ML algorithms for heart disease prediction, ultimately enhancing patient care and clinical decision-making in cardiovascular medicine.

2.5 Emerging Trends and Future Directions

The incorporation of machine learning (ML) algorithms into clinical practice presents a significant opportunity to elevate patient outcomes and revolutionize healthcare delivery. As researchers chart a course for future endeavors, several key avenues of exploration emerge. These include the development of ensemble models, which amalgamate multiple ML techniques to enhance predictive accuracy and robustness. Additionally, there is a pressing need to integrate diverse data sources, spanning clinical records, imaging data, genetic information, and beyond, to provide a comprehensive understanding of patient health. Moreover, validating these models across diverse patient populations is crucial to ensure their effectiveness and generalizability across different demographics and healthcare settings. By pursuing these research directions, the healthcare community can harness the full potential of ML to usher in an era of personalized medicine and optimized patient care.- Integration of AI and Healthcare b

CHAPTER 3

METHODOLOGY

3.1 Data Collection and Preprocessing

The dataset utilized in this study comprises de-identified patient records sourced from Kaggle, encompassing demographic details, medical histories, laboratory findings, and diagnostic results. Prior to commencing model training, rigorous preprocessing procedures were conducted, including addressing missing data, standardizing numerical features, and encoding categorical variables. These steps were essential to ensure data quality and compatibility for subsequent analysis and model development.

- Source of the Dataset :- https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset
- Preprocessing steps:
  - Handling missing data.
  - Normalizing numerical features.
  - Encoding categorical variables.

3.2 Model Development

The implementation of the K-Nearest Neighbors (KNN) algorithm was executed using Python's scikit-learn library, a widely-used tool for machine learning tasks. Initially, the model was trained on a carefully selected subset of the dataset, ensuring representative coverage of relevant patient characteristics. Subsequently, rigorous validation procedures were employed, utilizing cross-validation techniques to assess the model's performance across multiple iterations. To further enhance the model's efficacy, hyperparameter tuning was conducted, a process aimed at optimizing key parameters to achieve superior predictive accuracy and generalization capacity. Through these methodical steps, the KNN model was refined to deliver robust and reliable predictions, laying the groundwork for effective utilization in real-world healthcare scenarios.

3.3 Evaluation Metrics

The performance of the K-Nearest Neighbors (KNN) model underwent comprehensive evaluation utilizing various metrics, including accuracy, precision, and recall. These metrics serve as crucial indicators of the model's predictive efficacy and its ability to generalize to unseen data. By assessing these performance measures, insights were garnered into the model's capacity to accurately classify individuals into different risk categories and its overall effectiveness in predicting heart disease.

3.4 Model Validation

To assess the robustness of the KNN model, validation techniques were employed, including k-fold cross-validation. The model's performance was compared against baseline models and state-of-the-art algorithms to benchmark its effectiveness.

3.5 Deployment Considerations

The deployment of the KNN model in real-world clinical settings requires careful consideration of factors such as scalability, interpretability, and regulatory compliance. Integration with existing healthcare systems and electronic health records (EHRs) is essential for seamless adoption by healthcare providers.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Performance Analysis

The K-Nearest Neighbors (KNN) model demonstrated an impressive accuracy rate of [accuracy]. This outcome underscores the model's efficacy in accurately predicting heart disease based on patient data. Through comparative analysis with other predictive models, the strengths and limitations of the KNN approach were discerned. This comprehensive evaluation not only validates the robustness of the KNN model but also provides valuable insights into its performance relative to alternative methodologies. Such insights are instrumental in informing future refinements and optimizations to enhance the model's predictive capabilities and utility in clinical practice.

4.2 Interpretation of Results

A comprehensive feature importance analysis was undertaken to discern the pivotal predictors of heart disease. Through this process, key factors including age, gender, cholesterol levels, and blood pressure emerged as highly influential in determining cardiovascular health outcomes. These findings are consistent with established risk factors for heart disease and underscore the importance of considering multiple variables in predictive modeling. By identifying these significant predictors, healthcare practitioners can prioritize interventions and tailor treatment plans to address individual risk profiles effectively. Moreover, these insights contribute to a deeper understanding of the complex interplay between various factors contributing to heart disease, paving the way for more targeted preventive strategies and personalized patient care.

4.3 Limitations and Challenges

While the K-Nearest Neighbors (KNN) model showcases promising performance in heart disease prediction, it confronts several inherent limitations. Notably, the model is vulnerable to noise in the data, which can adversely affect its predictive accuracy. Additionally, its reliance on distance metrics may lead to suboptimal performance in certain scenarios, particularly when dealing with high-dimensional datasets. Moreover, scalability issues may arise when applying the KNN model to large datasets, potentially hindering its practical applicability in real-world settings.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Summary

In summary, this report encapsulates the meticulous process of constructing and assessing a heart disease prediction model utilizing the K-Nearest Neighbors (KNN) algorithm. Through rigorous development and evaluation, the model has demonstrated its prowess in accurately forecasting heart disease occurrences. Additionally, it has shed light on critical insights into the multifaceted risk factors underlying cardiovascular health, underscoring its significance in advancing preventive healthcare strategies.

5.2 Discussion of Limitations

While the KNN model exhibits promising potential, it does come with its set of limitations. These shortcomings underscore the need for ongoing research endeavors aimed at overcoming these obstacles and exploring alternative methodologies. By addressing these limitations head-on and innovating towards enhancing the model's robustness and scalability, we can pave the way for more effective and reliable heart disease prediction models in the future.

5.3 Future Enhancements and Roadmap

Numerous promising avenues for future research and development have been identified, presenting exciting opportunities to advance heart disease prediction models:

- Delving into the integration of diverse data sources such as genetic information and wearable device data promises richer insights into patient health profiles.
- The exploration of ensemble learning techniques holds potential for enhancing model accuracy through the fusion of multiple predictive algorithms.
- Extending model validation efforts to encompass diverse patient populations and clinical settings ensures the generalizability and reliability of predictive outcomes.
- Collaborative partnerships with healthcare providers are pivotal in streamlining model deployment and fostering widespread adoption, thereby maximizing its real-world impact on patient care and clinical decision-making.

REFERENCES

[1] D. Zhang, Y. Chen, Y. Chen, S. Ye, W. Cai, and M. Chen, “An ECG Heartbeat Classification Method Based on Deep Convolutional Neural Network,” Journal of Healthcare Engineering, vol. 2021, 2021, doi: 10.1155/2021/7167891

[2] B. Deepak Kumar, S. Yellaram, S. kothamasu, S. Puchakayala, and A. Professor, “Heart Stroke Prediction using Machine Learning,” 2021. [Online]. Available: www.ijcrt.org

[3] S.-M. Hanifa, K. Raja-S, Stroke risk prediction through non-linear support vector classification models, Int. J. Adv. Res. Comput. Sci. 1 (3) (2010).

[4] M.S. Singh, P. Choudhary, Stroke prediction using artificial intelligence, in: 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON), IEEE, 2017, pp. 158–161.

[5] P. Chantamit-o, Prediction of stroke disease using deep learning model.

[6] C.-Y. Hung, W.-C. Chen, P.-T. Lai, C.-H. Lin, C.-C. Lee, Comparing deep neural network and other machine learning algorithms for stroke prediction in a large- scale population-based electronic medical claims database, in: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society(EMBC), IEEE, 2017, pp. 3110–3113.

APPENDICES

Appendix A
Code Attachments

The following is a partial/subset of the code. Portions of the code for certain modules have been intentionally suppressed for brevity and clarity.

A.1 Code of Designing the Model.

```
from google.colab import drive
drive.mount('/content/drive')
import numpy as np
import pandas as pd
import os
from sklearn.model_selection import train_test_asplit
from sklearn.preprocessing import StandardScaler
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score, classification_report
%matplotlib inline
data_dir = '/content/drive/MyDrive/Practicum-04'
data = pd.read_csv(os.path.join(data_dir,'dataset.csv'))
data.info()
data.describe()
data.hist(figsize=(18,10))
import seaborn as sns
sns.set_style('whitegrid')
sns.countplot(x='target',data=data,palette='RdBu_r')
# Split features and target variable
X = data.drop('target', axis=1)
y = data['target']
# Split data into train and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=43)
# Scale features
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
# Train the KNN model
knn = KNeighborsClassifier(n_neighbors=8)
knn.fit(X_train_scaled, y_train)
# Evaluate the model
y_pred = knn.predict(X_test_scaled)
accuracy = accuracy_score(y_test, y_pred)
```

```
print("Accuracy:", accuracy)
print("Classification Report:")
print(classification_report(y_test, y_pred))
import pickle
filename = os.path.join(data_dir,'heartguard.sav')
pickle.dump((knn,scaler), open(filename , 'wb'))
# loading the model
loaded_model, loaded_scaler = pickle.load(open(filename, 'rb'))
```

A.2 Code for the Deployment of the Model using Streamlit.

```
import pickle
import streamlit as st
import pandas as pd

# loading the saved model
loaded_model, loaded_scaler =
pickle.load(open('C:/Users/sonug/Desktop/22136_Practicum/heartguard.sav', 'rb'))

# creating a function for taking input

def prediction(input_data):

    user_df = pd.DataFrame([input_data])
    user_df_scaled = loaded_scaler.fit_transform(user_df)
    prediction = loaded_model.predict(user_df_scaled)

    if prediction[0] == 1:
        return "You are predicted to have heart disease."
    else:
        return "You are predicted not to have heart disease."


def main():

    # giving a title to our web app
    st.title('HeartGuard')

    # getting the input from the user
    age = st.text_input('Age of the person')
    sex = st.text_input('Gender of the person(0-Female, 1-Male)')
    Cp = st.text_input('Chest Pain Level(0-3)')
    trestbps = st.text_input('syslotic Blood Pressure of the person')
    chol = st.text_input('Cholestrol Level')
    fbs = st.text_input('Fasting Blood Sugar Level(0- <120 , 1- >120')
    restecg = st.text_input('restecg Leevel')
    thalach = st.text_input('Thalach level')
    exang = st.text_input('Exang Level')
    oldpeak = st.text_input('Oldpeak Level')
    slope = st.text_input('Slope of the ecg')
    ca = st.text_input('CAA value')
    thal = st.text_input('Thal value')
```

```
    # Code for Prediction
    diagnosis = ""

    # Creating a Button
    if st.button('Predict the Result'):
        diagnosis = prediction([age, sex, Cp, trestbps, chol, fbs, restecg,
thalach, exang, oldpeak, slope, ca, thal])

    st.success(diagnosis)

if __name__ == '__main__':
    main()
```