Heart Condition Detector

Dissertation submitted in fulfilment of the requirements for the Degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING WITH DATA SCIENCE AND MACHINE LEARNING

By

PRANSHUL THAKUR

12219336

Supervisor

VED PRAKASH CHAUBEY

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

November 2023

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation/dissertation proposal entitled "HEART CONDITION DETECTOR” in partial fulfilment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Mr./Mrs. Ved Prakash Chaubey. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University’s Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

Signature of Candidate

Pranshul Thakur

R. No 10

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the B. Tech Dissertation/dissertation proposal entitled "**HEART CONDITION DETECTOR**", submitted by **Pranshul Thakur** at **Lovely Professional University, Phagwara, India** is a bonafide record of his / her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor

Ved Prakash Chaubey

Date:

Counter Signed by:

1) **Concerned HOD:**
   HoD's Signature: ________________

   HoD Name: ____________________

   Date: ___________________

2) **Neutral Examiners:**

   External Examiner

   Signature: _______________

   Name: __________________

   Affiliation: ______________

   Date: ___________________

   Internal Examiner

   Signature: _______________

   Name: __________________

   Date: ___________________

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to all those who have supported and contributed to the successful completion of this college project. Their guidance, assistance, and encouragement have been invaluable throughout this journey.

First and foremost, I would like to thank my professor, Ved Prakash Chaubey, for their guidance and expertise. Their insights and feedback were instrumental in shaping the direction of this project. I am grateful for their patience and dedication in helping me navigate through the complexities of the subject matter.

I want to acknowledge the resources at Lovely Professional University, without which this project would not have been possible. The library, research facilities, and technical support played a crucial role in gathering the necessary information and data.

In conclusion, I am thankful to all those mentioned above and anyone else who may have indirectly contributed to this project. Your support has been invaluable, and I am grateful for your assistance in making this project a reality.

Pranshul Thakur

12219336

Roll no 10

TABLE OF CONTENTS

CONTENTS	PAGE NO.
Inner first page – Same as cover	i
Declaration by the Scholar	ii
Supervisor’s Certificate	iii
Acknowledgement	iv
Table of Contents	v
Abstract	vi
List of Acronyms / Abbreviations	vii
List of Dataset	viii

CHAPTER1: INTRODUCTION

CHAPTER2: PROBLEM STATEMENT

CHAPTER3: REVIEW OF LITERATURE

CHAPTER4: MY WORK

CHAPTER5: CODE SNIPPETS

CHAPTER6: GITHUB LINK

CHAPTER7: FUTURE SCOPE

CHAPTER8: IMPROVEMENTS THAT CAN BE ADDED

CHAPTER9: CONCLUSION

BIBLIOGRAPHY

ABSTRACT

Heart disease is a major public health concern, accounting for a significant portion of deaths worldwide. Early detection and treatment of heart disease can improve patient outcomes and reduce mortality. Machine learning algorithms can be used to develop heart disease detectors that can identify individuals at risk of developing the disease.

This project develops a heart disease detector using Google Collaboratory, a cloud-based platform for machine learning. The detector uses Logistic Regression, Nearest Neighbour Algorithm (KNN), and Random Forest as its algorithm models. The detector is trained on the Kaggle UCI Heart Disease dataset, which contains data on over 300 patients.

The detector achieves an accuracy of 83% using Logistic Regression, 84% using KNN, and 85% using Random Forest. These results suggest that machine learning algorithms can be used to develop effective heart disease detectors.

The project also discusses the challenges of developing heart disease detectors, such as the need for large datasets and the difficulty of interpreting the results of machine learning algorithms.

Overall, this project demonstrates the potential of machine learning to develop effective heart disease detectors. Further research is needed to improve the accuracy of these detectors and to make them more interpretable.

LIST OF ACRONYMS

- Nearest Neighbor Algorithm = KNN
- Unique Client Identifier = UCI
- Fasting blood sugar = FBS
- Thalassemia = THAL
- Cancer and Carcinoma = CA
- Exchange Transfusion = Exang
- Resting Blood Pressure = TrestBPS
- Constrictive Pericarditis = CP
- Resting Electrocardiographic = Restecg
- Electrocardiogram = ECG

LIST OF DATASETS

The dataset that is being used is downloaded from UCI HEART DISEASE DATA from Kaggle. The link is given below along with the first 10 rows of the dataset:

https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1

INTRODUCTION

With the rapid progression of medical technology, the detection of heart disease has become increasingly crucial in the field of cardiology. The advancement of machine learning has provided a revolutionary approach to addressing medical challenges, and heart disease detection is no exception. This project aims to develop a heart disease detector using Google Collab, a cloud-based platform for machine learning, leveraging the power of Logistic Regression, Nearest Neighbour Algorithm (KNN), and Random Forest as its algorithm models. The detector will be trained on the Kaggle UCI Heart Disease dataset, a comprehensive collection of data on over 300 patients.

The significance of this project lies in its potential to contribute to the early identification of heart disease, enabling timely intervention and improved patient outcomes. By utilizing machine learning algorithms, we can effectively analyse medical data and identify patterns that may indicate the presence of heart disease. The project also highlights the versatility of Google Collab as a platform for developing and deploying machine learning models, making it accessible to a broader audience.

In the following sections, we will delve into the methodology of developing the heart disease detector, exploring the data preprocessing, feature engineering, and model selection processes. We will also present the evaluation results of the detector using various performance metrics, such as accuracy, precision, and recall. Furthermore, we will discuss the challenges encountered during the project and explore avenues for future research.

PROBLEM STATEMENT

Objective:

The objective of this project is to develop a heart disease detector using Google Collab, a cloud-based platform for machine learning. The detector will use Logistic Regression, Nearest Neighbour Algorithm (KNN), and Random Forest as its algorithm models. The detector will be trained on the Kaggle UCI Heart Disease dataset, which contains data on over 300 patients.

Evaluation Criteria:

The detector will be evaluated using the following performance metrics:

Accuracy

Precision

Recall

Expected Outcomes:

It is expected that the heart disease detector will achieve an accuracy of at least 80%. The detector should also have a high precision and recall, which means that it should correctly identify individuals with heart disease and correctly identify individuals without heart disease.

Challenges:

Developing a heart disease detector is a challenging task due to the following reasons:

The data is high-dimensional and noisy.

There are a large number of features in the dataset.

The relationship between the features and the target variable is complex.

Despite these challenges, it is believed that machine learning algorithms can be used to develop effective heart disease detectors.

REVIEW OF LITERATURE

A multitude of research studies have delved into the realm of heart disease prediction, employing a spectrum of data mining and machine learning techniques. Among these, K. Polaraju et al. utilized Multiple Linear Regression, achieving superior accuracy with a 70-30 data split. Marjia et al. harnessed KStar, J48, SMO, and Bayes Net within the WEKA software, where SMO and Bayes Net exhibited optimal performance. In another vein, S. Seema et al. centred their focus on the prediction of chronic diseases, with SVM standing out as the most accurate classifier. Ashok Kumar Dwivedi et al. recommended several algorithms, including Naive Bayes, Classification Tree, KNN, Logistic Regression, and SVM, ultimately favoring Logistic Regression for its precision. In parallel, Megha Shahi et al. explored SVM, Naive Bayes, Association Rule, KNN, ANN, and Decision Tree using WEKA, with SVM emerging as effective and delivering high accuracy.

Chala Beyene et al. concentrated on heart disease prediction by leveraging medical attributes and the WEKA software, culminating in valuable outcomes. Furthermore, R. Sharmila et al. introduced the concept of non-linear classification algorithms and the utilization of big data tools like HDFS, MapReduce, and SVM for heart disease prediction. Meanwhile, Jayami Patel et al. employed data mining techniques, with J48 emerging as the top performer. In addition, Purushottam et al. developed an efficient heart disease prediction system with commendable accuracy.

Jaymin Patel et al. employed data mining techniques within the WEKA software, offering a multifaceted approach to heart disease prediction. Boshra Brahmi et al. delved into evaluating various classification techniques, singling out J48 and Decision Tree as the top performers. In a different avenue, Noura Ajam recommended artificial neural networks for heart disease diagnosis, attaining an impressive 88% classification accuracy. Furthermore, Prajakta Ghadge proposed the utilization of big data for heart attack prediction through data modelling techniques.

S. Prabhavathi ventured into the application of a Decision Tree-based Neural Fuzzy System (DNFS) for the analysis and prediction of various heart diseases. Simultaneously, Sairabi H. Mujawar delved into the use of k-means and Naïve Bayes for heart disease prediction based on historical heart data. Sharan Monica.L delved into the realm of cardiovascular disease, leveraging data mining techniques to predict these ailments.

MY WORK

A simple heart condition detector which uses, Kaggle for its dataset and google collab to give a description about the condition of a person's heart

The code uses the following libraries:

NumPy

pandas

train tester

logistic regression

accuracy score

It uses Logistic Regression, Nearest Neighbour Algorithm (KNN) and Random Forest as its algorithm models.

We first:

1. Load the csv data to a Pandas Data Frame.

2. Then get the number of rows and columns in the dataset and some info about the data.

3. Then we check for missing values and check the distribution of Target Variable.

4. Then we split the targe and the feature.

5. Then we split the data into training data and test data and use the model training and data regression to create and algorithm and check its accuracy.

6. Finally, we build the predictive system using all the given resources we have.

CODE SNIPPETS

```
[ ] input_data = (62,0,0,140,268,0,0,160,0,3.6,0,2,2)

    # change the input data to a numpy array
    input_data_as_numpy_array= np.asarray(input_data)

    # reshape the numpy array as we are predicting for only on instance
    input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

    prediction = model.predict(input_data_reshaped)
    print(prediction)

    if (prediction[0]== 0):
      print('The Person does not have a Heart Disease')
    else:
      print('The Person has Heart Disease')

    [0]
    The Person does not have a Heart Disease
```

Model Evaluation

Accuracy Score

```
[ ] # accuracy on training data
    X_train_prediction = model.predict(X_train)
    training_data_accuracy = accuracy_score(X_train_prediction, Y_train)

[ ] print('Accuracy on Training data : ', training_data_accuracy)

    Accuracy on Training data :  0.8512396694214877

[ ] # accuracy on test data
    X_test_prediction = model.predict(X_test)
    test_data_accuracy = accuracy_score(X_test_prediction, Y_test)

[ ] print('Accuracy on Test data : ', test_data_accuracy)

    Accuracy on Test data :  0.819672131147541
```

Model Training

Logistic Regression

```
model = LogisticRegression()

[ ] # training the LogisticRegression model with Training data
    model.fit(X_train, Y_train)

    /usr/local/lib/python3.7/dist-packages/sklearn/linear_model/_logistic.py:940: ConvergenceWarning: lbfgs failed to converge (status=1):
    STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

    Increase the number of iterations (max_iter) or scale the data as shown in:
        https://scikit-learn.org/stable/modules/preprocessing.html
    Please also refer to the documentation for alternative solver options:
        https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
      extra_warning_msg=_LOGISTIC_SOLVER_CONVERGENCE_MSG)
    LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                       intercept_scaling=1, l1_ratio=None, max_iter=100,
                       multi_class='auto', n_jobs=None, penalty='l2',
                       random_state=None, solver='lbfgs', tol=0.0001, verbose=0,
                       warm_start=False)
```

Splitting the Features and Target

```
[ ] X = heart_data.drop(columns='target', axis=1)
    Y = heart_data['target']
```

```
[ ] print(X)
```

Importing the Dependencies

```
[ ] import numpy as np
    import pandas as pd
    from sklearn.model_selection import train_test_split
    from sklearn.linear_model import LogisticRegression
    from sklearn.metrics import accuracy_score
```

Data Collection and Processing

```
[ ] # loading the csv data to a Pandas DataFrame
    heart_data = pd.read_csv('/content/data.csv')
```

```
[ ] # print first 5 rows of the dataset
    heart_data.head()
```

GITHUB LINK

https://github.com/Pranshul-Thakur/Heart-Condition-Detector

FUTURE SCOPE

The heart disease detector project holds promising potential to revolutionize the field of cardiology and enhance the lives of individuals grappling with heart disease. Looking ahead, there are several compelling future directions for the project. One notable avenue involves the integration of the heart disease detector with wearable devices, such as smartwatches or fitness trackers. This integration would enable the continuous collection of real-time data on a patient's heart health, facilitating ongoing monitoring and the early detection of potential heart disease indicators.

Furthermore, the development of a dedicated mobile application could further empower patients, allowing them easy access to the heart disease detector and personalized risk assessments. Such an app could also serve as an invaluable resource, offering educational materials on heart disease prevention and treatment.

Additionally, collaboration with medical professionals is a vital next step. By seamlessly integrating the heart disease detector into electronic health records (EHRs), clinicians could gain access to real-time risk assessments for their patients. This invaluable information could guide treatment decisions, enhance the quality of care, and ultimately lead to improved patient outcomes. In sum, the future of the heart disease detector project holds the promise of more accessible and effective heart disease management and prevention.

IMPROVEMENTS THAT CAN BE ADDED

To further enhance the heart disease detector's technical capabilities, several key improvements can be considered. Firstly, the augmentation of available data for training the detector is paramount. Techniques such as synthetic data generation and data augmentation can significantly bolster the quantity and diversity of the dataset. This expansion would notably boost the accuracy of the detector, particularly in detecting rare diseases that may have limited available data.

In addition, the application of ensemble methods presents a compelling opportunity. Ensemble methods involve the amalgamation of multiple machine learning models, and they have a track record of delivering superior performance due to their ability to mitigate overfitting. By employing ensemble methods, the heart disease detector can attain greater accuracy and reliability.

Furthermore, the adoption of transfer learning is a noteworthy avenue for improvement. This technique leverages pre-trained models to address new challenges, which can be particularly advantageous when the dataset for training the detector is limited. By incorporating transfer learning, the heart disease detector can benefit from the knowledge acquired in broader domains, resulting in improved accuracy and effectiveness.

Incorporating these technical enhancements holds the potential to elevate the heart disease detector's precision and overall performance, making it an even more valuable tool for early diagnosis and management of heart-related conditions.

CONCLUSION

In conclusion, heart disease detection using machine learning has the potential to revolutionize the way we diagnose and manage this critical condition. By leveraging the power of algorithms to analyse vast amounts of patient data, we can identify patterns and risk factors that may indicate the presence of heart disease, even in its early stages. This information can then be used to guide treatment decisions, improve patient outcomes, and ultimately reduce the burden of heart disease on individuals and healthcare systems worldwide.

As machine learning techniques continue to evolve, we can expect even more sophisticated and accurate heart disease detection models to emerge. These models will likely incorporate a wider range of data sources, including wearable devices, genetic data, and electronic health records. They may also employ more complex algorithms, such as deep learning, to better capture the nuances of heart disease risk assessment.

With continued research and development, heart disease detection using machine learning has the potential to transform the way we approach this devastating disease. By empowering clinicians with powerful diagnostic tools, we can work towards a future where heart disease is detected early, treated effectively, and ultimately prevented

BIBLIOGRAPHY

KAGGLE UCI DATASET

REVIEW OF LITREATURE

MACHINE LEARNING

VARIOUS PARAMETERS IN AN ECG WAVE

FOR STEPS REFRENCE

Checklist for Dissertation-III Supervisor

Name: ____________________________ UID: ________ Domain: _____
Registration No: ____________Name of student:__________________

Title of Dissertation:
____

- ☐ Front pages are as per the format.
- ☐ Topic on the PAC form and title page are same.
- ☐ Front page numbers are in roman and for report, it is like 1, 2, 3…….
- ☐ TOC, List of Figures, etc. are matching with the actual page numbers in the report.
- ☐ Font, Font Size, Margins, line Spacing, Alignment, etc. are as per the guidelines.
- ☐ Color prints are used for images and implementation snapshots.
- ☐ Captions and citations are provided for all the figures, tables etc. and are numbered and center aligned.
- ☐ All the equations used in the report are numbered.
- ☐ Citations are provided for all the references.
- ☐ **Objectives are clearly defined.**
- ☐ Minimum total number of pages of report is 30.
- ☐ Minimum references in report are 5.

Here by, I declare that I had verified the above mentioned points in the final dissertation report.

Signature of Supervisor with UID