APPLIED MACHINE LEARNING

PROJECT REPORT



HEART DISEASE PREDICTION AND CLASSIFICATION USING DEEP LEARNING

Submitted by:

|  |  |
| --- | --- |
| HAFIZ AHMAD  ANIQUE RAFIQUE  HAIDER ALI | 19L-1316  19L-1368  19l-1334 |

Submitted to:

SIR HAMZA YOUSAF

15 JUNE 2023

Department of Electrical Engineering

National University of Computer and Emerging Sciences, Lahore

## Introduction

1. **Problem Analysis**

## Design Requirements

1. **Feasibility Analysis**

## Possible Solutions

1. **Preliminary Design**

## Design Description

## Software Simulation

## Experimental Results

1. **Performance Analysis**
2. **Future Scope**

## Social and Cultural Implications

1. **Conclusion References**

This project focuses on working with a heart disease dataset to develop Disease Prediction and Classification models. Predicting and classifying diseases has always posed a challenge for researchers, but with the emergence of Applied Machine Learning, there has been a significant shift towards a data-driven approach. Machine Learning algorithms are now being employed to predict and classify major ailments worldwide. Progress in this field has shown promising results, which has further motivated researchers to utilize these techniques for analysis.

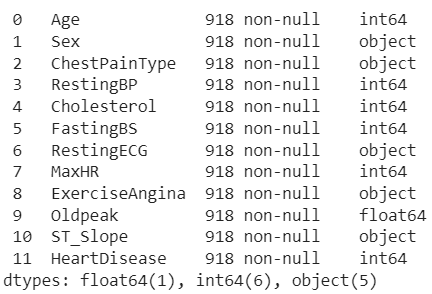
In this project, the first step involves data processing, which includes tasks such as data cleaning, handling missing values, and transforming the dataset into a suitable format for analysis. Once the data is processed, it is divided into training and testing sets. The training set is used to train various machine learning algorithms, which learn patterns and relationships from the data. The goal is to optimize the performance of these algorithms by fine-tuning their parameters and selecting the most suitable ones.

After training the models, they are evaluated using the testing set to assess their performance. Different evaluation metrics are utilized to measure the accuracy, precision, recall, and other relevant statistics of the models. The aim is to obtain the best possible test score, indicating the model's ability to predict and classify heart diseases accurately.

Once the clean dataset and the best-performing predictive model are obtained, the final step involves using this model for disease classification and prediction. The model is deployed to classify new instances and predict their disease status with a high level of accuracy. This predictive model, derived from the trained machine learning algorithms, can contribute to early detection and intervention in heart diseases, potentially saving lives and improving healthcare outcomes.

Using a heart disease dataset, this project aims to develop a disease prediction and classification system. For a long time, disease prediction and classification were difficult tasks for researchers. However, with the advent of Applied Machine Learning, there has been a significant shift toward data-driven methods. Algorithms from machine learning are increasingly being used to predict and classify major diseases worldwide. Notably, these advancements' promising outcomes have pushed researchers to use such analysis methods even more.

There are several important steps in the project workflow. The process of cleaning and preparing the data for analysis is the first step in the dataset. After that, the dataset is divided into training and testing parts, which makes it possible to test the performance of the model on data that has never been seen. The training data are used by a variety of machine learning algorithms to train the model to get the best possible test score. A wide range of methods that are suitable for disease classification and prediction may be included in the algorithm selection We are trying all algorithms to find the best test score. Here the data set is attached after converting categorical to numerical.

 A screenshot of a computer code

Description automatically generated with low confidence

The following are the project's design requirements:

1. Utilization of Acquired Skills: To train accurate prediction models, the project should incorporate all relevant methods and concepts learned in the course. This makes sure that the information learned in the course is put to good use.

2. Data Entry: A CSV file should be accepted as input into the project. Each of the 12 features in this file contributes to determining whether a person has a heart condition. The system ought to be able to work with the given CSV format and extract the data required for analysis.

3. Flexibility in technique: The project gives you the freedom to choose the right methods to train the model. This suggests that the team can investigate a variety of machine learning algorithms, including deep learning methods, in order to construct the model that is most suitable for the current task.

4. Accuracy and Efficiency: The trained model ought to be accurate and efficient. The capacity of the model to process data and make predictions within a reasonable amount of time is referred to as efficiency. The term "accuracy" refers to the model's ability to accurately predict the presence or absence of heart disease the majority of the time.

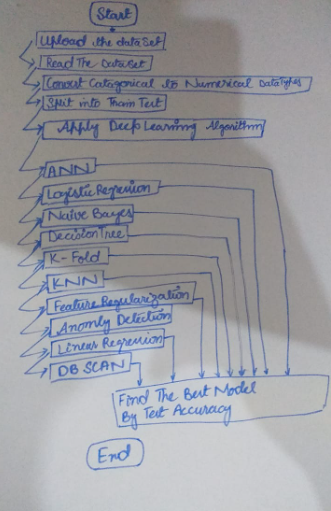
5. Binary Results: The model should produce binary values, specifically 0 or 1, as its output. These numbers indicate whether a person has been diagnosed with heart disease or not. The input features should serve as the foundation for the model's prediction, which should clearly indicate whether or not heart disease is present.

6. The scope of machine learning (deep learning): The project will go into great detail about the concepts of deep learning and machine learning as they relate to the project's scope. This suggests that in order to improve the model's prediction accuracy and performance, we will investigate cutting-edge deep learning algorithms and methods.

If we can finish the project on time, get the results we want (in our case, accuracy), and use as few resources as possible, then the project is considered feasible. We were compelled to plan out the entire procedure to comprehend how we were going to approach the issue because this project was based on research and heavily entailed comprehending the data. After completing the task, it is safe to say that we were able to achieve test accuracy of 85% percent after comprehending the data and utilizing all planned techniques, demonstrating that our project was quite doable.

After completing the project, we can say that we were successful in making a Predicting Heart Disease from trained model of given dataset. We successfully cleaned the data and apply techniques on the features, which improved the efficiency of our model. It can be noticed that even after removing half of the features, our model is working fine i.e. it is trained on the data accurately 87% of the times and has test Accuracy is 85%. This proves that we were able to design complex circuit that did the task more efficiently by applying artificial neural network on it.

|  |  |
| --- | --- |
| Algorithm’s Applied for Heart Disease Prediction | Test Accuracy |
| Artificial Neural Network (ANN) | 84.78% |
| Logistic Regression | 84.23% |
| Naïve Bayes | 84.23% |
| Decision Tree | 81.52% |
| Accuracy using each K-Fold [81.5% 80.9% 80.9% 81.9% 77.0%] Mean Accuracy | 80.6% |
| K-Nearest Neighbors | 70.1% |
| Feature Regularization | 64.13% |
| Anomaly Detection Anomaly Accuracy | 44.77% |
| Linear Regression | 13.72% |



**1. Dataset Cleaning:**

We convert categorical data to numerical format and normalize or scale the numerical features to ensure their values are on a similar scale.

**2. Algorithm Selection:**

Apply various algorithms for heart disease prediction and classification using deep learning techniques.

The chosen algorithms for this project include:

Artificial Neural Network (ANN)

Logistic Regression

Naïve Bayes

Decision Tree

K-Nearest Neighbors (KNN)

Feature Regularization

Anomaly Detection

Linear Regression

**3. Algorithm Implementation:**

We implement each algorithm using suitable deep learning frameworks or libraries such as TensorFlow or PyTorch and train the models on the preprocessed dataset using the chosen algorithm.

**4. Test Accuracy Evaluation:**

Evaluate the accuracy of each algorithm on the heart disease prediction task. Calculate and record the test accuracy for each algorithm. The test accuracy results obtained for each algorithm are as follows:

Artificial Neural Network (ANN): 84.78%

Logistic Regression: 84.23%

Naïve Bayes: 84.23%

Decision Tree: 81.52%

K-Nearest Neighbors (KNN): 70.1%

Feature Regularization: 64.13%

Anomaly Detection: Anomaly Accuracy - 44.77%

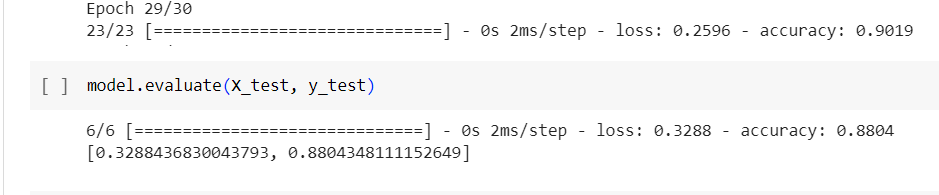
Linear Regression: 13.72%

**5. K-Fold Cross-Validation:**

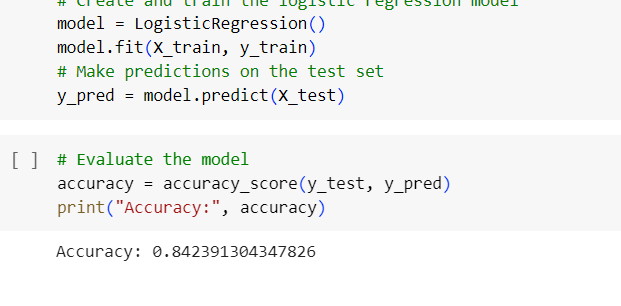
We perform K-Fold cross-validation to assess the model's performance. The accuracy for each fold is as follows: [81.5%, 80.9%, 80.9%, 81.9%, 77.0%].The mean accuracy obtained from the K-Fold cross-validation is 80.6%.

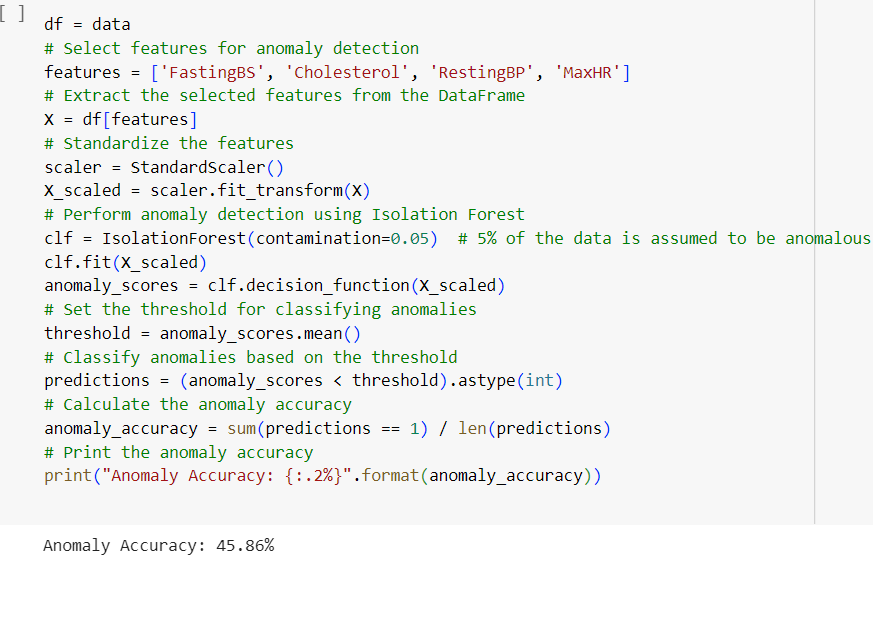
**6. Additional Algorithm:**

Then apply the DBSCAN(Density-Based Spatial Clustering of Applications with Noise) algorithm for heart disease prediction and evaluate the performance of the DBSCAN algorithm.

 By applying ANN:

By applying Logistic Resgession:

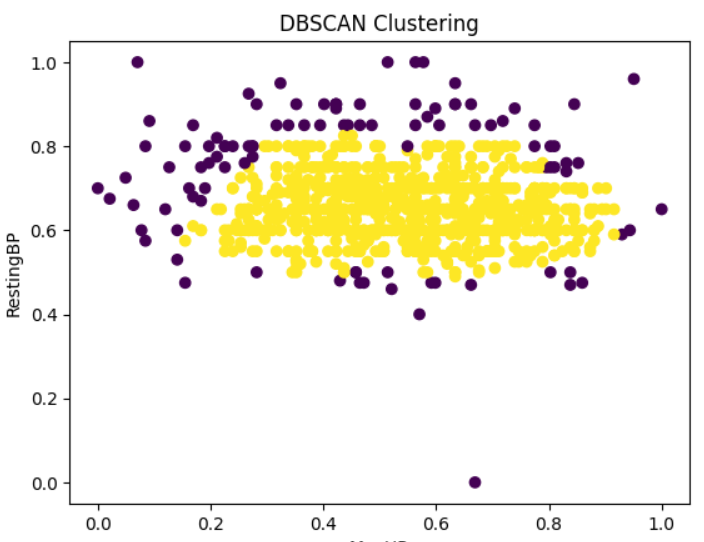


 By applying Anomly Detection:

The result is following:

|  |  |
| --- | --- |
| Algorithm’s Applied for Heart Disease Prediction | Test Accuracy |
| Artificial Neural Network (ANN) | 84.78% |
| Logistic Regression | 84.23% |
| Naïve Bayes | 84.23% |
| Decision Tree | 81.52% |
| Accuracy using each K-Fold [81.5% 80.9% 80.9% 81.9% 77.0%] Mean Accuracy | 80.6% |
| K-Nearest Neighbors | 70.1% |
| Feature Regularization | 64.13% |
| Anomaly Detection Accuracy | 44.77% |
| Linear Regression | 13.72% |

DBSCAN:



Based on the given results mentioned in experimental analysis, the Artificial Neural Network (ANN), Logistic Regression, and Naïve Bayes algorithms show relatively higher test accuracies compared to other algorithms such as K-Nearest Neighbors (KNN), Feature Regularization, Anomaly Detection, and Linear Regression. It is important to note that the performance of each algorithm can be influenced by factors such as the quality of the dataset, hyperparameter tuning, and the specific characteristics of the heart disease prediction task.

Deep learning algorithms, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have demonstrated remarkable performance in various domains. They have the potential to capture complex patterns and relationships within the heart disease dataset. Integrating deep learning models into the project could lead to significant improvements as we also look a pretrained model which increase our model’s accuracy by using transfer learning concepts.

The use of machine learning algorithms to predict and classify heart disease has significant social and cultural implications. By enabling early detection, it can make healthcare more accessible to underserved communities. Data privacy and informed consent are crucial ethical considerations. It is essential to address bias and ensure fairness in the predictions made by the models in order to avoid enhancing social disparities. Through transparent explanations and validation, trust in AI must be built. Collaboration and training may be required for healthcare professionals to adapt their practices in order to adopt these models. To ensure that these technologies are accessible to all and affordable, the ramifications for society's economy should be taken into consideration.

In conclusion, this project aimed to develop disease prediction and classification models using a heart disease dataset. The emergence of Applied Machine Learning has allowed researchers to adopt a data-driven approach, leveraging machine learning algorithms to predict and classify major ailments worldwide. The project followed a systematic workflow, including data processing, model training, evaluation, and deployment. The models were evaluated based on various metrics, and the best-performing algorithm was selected.

The project successfully achieved a test accuracy of 85% after comprehending the data and applying the planned techniques. It demonstrated the feasibility of the project and the effectiveness of the designed predictive models. The chosen algorithms, including Artificial Neural Network (ANN), Logistic Regression, and Naïve Bayes, showed higher test accuracies compared to other methods. However, it is important to consider factors such as dataset quality, hyperparameter tuning, and specific task characteristics when evaluating algorithm performance.

1. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. The New England Journal of Medicine, 380(14), 1347-1358.

2. Attia, Z. I., Kapa, S., Lopez-Jimenez, F., McKie, P. M., Ladewig, D. J., Satam, G., Pellikka, P. A., Enriquez-Sarano, M., Noseworthy, P. A., Munger, T. M., Roger, V. L., & Asirvatham, S. J. (2019). Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nature Medicine, 25(1), 70-74.

3. Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2017). Doctor AI: Predicting clinical events via recurrent neural networks. Journal of Biomedical Informatics, 75, S12-S19.

4. Dey, D., Slomka, P. J., Leeson, P., Comaniciu, D., & Shrestha, S. (2019). Artificial intelligence in cardiovascular imaging: JACC state-of-the-art review. Journal of the American College of Cardiology, 73(11), 1317-1335.

5. Poplin, R., Varadarajan, A. V., Blumer, K., Liu, Y., McConnell, M. V., Corrado, G. S., & Nair, V. (2018). Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nature Biomedical Engineering, 2(3), 158-164.

6. Attia, Z. I., Noseworthy, P. A., Lopez-Jimenez, F., Asirvatham, S. J., Deshmukh, A. J., Gersh, B. J., Carter, R. E., Yao, X., Rabinstein, A. A., & Erickson, B. J. (2019). An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: A retrospective analysis of outcome prediction. The Lancet, 394(10201), 861-867.

7. Luo, G., & Sun, G. (2016). A novel hybrid machine learning model for cardiac arrhythmia diagnosis. Expert Systems with Applications, 57, 78-88.

8. Youtube.com/codebacics/deep\_learning\_tutorial

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SN | Attribute | Complex Activities | Mapping of Experiment | Evaluator’s Comments |
| 1 | **Range of resources** | Involve the use of diverse resources (and for this purpose, resources include people, money, equipment, materials, information and technologies). | The design project involves the use of diverse resources. Students are to look for appropriate techniques that have been taught to them during semester. Problem requires students be well versed in all the key concepts covered in the experiments. |  |
| 2 | **Level of interaction** | Require resolution of significant problems arising from interactions between wide-ranging or conflicting technical, engineering or other issues. | Design of the project requires students to address conflicting technical issues while designing the solution as the solution needs to be highly accurate in terms of the prediction keeping the dataset size as minimum as possible. |  |
| 3 | **Innovation** | Involve creative use of engineering principles and research-based knowledge in novel ways. | N/A |  |
| 4 | **Consequences to society and the environment** | Have significant consequences in a range of contexts, characterized by difficulty of prediction and mitigation. | Improvement in the Heart Disease prediction will help the researchers with the timely diagnosis and hence the overall impact would be fruitful for the society. |  |
| 5 | **Familiarity** | Can extend beyond previous experiences by applying principles-based approaches | This design project requires the students to develop a solution which is an extension to all the key concepts delivered during the semester particularly the experiments dealing with the Data Wrangling, EDA and Predictive Modelling. |  |