**Disease Prediction System**

****

ITS307 DATA ANALYTICS

BACHELOR OF SCIENCE IN INFORMATION TECHNOLOGY (YEAR III, SEMESTER I)

**RESEARCHER (S)**

PEMA DENDUP (12200070)

SONAM PELKI (12200081)

UGYEN KEZANG (12200094)

UGYEN LHAMO (12200095)

**GUIDED BY**

NIMA DEMA

*Gyalpozhing College of Information Technology Gyalpozhing, Mongar*



**Abstract**

This report represents the mini-project assigned to fifth semester students for fulfillment of ITS307, Data Analytics, given by the department of IT, GCIT. Cardiovascular disease and diabetes are the most common cause of death worldwide over the last few decades in the developed as well as underdeveloped and developing countries. Early detection of these diseases can reduce the mortality rate. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by doctor is not available since it requires more sapience, time and expertise.

In this project, we have developed and researched about models for heart and diabetes disease prediction through techniques like logistic regression and SVM on the dataset available publicly in Kaggle Website, further evaluating the results using accuracy score. The early prognosis of cardiovascular and diabetes disease can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine.

Keywords: *Machine Learning; Logistic Regression; SVM; Cardiovascular Diseases; Diabetes Disease;*

**Table of Contents**

| **SI.NO** | **Topics** | **Page No.** |
| --- | --- | --- |
| **1** | **Introduction**  1.1. Problem Statement  1.2. Aim  1.3. Objective  1.4. Scope & Limitations  1.5. Background | 3-4 |
| **2** | **Related Work** | 5-6 |
| **3** | **Methodology**  3.1. Proposed Overview  3.2. Algorithm  3.3. Dataset  3.4. Evaluation Metrics  3.5. Experimental Setup | 7-9 |
| **4** | **Result and Discussion** | 10 |
| **5** | **Conclusion** | 11 |
| **6** | **Reference** | 12 |

1. **Introduction**

According to the World Health Organization, every year 12 million deaths occur worldwide due to Heart Disease as well as Diabetes. The load of cardiovascular disease and diabetes is rapidly increasing all over the world from the past few years. Many researches have been conducted in an attempt to pinpoint the most influential factors of the heart and diabetes as well as accurately predict the overall risk. Heart disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The early diagnosis of heart disease as well as diabetes plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reducing the complications. This project aims to predict future Heart Disease and Diabetes Disease by analyzing data of patients which classifies whether they have the disease or not by using machine-learning algorithms.

**1.1. Problem statement**

The traditional approach of diagnosis entails a patient visiting a doctor, undergoing many medical tests, and then reaching a conclusion. This process is time-consuming and indolence people may ignore the checkup.

**1.2. Aims**

To propose an automated disease prediction system.

**1.3. Objective**

The main objective of developing this project are:

1. To develop machine learning models to predict future possibility of heart and diabetes diseases by implementing Logistic Regression and Support Vector Machine.
2. To determine significant risk factors based on medical dataset which may lead to heart and diabetes diseases.

**1.4. Scope and Limitations**

A. User Scope:

* World

B. System Scope:

* Website (online)
* Phone
* Laptop

C. Limitations

* Limited Disease Prediction:
* The model predicts Diabetes and Heart Disease. All diseases cannot be predicted.

D. No Recommendation:

* The model just predicts the disease. There will be no recommendation of what to do after predicting the disease

**1.5 Background**

This project consists of two models. One model is for heart disease prediction and another model is for diabetes disease prediction.

As for the heart disease model, we have used logistic regression and for diabetes, a support vector machine classifier is used to train the models. Both the models are used to detect the diseases accordingly, and help them reduce the complications.

1. **Related Work**

With growing development in the field of medical science alongside machine learning, various experiments and research has been carried out in recent years releasing the relevant significant papers.

[1]Research has been done in this field and people have produced methods to predict heart disease using supervised machine learning algorithms. Several research papers have been written on this topic. In the journal done by the IJCRT organization, prediction of the user’s disease is done using a random forest classifier algorithm. The accuracy achieved using Random forest classifiers for each disease such as Diabetes Model 98.25%, Breast Cancer Model 98.25 %, Heart Disease Model 85.25%, Kidney Disease Model 99% and Liver Disease Model 78%.

[2]In another article, machine learning is used in detecting if a person has a heart disease or not. Machine learning has been used to detect whether a person is suffering from a cardiovascular disease by considering certain attributes like chest pain, cholesterol level, age of the person and some other attributes. In this paper, they have used KNN and random forest algorithms to classify people who have a heart disease from people who do not. The prediction accuracy obtained by K-Nearest Neighbor is 86.885% and the prediction accuracy obtained by the Random Forest algorithm is 81.967%.

[3] 2019 3rd International Conference on Computing Methodology Communication (ICCMC) states that to overcome the problem of not being able to predict the disease at an early stage on the basis of symptoms by the doctor, data mining was introduced. It has used K-Nearest Neighbors (KNN) and Convolutional Neural Network (CNN) machine learning algorithms for accurate prediction of disease. The accuracy of general disease prediction by using CNN is 84.5% which is more than KNN. Not only that, the time and the memory requirement for KNN was more compared to CNN. The system was able to give the risk associated with general disease after the prediction.

[4] At the 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), it was stated that cardiovascular disease (heart) is a crucial reason for mortality. So, early recognition of cardiovascular infections and continual management of clinician can decrease the death rate. The objective of this paper was to sum up the new examination along relative outcomes on coronary illness expectation and furthermore construct scientific ends by using techniques of data mining and classification using machine learning.

[5] “Multi Disease Prediction Using Data Mining Techniques” (2017).  
In this work, the performance of two distinct data mining classification algorithms was evaluated in order to determine the best classifier for the prediction of various diseases. The development of accurate and computationally effective classifiers for medical applications is a significant challenge in the fields of data mining and machine learning. Moreover this research evaluates the effectiveness of data mining-based disease prediction in medicine. The classifier categorized the data on medical diagnoses for diseases including cancer, liver issues, heart illness, and so forth. In terms of data classification, SVM method outperforms traditional cluster ensemble technique.

1. **Methodology**

**Proposed Methods**

**3.1. System Overview**

For the disease prediction, there are going to be two models; one for Heart Disease Prediction and another for Diabetes Disease Prediction.

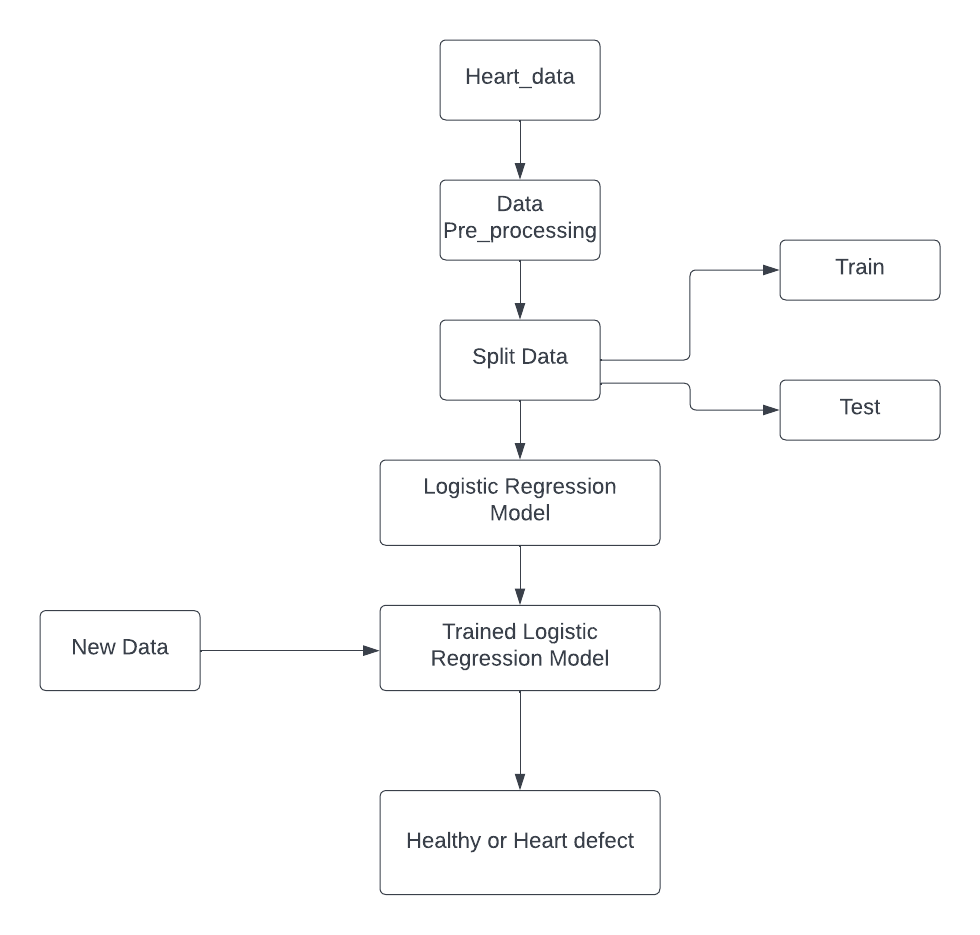
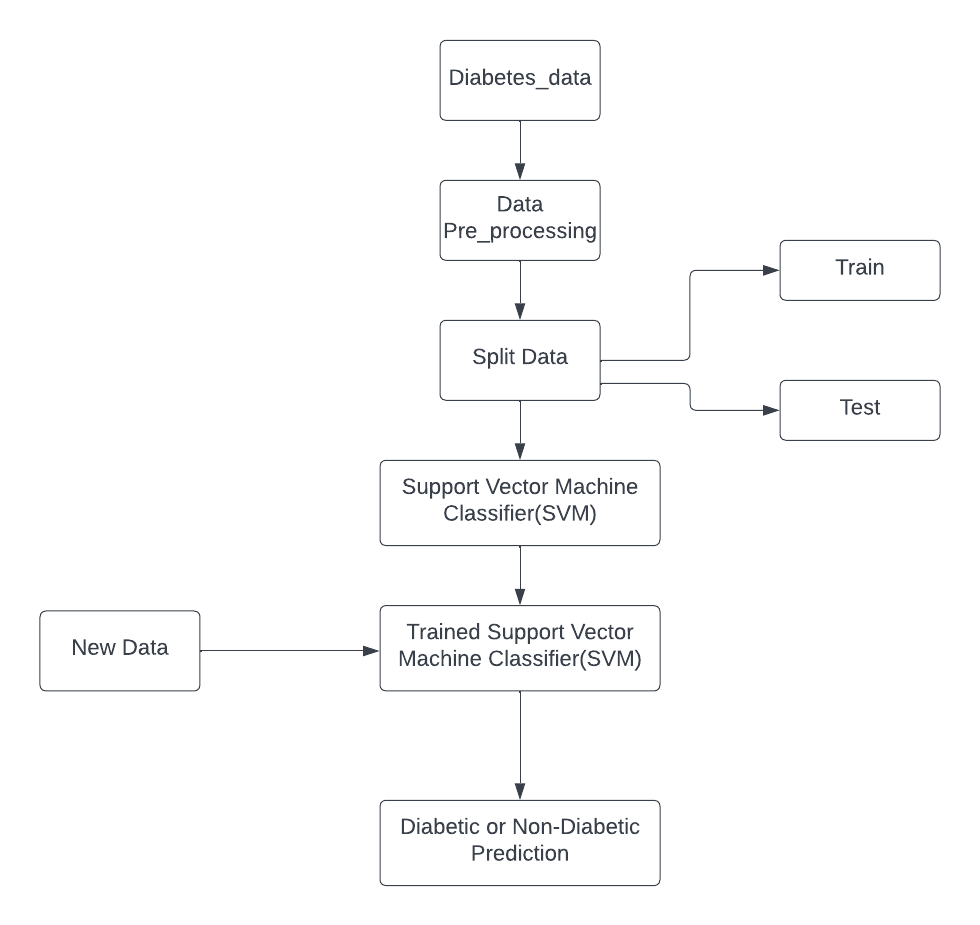
 

Figure 1: Heart Disease Figure 2: Diabetes Disease

**3.2. Algorithm**

Some of the algorithm used in our projects:

I. Logistic Regression

Logistic Regression is a supervised classification algorithm. It is a predictive analysis algorithm based on the concept of probability. For the heart disease prediction system we used Logistic Regression. Because our data is binary classification, we are going to classify whether the person has heart disease or not. Since Logistic Regression is best algorithm when it comes to binary classification

II. Support Vector Machine Classifier (SVM)

For the diabetes prediction system, the SVM algorithm is used. The SVM algorithm primarily creates a line to separate the dataset into classes, enabling it to decide the test data into which classes it belongs. The line or decision boundary is called a hyperplane. The algorithm works on two types: linear and nonlinear.   
Linear SVM is used when the dataset comprises two classes and is separable. In case of inseparable dataset, a nonlinear SVM is applied, where the algorithm converts the original coordinate area into a separable space. There can be multiple hyperplanes, and the best hyperplane is chosen with the maximum margin between data points. The dataset closest to the hyperplane is called a support vector.

**3.3. Dataset**

The dataset is from the Kaggle, for both the models (i.e., heart disease prediction and diabetes prediction system).

Both the dataset are structured as it is in csv file format.

1. Heart Disease Prediction

The dataset comprises 10 numerical features including the target.

The features include:

i. age

ii. Sex

iii. Chest pain type (cp) with four values

iv. Maximum Heart Rate achieved (thalach)

v. Exercise induced angina (exang)

vi. Oldpeak(oldpeak) = ST depression induced by exercise relative to rest.

vii. Slope of the peak exercise ST segment.

viii. Number of Major Vessels (0-3) colored by fluoroscopy (ca)

ix. Thal: 0=normal; 1=fixed defect; 2=reversal defect.

x. Target: 0=normal; 1=heart disease.

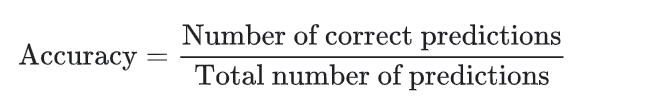
1. Diabetes Disease Prediction

The dataset comprises 9 features including the feature.

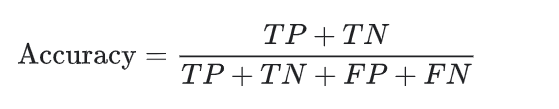
Features include Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, Diabetes Pedigree Function, Age, and Outcome (Target).

**3.4. Evaluation Metrics**

To evaluate the performance of the model we have used an accuracy score. It is calculated by dividing the number of correct predictions by the total prediction number.



For binary classification, accuracy can also be calculated in terms of positive and negative as:



Where TP=True Positives, TN=True Negatives, FP=False Positives, FN=False Negatives.

**3.5. Experimental Setup**

* Programming language: Python 3
* Platform for training a model: Jupyter Notebook
* Machine Learning Libraries: Pandas, NumPy, Sci-Kit Learn.
* Python Libraries for Visualization: Matplotlib and Seaborn.

1. **Result and Discussion**
2. Heart Disease Model

Initially, the dataset consisted of 13 features excluding the target, depending on the correlation of the features, 4 features were dropped and only 9 features were considered while developing the model.

On using the Logistic Regression on Heart Disease Model, we got the training data accuracy of 85.95% and testing data accuracy of 78.68%.

1. Diabetes Disease Model

There are 8 features considered excluding the outcome. On using the Support Vector Machine classifier, the training data accuracy was of 77.83% and testing data accuracy was of 81.07%.

When it comes to understanding the accuracy scores: if the accuracy is over 90%, it is very good; 70%-90% is good; between 60% and 70% is OK and below 60% is poor.

Since both of our models have both its training as well as testing accuracy above 70%, our model is good to go and is realistic as well.

**Conclusion**

The early prognosis of cardiovascular and diabetes diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine. This project consisted of two models; one for heart using logistic regression and another for diabetes using Support Vector Machine. Further for its enhancement, we can train on models and predict the types of cardiovascular disease as well as type of diabetes, providing recommendations to the user and also use more enhanced models.

**References**

R. Alanazi, “Identification and prediction of chronic diseases using machine learning approach,” *Journal of Healthcare Engineering*, vol. 2022, pp. 1–9, 2022. Retrieved from <https://www.hindawi.com/journals/jhe/2022/2826127/>

D. Dahiwade, G. Patle and E. Meshram, “Designing disease prediction model using machine learning approach - IEEE xplore.” . Retrieved from <https://ieeexplore.ieee.org/document/8819782>.

“Heart disease prediction using machine learning techniques - iopscience.” Available: <https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012046>.

“Multi disease prediction using data mining techniques - researchgate.” Available: <https://www.researchgate.net/publication/319851535_Multi_Disease_Prediction_using_Data_Mining_Techniques>.

S. Verma and A. Gupta, "Effective Prediction of Heart Disease Using Data Mining and Machine Learning: A Review," *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, 2021, pp. 249-253, doi: 10.1109/ICAIS50930.2021.9395963. Retrieved from <https://ieeexplore.ieee.org/document/9395963>