IIT School of Applied Technology
ILLINOIS INSTITUTE OF TECHNOLOGY

ITMS 525 Project Proposal
Data Mining
Department of Information Technology Management

Group Number: Cardio Disease Prediction

First Name	Last Name	Email (hawk.iit.edu)	Student ID
Vaishnavi	Chaudhari	vchaudhari@hawk.iit.edu	A20446094
Megha	Shrivastava	mshrivastava1@hawk.iit.edu	A20450886

Table of Contents

1. Introduction

In today's world, there is a substantial increase in heart diseases. The data set that we have selected predicts whether an individual will suffer from cardio disease or not. The prediction of the disease is impacted by various features. The data set used here has information related to the cardio disease. According to the dataset, the presence of the cardio disease is predicted using id number, age, gender, height, weight, systolic blood pressure, diastolic blood pressure, cholesterol level, glucose, smoking habits, alcohol consumption, and day to day activities. The model will be predicting, classifying, and clustering the data according to the problem statement. The model presented can be useful to detect if an individual following certain trends and habit is prone to cardio disease or not.

2. Data Sets

The dataset is a cardio disease dataset. The data can be referred from:
https://www.kaggle.com/raminhashimzade/cardio-disease.
The data frame with 70000 rows and 14 variables- 13 feature variables and 1 predict variable. The variables present in the data frame are: **id** (number of observations) , **age** (age of the patient in days), **age** (age of the patient in years), **gender** (gender of the patient where 1 – women, 2 - men), **height** (height of the patient in cm), **weight** (weight of the patient in kg), **ap_hi** (systolic blood pressure of the patient), **ap_lo** (diastolic blood pressure of the patient), **cholesterol** (cholesterol level of the patient where 1:normal, 2:above normal, 3:well above normal), **gluc** (glucose level of the patient where 1:normal, 2:above normal, 3:well above normal), **smoke** (whether patient smokes or not where 0 = no, 1= yes), **alco** (amount of alcohol consumption of the patient which is a binary feature where 0 = no, 1 = yes), **active** (Active live represented by binary feature where 0 = if the patient have passive life, 1 = if the patient have active life), **cardio** (It is the target variable represented by 0 and 1, 0 = if the patient does not have cardio disease and 1 = if the patient have cardio disease).

The data type of the variables
Id, age, height, weight, ap_hi, ap_lo: Discrete variables (Quantitative datatype)
Gender, cholesterol, gluc: Ordinal variable (Qualitative datatype)
Smoke, alco, active, cardio: Binary variable (Qualitative datatype)

3. Research Problems

The problem statement is: **To detect whether the patient will have a cardio disease or not based on the attribute or the features associated with it. Classifying the patient's data into patients with the cardio disease and without cardio disease and grouping them together. Trying out different classification algorithms to find the model with the best classifying abilities.** If it is predicted that the patient will suffer from the cardio disease, then the value is 1 and if it is predicted that the patient will not suffer from the cardio disease then the value is 0. In the given dataset we have several factors that affect the cause of cardio disease, so it is important to find out how much each attribute has a contribution to the changing value of the predicted variable.

4. Potential Solutions

To address the problem mentioned in the problem statement we need **a predictive model, classification models, and clustering models** that can be used to predict, classify and cluster the values of the cardio variable with changing values of the x variables (attribute values). To build a **predictive** model we initially need to have a regression model of the data set selected. For building this model we need to preprocess the data variables. Build a multiple regression model, validate the model. To get the best model we need to select those features which have the maximum effect on the cardio variable.
To **classify** the patients, we need to build a classification of the data set selected. For building these models we need to preprocess the data variables according to each model. To get the best model we need to select those features which have the maximum effect on the cardio variable.

For this dataset, we have one **dependent** variable which will be predicting the values of the possibility of the patient having the cardio disease. This variable is the **cardio** variable.
The **independent variables** for this predictive model will be the attributes or the features used by one to determine the condition of the patient. The independent variables present in the dataset are – Id, age, height, weight, ap_hi, ap_lo, Gender, cholesterol, gluc, Smoke, alco, active.

5. Evaluations

5.1 Methods and Process

1. Prediction:

As the data set has inconsistencies in the data values, we need to first preprocess the data values to use them for building a predictive model. We will check the correlations of the independent variables with the dependent variable. The plots can be used to analyze the correlation as well. If the correlation is not up to the mark, we must perform transformations on the data variable. If there are some missing values in the data set, then these values will be added by using the average of the numerical data. The multicollinearity, influential point, and interaction terms must be addressed.
The next step will be building a model based on N fold validation as the data set has more than 5000 entries. The dataset here must be split into the training dataset and the testing dataset. The model will be made using the training dataset with a feature selection technique Backward elimination based on the p-value. The model will be validated by checking if it passes the goodness of fit test and by residual analysis. The goodness of fit will have: Goodness of fit test (F-test), Individual Parameter Test, Coefficient of determination R2 the plot analysis for constant variance, linearity and normal distribution will be done. To check the normality Shapiro-Wilk test will be used. Potential outliners have to be checked. We will be using **the multiple regression model.**

2. Classification:

This dataset has some redundant columns and some missing values. Performing classification on unclean data can reduce the accuracy of the model. Hence, we need to perform data cleaning. Data must be data preprocessed based on the requirements of each model. For example, KNN model requires the data to be scaled for accurate prediction. If there are some missing values in the data set, then these values will be added by using the average of the numerical data. The multicollinearity, influential point and interaction terms must be addressed.

To increase the accuracy of the model we need to perform feature selection or feature reduction technique. We will be using the PCA technique to perform feature reduction. The dataset here must be split into training dataset and the testing dataset. The model will be trained using the train dataset and its accuracy will be tested on the test dataset. The models will have different accuracy and the model with the highest accuracy will be selected. We will be using **KNN, Naïve-bayes, Decision tree, SVM, Random forest and Neural networks.**

5.2 Evaluations and Results

The model built above will be tested on the test data set. The evaluation of the model will be done based on the value of **the root mean squared error** (RMSE) for the prediction part and different **accuracy matrices** for the classification part. RMSE is the measure of how much the data points are aligned with the line. This will determine how dependable our model is and how well can it predict the values of the dependent variable. The accuracy matrices for different models will measure how accurate our model can predict the target variable. A model is expected to have high accuracy to be selected.

5.3 Findings

The finding of this project is to monitor if the patient is prone to cardio disease or not based on the changing attributes. The model must be able to predict, classify and cluster the accurate value of the target variable. For this, the model must be accurate with minimum errors.

6. Expected Outcomes

6.1 Conclusions

The prediction model build would be used to predict the possibility of the disease based on the independent attributes listed. This model will be useful to predict and classify patients with all different combinations of values for feature variables. This model can be used in the market to predict and classify patients with or without the disease based on their daily habits.
Predict variable will change with respect to the changing features in the dataset. This forecasting of the patient's state of health will be very useful if it could be predicted beforehand.

6.2 Limitations

The limitation of this model is that there are other features that will affect whether the patient will suffer from cardio disease or not, these features are not present in the current dataset. Sometimes there are exceptions of patients having all positive signs of cardio disease but do not have it. This can be misleading and hence is responsible for reducing the accuracy of the model.

6.3 Potential Improvements or Future Work

The classified data can be clustered together into clusters of patients having the disease or not. For clustering we will be using K-Means Clustering, Mean shift Clustering, Centroid based Clustering. The visualization can be performed on the dataset to help better understand the pattern in the data and can be useful for instant insights into it.