ABSTRACT

Nowadays, health disease are increasing day by day due to lifestyle, hereditary. Especially, heart disease has become more common these days, i.e. life of people is at risk. Each individual has different values for Blood pressure, cholesterol and pulse rate. But according to medically proven results the normal values of Blood pressure is 120/90, Cholesterol is 100-129 mg/dL ,Pulse rate is 72, Fasting Blood Sugar level is 100 mg/dL ,Heart rate is 60-100 bpm, ECG is normal, Width of major vessels is 25 mm (1 inch) in the aorta to only 8 μm in the capillaries. Here we do the survey about different classification techniques used for predicting the risk level of each person based on age, gender, Blood pressure, cholesterol, pulse rate.

“Disease Prediction” system based on predictive modeling predicts the disease of the user on the basis of the symptoms that user provides as an input to the system. The system analyzes the symptoms provided by the user as input and gives the probability of the disease as an output. Disease Prediction is done by implementing 5 techniques such as Naïve Bayes, KNN, Decision Tree, Linear Regression and Random Forest Algorithms. These techniques calculate the probability of the disease. Therefore, average prediction accuracy probability 83% is obtained.

***Existing Model:***

**Prediction system for heart disease using Naive Bayes and particle swarm**

* Particle swarm optimization (PSO)

PSO is an Evolutionary Computation technique is proposed by Kennedy et al. in 1995. PSO is motivated by social behaviors such as bird flocking and fish schooling. In PSO population swarm consists of “n” particles, and the position of each particle stands for the potential solution in D-dimensional space. The particles change its condition based on three aspects: To keep its inertia; To change the condition according to its most optimist position; To change the condition according to the swarm’s most optimist position[30]. In PSO, a population are encoded as particles in the search space dimensionality D.PSO starts with the random initialization of a population of particles. Based on the best experience of one particle (pbest) and its neighboring particles (gbest), PSO searches for the optimal solution by updating the velocity and the position of each particle; PSO is used as feature subset selection method due to its advantages:

* Simple and easy to implement.
* Continuous optimization approach.
* Naïve Bayes’ Classifier

Naive Bayes classifiers are a family of simple probabilistic classifiers based by using Bayes theorem with s t r o n g (Naive) independence assumptions between the features. Naive Bayes classifiers are highly scalable by requiring several parameters linear for the number of features or predictors as variable in a learning problem.[112] It is the simplest and the fastest probabilistic classifier especially for the training phase.

**Feature selection** - It is a process of removing the irrelevant and redundant features from dataset based on evaluation criterion which is used to improve accuracy. There are two approaches as individual evaluation and other one is subset evaluation. The process of feature selection is classified into three broad classes. One is filter and another one is wrapper and third one is embedded method based on how the feature selection is deployed by supervised learning algorithm. In this paper, they proposed a model which uses Naive Bayes as classifier and PSO as Feature subset selection measure for prediction of heart disease.[113]

**Proposed system** - In this section, we propose a methodology to improve the performance of Bayesian classifier for prediction of heart disease. Algorithm for our proposed model is shown below:

**Algorithm 1**: Heart disease prediction by using Bayes classifier and PSO.

* Input: Heart disease dataset.
* Output: Classify patient dataset into heart disease or not (normal).
* Step 1: Read the dataset.
* Step 2: Apply particle swarm optimization for f ea ture selection.
* Step 3: Remove the features with low value of PSO.
* Step 4: Apply Naive Bayes classifier on relevant features. Step 5: Evaluate the performance of NB+PSO model.

The above algorithm divided into two sections, section 1 (step 2 and step 3) performs processing and feature subset selection. In section 2 (step 4 and step 5) Naive Bayes is applied on relevant features data and evaluate the performance in terms of accuracy.[114]

Accuracy= (No. of objects correctly classified/Total no. of objects in test set)

Cross validation technique used to split into training and testing data.

**Predictive Data Mining for Medical Diagnosis: Heart Disease Prediction**

The successful application of data mining in highly visible fields like e-business, marketing and retail has led to its application in other industries and sectors However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. Number of experiment has been conducted to compare the performance of predictive data mining technique on the same dataset and the outcome reveals that Decision Tree outperforms and sometime Bayesian classification is having similar accuracy as of decision tree but other predictive methods like KNN, Neural Networks, Classification based on clustering are not performing well. The second conclusion is that the accuracy of the Decision Tree and Bayesian Classification further improves after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute enough for heart disease prediction.[9][115]

* Data Mining in the Heart Disease Prediction.

Three different supervised machine learning algorithms i.e. Naive Bayes, K-NN, Decision List algorithm have been used for analyzing the dataset in. Tanagra tool is used to classify the data and the data is evaluated using 10-fold cross validation and the results are compared. Tanagra is a data mining suite build around graphical user interface algorithms. Decision Tree is a popular classifier which is simple and easy to implement. It requires no domain knowledge or parameter setting and can handle high dimensional data. The results obtained from Decision Trees are easier to read and interpret. The drill through feature to access detailed patients‟ profiles is only available in Decision Trees. Naïve Bayes is a statistical classifier which assumes no dependency between attributes. It attempts to maximize the posterior probability in determining the class. The advantage of using naive Bayes is that one can work with the naive Bayes model without using any Bayesian methods. Naive Bayes classifiers have works well in many complex real-world situations. The k-nearest neighbor’s algorithm (k-NN) is a method for classifying objects based on closest training data in the feature space. k-NN is a type of instance-based learning. The k-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms. But the accuracy of the k-NN algorithm can be severely degraded by the presence of noisy or irrelevant features, or if the feature scales are not consistent with their importance. The experiment is performed using training data set consists of 3000 instances with 14 different attributes. The dataset is divided into two parts that is 70% of the data are used for training and 30% are used for testing.

* Data Mining and Artificial Neural network

Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques, namely, Decision Trees, Naïve Bayes and Neural Network. is implemented in using .NET platform. IHDPS i s Web-based, user-friendly, scalable, reliable and expandable system. It can also answer complex “what if” queries which traditional decision support systems cannot. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. It enables significant knowledge, e.g. patterns, relationships between medical factors related to heart disease. As a Data source a total of 909 records with 15 medical attributes (factors) were obtained from the Cleveland Heart Disease database. Figure 1 lists the attributes. The records were split equally into two datasets: training dataset (455 records) and testing dataset (454 records). Initially, the data warehouse is pre- processed in order to make it suitable for the mining process. Once the preprocessing gets over, the heart disease warehouse is clustered with the aid of the K-means clustering algorithm, which will extract the data appropriate to heart attack from the warehouse. Consequently, the frequent patterns applicable to heart disease are mined with the aid of the MAFIA algorithm from the data extracted. In addition, the patterns vital to heart attack prediction are selected on basis of the computed significant weightage. The neural network is trained with the selected significant patterns for the effective prediction of heart attack. Multi-layer Perceptron Neural Network with Back-propagation is being used as the training algorithm. In feed-forward neural networks the neurons of the first layer forward their output to the neurons of the second layer, in a unidirectional fashion, which explains that the neurons are not received from the reverse direction.[116][24]

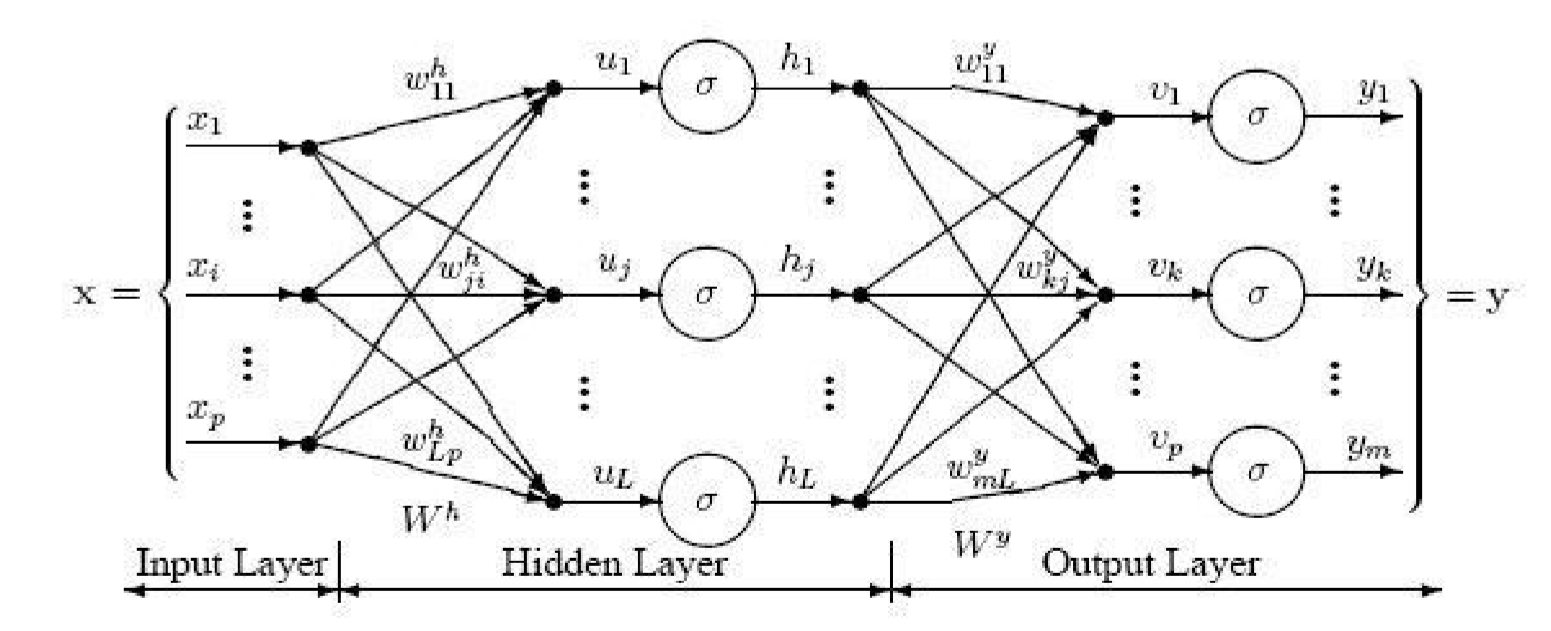
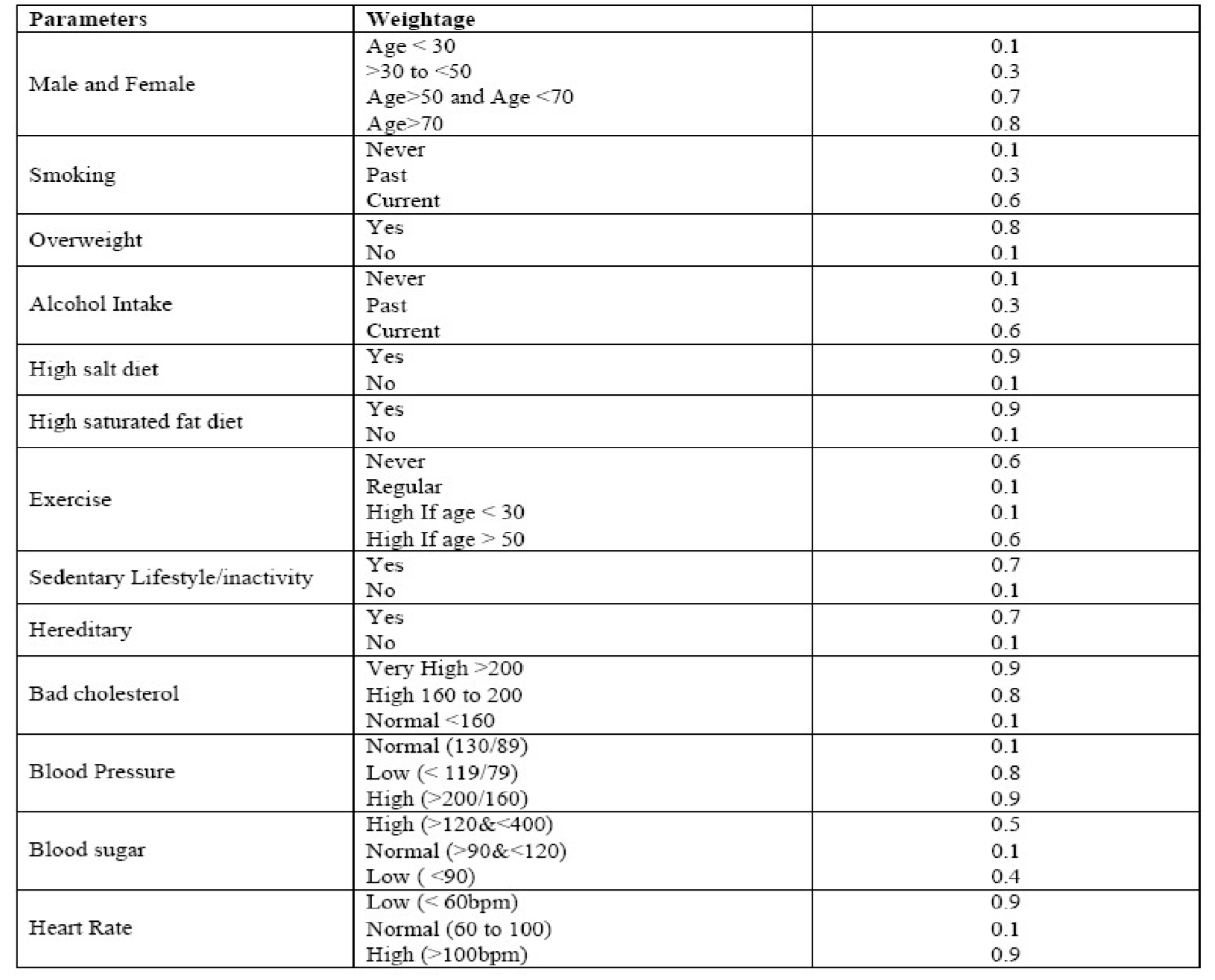


Figure 12 The sample combinations of heart attack parameters for normal and risk level along with their values and weightages

* Data Mining and Genetic Algorithm

Genetic algorithm has been used in, to reduce the actual data size to get the optimal subset of attributed sufficient for heart disease prediction. Classification is a supervised learning method to extract models describing important data classes or to predict future trends. Three classifiers e.g. Decision Tree, Naïve Bayes and Classification via clustering have been used to diagnose the presence of heart disease in patients. Classification via clustering: Clustering is the process of grouping similar elements. This technique may be used as a preprocessing step before feeding the data to the classifying model. The attribute values need to be normalized before clustering to avoid high value attributes dominating the low value attributes. Further, classification is performed based on clustering. Experiments were conducted with Weka 3.6.0 tool. Data set of 909 records with 13 attributes. All attributes are made categorical and inconsistencies are resolved for simplicity. To enhance the prediction of classifiers, genetic search is incorporated. Observations exhibit that the Decision Tree data mining technique outperforms other two data mining techniques after incorporating feature subset selection but with high model construction time. Naïve Bayes performs consistently before and after reduction of attributes with the same model construction time.[117]

* Association Rule Discovery

Association rules represent a promising technique to improve heart disease prediction. Unfortunately, when association rules are applied on a medical data set, they produce an extremely large number of rules. Most of such rules are medically irrelevant and the time required to find them can be impractical. In, four constraints were proposed to reduce the number of rules: item filtering, attribute grouping, maximum itemset size, and antecedent/consequent rule filtering. When association rules are applied on a medical data set, they produce an extremely large number of rules. Most of such rules are medically irrelevant and the time required to find them can be impractical. A more important issue is that, in general, association rules are mined on the entire data set without validation on an independent sample. To solve these limitations, the author has introduced an algorithm that uses search constraints to reduce the number of rules, searches for association rules on a training set, and finally validates them on an independent test set. Instead of using only Support and confidence, one more parameter i.e. lift have been used as the metrics to evaluate the medical significance and reliability of association rules. Medical doctors use sensitivity and specificity as two basic statistics to validate results. Sensitivity is defined as the probability of correctly identifying sick patients, whereas specificity is defined as the probability of correctly identifying healthy individuals. Lift was used together with confidence to understand sensitivity and specificity. To find predictive association rules in a medical data set the algorithm has three major steps. First, a medical data set with categorical and numeric attributes is transformed into a transaction data set. Second, four constraints mentioned above are incorporated into the search process to find predictive association rules with medically relevant attribute combinations. Third, a train and test approach are used to validate association rules.[118][119]

* **Issues and Challenges**

Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial. Clinical decisions are often made based on doctor’s intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Data mining have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions.

***Proposed System:***

In this project we have used 5 algorithms to find out the reasons of heart disease and create a model to get the maximum accuracy possible. Data set is taken from UCI repository. From this dataset we have used 5 widely used algorithms Logical Regression, Random forest, K-Nearest Neighbors and Decision Tree to create the model with the maximum accuracy possible. We have also explored precision score, recall score, F-score, false negative using confusion matrix for every algorithm used.

# ***System Requirements***

1. Intel® Core™ i5 processor 8250U at 1.60 GHz or 1.80 GHz, 8 GB of DRAM.
2. Disk space 2TB.
3. Operating System: 64-bit Windows 10 Pro.

Recommended System Requirements

* Processors: Intel® Core™ i5 processor 4300M at 2.60 GHz or 2.59 GHz (1 socket, 2 cores, 2 threads per core), 8 GB of DRAM.
* Operating systems: Windows® 10, macOS\*, and Linux\*

Minimum System Requirements

* Processors: Intel Atom® processor or Intel® Core™ i3 processor
* Disk space: 1 GB
* Operating systems: Windows\* 7 or later, macOS, and Linux
* Python\* versions: 2.7.X, 3.6.X
* Included development tools: conda\*, conda-env, Jupyter Notebook\* (IPython)
* Compatible tools: Microsoft Visual Studio\*, PyCharm\*Included Python packages - NumPy, SciPy, scikit-learn\*, pandas, Matplotlib, Numba\*, Intel® Threading Building Blocks, pyDAAL, Jupyter, mpi4py, PIP\*, and others.

Software

* PIP and NumPy: Ubuntu\*, Python 3.6.2, NumPy 1.13.1, scikit-learn 0.18.2
* Windows: Python 3.6.2, PIP and NumPy 1.13.1, scikit-learn 0.18.2
* Intel® Distribution for Python\* 2018

Modifications

* Scikit-learn: Conda\*-installed NumPy with Intel® Math Kernel Library (Intel® MKL) on Windows (PIP-installed SciPy on Windows contains Intel MKL dependency).