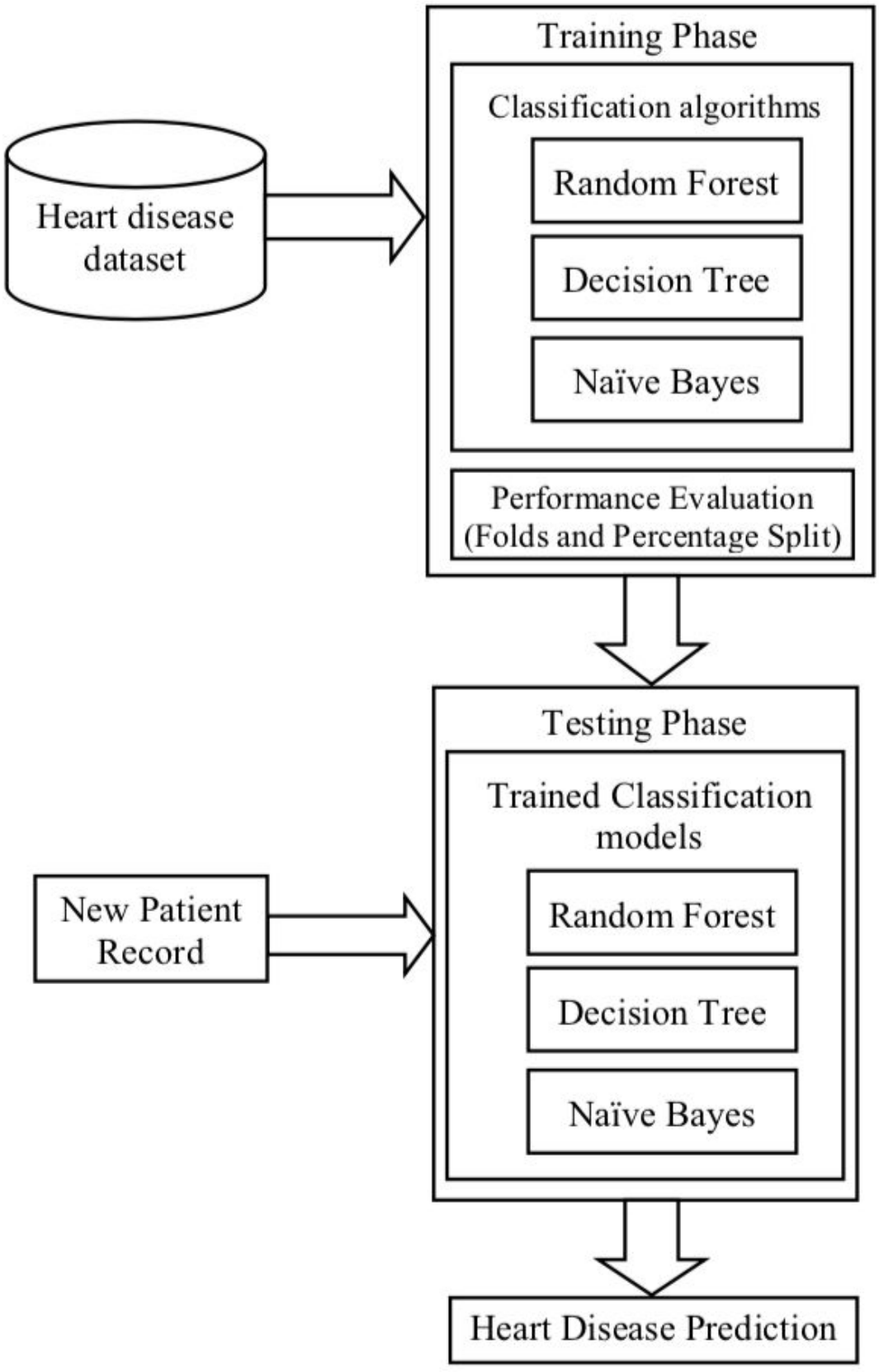
Geraldo Braho
North American University
Comp 5353

HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES

In today's world of stress Heart, being an essential organ in a human body which pumps blood through the body for the blood circulation is essential and its health is to be conserved for a healthy living.The health of a human heart is based on the experiences in a person's life and is completely dependent on professional and personal behaviours of a person.
According to the World Health Organization, every year more than 12 million deaths are occurring worldwide due to the various types of heart diseases which is also known by the term cardiovascular disease.The increase in the possibility of heart disease among young may be due to the bad eating habits, lack of sleep, restless nature, depression and numerous other factors such as obesity, poor diet, family history, high blood pressure, high blood cholesterol, idle behaviour, family history, smoking and hypertension. A risk of a heart attack or the possibility of the heart disease if identified early, can help the patients take precautions and take regulatory measures. Recently, the healthcare industry has been generating huge amounts of data about patients and their disease diagnosis reports are being especially taken for the prediction of heart attacks worldwide. When the data about heart disease is huge, the machine learning techniques can be implemented for the analysis.

Data Mining is a task of extracting the vital decision making information from a collective of past records for future analysis or prediction. Classification is one data mining technique through which the future outcome or predictions can be made based on the historical data that is available. In this research work that I read , the supervised machine learning concept is utilized for making the predictions. A comparative analysis of the three data mining classification algorithms namely Random Forest, Decision Tree and Naïve Bayes are used to make predictions.

5. METHODOLOGY



CLASSIFICATION USING RANDOM FOREST

Random forests (RF) are combination of tree predictors using decision tree such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest It is a supervised classification algorithm used for the prediction and it is considered as the superior due to its large number of trees in the forest giving improved accuracy than decision trees.

The algorithm for random forest is given below:

- Randomly select k features from where $k << m$.
- Surrounded by the k features, node "d" using the best split point.
- Split the node into daughter nodes using the best split.
- Repeat 1 to 3 steps until l number of nodes has been reached.
- Construct forest by repeating steps 1 to 4 for n number times to create n number of trees.

CLASSIFICATION USING DECISION TREE

Decision tree builds classification or regression models in the structure of a tree making it simple to debug and handle. Decision trees can handle both categorical and numerical data

The algorithm for the decision tree is given below:

- Identify the information gain for the attributes in the dataset.
- Sort the information gain for the heart disease datasets in descending order.
- After the identification of the information gain assign the best attribute of the dataset at the root of the tree.
- Then calculate the information gain using the same formula.
- Split the nodes based on the highest information gain value.
- Repeat the process until each attributes are set as leaf nodes in all the branches of the tree.

CLASSIFICATION USING NAÏVE BAYES

Naïve Bayes (NB) is a statistical classifier which assumes no enslavement between attributes

The working principle of naïve Bayes classifier is as follows:

• *Training Step*:
• *Prediction Step*

Table.3. Classification of heart disease using Percentage Split

(A)	(F)	Metrics							
		TP Rate	**FP Rate**	**Precision**	**Recall**	**F-measure**	**MCC**	**ROC Area**	**PRC Area**
NB	2	0.56	0.56	0.313	0.56	0.402	0	0.498	0.506
	5	0.56	0.56	0.313	0.56	0.402	0	0.488	0.501
	8	0.56	0.56	0.313	0.56	0.402	0	0.481	0.497
	10	0.56	0.56	0.313	0.56	0.402	0	0.482	0.498
DT	2	0.775	0.24	0.775	0.775	0.774	0.541	0.787	0.747
	5	0.785	0.241	0.788	0.785	0.781	0.562	0.821	0.797
	8	0.775	0.23	0.775	0.775	0.775	0.544	0.802	0.751
	10	0.77	0.241	0.77	0.77	0.77	0.532	0.819	0.774
RF	2	0.789	0.224	0.789	0.789	0.789	0.571	0.86	0.842
	5	0.804	0.201	0.804	0.804	0.804	0.602	0.864	0.847
	8	0.799	0.207	0.799	0.799	0.799	0.592	0.861	0.841
	10	0.809	0.192	0.81	0.809	0.809	0.614	0.864	0.848

The overall objective of the work is to predict more exactly the occurrence of heart disease using data mining techniques. In this research work, the UCI data repository is used for performing the comparative analysis of three algorithms such as Random Forest, Decision trees and Naive Bayes. From the research work, it has been experimentally proven that Random Forest provides perfect results as compare to Decision tree and Naive Bayes.

References:

David, H. F., & Belcy, S. (n.d.). *HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES.*[PDF]. Http://web.b.ebscohost.com.

Synced. (2017, October 24). How Random Forest Algorithm Works in Machine Learning. Retrieved from https://medium.com/@Synced/how-random-forest-algorithm-works-in-machine-learning-3c0fe15b6674