Predictive Diagnosis through Data Mining: A Secondary Review

Guided By: Dr. Sandesh Kumar Sharma
(Associate Professor)

Presented By: Ritwik Chawla
(Roll No. 0335)

Contents

A LOT OF DATA YET TO BE CONVERTED INTO INFORMATION

Improve Healthcare, Win $3,000,000.

Identify patients who will be admitted to a hospital within the next year using historical claims data.

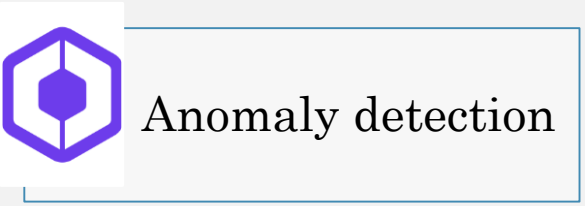
kaggle

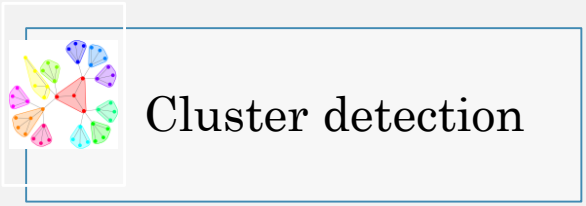
The Heritage Health Prize: Bringing Data Science to Preventative Medicine

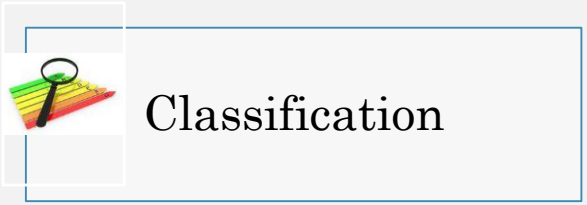
More than 71 million people are hospitalized every year in the United States, creating at least $30 billion in annual avoidable costs. Thinking big, the Heritage Provider Network (HPN) launched the **Heritage Health Prize** to spur new algorithms that could allow care providers to reach patients before emergencies occur. Such models would revolutionize

Learn More About GE NFL Head Health Initiative

Types of pattern detection







Knowledge Discovery of Data (KDD)

Data

Stored in Database (millions of Data Points)

Selection

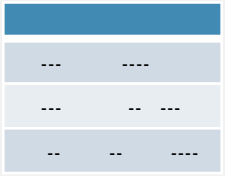
Right data from Database

Target Data

Preprocessing

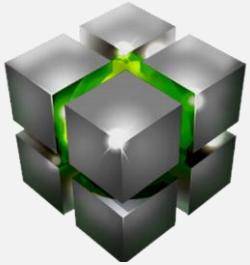
Data preprocessing and cleaning

Preprocessed Data



Transformation

Decreased dimensionality to increase efficiency



Transformed Data

Data Mining

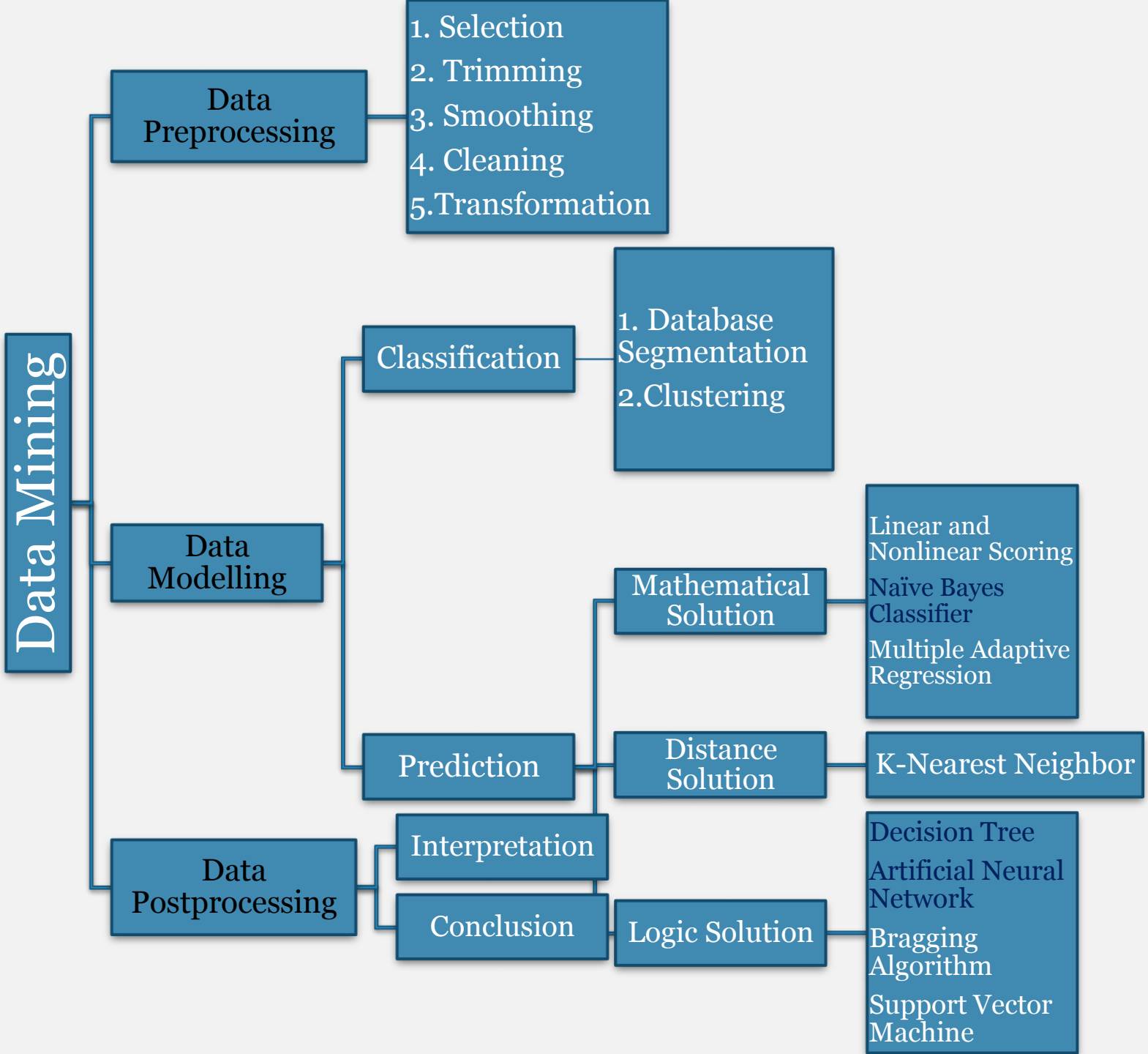
Intelligent algorithms for classification

Patterns

Interpretation

Data visualization and understanding

Knowledge


Data Mining
Data Preprocessing
1. Selection
2. Trimming
3. Smoothing
4. Cleaning
5.Transformation
Data Modelling
Classification
1. Database Segmentation
2.Clustering
Prediction
Mathematical Solution
Linear and Nonlinear Scoring
Naïve Bayes Classifier
Multiple Adaptive Regression
Distance Solution
K-Nearest Neighbor
Data Postprocessing
Interpretation
Conclusion
Logic Solution
Decision Tree
Artificial Neural Network
Bragging Algorithm
Support Vector Machine

Objectives

1. To study the various Data mining techniques of diagnosis and prediction system.
2. To study the applications and compare the performances of three predictive data mining techniques of heart disease prediction.
3. To study the global market and future prospects of predictive diagnosis through data mining.

Research Questions

1. How different data mining techniques of heart disease prediction system are compared based on existing performance measures?
2. How data mining in healthcare is a novel predictive diagnosis system for the future?

Methodology

A scientific literature review search was conducted to know the various predictive diagnosis technique using data mining and their performances.

- **Study Design:** Exploratory design with Secondary review
- **Study Settings:** NTT Data Services, Bengaluru
- **Study Duration:** Three months (February 2017 – April 2017)
- **Data Sources:** Various leading articles and journal publications.
- **Inclusion Criteria:** Techniques include Naïve Bayes Classifier, Decision Tree, and Artificial Neural Networks for prediction system of Heart Disease. Various articles from leading international journals, web pages, published newspaper articles, published books, conference papers, thesis, etc.
- **Exclusion Criteria:** diseases, like Diabetes, Cancer, Genetic disorders, and other major diseases. Unpublished study material, blogs and articles in language other than English.

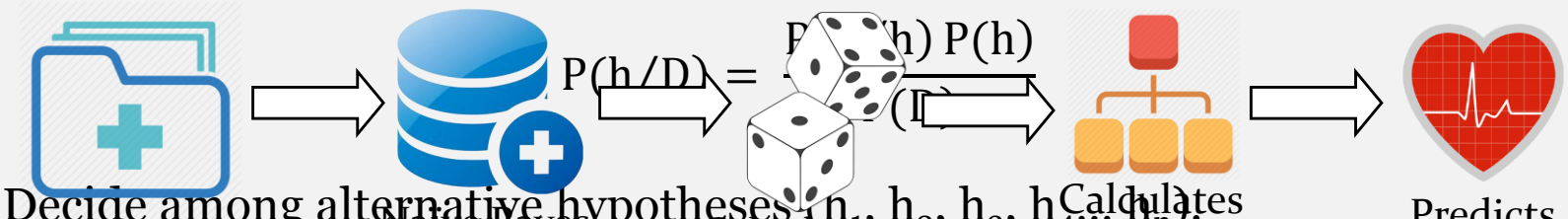
Heart Disease Prediction

- According to World Health Organization (WHO) Cardiovascular Diseases are the number one cause of death globally.
- An estimated 17.5 million people died from CVDs in 2012, representing 31% of all global deaths.

Risk Factors	Heart Diseases
	
Hypertension	Angina
Diabetes	Heart Attack
Hyperlipidemia	 Atherosclerosis
Smoking	Coronary Artery Disease
Excessive Alcohol use	Cardiac Arrhythmias

- Data mining can cater to all the countries where there is a **shortage of specialists** and **misdiagnosed cases** are high.

Naïve Bayes Technique



- Decide among alternative hypotheses (h_1, h_2, h_3, $h_4 \ldots h_n$).
- The hypothesis with the highest probability is called the maximum a posteriori hypothesis, or h_{MAP}

$$h_{MAP} = \text{argmax}_{h \in H} P(h/D)$$

Example

Finding the Maximum a posteriori probability to find if patient has the disease or not.

In the U.S. 0.8% of people have a unique form of heart disease.

- When the disease is present a blood, test returns a correct POS result 98% of the time.
- It returns a correct NEG result 97% of the time in cases when the disease is not present.
- A patient undergoes a blood test for the disease and the test result is POS.

P(disease) = .008, P(not disease) = .992

P(POS | disease) = .98, P(POS | not disease) = .03

Selecting h_{MAP} = disease

To know the exact probability,

P(disease | POS) = 0.0078 / (0.0078 + 0.0298) = 0.21

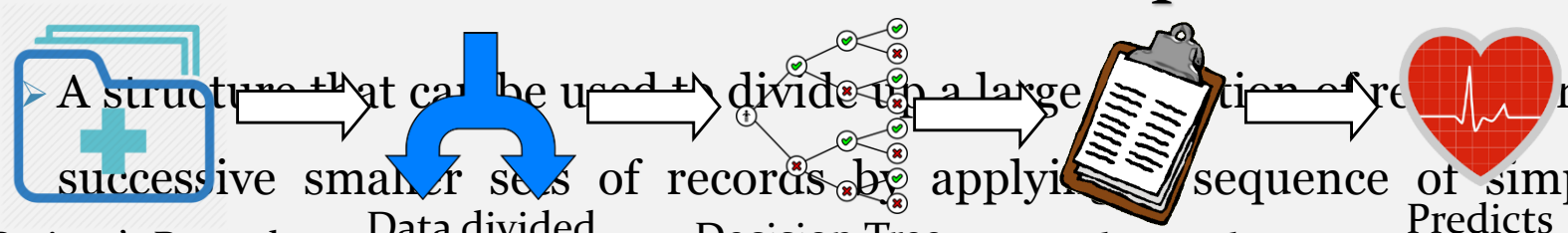
The patient has a 21% chance of having disease.

P(disease) = .008(.98) = .0078

P(not disease) = .992(.03) = .0298

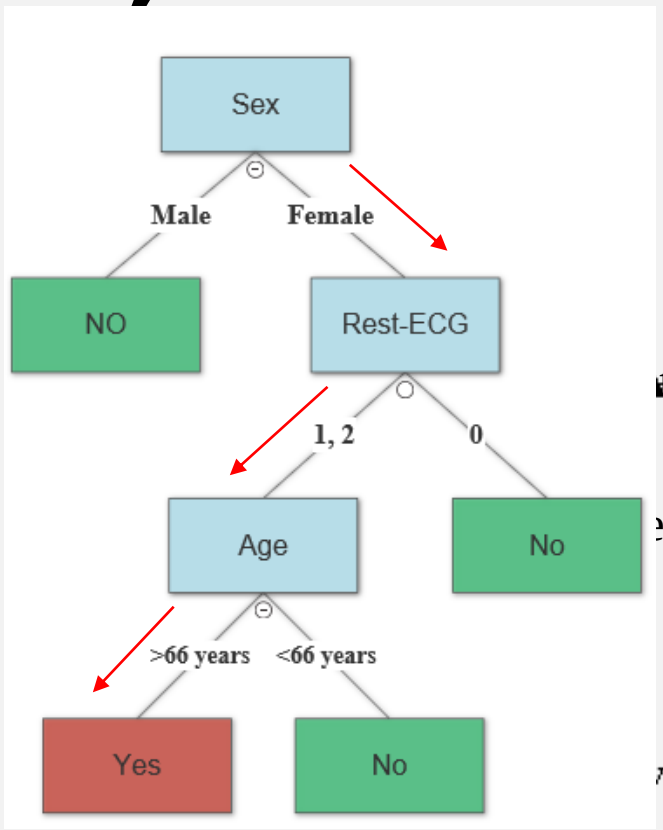
Therefore, patient can be classified as **not having heart disease.**

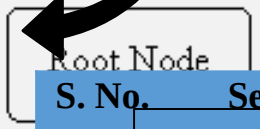
Decision Tree Technique



A structure that can be used to divide up a large collection of records into successive smaller sets of records by applying a sequence of simple decision rules. With each successive division, the members of the resulting sets become more and more similar to one another.

-Berry and Linoff





S. No.	Sex	Rest-ECG	Age	Diagnosis
1	Male	0	63	No
2	Female	1	67	Yes
3	Female	0	62	No
4	Male	1	56	No
5	Female	2	66	Yes
6	Female	1	69	Yes
7	Male	2	68	Yes
8	Female	0	64	No
9	Female	1	70	Yes
10	Female	0	51	No

IF Sex = Female **AND** RestECG = having ST_T wave abnormal **OR** left ventricular hypertrophy **AND** Age ≥ 66 years **THEN** Diagnosis = YES

Rules helps to decide how much the patient is likely to have heart disease

Set of possible answers

1 = having ST_T wave abnormal

2 = left ventricular hypertrophy

Neural Network Technique

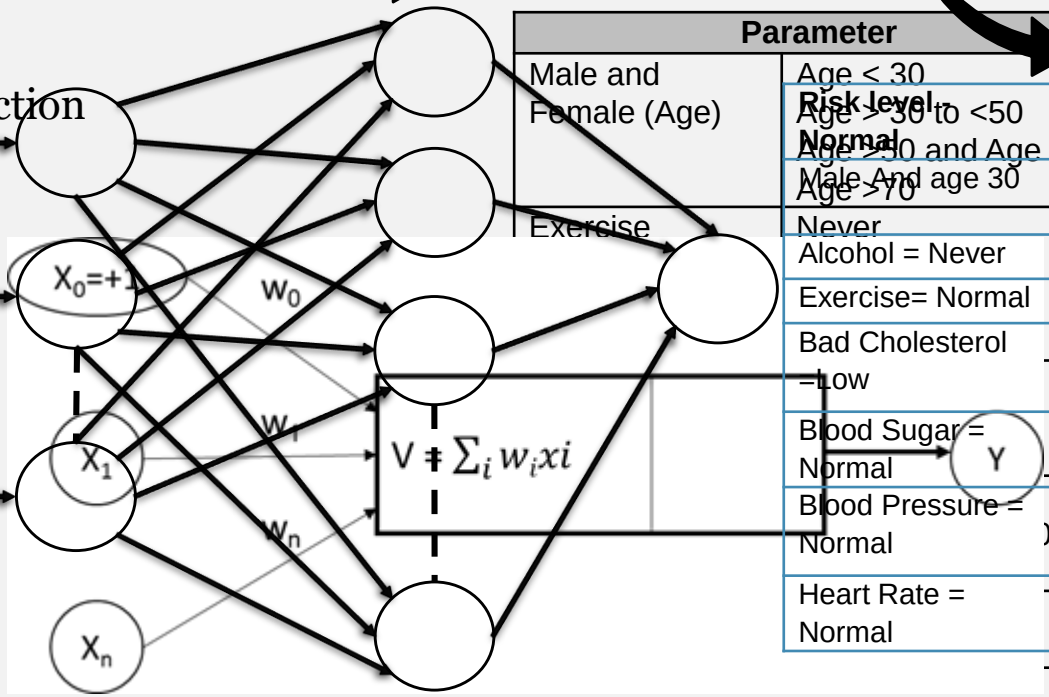
Artificial neural networks [illegible] inspir[illegible] wledge of biological nervou[illegible] system, [illegible]e brain.

- The first layer ,called input layer is constituted by the inputs x_n,
- The intermediary layers are called hidden layers,
- The last layer is the output layer, constituting output, *Y*.

Three basic elements:

1. A set of synapses
2. Summation
3. Activation function

Age
Exercise
Overweight
Blood Pressure
Heart Rate
Sedentary Lifestyle
Alcohol Intake



Parameter		Weightage
Male and Female (Age)	Age < 30	0.1
	Age > 30 to <50	0.5
	Age >50 and Age <70	0.7
	Age >70	0.8
Exercise	Never	0.6
Alcohol = Never	Alcohol = Past	0.1
Exercise= Normal	Exercise = Never	0.1
Bad Cholesterol =Low	Bad Cholesterol = High	0.1
Blood Sugar = Normal	Blood Sugar = High	0.9
Blood Pressure = Normal	Blood Pressure = High	0.1
Heart Rate = Normal	Heart Rate = High	0.7

Risk level = Normal	Risk level = High
Male And age 30	Female And Age >70
Alcohol = Never	Alcohol = Past
Exercise= Normal	Exercise = Never
Bad Cholesterol =Low	Bad Cholesterol = High
Blood Sugar = Normal	Blood Sugar = High
Blood Pressure = Normal	Blood Pressure = High
Heart Rate = Normal	Heart Rate = High

0.9
0.9
0.1
0.9
0.1

Results

- Comparison of performances has been made on the lines of **Accuracy, Sensitivity and Specificity.**
- Most used **Attributes**, **big players** and **future trends** have also been identified.

	Predicted Heart Disease: YES	**Predicted Heart Disease: NO**
Actual Heart Disease: YES	True Positive (TP)	False Negative (FN)
Actual Heart Disease: NO	False Positive (FP)	True Negative (TN)

Confusion Matrix Table

Performance Indicators

$$Sensititivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

Results: Attributes

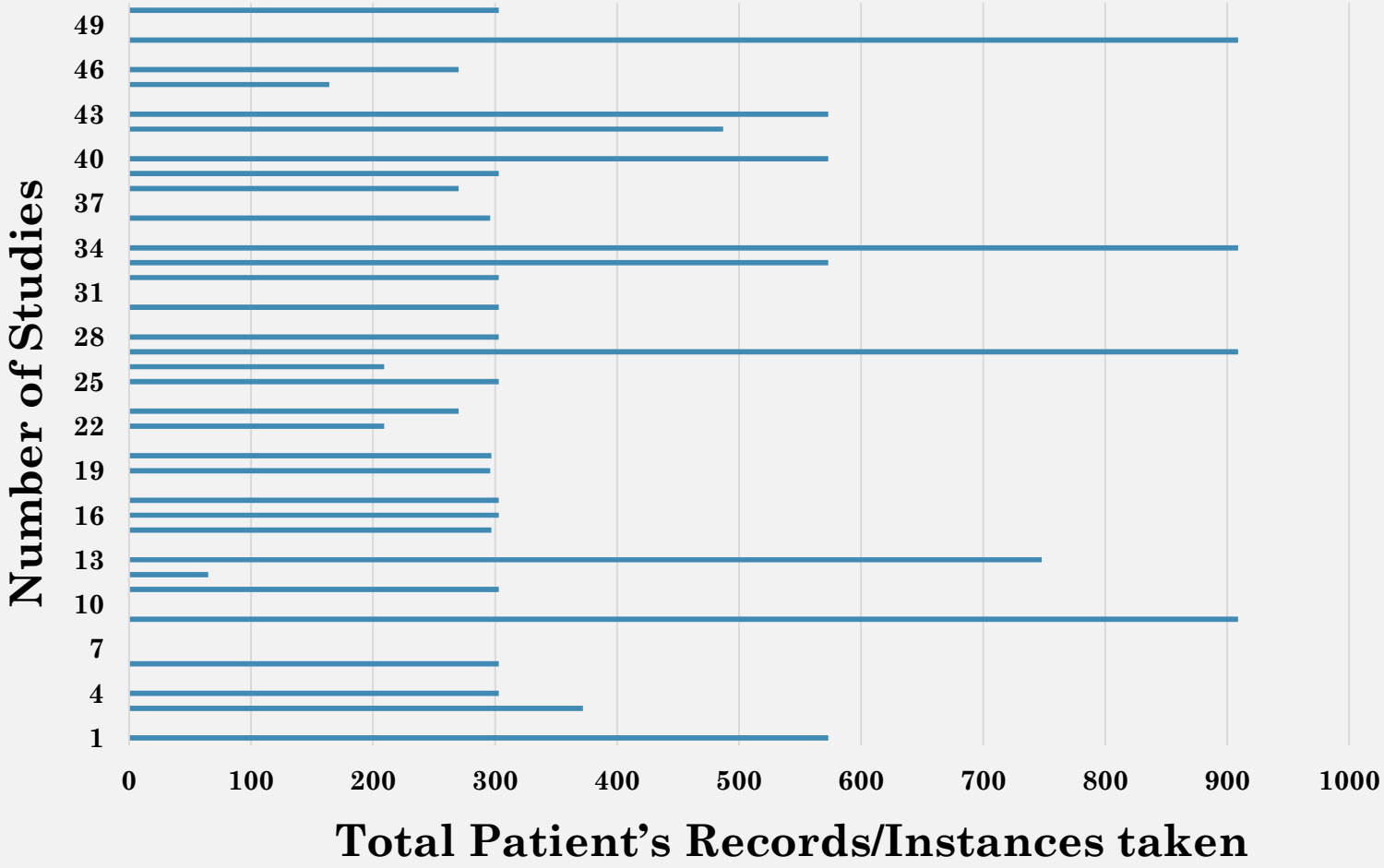
Most Commonly used Input Attributes/Features:

1. **Age** - age in years
2. **Sex** - sex
3. **Cpt** - chest pain type
4. **Trestbps** - resting blood pressure
5. **Chol** - serum cholesterol in mg/dl
6. **Fbs** - fasting blood sugar
7. **Restecg** - resting electrocardiographic results
8. **Thalach** - maximum heart rate achieved
9. **Exang** - Exercise induced angina
10. **Oldpeak** - ST depression induced by exercise relative to rest
11. **Slope** - Slope of the peak exercise ST segment
12. **CA** - Number of major vessels coloured by fluoroscopy
13. **Thal** - Heart defect

Most Commonly used Output Attribute:

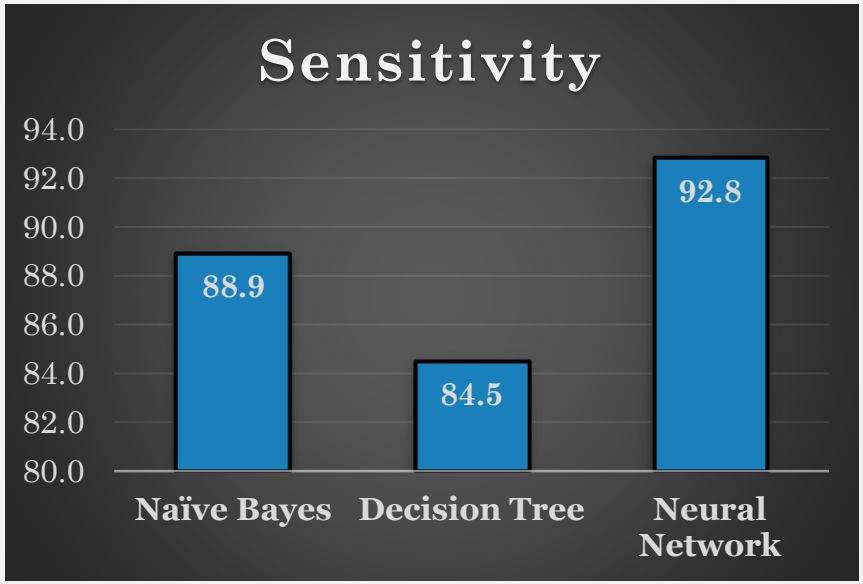
num - the predicted attribute - diagnosis of heart disease (angiographic disease status)

Results: Records

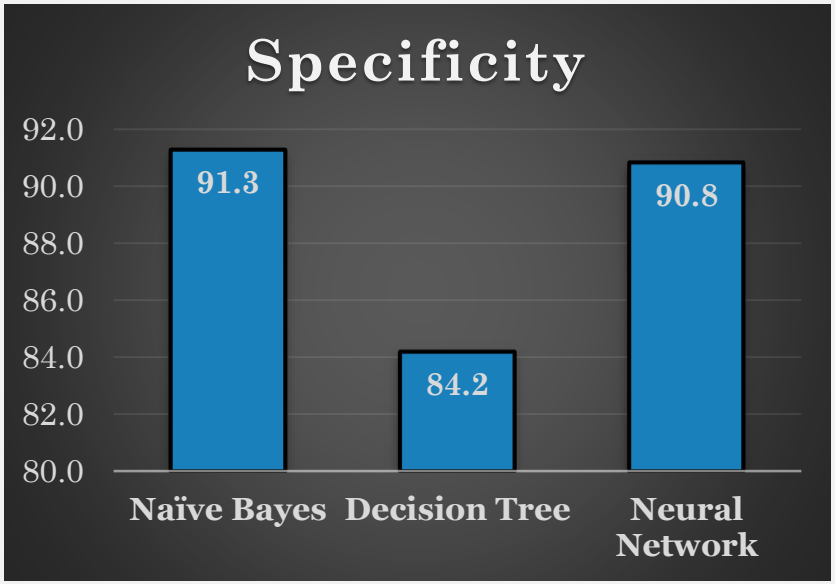


Range of Records	
0-200	4
200-400	32
Above 400	13

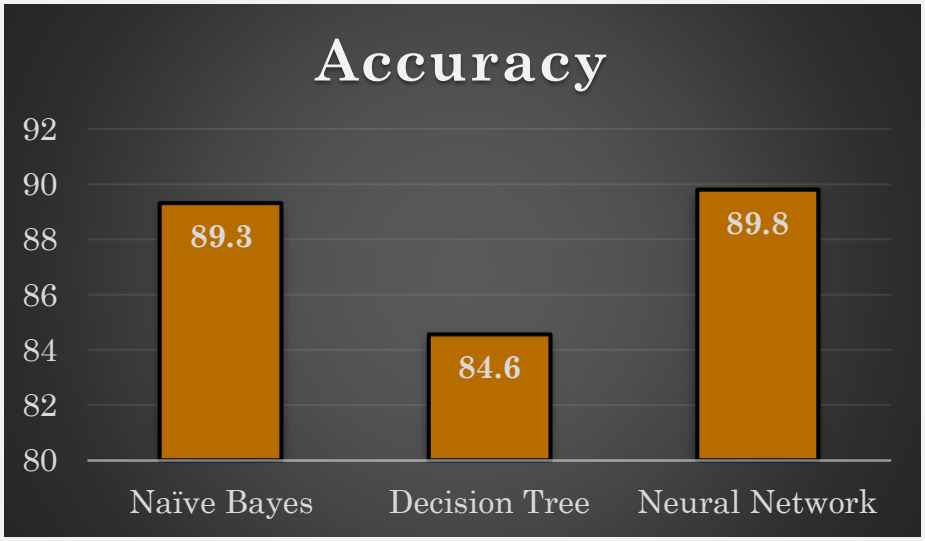
Results: Performance



Highest Average Sensitivity – Neural Network



Highest Average Specificity – Naïve Bayes



Highest Average Accuracy – Neural Network

Big Players in Data Mining: Global Market

Health Informatics

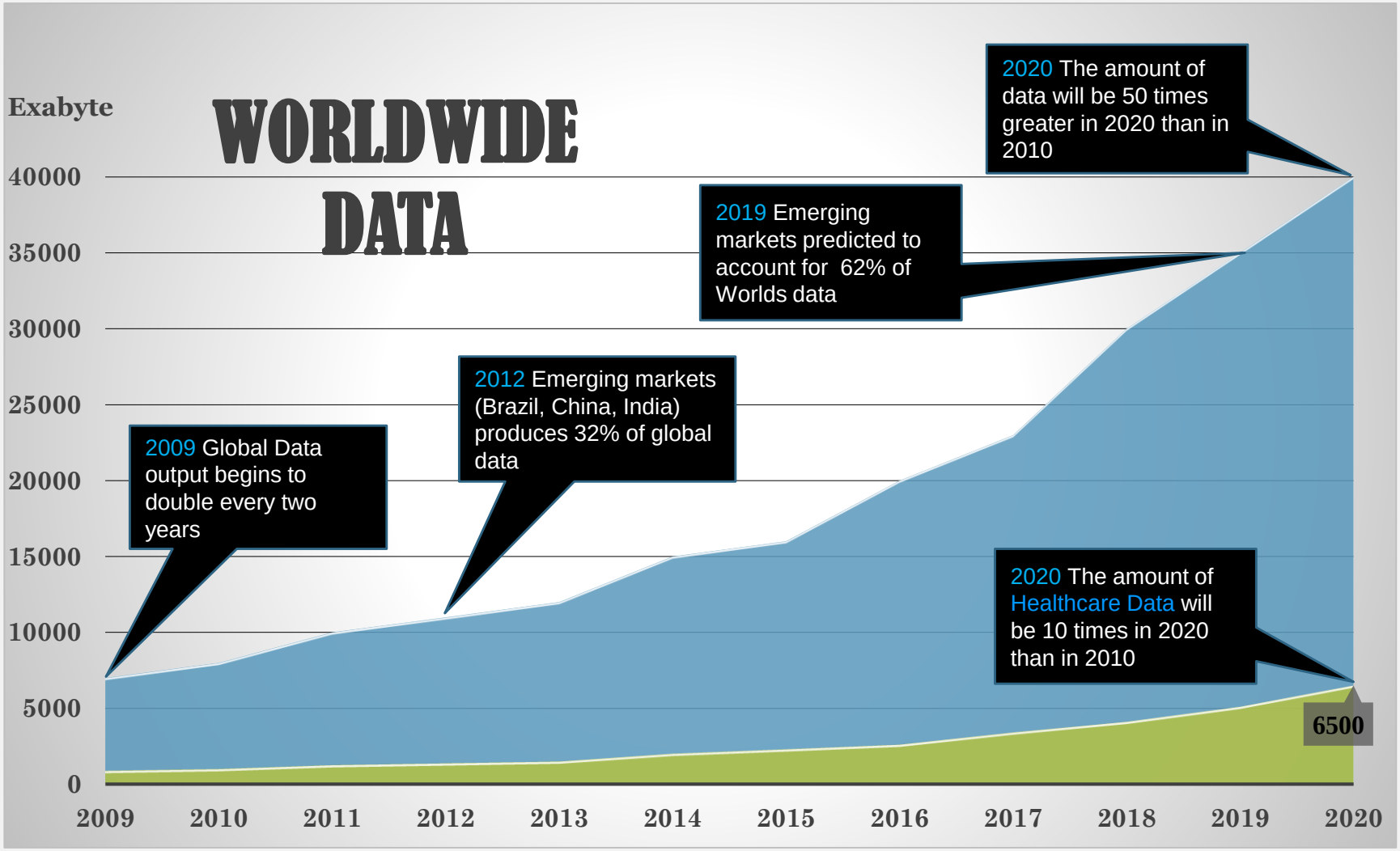
Advance Analytics

Oracle Data Mining (ODM)

Enterprise Analytics

Analytics Explorer

Future of Data Mining



Source: http://www.popularmechanics.com/military/a9465/nsa-data-mining-how-it-works-15910146/

Conclusions

- This study reviewed how three different data mining techniques for heart disease prediction can be used efficiently to model and predict heart disease cases by identifying the common types of attribute and number of patient records required. The outcome of this study can be used as an assistant tool by cardiologists to help them to make more consistent diagnosis of heart disease to bring out more clarity on the subject and how the experiments work.
- The Neural Network model shows the maximum classification accuracy of 89.8 per cent. Decision tree on the other hand could not exceed 84.6 per cent.
- The market for Predictive analytics in healthcare has been increasing many folds with big players contributing and emerging economies like Brazil, China and India predicted to contribute a greater share in this segment.

Recommendations

1. To improve the classification accuracy of the models further researches can be conducted using different classification algorithms like support vector machine, Rule Induction, etc.
2. Further investigations can be performed with different parameter settings to enhance and expand the capabilities of the prediction models.
3. The patterns found can be used to design a Knowledge Based System (KBS) with the help of domain experts who have years of experience in the problem domain.
4. A researcher can plan to perform additional experiments with more dataset and algorithms to improve the classification accuracy and to build a model that can predict specific heart disease types and more accurate results.

Thank You

"If you mine the data hard enough, you can also find messages from God."

— Scott Adams, Dilbert