Al-Imam Al-Adham University College
Department of Advocacy and rhetoric
Baghdad- Iraq
E-mail: zahraa_raji_cs@yahoo.com

Orange Data Mining as a tool to compare Classification Algorithms

Zahraa Raji Mohi

Abstract

At the present time, there are many data available and saved sets on the web, medical data is one of them; because of its huge volume it is importance I'll classify it with effective techniques to use it more clearly with the best way. Data mining techniques is one of the most popular and effective classification tools that is used , a lot of these tools are available for free such as Weka, Rapid miner, Knime and Orange which are easy to use.

In this research we choose Orange as data mining tool to classify two types of selected medical data for testing (Breast cancer and heart-disease) depending on previous medical tests by taking a set of information such as cholesterol, chest pain, blood pressure, blood sugar, age and sex to find if the patient with or without disease then find the best algorithm for classification using accurate measures as best performance criteria by applying decision tree, Naïve Bayes and K-nearest neighbor classification algorithms on the chosen data and comparing them.

Key words: Data mining, Orange mining tool, classification algorithms.

1- Introduction

Now-a-days data increased so fast, this causes difficulty for the user to analyze and classify data manually. This data mining helps to mine useful knowledge and information from specific data. Data mining is the process of interesting knowledge discovering from huge data tanked in databases, information warehouses or data repositories. Data mining has a problem in the categorization to find rules to classify the data that is used in prerecorded categories. When data mining uses huge data, time of the execution algorithms may be amortize time. This fixture automates tools that assist people to convert the huge data into beneficial information. Many data mining open-source tools are available for free use like Orange which is an available corpus of procedures and methods for best data parsing and which assists class parsing, Decision trees, Predictively parsing, data mining, etc. [1,2].

2- Related Work

Data mining classification algorithm which have been chosen are: Decision tree, K-Nearest Neighbor and Naïve Bayes to compared using Orange mining tool by applying accuracy to measure the performance of the selected algorithm which symbolizes the ratio of the truly classified data.

Below are some related works:

- **Abdullah H. Wahbeh, *et.* Al (2011)[3]:** They refer to a similar report between a number of information mining instruments and programming bundles. The paper results demonstrate the exhibition of the apparatuses because the errand character is usefully effects by utilize dataset and by the grouping calculations that actualize inside the toolbox. As far as for the material issue, the WEKA toolbox reaches the peak than follows up by Orange, Tanagra, and KNIME separately. At last; WEKA toolbox has accomplished the most amazingly improvement in grouping execution; when transfer from the ratio split to the Cross Validation test mode, trailed by KNIME, Orange and Tanagra separately.
- **Beant Kaur, et. Al (2014) [7]:** they assert that Data mining instruments can address business addresses that customarily take much tedious to determine. Utilizing information mining methods require some investment for the forecast of the sickness with more precision. It studies various papers including information for the forecast of coronary illness because of utilizing neural systems is almost 100%. Utilizing information mining calculation gives effective outcomes. Applying information mining systems to coronary illness treatment information can give as dependable execution as that accomplished in diagnosing coronary illness.
- **Kalpana Rangra, et. Al (2014) [8]:** Mention how the rapid data growth make revolution in business and many application, as data development enthusiasm for business usage devising has gained the improvement of data mining fields that is important portion for available data. Data mining application is an important innovation that must be developed to get various sorts of information and data reaching all over the world. The research gives inclusive and virtual checkup of six free data mining tools. The checkup illustrates the specific techniques, founders, and organization for all used device with its applications. By using the checkup, ruling and design of facility can be simplex.
- **Amrita Naik, et al (2016) [1]:** Comment on the a relationship between information mining characterization calculation utilizing some open sources information mining instruments, for example, Rapid excavator, WEKA, Orange, Knime and Tangaro. The effect of the five instruments is seen thruogh utilizing the precision of chose order calculation like Decision tree, K-Nearest Neighbor and Naïve Bayes. Testing dataset used in an Indian Liver Patient in the Classification calculation so as to control the general population.
- **Slater S, et. Al (2016) [13]:** this article, highlights some of the most widely used, most accessible, and most powerful tools available for the researchers who are interested in conducting EDM/LA research. It will highlight the utility that these tools have with respect to common data preprocessing and analysis steps in a typical research project as well as more descriptive information such as price point and user friendliness. Also it highlights niche tools in the field, such as those used for Bayesian knowledge tracing (BKT), data visualization, text

analysis, and social network analysis. It also discusses the importance of familiarizing oneself with multiple tools a data analysis toolbox for the practice of EDM/LA research.

- **Sarangam Kodati , et. Al (2018) [12]:** They refer to Health care industry that contains large amount of data and hidden information. Effective decisions are made with this hidden information by applying patient; however, data mining these tests could be reduced. But there is a lack of relevant analyzing tool according to provide effective test outcomes together with the hidden information. As a result, the system is developed using data mining algorithms for classifying the data and to detect the heart diseases. Data mining acts a solution by many healthcare problems. Naïve Bayes, SVM, Random Forest, KNN algorithm is one such data mining method which serves the diagnosis regarding heart diseases. Analyzing few parameters also predicts heart diseases, and devices heart diseases prediction system (HDPS) based total on the data mining approaches.

3- Data Mining Classification Algorithms

This study focuses on the following classification algorithm for comparison in Orange data mining tool:

A. **Decision Tree:** It is a classifier with a multistage basic leadership, its fundamental thought engages any multistage approach to separate a mind boggling choice into an association of a few less complex choices, trusting the last arrangement acquired along these lines would take after the planned wanted arrangement. In contrast to customary measurable classifiers, which utilize every accessible component at the same time and settle on a solitary participation choice for every pixel, it utilizes a multistage or successive way to deal with the issue of mark task. The naming procedure is viewed as a chain of basic choices dependent on the after effects of successive tests as opposed to a solitary, complex choice. Sets of choice successions structure the parts of the decision tree, with tests being connected at the hubs. Decision tree development includes the recursive parceling of a lot of preparing information, which is part into progressively homogeneous subsets based on tests connected to at least one of the element esteems. These tests are spoken to by hubs. Marks are doled out to terminal (leaf) hubs by methods for a portion procedure, for example, greater part casting a ballot [3,4].

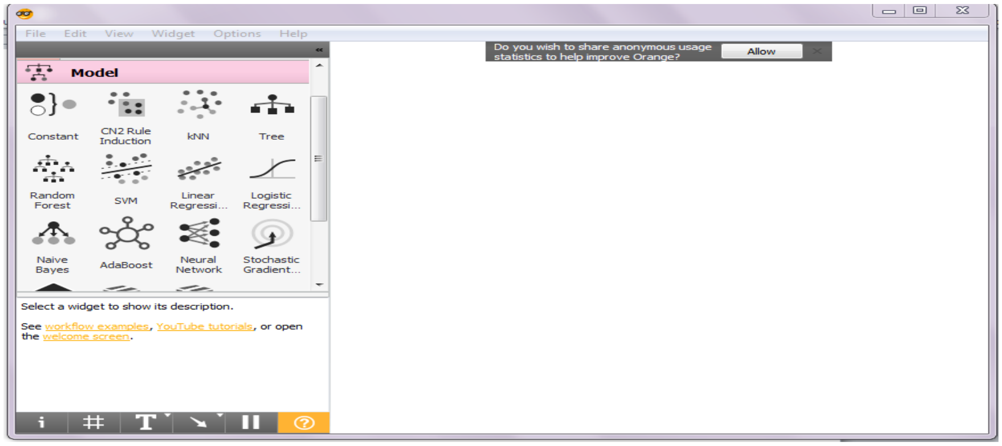
B. **Naïve Bayes:** It is a classifier with term in Bayesian measurements managing a basic probabilistic classifier dependent on applying Bayes' hypothesis with solid (naive) freedom suppositions. A progressively spellbinding term for the fundamental likelihood model would be "autonomous element model". In straightforward terms, a Guileless Bayes classifier expects that the nearness (or nonattendance) of a specific component of a class is random to the nearness (or nonappearance) of some other element. For instance, an organic product might be viewed as an apple in the event that it is red, round, and around 4" in measurement. Despite the fact that these highlights rely upon the presence of different highlights, a guileless Bayes classifier considers these properties to independently add to the probability that this natural item is an apple. Dependent upon the accurate thought of the probability model, Gullible Bayes classifiers can be arranged all around successfully in a controlled getting the hang of setting. In various valuable applications, parameter estimation for unsuspecting Bayes models uses the strategy for most extraordinary likelihood; figuratively speaking, one can work with the guileless Bayes model without confiding in Bayesian probability or using any Bayesian procedures [5].

C. **K-Nearest Neighbor:** The *k*-nearest neighbor classification algorithm (KNN) classifies new articles as indicated by the result of the nearest object or the results of a few nearest questions in the component space of the preparation set and it is the least difficult. AI calculation: an item classified by a larger part vote of its neighbors, with the new item being apportioned to the class that is most essential among its k nearest neighbors (k is a positive entire number, and normally little). The best choice of k depends on the data. Greater estimations of k will all in all lessen the effect of clatter on the classification, yet will make constrain between classes less undeniable. A better than average estimation of k can be picked by cross-endorsement. In twofold (two class) classification issues, it is helpful to pick k to be an odd number as this avoids tied votes. In the classification organization, k is a customer defined enduring; another thing with given features (often suggested as a request or test point) which is classified by doling out to it the imprint that is most progressive among the k planning tests nearest to that new protest [6].

4- Orange Data Mining tool:

Orange is an open source data burrowing available in vain at (https://orange.biolab.si) which is computer-based intelligence and data mining programming (written in Python). It has a visual programming front-end for explorative data examination and portrayal, like manner can be used as a Python library. The program is kept up and made by the Bioinformatics Research facility of the Workforce of PC and Data Science at College of Ljubljana. Orange is a section-based visual programming for data mining, computer based intelligence and data examination. Parts are called contraptions and they go from fundamental, subset decision and preprocessing, to observational appraisal of learning computations and perceptive illustrating [9, 7].

Orange moreover joins a ton of graphical contraptions that use methodologies from focus library and Orange modules. Through visual programming, devices can be assembled into an application by a visual programming instrument called Orange Canvas [3].



Do you wish to share anonymous usage statistics to help improve Orange?

Select a widget to show its description.

See workflow examples, YouTube tutorials, or open the welcome screen.

Figure1: Main Interface of Orange that shows the models can perform KNN, Trees Naïve Bayes and etc.

5- Experimental and Evaluation Data:

The experiment of this study begins by opening orange data mining tool then choosing the test data set used. Secondly, applying selected classification algorithm , and finally viewing the evaluation results as shown in Figure 2.

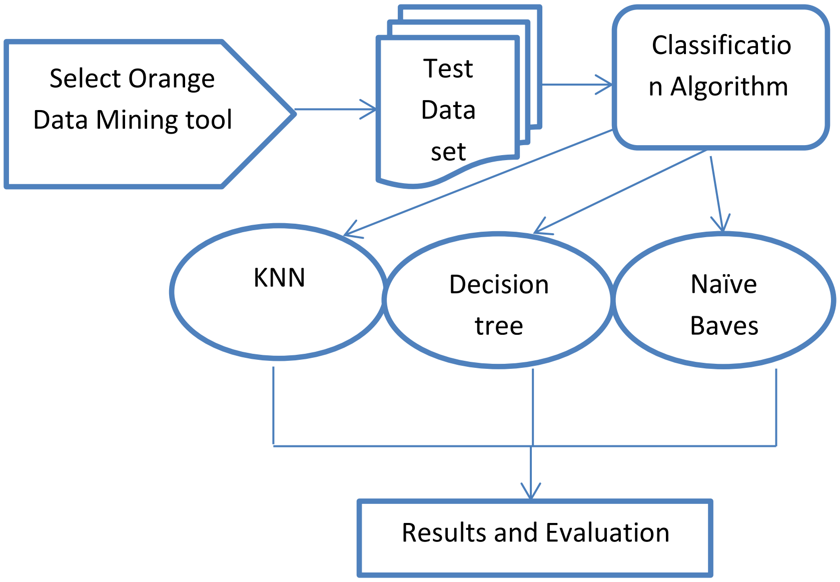


Figure2: Work Methodology Study

5.1 *Data Description*

We have chosen the informational collection from the UCI (Unique Client Identifier) store [11]. Two types of heart disease and breast cancer contain 160 distinct patients record where, the quantity of cases is 303. The informational index is gathered from

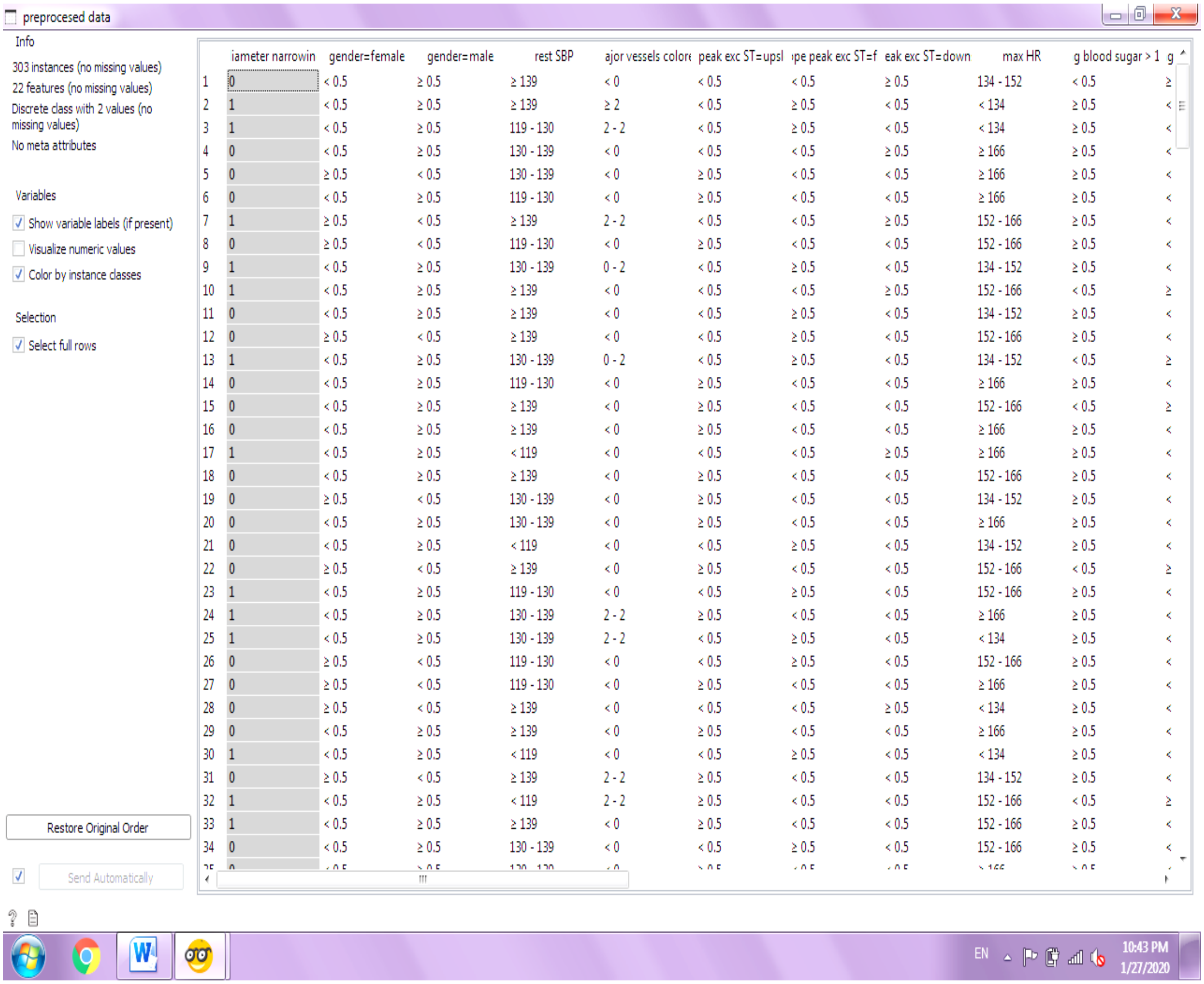
investigating patients utilizing https://www.simplehearttest.com/site. This informational index contains 102 male patient records and 58 female patient records. For heart disease 10 information characteristics as following:
1- Age: Age of the patient 2- Sex: Sexual orientation of the patient 3- Chest torment: yes/no 4- Resting blood pressure SBP 5- Blood cholesterol 6- Fasting blood sugar>120 7- Resting electrocardiographic result 8- Maximum rate heart achieved 9- Exercise induced angina 10- Depression induced by exercise relative to rest 10- The slope of the peak exercise ST segment [10].

Second breast cancer disease informational indexes contain 200 unique patients' records where, the quantity of occurrences is 683. 10 information properties as following: 1- Cluster thickness 2-Unif cell size 3-Unif cell shape 4-Minimal grip 5-Single cell size 6-Exposed cores 7-Insipid chromatin 8-Ordinary nucleoli 9-Mitoses 10- Type.

If we take the male and female data separately we must add information to analyze the data that may increase woman infection such as high blood pressure during the pregnancy, having certain disease, Endometriosis, chest pain, has father or brother with heart attack. This could be a further study.

5.2 Data Preprocessing

This step is a preprocessing of data by transforming data into most suitable form by attribute selection, handling a missing data and eliminating the outlier for mining algorithm the 10 characteristic selected for preprocessing data with 272 case for heart disease and 386 for breast cancer as shown in Fig(3) for heart disease and Fig(4) for breast cancer.



Info

303 instances (no missing values)
22 features (no missing values)
Discrete class with 2 values (no missing values)
No meta attributes

Variables

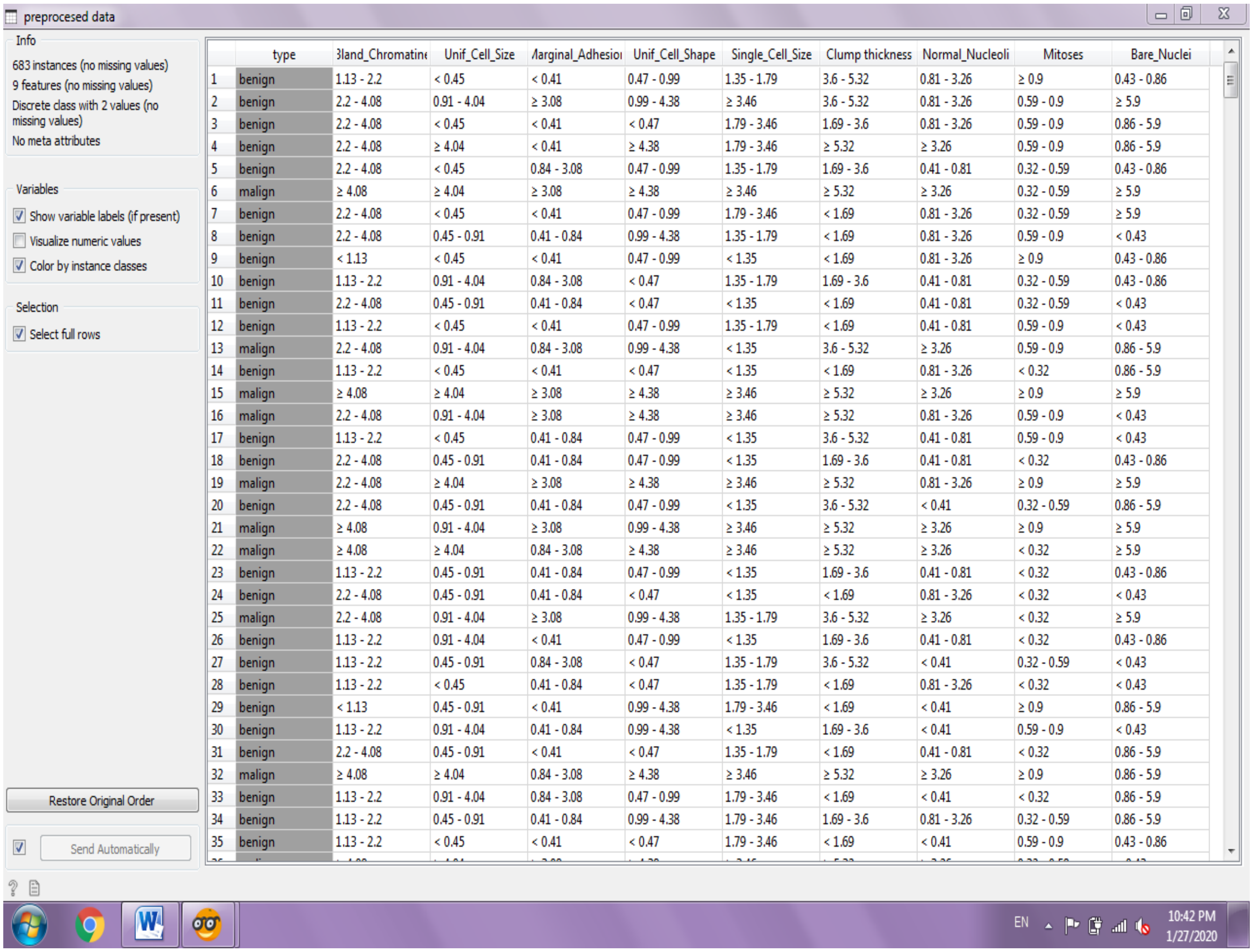
☑ Show variable labels (if present)
☐ Visualize numeric values
☑ Color by instance classes

Selection

☑ Select full rows

	iameter narrowin	gender=female	gender=male	rest SBP	ajor vessels colore	peak exc ST=upsl	pe peak exc ST=f	eak exc ST=down	max HR	g blood sugar > 1	g
1	0	< 0.5	≥ 0.5	≥ 139	< 0	< 0.5	< 0.5	≥ 0.5	134 - 152	< 0.5	≥
2	1	< 0.5	≥ 0.5	≥ 139	≥ 2	< 0.5	≥ 0.5	< 0.5	< 134	≥ 0.5	<
3	1	< 0.5	≥ 0.5	119 - 130	2 - 2	< 0.5	≥ 0.5	< 0.5	< 134	≥ 0.5	<
4	0	< 0.5	≥ 0.5	130 - 139	< 0	< 0.5	< 0.5	≥ 0.5	≥ 166	≥ 0.5	<
5	0	≥ 0.5	< 0.5	130 - 139	< 0	≥ 0.5	< 0.5	< 0.5	≥ 166	≥ 0.5	<
6	0	< 0.5	≥ 0.5	119 - 130	< 0	≥ 0.5	< 0.5	< 0.5	≥ 166	≥ 0.5	<
7	1	≥ 0.5	< 0.5	≥ 139	2 - 2	< 0.5	< 0.5	≥ 0.5	152 - 166	≥ 0.5	<
8	0	≥ 0.5	< 0.5	119 - 130	< 0	≥ 0.5	< 0.5	< 0.5	152 - 166	≥ 0.5	<
9	1	< 0.5	≥ 0.5	130 - 139	0 - 2	< 0.5	≥ 0.5	< 0.5	134 - 152	≥ 0.5	<
10	1	< 0.5	≥ 0.5	≥ 139	< 0	< 0.5	< 0.5	≥ 0.5	152 - 166	< 0.5	≥
11	0	< 0.5	≥ 0.5	≥ 139	< 0	< 0.5	≥ 0.5	< 0.5	134 - 152	≥ 0.5	<
12	0	≥ 0.5	< 0.5	≥ 139	< 0	< 0.5	≥ 0.5	< 0.5	152 - 166	≥ 0.5	<
13	1	< 0.5	≥ 0.5	130 - 139	0 - 2	< 0.5	≥ 0.5	< 0.5	134 - 152	< 0.5	≥
14	0	< 0.5	≥ 0.5	119 - 130	< 0	≥ 0.5	< 0.5	< 0.5	≥ 166	≥ 0.5	<
15	0	< 0.5	≥ 0.5	≥ 139	< 0	≥ 0.5	< 0.5	< 0.5	152 - 166	< 0.5	≥
16	0	< 0.5	≥ 0.5	≥ 139	< 0	≥ 0.5	< 0.5	< 0.5	≥ 166	≥ 0.5	<
17	1	< 0.5	≥ 0.5	< 119	< 0	< 0.5	< 0.5	≥ 0.5	≥ 166	≥ 0.5	<
18	0	< 0.5	≥ 0.5	≥ 139	< 0	≥ 0.5	< 0.5	< 0.5	152 - 166	≥ 0.5	<
19	0	≥ 0.5	< 0.5	130 - 139	< 0	≥ 0.5	< 0.5	< 0.5	134 - 152	≥ 0.5	<
20	0	< 0.5	≥ 0.5	130 - 139	< 0	≥ 0.5	< 0.5	< 0.5	≥ 166	≥ 0.5	<
21	0	< 0.5	≥ 0.5	< 119	< 0	< 0.5	≥ 0.5	< 0.5	134 - 152	≥ 0.5	<
22	0	≥ 0.5	< 0.5	≥ 139	< 0	≥ 0.5	< 0.5	< 0.5	152 - 166	< 0.5	≥
23	1	< 0.5	≥ 0.5	119 - 130	< 0	< 0.5	≥ 0.5	< 0.5	152 - 166	≥ 0.5	<
24	1	< 0.5	≥ 0.5	130 - 139	2 - 2	≥ 0.5	< 0.5	< 0.5	≥ 166	≥ 0.5	<
25	1	< 0.5	≥ 0.5	130 - 139	2 - 2	< 0.5	≥ 0.5	< 0.5	< 134	≥ 0.5	<
26	0	≥ 0.5	< 0.5	119 - 130	< 0	< 0.5	≥ 0.5	< 0.5	152 - 166	≥ 0.5	<
27	0	≥ 0.5	< 0.5	119 - 130	< 0	≥ 0.5	< 0.5	< 0.5	≥ 166	≥ 0.5	<
28	0	≥ 0.5	< 0.5	≥ 139	< 0	< 0.5	< 0.5	≥ 0.5	< 134	≥ 0.5	<
29	0	< 0.5	≥ 0.5	≥ 139	< 0	≥ 0.5	< 0.5	< 0.5	≥ 166	≥ 0.5	<
30	1	< 0.5	≥ 0.5	< 119	< 0	< 0.5	≥ 0.5	< 0.5	< 134	≥ 0.5	<
31	0	≥ 0.5	< 0.5	≥ 139	2 - 2	≥ 0.5	< 0.5	< 0.5	134 - 152	≥ 0.5	<
32	1	< 0.5	≥ 0.5	< 119	2 - 2	≥ 0.5	< 0.5	< 0.5	152 - 166	< 0.5	≥
33	1	< 0.5	≥ 0.5	≥ 139	< 0	≥ 0.5	< 0.5	< 0.5	152 - 166	≥ 0.5	<
34	0	< 0.5	≥ 0.5	130 - 139	< 0	< 0.5	≥ 0.5	< 0.5	152 - 166	≥ 0.5	<

Figure 3: Heart Disease Preprocessing Data



Info

683 instances (no missing values)
9 features (no missing values)
Discrete class with 2 values (no missing values)
No meta attributes

Variables

☑ Show variable labels (if present)
☐ Visualize numeric values
☑ Color by instance classes

Selection

☑ Select full rows

☑ Send Automatically

	type	Bland_Chromatine	Unif_Cell_Size	Marginal_Adhesion	Unif_Cell_Shape	Single_Cell_Size	Clump thickness	Normal_Nucleoli	Mitoses	Bare_Nuclei
1	benign	1.13 - 2.2	< 0.45	< 0.41	0.47 - 0.99	1.35 - 1.79	3.6 - 5.32	0.81 - 3.26	≥ 0.9	0.43 - 0.86
2	benign	2.2 - 4.08	0.91 - 4.04	≥ 3.08	0.99 - 4.38	≥ 3.46	3.6 - 5.32	0.81 - 3.26	0.59 - 0.9	≥ 5.9
3	benign	2.2 - 4.08	< 0.45	< 0.41	< 0.47	1.79 - 3.46	1.69 - 3.6	0.81 - 3.26	0.59 - 0.9	0.86 - 5.9
4	benign	2.2 - 4.08	≥ 4.04	< 0.41	≥ 4.38	1.79 - 3.46	≥ 5.32	≥ 3.26	0.59 - 0.9	0.86 - 5.9
5	benign	2.2 - 4.08	< 0.45	0.84 - 3.08	0.47 - 0.99	1.35 - 1.79	1.69 - 3.6	0.41 - 0.81	0.32 - 0.59	0.43 - 0.86
6	malign	≥ 4.08	≥ 4.04	≥ 3.08	≥ 4.38	≥ 3.46	≥ 5.32	≥ 3.26	0.32 - 0.59	≥ 5.9
7	benign	2.2 - 4.08	< 0.45	< 0.41	0.47 - 0.99	1.79 - 3.46	< 1.69	0.81 - 3.26	0.32 - 0.59	≥ 5.9
8	benign	2.2 - 4.08	0.45 - 0.91	0.41 - 0.84	0.99 - 4.38	1.35 - 1.79	< 1.69	0.81 - 3.26	0.59 - 0.9	< 0.43
9	benign	< 1.13	< 0.45	< 0.41	0.47 - 0.99	< 1.35	< 1.69	0.81 - 3.26	≥ 0.9	0.43 - 0.86
10	benign	1.13 - 2.2	0.91 - 4.04	0.84 - 3.08	< 0.47	1.35 - 1.79	1.69 - 3.6	0.41 - 0.81	0.32 - 0.59	0.43 - 0.86
11	benign	2.2 - 4.08	0.45 - 0.91	0.41 - 0.84	< 0.47	< 1.35	< 1.69	0.41 - 0.81	0.32 - 0.59	< 0.43
12	benign	1.13 - 2.2	< 0.45	< 0.41	0.47 - 0.99	1.35 - 1.79	< 1.69	0.41 - 0.81	0.59 - 0.9	< 0.43
13	malign	2.2 - 4.08	0.91 - 4.04	0.84 - 3.08	0.99 - 4.38	< 1.35	3.6 - 5.32	≥ 3.26	0.59 - 0.9	0.86 - 5.9
14	benign	1.13 - 2.2	< 0.45	< 0.41	< 0.47	< 1.35	< 1.69	0.81 - 3.26	< 0.32	0.86 - 5.9
15	malign	≥ 4.08	≥ 4.04	≥ 3.08	≥ 4.38	≥ 3.46	≥ 5.32	≥ 3.26	≥ 0.9	≥ 5.9
16	malign	2.2 - 4.08	0.91 - 4.04	≥ 3.08	≥ 4.38	≥ 3.46	≥ 5.32	0.81 - 3.26	0.59 - 0.9	< 0.43
17	benign	1.13 - 2.2	< 0.45	0.41 - 0.84	0.47 - 0.99	< 1.35	3.6 - 5.32	0.41 - 0.81	0.59 - 0.9	< 0.43
18	benign	2.2 - 4.08	0.45 - 0.91	0.41 - 0.84	0.47 - 0.99	< 1.35	1.69 - 3.6	0.41 - 0.81	< 0.32	0.43 - 0.86
19	malign	2.2 - 4.08	≥ 4.04	≥ 3.08	≥ 4.38	≥ 3.46	≥ 5.32	0.81 - 3.26	≥ 0.9	≥ 5.9
20	benign	2.2 - 4.08	0.45 - 0.91	0.41 - 0.84	0.47 - 0.99	< 1.35	3.6 - 5.32	< 0.41	0.32 - 0.59	0.86 - 5.9
21	malign	≥ 4.08	0.91 - 4.04	≥ 3.08	0.99 - 4.38	≥ 3.46	≥ 5.32	≥ 3.26	≥ 0.9	≥ 5.9
22	malign	≥ 4.08	≥ 4.04	0.84 - 3.08	≥ 4.38	≥ 3.46	≥ 5.32	≥ 3.26	< 0.32	≥ 5.9
23	benign	1.13 - 2.2	0.45 - 0.91	0.41 - 0.84	0.47 - 0.99	< 1.35	1.69 - 3.6	0.41 - 0.81	< 0.32	0.43 - 0.86
24	benign	2.2 - 4.08	0.45 - 0.91	0.41 - 0.84	< 0.47	< 1.35	< 1.69	0.81 - 3.26	< 0.32	< 0.43
25	malign	2.2 - 4.08	0.91 - 4.04	≥ 3.08	0.99 - 4.38	1.35 - 1.79	3.6 - 5.32	≥ 3.26	< 0.32	≥ 5.9
26	benign	1.13 - 2.2	0.91 - 4.04	< 0.41	0.47 - 0.99	< 1.35	1.69 - 3.6	0.41 - 0.81	< 0.32	0.43 - 0.86
27	benign	1.13 - 2.2	0.45 - 0.91	0.84 - 3.08	< 0.47	1.35 - 1.79	3.6 - 5.32	< 0.41	0.32 - 0.59	< 0.43
28	benign	1.13 - 2.2	< 0.45	0.41 - 0.84	< 0.47	1.35 - 1.79	< 1.69	0.81 - 3.26	< 0.32	< 0.43
29	benign	< 1.13	0.45 - 0.91	< 0.41	0.99 - 4.38	1.79 - 3.46	< 1.69	< 0.41	≥ 0.9	0.86 - 5.9
30	benign	1.13 - 2.2	0.91 - 4.04	0.41 - 0.84	0.99 - 4.38	< 1.35	1.69 - 3.6	< 0.41	0.59 - 0.9	< 0.43
31	benign	2.2 - 4.08	0.45 - 0.91	< 0.41	< 0.47	1.35 - 1.79	< 1.69	0.41 - 0.81	< 0.32	0.86 - 5.9
32	malign	≥ 4.08	≥ 4.04	0.84 - 3.08	≥ 4.38	≥ 3.46	≥ 5.32	≥ 3.26	≥ 0.9	0.86 - 5.9
33	benign	1.13 - 2.2	0.91 - 4.04	0.84 - 3.08	0.47 - 0.99	1.79 - 3.46	< 1.69	< 0.41	< 0.32	0.86 - 5.9
34	benign	1.13 - 2.2	0.45 - 0.91	0.41 - 0.84	0.99 - 4.38	1.79 - 3.46	1.69 - 3.6	0.81 - 3.26	0.32 - 0.59	0.86 - 5.9
35	benign	1.13 - 2.2	< 0.45	< 0.41	< 0.47	1.79 - 3.46	< 1.69	< 0.41	0.59 - 0.9	0.43 - 0.86

Figure 4: Breast Cancer Preprocessing Data

5.3 Data Accuracy

Accuracy is not really a reliable metric for the real performance of a classifier when the number of samples in different classes vary greatly (unbalanced target) because it will yield misleading result. The accuracy is measured as follows in equation 1 as shown in Fig(5):

$$\textbf{Accuracy} = \frac{number\ of\ correct\ predictions}{total\ of\ all\ cases\ to\ be\ predicted} \qquad \textbf{...... (1)}$$

$$= \frac{a+d}{a+b+c+d}$$

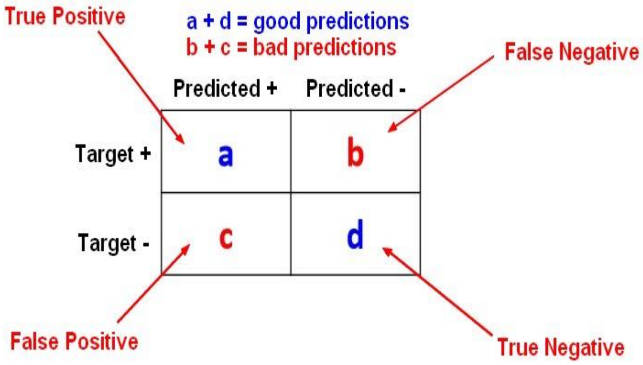


Figure 5: Accuracy Measurement

For orange data mining tool, the accuracy of decision tree, naïve Bayes, KNN classifiers results are shown in table1.The achieved accuracy measure ranged between 55% and 93% for breast cancer data the heights for KNN classifier and the lowest for NB classifier, while the achieved accuracy measure ranged between 51% and 96% for the heart disease data the heights were for KNN classifier and the lowest for decision tree classifier this is more advanced result from the other.

Table 1: The Accuracy Results for the Selected Data across the Classifiers

Algorithm	Breast cancer accuracy	Heart disease accuracy
Decision tree	74%	51%
Naïve Bayes	55%	89%
KNN	93%	96%

The above table shows the effectiveness of each classifier based prediction model for each data patient. In the following, we have shown the experimental results Naive Bayes (NB), Decision Tree (DT), K-Nearest Neighbors (KNN), classifiers based on the models. For each model, we utilized the same datasets in order to compare the techniques fairly. For the purpose of evaluating our approach, it trains each model on 9 sets and tests it using the remaining one set. According to the procedure of cross validation, this has been repeated 10 times and we got a mean prediction results for each model. To show the effectiveness of each classifier based model, we calculated and compared the prediction results in terms of accuracy measure as defined above.

5.4 Performance Improvement:

It is critical to quantify the impact of utilizing various informational indexes to assess the classifier on the proposed information mining device. As shown in Figure3. The highest improvement performance accuracy of the selected classifier for both different data type shows the more efficient is KNN classifier accuracy percentage as shown in Fig (6).

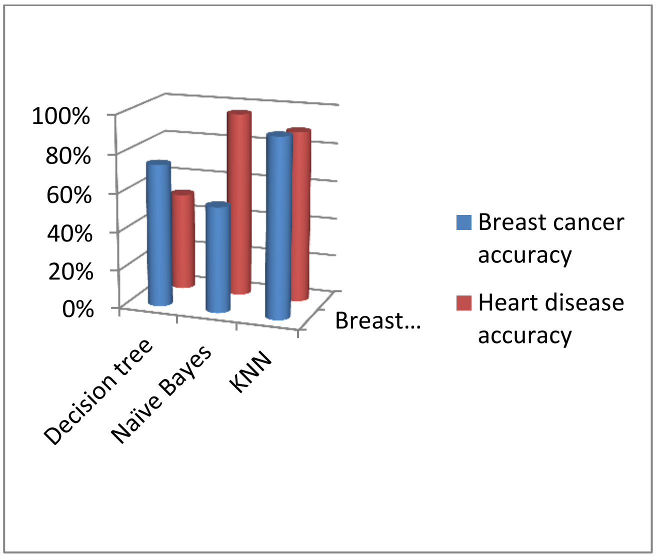


Figure 6: Performance Improvement Classifeirs

6- Conclusion:

After applying the three algorithms on the heart disease and breast cancer the results are as below:

Table 2: Comparison result of using Naïve Bayes, Decision tree and KNN on breast cancer data.

Evaluation Criteria	Naïve Bayes	Decision tree	KNN
Time to achieve model (in sec)	0.22	0.13	0.11
Correctly classified instances (CCI)	226	300	314
Incorrectly classified instances (ICI)	160	86	72

Table 3: Comparison result of using Naïve Bayes, Decision tree and KNN on heart disease data.

Evaluation Criteria	Naïve Bayes	Decision tree	KNN
Time to achieve model (in sec)	0.20	0.10	0.13
Correctly classified instances (CCI)	233	213	241
Incorrectly classified instances(ICI)	39	49	21

In Tables 2 and 3, we have shown the prediction results of each classifier based model in terms of CCI (correctly classified instances) rate, ICI (incorrectly classified instances) rate, value of two different datasets. For example, as shown in Table 2, we see that the correctly classified instances of are 300 on decision tree are higher than other classifier based models. Similarly, according to the experimental results shown in Table 3, we see that the correctly classified instances are 233 on Naïve Bayes that are also higher than other classifier based models.

From the above results we notice that time to build model in KNN is less than others on breast cancer data and is lesser in decision tree on heart disease data and as shown in table1 the accuracy criteria gave highest value on the KNN algorithm. Hence it was superior and more efficient according to the Accuracy measure result. From table1 the achieved results show accurately the KNN classifier was more efficient in accuracy for the both given data set while the NB classifier was the lowest efficient from the selected data classifier on the given data set.

References:

1. Amrita Naik, Lilavati Samant , "Correlation review of classification algorithm using data mining tool: WEKA , Rapidminer , Tanagra ,Orange and Knime" , Elsevier, vol.85,2016,pp.662-668.

2. survey on prediction and analysis the occurrence of heart disease using data mining techniques, international journal of pure and applied mathematics, vol 118, no.8 , 2018,165-174, ISNN:1311-8080 .

3. Abdullah H Wahbeh, Qasem A. Al-Radaideh, Mohammed Nat. Al-Kabi, Emad Al-Shawakfa, "A Comparison Study between Data Mining Tools over some Classification Methods", IJACSA, special issue on artificial intelligence, 2011, pp.18-26.

4. Mahesh Pal, Paul M. Mather, "An assessment of the effectiveness of decision tree methods for land cover classification", Elsevier, vol.86, 2003, pp. 554-565.

5. P.Bhargavi, Dr.S.Jyothi, "Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils", IJCSNS, vol.9,2019,pp.117-122.

6. Johannes Ledolter," Data mining and business analytics with *R*", John Wiley & Sons,Inc, 2013, pp.116.

7. Beant Kaur, Williamjeet Singh, "Review on Heart Disease Prediction System using Data Mining Techniques", IJRITCC |, vol.2, 2014, pp.3003-3008.

8. Kalpana Rangra, Dr. K. L. Bansal, "Comparative Study of Data Mining Tools", IJARCSSE, vol.4, 2014, pp.216-223.

9. Janez Demˇsar, ,Blaˇz Zupan,, Gregor Leban, and Tomaz Curk, "Orange: From Experimental Machine Learning to Interactive Data Mining", Springer, Verlag Berlin Heidelberg, 2004, pp.537-539.

10. http://eric.univ-lyon2.fr/~ricco/.

11. https://www.simplehearttest.com/site.

12. Sarangam Kodati , R. Vivekanandam, "Analysis of Heart Disease using in Data Mining Tools Orange and Weka" , Global Journal of Computer Science and Technology, vol.18 ,ver.1,2018 ,pp.12-19.

13. Slater S., Joksimovic S., & Gasevic ," Tools for educational data mining", Journal of Educational and Behavioral Statistics, vol. 42, no. 1, 2016, pp. 85-106.