Assignment-1 Report

Data Mining
CSE-5334-004

Student Name:
Mehul Ganjude-1001990551,
Gagan Ujjini Mallikarjuna -1001851247,
Venkata Nagendar Shantiswaroop Adibhatla-1001862413

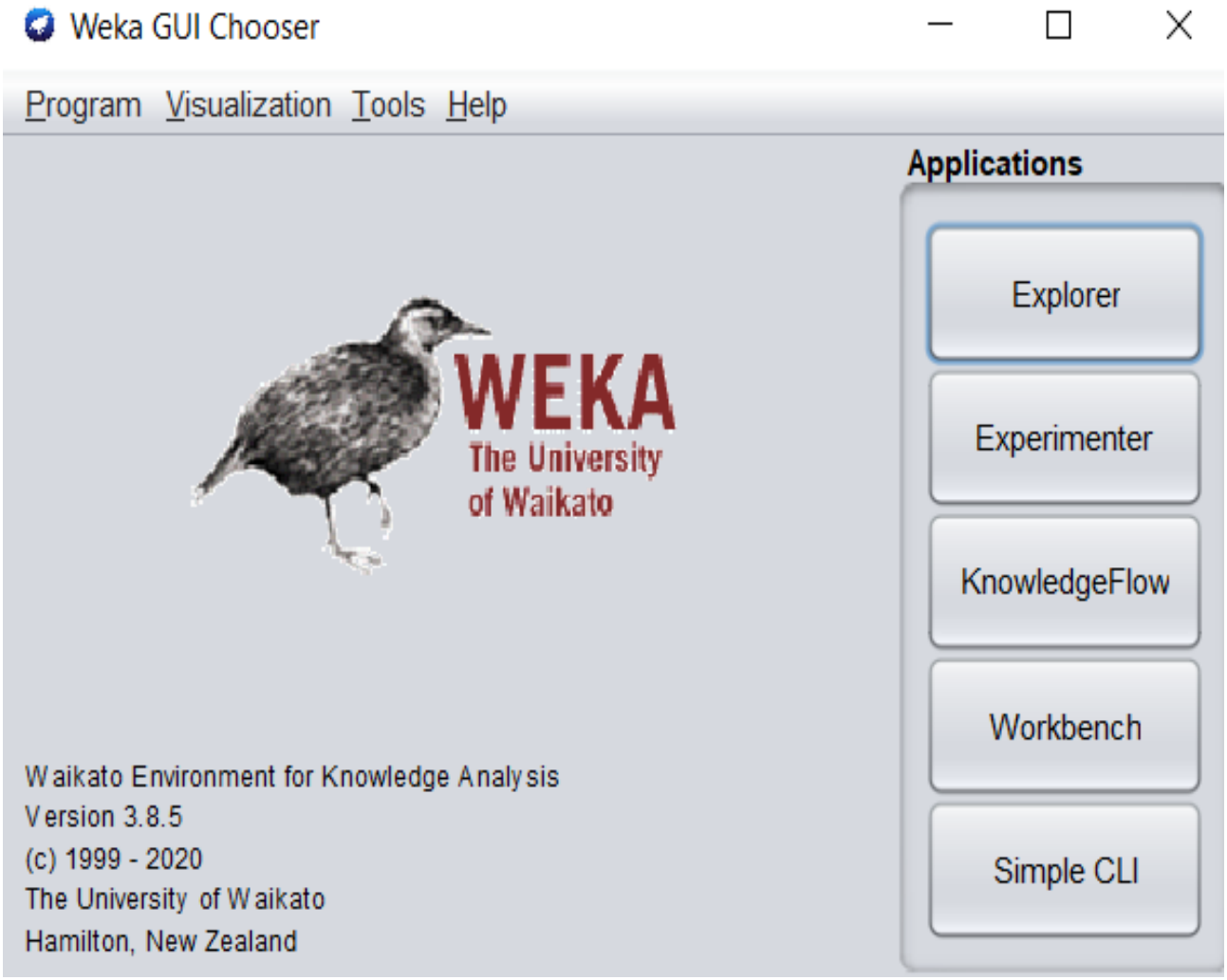
Professor: Dr Elizabeth D Diaz
Teaching Assistant: Pralobh Lokhande

Tool: Weka

WEKA

Introduction

WEKA (Waikato Environment for Knowledge Analysis) developed at the University of Waikato, New Zealand, is free software licensed under the GNU General Public License and is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization.



Retrieving the data

In Weka the data needs to be in ARFF format (Attribute-Relation file format). Here we have converted the dataset from CSV to ARFF. And the following screenshots of the dataset are attached below in CSV and ARFF formats respectively.

Surgical dataset in CSV format.

```
bmi,Age,asa_status,baseline_cancer,baseline_charlson,baseline_cvd,baseline_dementia,baseline_diabetes,baseline_digestive,baseline_osteoart,
19.31,59.2,1,1,0,0,0,0,0,0,0,0,19,0.18337045,0.00742391,-0.57,3,0,7.63,6,1,0,-0.43,1,no
18.73,59.1,0,0,0,0,0,0,0,0,0,0,1,0.31202858,0.01667328,0.21,0,0,12.93,0,1,0,-0.41,1,no
21.85,59,0,0,0,0,0,0,0,0,0,0,6,0.15070644,0.00196232,0,2,0,7.68,5,3,0,0.08,1,no
18.49,59,1,0,1,0,0,1,1,0,0,0,7,0.05616606,0,-0.65,2,1,7.58,4,3,0,-0.32,1,no
19.7,59,1,0,0,0,0,0,0,0,0,0,11,0.19730477,0.00276434,0,0,0,7.88,11,0,0,0,1,no
20.24,59,0,1,0,0,0,0,0,0,0,1,0,14,0.06847826,0,0,1,0,7.63,0,3,0,0.15,1,no
21.18,59,0,1,0,0,0,0,0,0,0,0,0,14,0.06847826,0,0,0,0,9.62,10,3,0,0,1,no
18.99,58.9,0,0,0,0,0,0,0,0,0,0,1,0.31202858,0.01667328,-0.38,4,0,8.6,6,1,0,0,1,no
22.2,58.9,1,0,0,0,0,0,0,0,0,0,1,0.31202858,0.01667328,0,4,0,13,10,0,0,0,1,no
20.83,58.9,1,1,6,0,0,0,1,0,0,0,18,0.46612903,0.01290323,1.87,4,1,10.05,5,1,0,2.08,1,no
22.37,58.8,1,0,0,0,0,0,0,1,0,0,0,0.08197692,0.00295946,-2.03,1,0,10.3,4,0,0,-2.48,1,no
23.46,58.8,0,1,8,0,0,0,0,0,0,0,1,0.31202858,0.01667328,7.56,3,1,7.9,2,0,1,2.86,1,no
22.75,58.8,0,0,0,0,0,0,0,0,0,0,6,0.15070644,0.00196232,-0.57,1,0,10.45,10,3,0,-0.43,1,no
21.35,58.8,1,1,3,1,0,1,0,0,0,0,10,0.04977376,0.00226244,-0.84,4,0,15.75,6,1,0,-2.13,1,no
23.08,58.8,0,0,0,0,0,0,0,0,0,0,11,0.19730477,0.00276434,0,3,1,7.77,6,1,0,-1.3,1,no
21.95,58.8,1,0,0,1,0,0,0,0,0,0,11,0.19730477,0.00276434,-0.99,2,0,7.7,8,1,0,-1.16,1,no
21.04,58.8,0,1,3,1,0,0,0,0,1,1,13,0.10936917,0.00037327,-1.34,0,1,7.48,3,3,0,-0.19,1,no
21.25,58.8,0,0,0,0,0,0,1,0,0,0,18,0.46612903,0.01290323,0,1,0,10.77,11,3,0,0.19,1,no
22.77,58.8,0,1,0,0,0,0,0,0,0,0,19,0.18337045,0.00742391,0,1,0,17.67,8,1,0,-1.33,1,no
23.05,58.7,1,0,0,1,0,0,1,0,0,0,1,0.31202858,0.01667328,-0.78,1,1,10.97,0,3,0,-0.41,1,no
21.25,58.7,0,0,0,0,0,0,0,1,0,0,5,0.09747607,0.00739774,0,4,0,8.5,1,1,0,-1.07,1,no
25.08,58.7,0,0,0,0,0,0,0,1,0,0,5,0.09747607,0.00739774,0.8,3,1,10.9,2,3,0,-1.06,1,no
23.13,58.7,1,1,3,0,0,1,0,0,1,0,6,0.15070644,0.00196232,-0.57,2,0,10.53,5,1,0,-1.21,1,no
23.82,58.7,0,0,0,0,0,0,0,0,0,0,7,0.05616606,0,-0.57,2,1,7.97,7,1,0,-0.43,1,no
24.26,58.7,0,1,3,0,0,0,0,0,0,1,11,0.19730477,0.00276434,-0.33,2,1,7.63,4,3,0,0.37,2,no
23.58,58.7,0,1,2,0,0,0,0,0,0,0,12,0.13541667,0.00173611,0,1,0,14.45,0,1,0,0,1,no
24.61,58.7,0,1,0,0,0,0,0,0,1,0,16,0.01978022,0.0021978,0,2,0,7.75,6,2,0,0,1,no
18.53,58.7,0,0,1,0,0,0,1,0,0,0,18,0.46612903,0.01290323,-0.32,2,0,11.47,9,3,0,0.37,1,no
23.08,58.7,1,1,2,1,0,0,1,0,0,0,20,0.0161182,0.00067159,-0.83,0,0,7.5,1,0,0,-0.57,1,no
```

Surgical dataset in ARFF format.

```
@relation Surgical-dataset

@attribute bmi numeric
@attribute Age numeric
@attribute asa_status numeric
@attribute baseline_cancer numeric
@attribute baseline_charlson numeric
@attribute baseline_cvd numeric
@attribute baseline_dementia numeric
@attribute baseline_diabetes numeric
@attribute baseline_digestive numeric
@attribute baseline_osteoart numeric
@attribute baseline_psych numeric
@attribute baseline_pulmonary numeric
@attribute ahrq_ccs numeric
@attribute ccsComplicationRate numeric
@attribute ccsMort30Rate numeric
@attribute complication_rsi numeric
@attribute dow numeric
@attribute gender numeric
@attribute hour numeric
@attribute month numeric
@attribute moonphase numeric
@attribute mort30 numeric
@attribute mortality_rsi numeric
@attribute race numeric
@attribute complication {no,yes}

@data
19.31,59.2,1,1,0,0,0,0,0,0,0,0,19,0.18337,0.007424,-0.57,3,0,7.63,6,1,0,-0.43,1,no
18.73,59.1,0,0,0,0,0,0,0,0,0,0,1,0.312029,0.016673,0.21,0,0,12.93,0,1,0,-0.41,1,no
21.85,59,0,0,0,0,0,0,0,0,0,0,6,0.150706,0.001962,0,2,0,7.68,5,3,0,0.08,1,no
```

Glimpse of Data

The Dataset chose is Surgical Dataset. It has 25 attributes few of them are bmi, age, gender, month, race, complication and the total number of instances are 14635. The majority of the data is in the binary format (like 0 and 1) and other columns provide more information about bmi, age, etc.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
18.73	59.1	0	0	0	0	0	0	0	0	0	0	1	0.312029	0.016673	0.21	0	0	12.93	0	1	0
21.85	59	0	0	0	0	0	0	0	0	0	0	6	0.150706	0.001962	0	2	0	7.68	5	3	0
18.49	59	1	0	1	0	0	1	1	0	0	0	7	0.056166	0	-0.65	2	1	7.58	4	3	0
19.7	59	1	0	0	0	0	0	0	0	0	0	11	0.197305	0.002764	0	0	0	7.88	11	0	0
20.24	59	0	1	0	0	0	0	0	0	1	0	14	0.068478	0	0	1	0	7.63	0	3	0
21.18	59	0	1	0	0	0	0	0	0	0	0	14	0.068478	0	0	0	0	9.62	10	3	0
18.99	58.9	0	0	0	0	0	0	0	0	0	0	1	0.312029	0.016673	-0.38	4	0	8.6	6	1	0
22.2	58.9	1	0	0	0	0	0	0	0	0	0	1	0.312029	0.016673	0	4	0	13	10	0	0
20.83	58.9	1	1	6	0	0	0	1	0	0	0	18	0.466129	0.012903	1.87	4	1	10.05	5	1	0
22.37	58.8	1	0	0	0	0	0	0	1	0	0	0	0.081977	0.002959	-2.03	1	0	10.3	4	0	0
23.46	58.8	0	1	8	0	0	0	0	0	0	0	1	0.312029	0.016673	7.56	3	1	7.9	2	0	1
22.75	58.8	0	0	0	0	0	0	0	0	0	0	6	0.150706	0.001962	-0.57	1	0	10.45	10	3	0
21.35	58.8	1	1	3	1	0	1	0	0	0	0	10	0.049774	0.002262	-0.84	4	0	15.75	6	1	0
23.08	58.8	0	0	0	0	0	0	0	0	0	0	11	0.197305	0.002764	0	3	1	7.77	6	1	0
21.95	58.8	1	0	0	1	0	0	0	0	0	0	11	0.197305	0.002764	-0.99	2	0	7.7	8	1	0
21.04	58.8	0	1	3	1	0	0	0	0	1	1	13	0.109369	0.000373	-1.34	0	1	7.48	3	3	0
21.25	58.8	0	0	0	0	0	0	1	0	0	0	18	0.466129	0.012903	0	1	0	10.77	11	3	0
22.77	58.8	0	1	0	0	0	0	0	0	0	0	19	0.18337	0.007424	0	1	0	17.67	8	1	0
23.05	58.7	1	0	0	1	0	0	1	0	0	0	1	0.312029	0.016673	-0.78	1	1	10.97	0	3	0
21.25	58.7	0	0	0	0	0	0	0	1	0	0	5	0.097476	0.007398	0	4	0	8.5	1	1	0
25.08	58.7	0	0	0	0	0	0	0	1	0	0	5	0.097476	0.007398	0.8	3	1	10.9	2	3	0
23.13	58.7	1	1	3	0	0	1	0	0	1	0	6	0.150706	0.001962	-0.57	2	0	10.53	5	1	0

Check for Missing Data

Missing data are the one whose data value is not present in the instance. These missing instances can some time cause significant impact on the final conclusion. Upon investigation of the dataset (Surgical Dataset) we did not find any instance of missing values in one or more attributes and most of the attributes have 0 or 1 as their values.

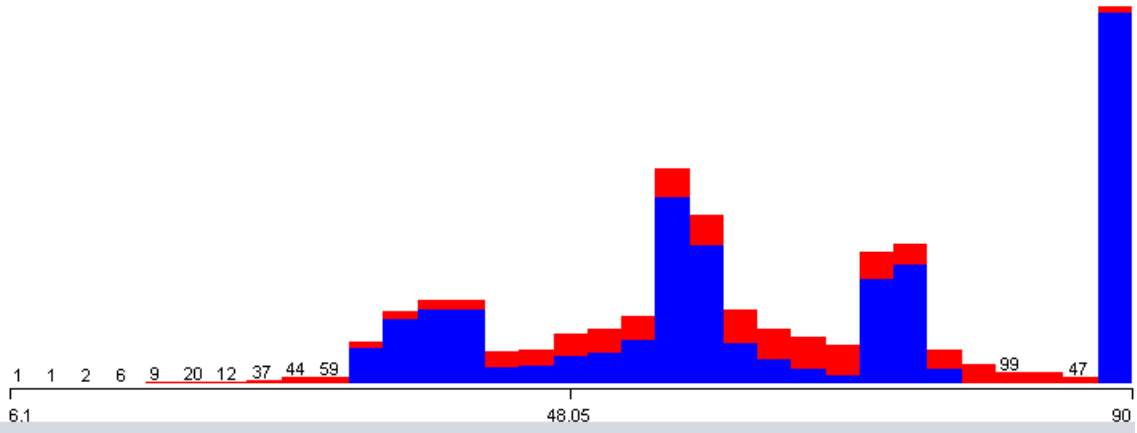
Relation: Surgical-dataset

No.	1: bmi Numeric	2: Age Numeric	3: asa_status Numeric	4: baseline_cancer Numeric	5: baseline_charlson Numeric	6: baseline_cvd Numeric	7: baseline_dementia Numeric	8: baseline_diabetes Numeric	9: baseline_digestive Numeric	10: baseline_o Numeric
1	19.31	59.2	1.0	1.0	0.0	0.0	0.0	0.0	0.0	
2	18.73	59.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
3	21.85	59.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
4	18.49	59.0	1.0	0.0	1.0	0.0	0.0	1.0	1.0	
5	19.7	59.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	
6	20.24	59.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	
7	21.18	59.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	
8	18.99	58.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
9	22.2	58.9	1.0	0.0	0.0	0.0	0.0	0.0	0.0	
10	20.83	58.9	1.0	1.0	6.0	0.0	0.0	0.0	1.0	
11	22.37	58.8	1.0	0.0	0.0	0.0	0.0	0.0	0.0	
12	23.46	58.8	0.0	1.0	8.0	0.0	0.0	0.0	0.0	
13	22.75	58.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
14	21.35	58.8	1.0	1.0	3.0	1.0	0.0	1.0	0.0	
15	23.08	58.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
16	21.95	58.8	1.0	0.0	0.0	1.0	0.0	0.0	0.0	
17	21.04	58.8	0.0	1.0	3.0	1.0	0.0	0.0	0.0	
18	21.25	58.8	0.0	0.0	0.0	0.0	0.0	0.0	1.0	
19	22.77	58.8	0.0	1.0	0.0	0.0	0.0	0.0	0.0	
20	23.05	58.7	1.0	0.0	0.0	1.0	0.0	0.0	1.0	
21	21.25	58.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
22	25.08	58.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
23	23.13	58.7	1.0	1.0	3.0	0.0	0.0	1.0	0.0	
24	23.82	58.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
25	24.26	58.7	0.0	1.0	3.0	0.0	0.0	0.0	0.0	
26	23.58	58.7	0.0	1.0	2.0	0.0	0.0	0.0	0.0	
27	24.61	58.7	0.0	1.0	0.0	0.0	0.0	0.0	0.0	
28	18.53	58.7	0.0	0.0	1.0	0.0	0.0	0.0	1.0	

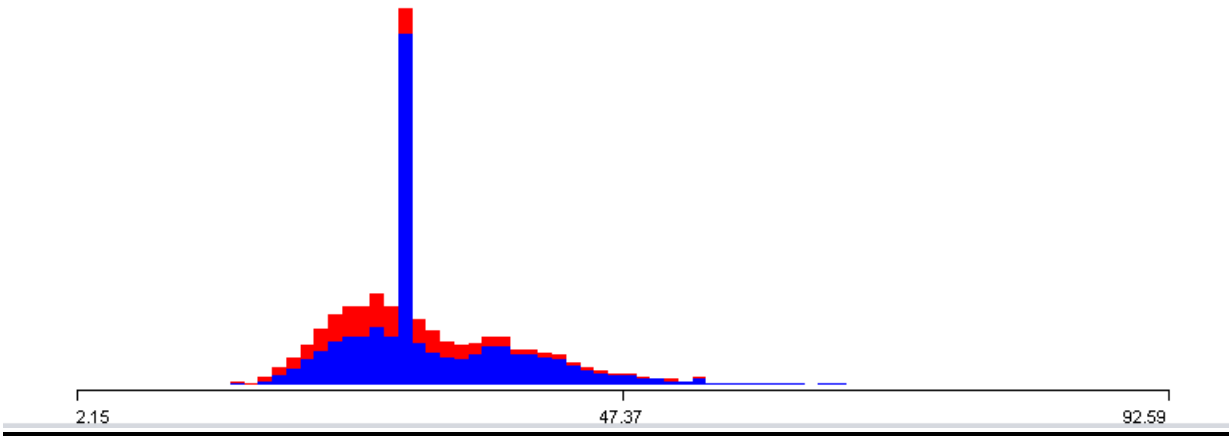
Data Exploration

Data Exploration is the process of performing investigation on data to discover pattern or get some useful insights on the data. Upon exploration of the Surgical dataset, we could get the following insights.

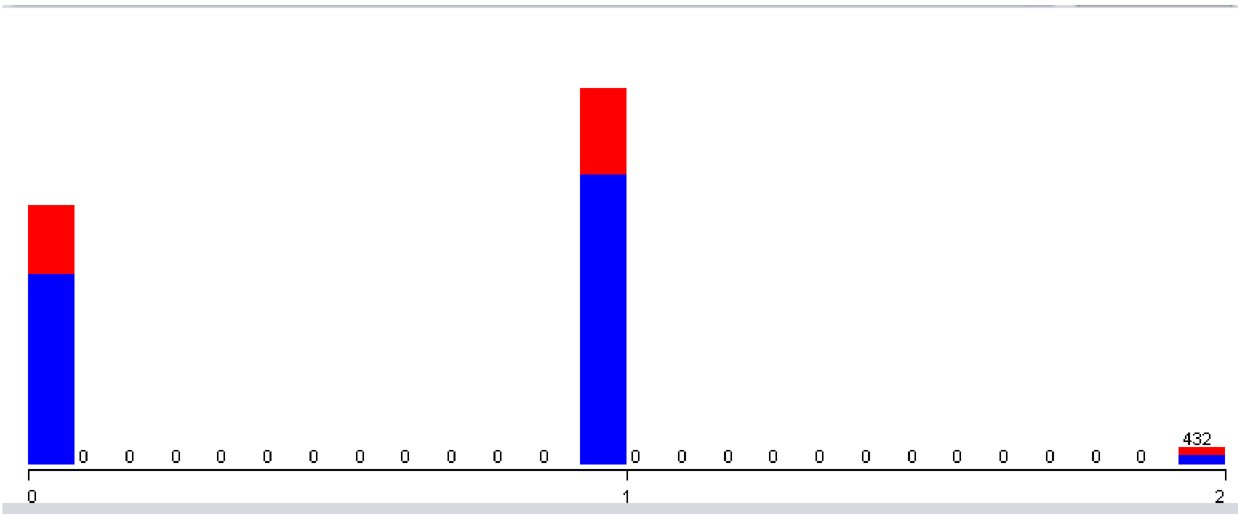
Attribute Age:



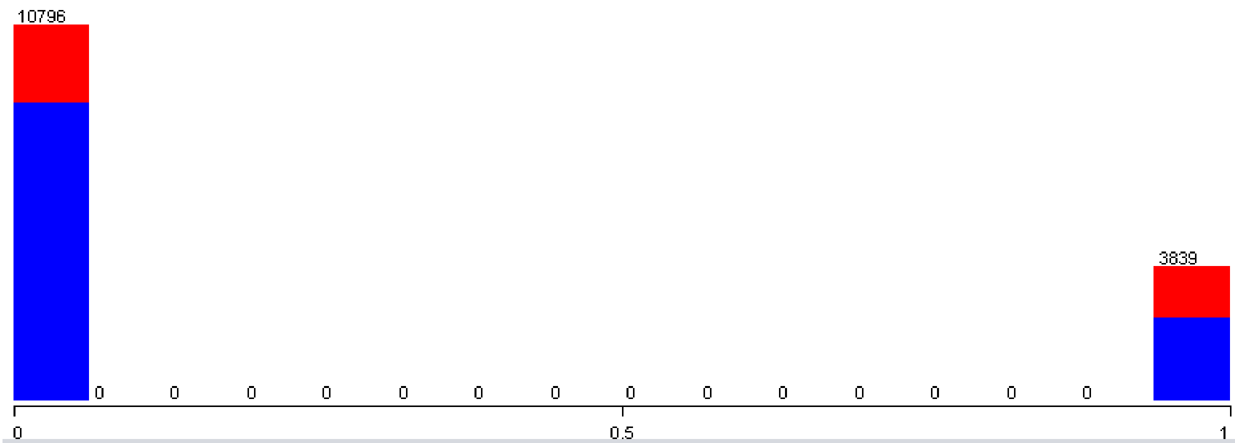
Attribute BMI:



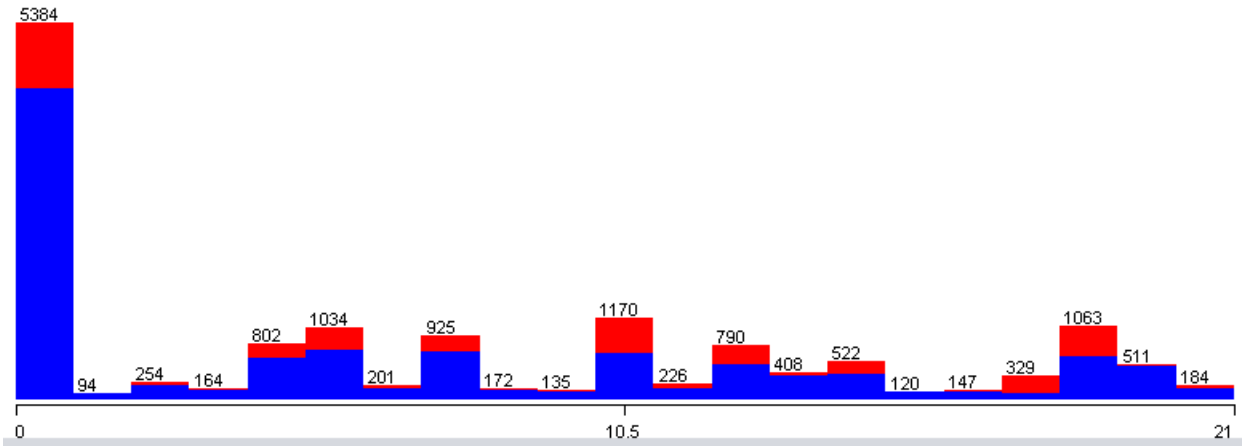
Attribute asa_status:



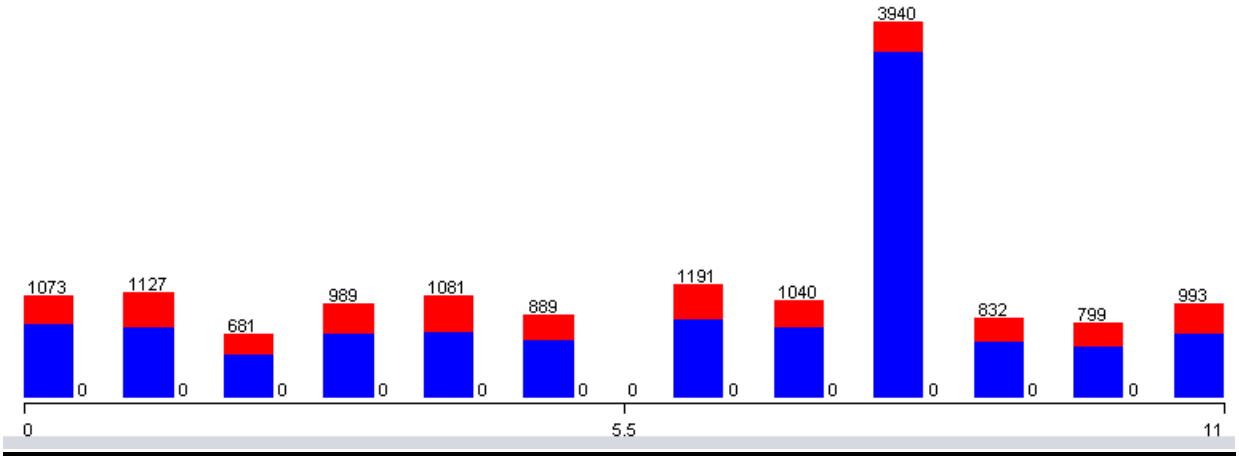
Attribute baseline_cancer:



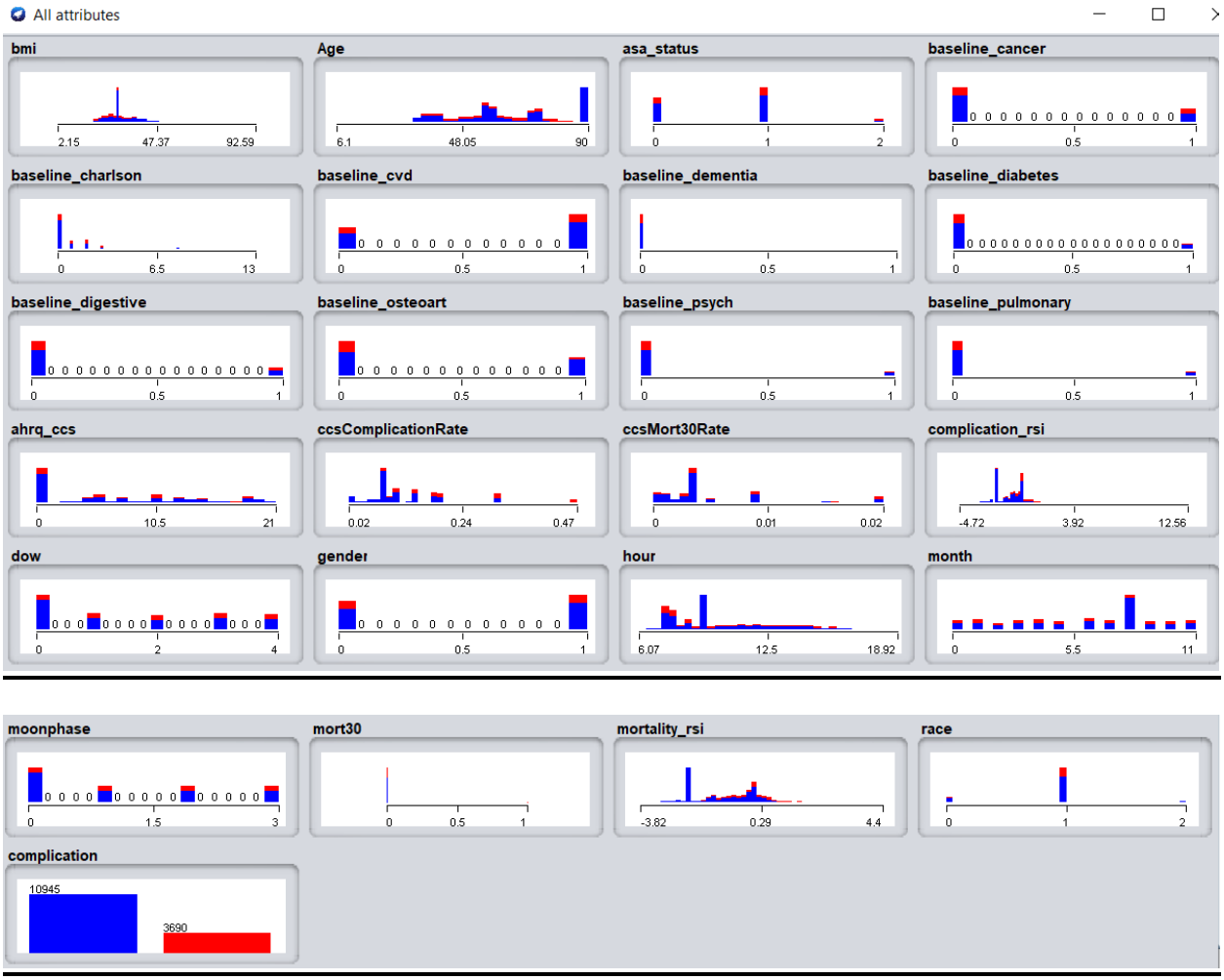
Attribute ahrq_ccs:



Attribute month:

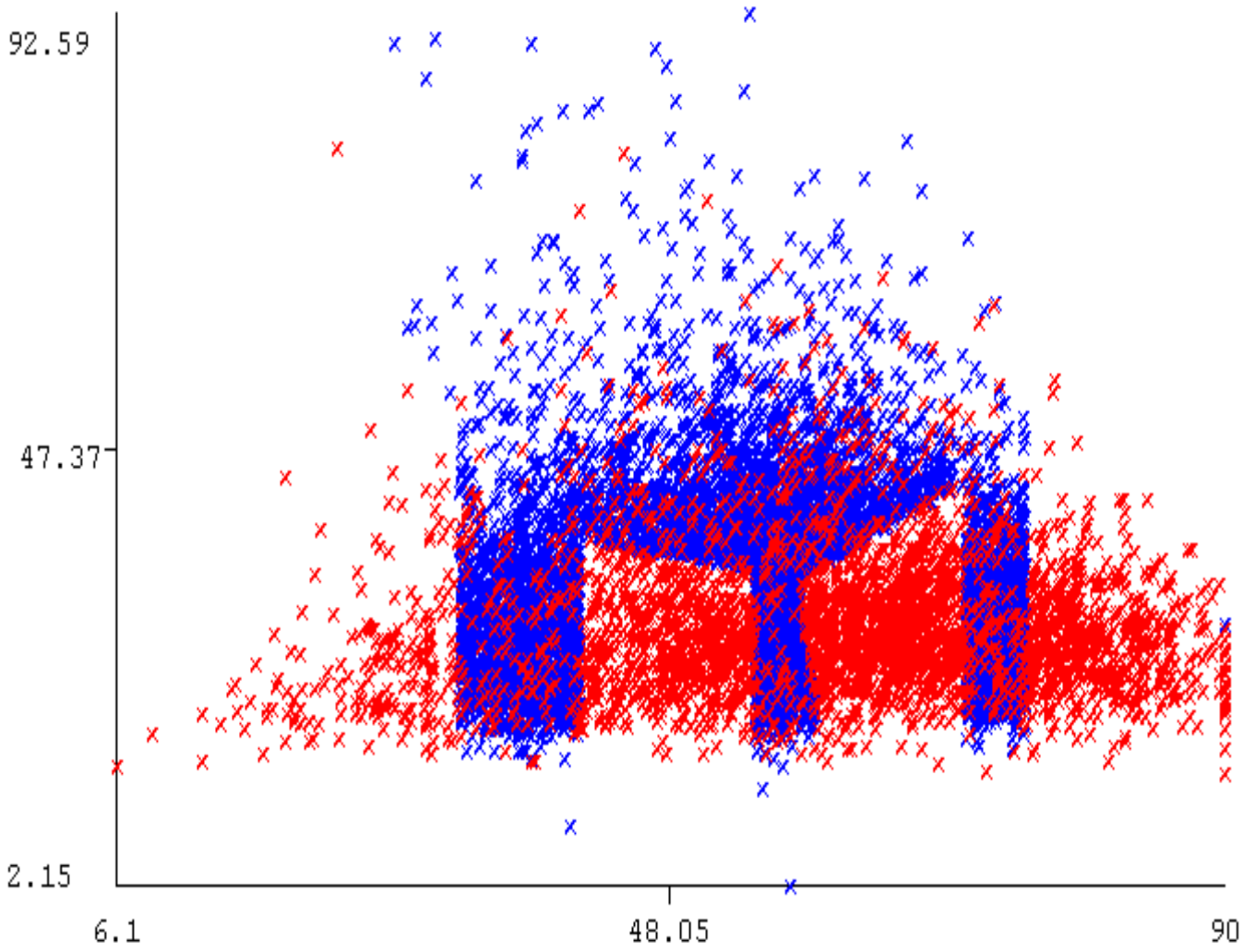


All Attributes:

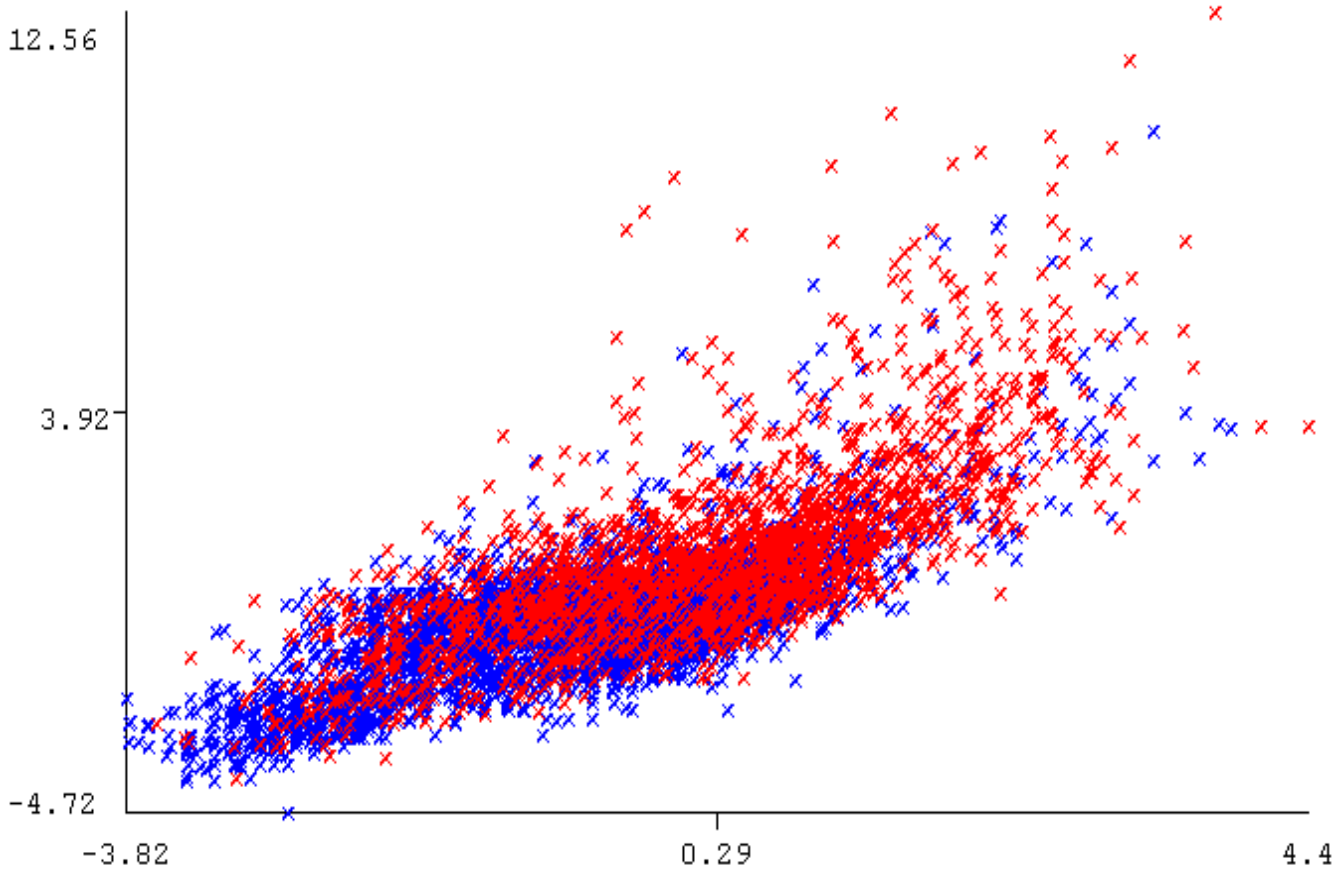


Tasks

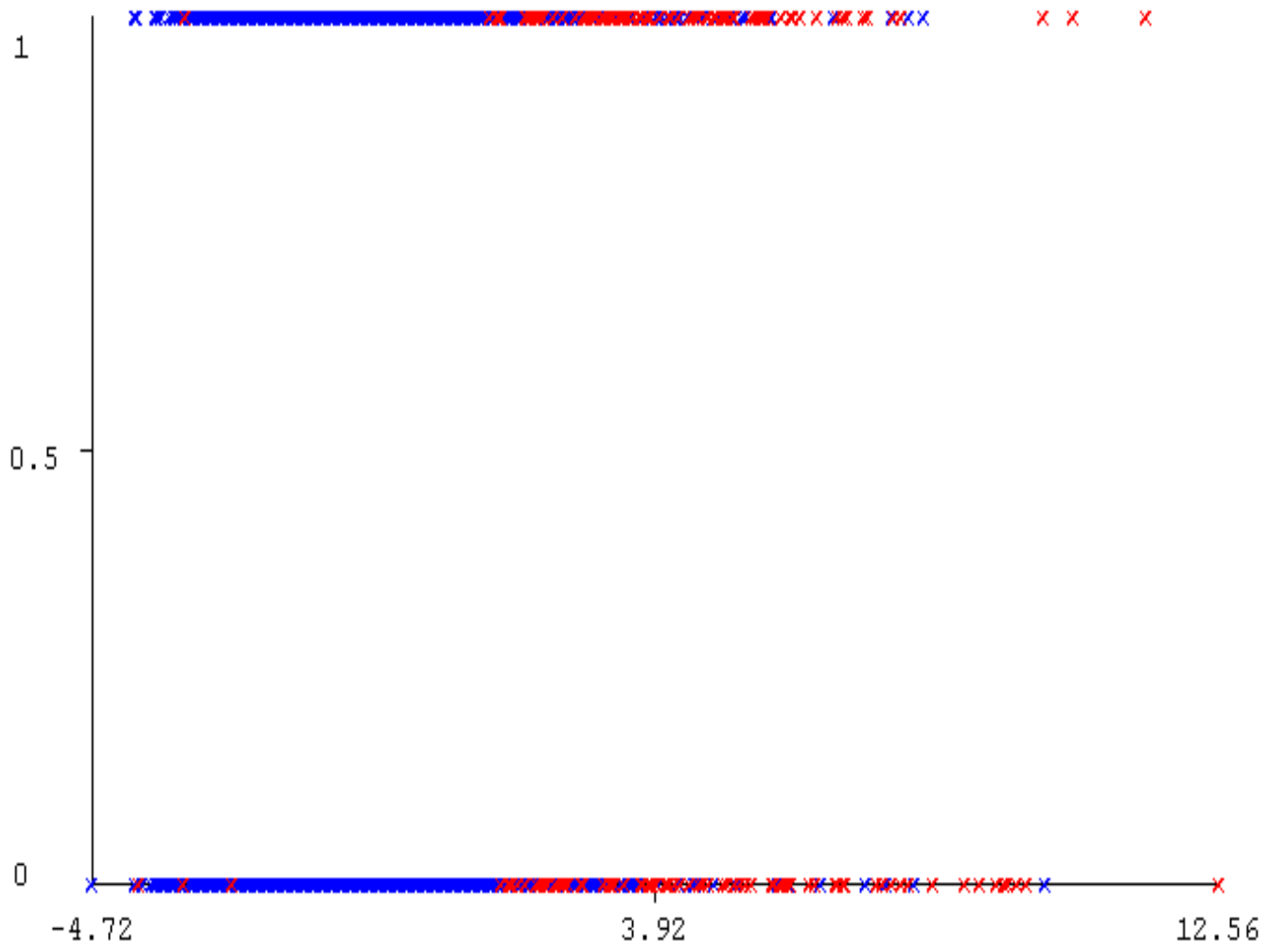
1. The Following below is a plot of Age (x-axis) and bmi (y-axis). From the graph below we can see that the graph is a scatter plot with most of the points appearing in the center and the bmi is less in the initial ages whereas in the mid ages bmi increases exponentially and reaches a peak after which the bmi decreases as the age increases.



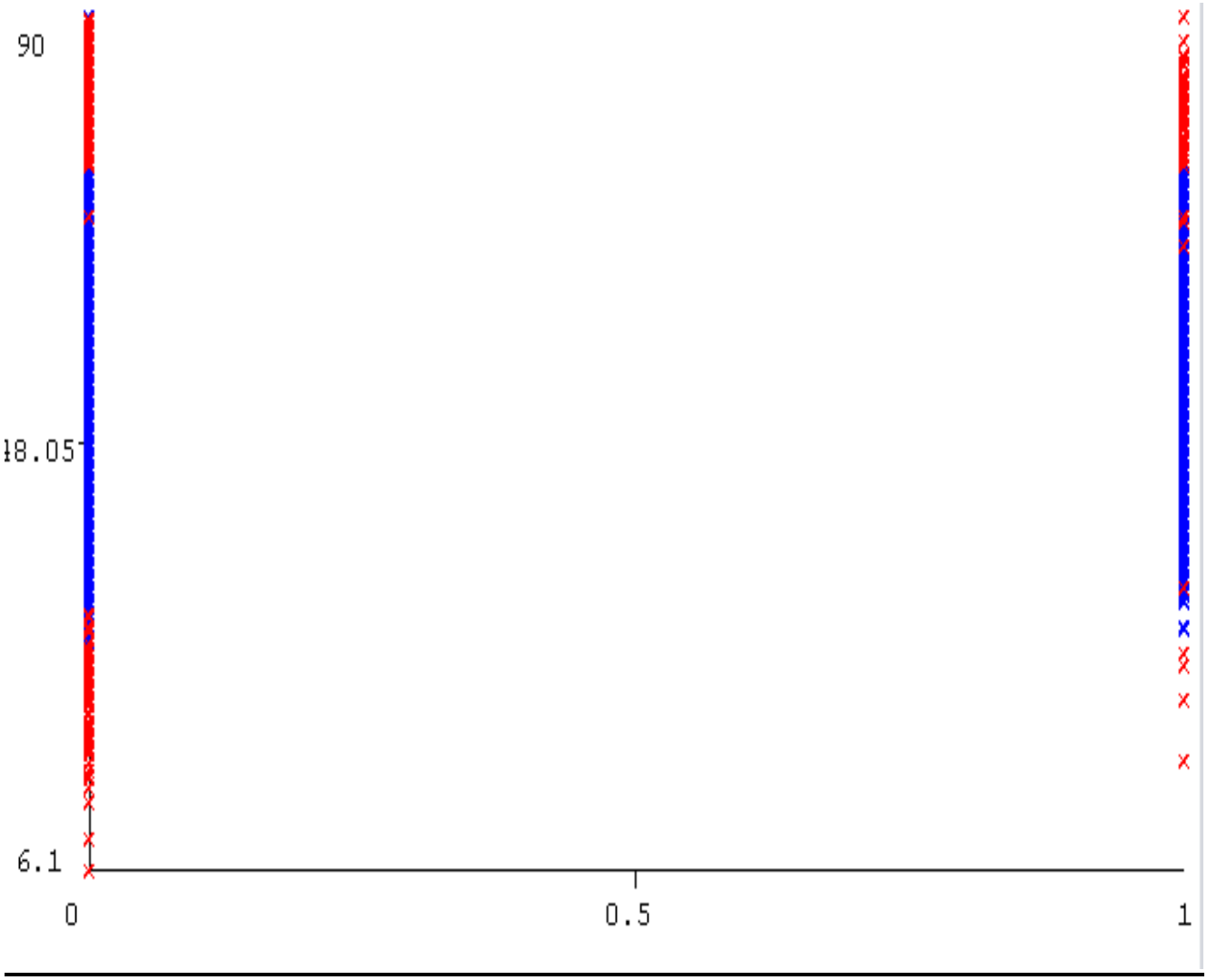
2. The Following below is a plot of mortality_rsi (x-axis) and complication_rsi (y-axis). From the graph below we can see that the graph is a scatter plot and the points are increasing linearly and the most of the points are dense in the center of the graph and as the mortality and complication increase the points are far apart.



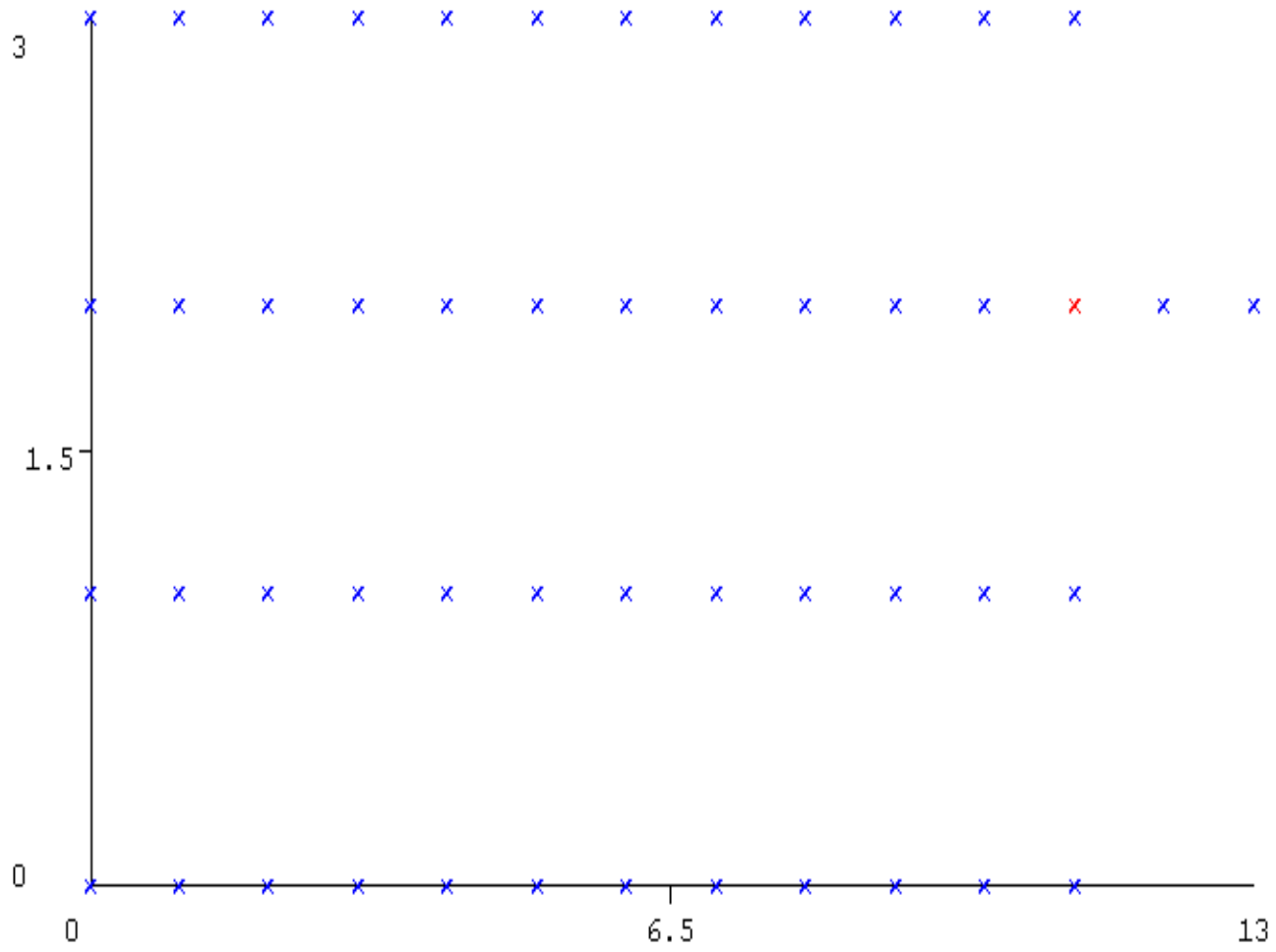
3. The Following below is a plot of complication_rsi (x-axis) and gender (y-axis). From the graph below we can see that the only two values that are taken are 0 and 1 which means that both the attributes(complication, gender) are taking binary values only and the graph is moving parallel to the x-axis.



4. The Following below is a plot of baseline_diabetes (x-axis) and age (y-axis). From the graph below we can see that there are only two values for baseline_diabetes which are 0 and 1 and the diabetes values are increasing as the age increases and the graph is perpendicular to x-axis.



5. The Following below is a plot of baseline_charlson (x-axis) and moonphase (y-axis). From the graph below we can see that there are very few values and all the values are parallel to the x-axis.



References:

- https://www.youtube.com/watch?v=U-1sTxmHE5U
- https://www.softwaretestinghelp.com/weka-explorer-tutorial/
- https://www.cs.auckland.ac.nz/courses/compsci367s1c/tutorials/IntroductionToWeka.pdf