**DATA MINING ASSINGMENT 1**

**Batch 14 - Contact Lenses**

The goal is to see if a patient should have soft or hard contact lenses versus no contact lens

Now we are going to show the data set in an Excel Sheet.

**1. Title: Database for fitting contact lenses**

This Data set display the attributes such as-

* Age
* Type of Spectacle
* Astigmation
* Tear-Prod Rate
* Type of Contact Lenses.

Notes:

* This database is complete (all possible combinations of attribute-value pairs are represented).
* Each instance is complete and correct.

**2. Relevant Information Paragraph:**

1. The examples are complete and noise free.
2. The examples highly simplified the problem.
3. The attributes do not fully describe all the factors affecting the decision as to which type, if any, to fit.

**3. Number of Instances:** 24

**4. Relevant Information Paragraph:**

1. The examples are complete and noise free.
2. The examples highly simplified the problem
3. The attributes do not fully describe all the factors affecting the decision as to which type, if any, to fit.

**5. Number of Attributes:** 4 (all nominal)

**6. Attribute Information:**

-- 3 Classes

1 : the patient should be fitted with hard contact lenses,

2 : the patient should be fitted with soft contact lenses,

3 : the patient should not be fitted with contact lenses.

-- 4 Attributes

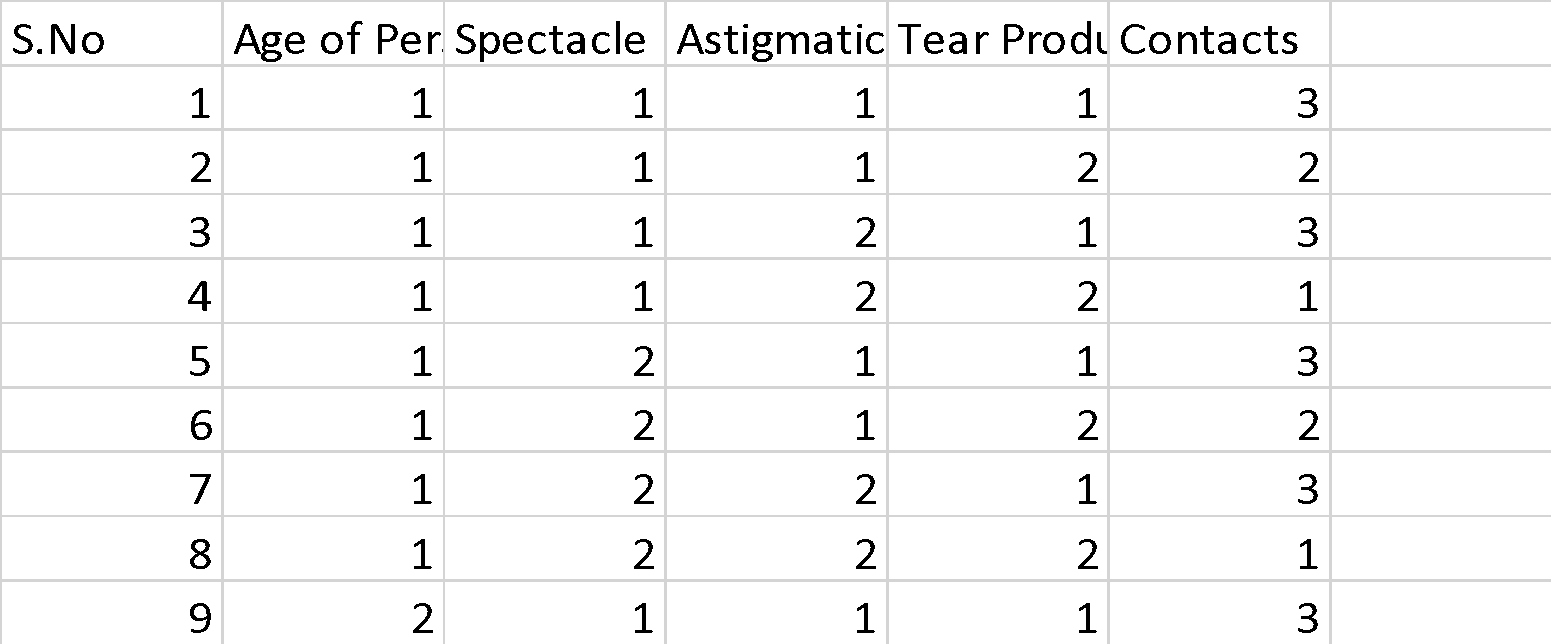
1. age of the patient: (1) young, (2) pre-presbyopic, (3) presbyopic

2. spectacle prescription: (1) myope, (2) hypermetrope

3. astigmatic: (1) no, (2) yes

4. tear production rate: (1) reduced, (2) normal

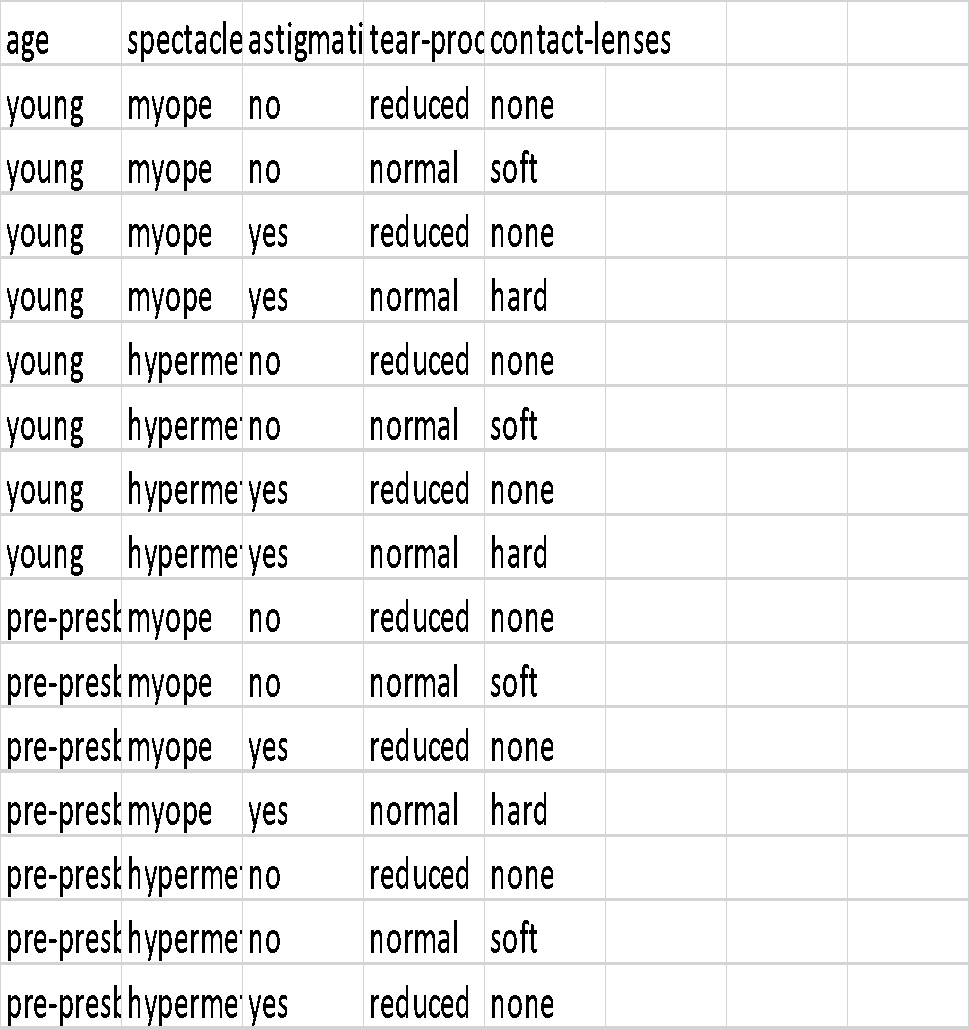
So now, let us look at the data that we have acquired –



**7. Number of Missing Attribute Values:**  0

Now based upon this data , we can generate the required information of Contact Lenses.

**Let us look at the Table-**

****

So, as we see , this is the following information we retrieve.

Based on this we will get the Class Distribution.

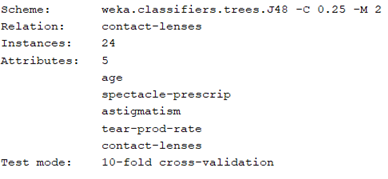
**8. Class Distribution-**

1. hard contact lenses: 4

2. soft contact lenses: 5

3. no contact lenses: 15

**Now this is how the dataset of contact lenses are classified by the Weka tool-**

****

So now we are going to talk about the Central Tendency and The Dispersion of the Data.

A measure of **central tendency** is a single value that attempts to describe a **set** of **data** by identifying the **central** position within that **set** of **data**.

So, for this dataset let us find the Central Tendency

In Central Tendency, we must find 3 main values they are-

* Mean
* Median
* Mode

So now for Each Attribute we shall find Mean first-

1. For Age of Person the Mean is : 2
2. The mean of Spectacle is :1.5
3. The mean for Astigmatic is :1.5
4. The mean of Tear Production Rate: 1.5
5. The mean of contacts : 2.4

So now , next is the Median-

1. Median of Age of Person:2
2. Median of Spectacle :1
3. Median of Astigmatic : 2
4. Median of Tear Production Rate : 1
5. Median of contacts-:3

Now next the Mode –

1. Mode of Age of Person- 1
2. Mode of Spectacle- 1
3. Mode of Astigmatic- 1
4. Mode of Tear Production rate- 1
5. Mode of Contacts- 3

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Age of Person** | **Spectacle** | **Astigmatic** | **Tear Production Rate** | **Contacts** |
| **Mean** | 2 | 1.5 | 1.5 | 1.5 | 2.458 |
| **Median** | 2 | 1.5 | 1.5 | 1.5 | 3 |
| **Mode** | 1 | 1 | 1 | 1 | 3 |

Now let us look at the dispersion of the dataset –

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Age of Person** | **Spectacle** | **Astigmatic** | **Tear Production Rate** | **Contacts** |
| **Dispersion** | 0.834057656 | 0.510753918 | 0.510753918 | 0.510753918 | 0.779027636 |

Now let us look whether our dataset has any data quality issue.

The 7 major Data Quality Issues Include-

* Poor Organization
* Too Much Data
* Inconsistent Data
* Poor Data Security
* Poorly Defined Data
* Incorrect Data
* Poor Data Recovery
* As we have already seen our dataset, we know that our values do not have a poor organization and they are arranged in a proper order with correct attribute names.
* Our Dataset contains Minimal amount of values.
* There is a No Inconsistency in our data as all rows have appropriate values.
* No evidence of Poor Data Security.
* All our data has been properly defined by the use of correct attributes and instances.
* There is no incorrect data in our dataset as the values have been logically allocated.
* Since All of our data is correct there is no need for Recovery of data.

Ok , Now let us look at the most suitable preprocessing technique for this dataset.

It is The Smoothing Technique.

Smoothing is the process of removing noisy data from the dataset.

Let’s look at it in detail.

In Smoothing the main Process is Binning.

There are mainly two types of binning , they are-

* Smoothing by bin means.
* Smoothing by bin boundaries.

Now let us perform Smoothing on our dataset.

**Binning by Mean**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Age of Person** | | | | | | | |  | **Mean** |  | **Age of Person- Binned by Mean** | | | | | | | |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |  | **1** |  | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |  | **2** |  | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |  | **3** |  | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Spectacle** | | | | | | | |  | **Mean** |  | **Spectacle- Binned by Mean** | | | | | | | |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |  | **1** |  | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 |  | **1.5** |  | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 |
| 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |  | **2** |  | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Astigmatic** | | | | | | | |  | **Mean** |  | **Astigmatic - Binned by Mean** | | | | | | | |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |  | **1** |  | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 |  | **1.5** |  | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 |
| 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |  | **2** |  | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Tear Production Rate** | | | | | | | |  | **Mean** |  | **Tear Production Rate - Binned by Mean** | | | | | | | |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |  | **1** |  | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 |  | **1.5** |  | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 |
| 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |  | **2** |  | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Contacts** | | | | | | | |  | **Mean** |  | **Contacts - Binned by Mean** | | | | | | | |
| 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 |  | **1.5** |  | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 |
| 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |  | **2.875** |  | 2.875 | 2.875 | 2.875 | 2.875 | 2.875 | 2.875 | 2.875 | 2.875 |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |  | **3** |  | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |

**Binning by Boundary:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Age of Person** | | | | | | | |  | **Low** | **High** |  | **Age of Person- Binned by Boundary** | | | | | | | |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |  | **1** | **1** |  | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |  | **2** | **2** |  | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |  | **3** | **3** |  | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Spectacle** | | | | | | | |  | **Low** | **High** |  | **Spectacle- Binned by Boundary** | | | | | | | |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |  | **1** | **1** |  | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 |  | **1** | **2** |  | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 |
| 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |  | **2** | **2** |  | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |

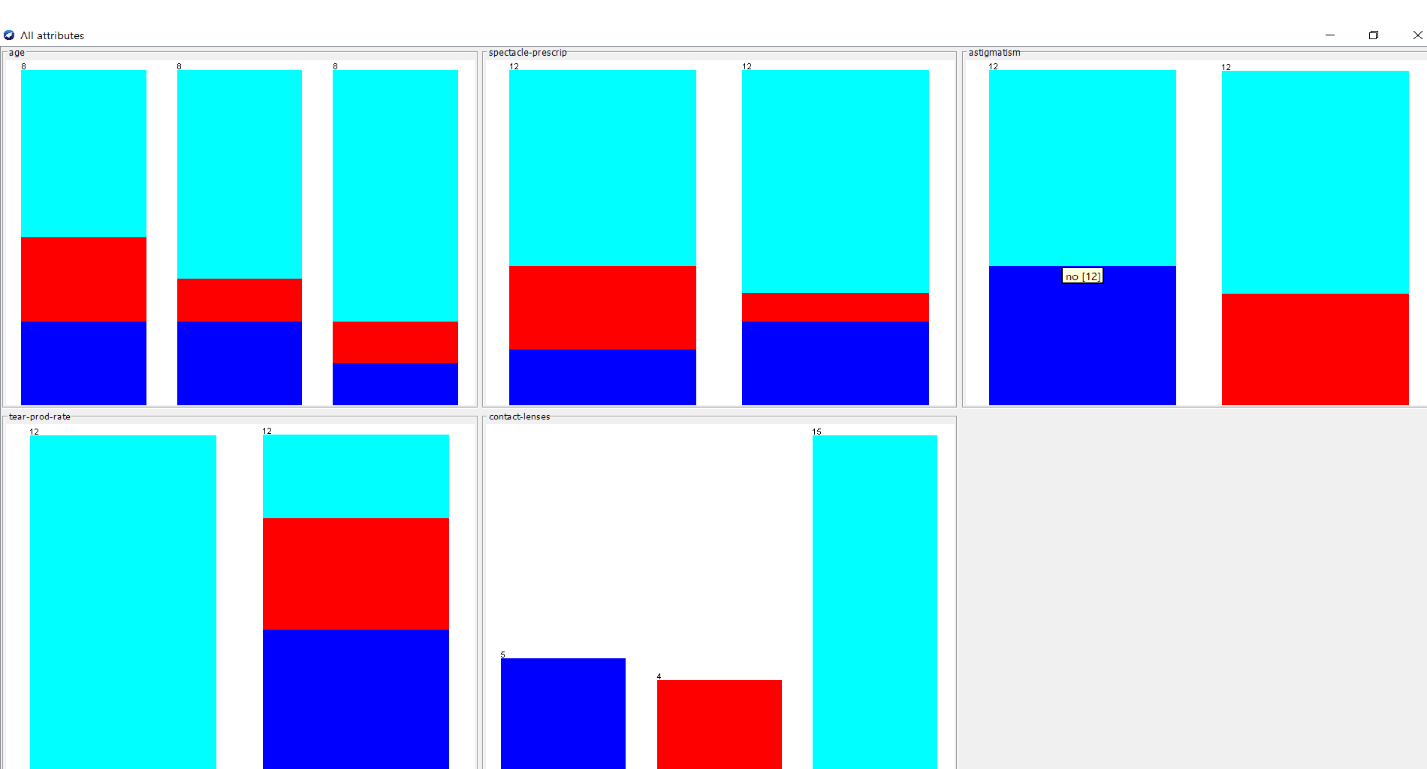
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Astigmatic** | | | | | | | |  | **Low** | **High** |  | **Astigmatic - Binned by Boundary** | | | | | | | |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |  | **1** | **1** |  | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 |  | **1** | **2** |  | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 |
| 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |  | **2** | **2** |  | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Tear Production Rate** | | | | | | | |  | **Low** | **High** |  | **Tear Production Rate - Binned by Boundary** | | | | | | | |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |  | **1** | **1** |  | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 |  | **1** | **2** |  | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 |
| 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |  | **2** | **2** |  | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Contacts** | | | | | | | |  | **Low** | **High** |  | **Contacts - Binned by Boundary** | | | | | | | |
| 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 |  | **1** | **2** |  | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 |
| 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |  | **2** | **3** |  | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |  | **3** | **3** |  | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |

**Now let us Visualize our Data in the WEKA tool.**

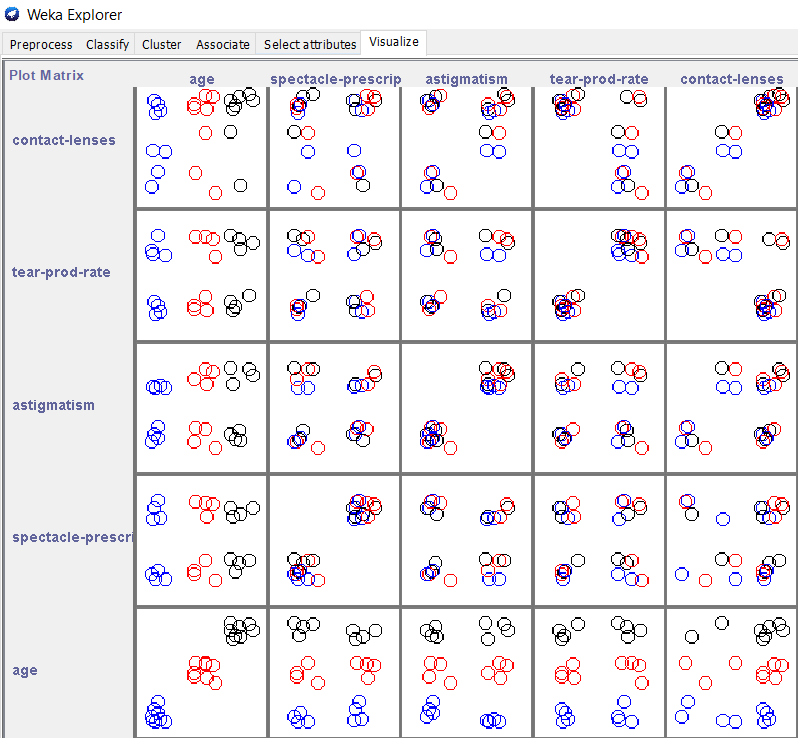
First let us look how the Dataset looks like in a Stack Chart



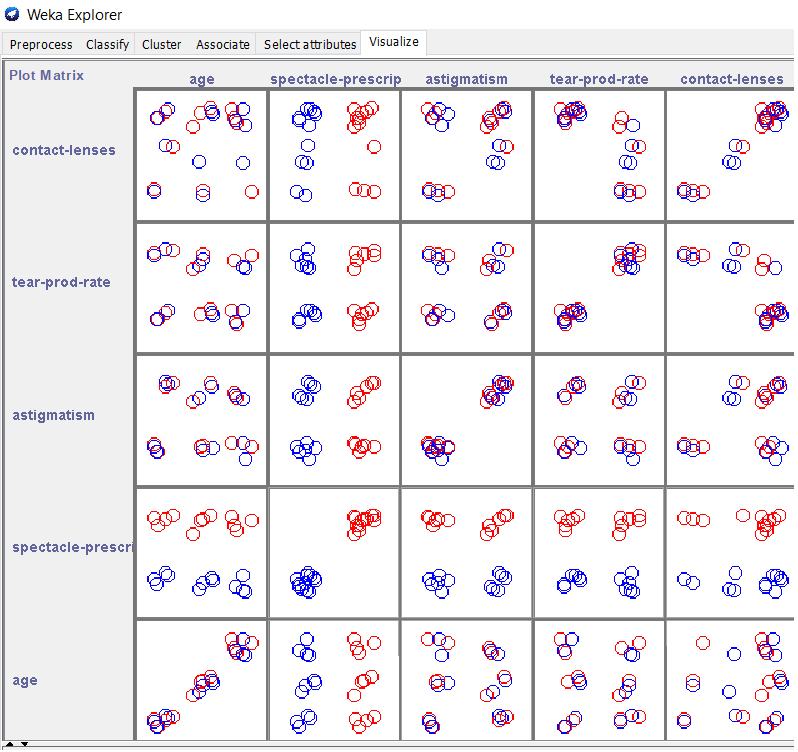
Now we will analyze how the dataset has been plotted in the PLOT MATRIX

Let us look at each Attribute Separately-

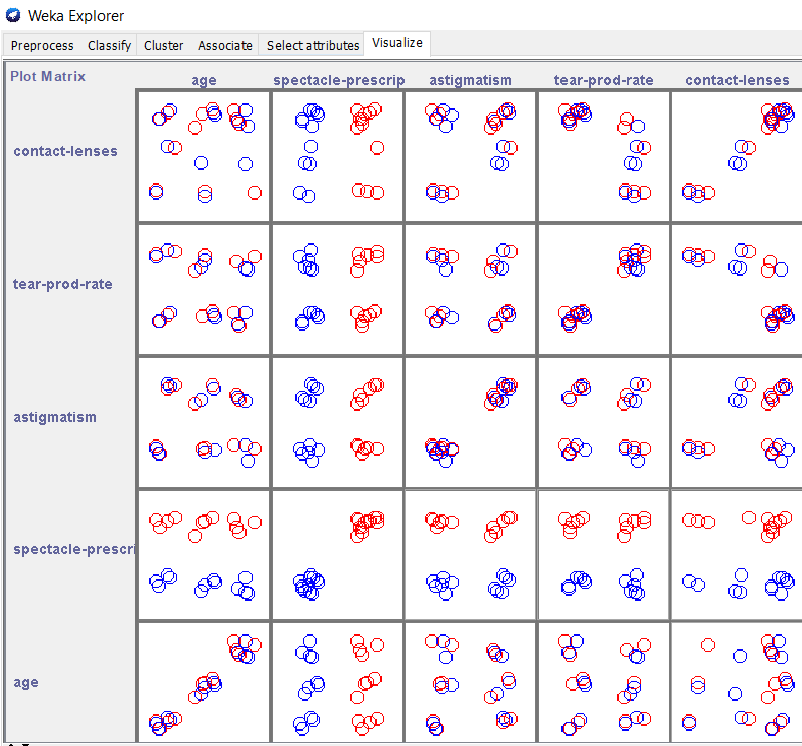
1. Age of Patient –



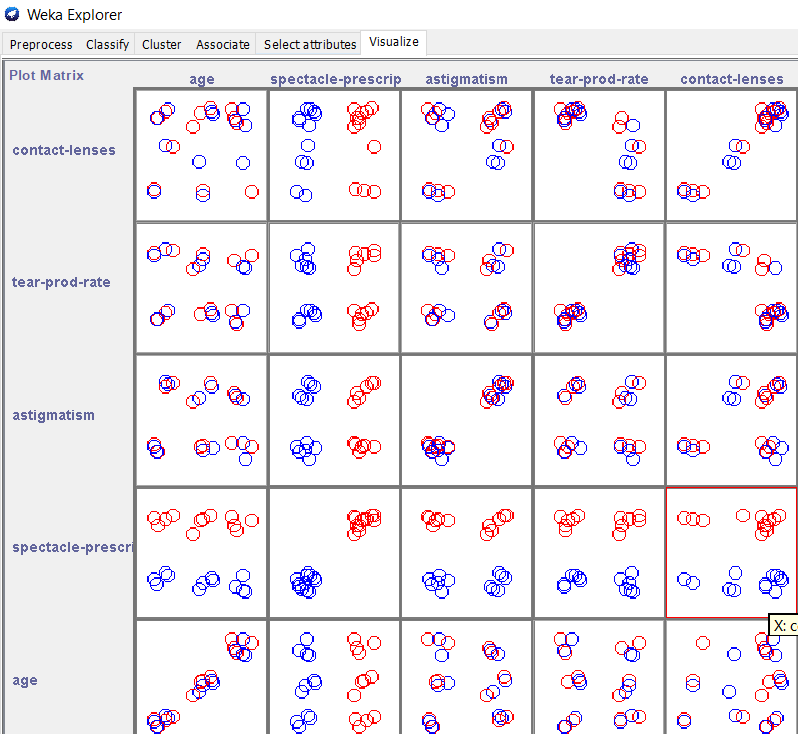
1. Spectacle-



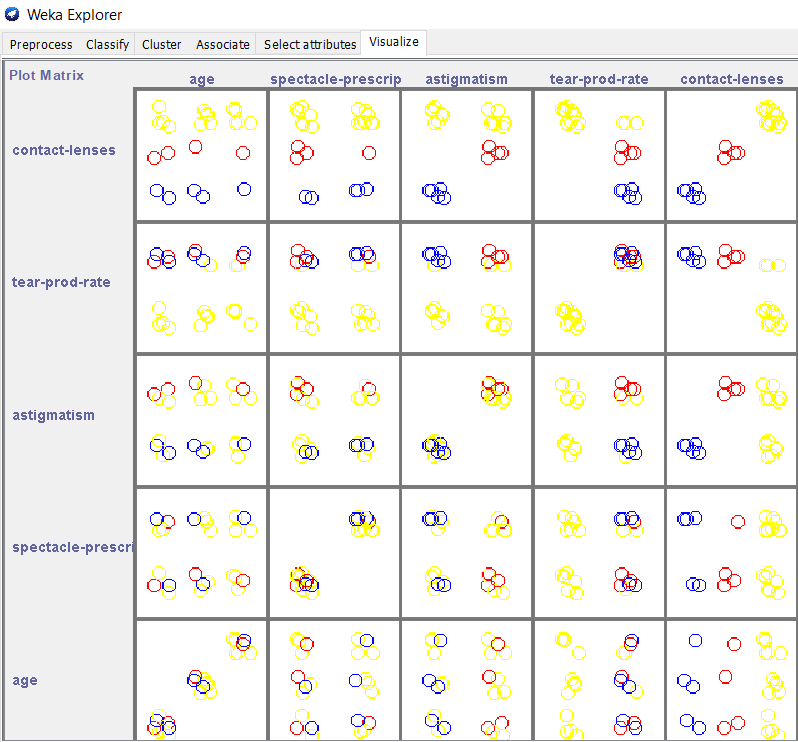
1. Astigmation-



1. Tear Rate Production-



1. Contact Lenses-



**Submitted By-**

Harshini M – RA1811003010216

Srihari Karthik – RA1811003010256

Raghu B – RA1811003010303