|  |  |
| --- | --- |
| Data MINING PROJECT  INVESTIGATION OF DIABETES DATA WITH WEKA | Nuriye Merve Tatlıdil  150212002 |

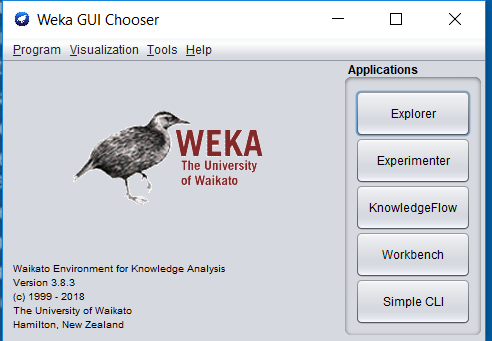
**What is Data Mining?**

Data mining is the process of accessing information from large-scale data and mining information. Or, in a sense, searching through the computer program the relations that may enable us to make predictions about the future from large heaps of data. This is; It includes a number of technical approaches, such as clustering, data summarization, analysis of changes, detection of deviations. It may be possible to see data mining as a series of statistical methods. However, data mining differs from traditional statistics in a few ways. The purpose of data mining is the elimination of qualitative models that can easily be converted into logical rules or visual presentations. In this context, data mining is human-centered, and sometimes the human-computer interface is combined. The data mining field also includes fundamentals such as statistics, machine information, databases, and high-performance computing.

**What is WEKA?**

Weka is an open-source program developed on the Java platform for developing a data mining application.

Weka has a completely modular design and is able to perform business intelligence , classification ,visualization and data mining operations on data sets with the features it contains.Files used in Weka .You must have ARFF files.



**Definition Of Project**

**What is Diabetes?**

Diabetes is a developing and lifelong disease in which the secretory gland in your body does not produce enough insulin hormone or the insulin hormone that is produced cannot be used effectively. As a result, a person can not use glucose that is passed from the foods they eat to blood glucose and blood sugar rises.

[Hypothesis](https://www.seslisozluk.net/hypothesis-nedir-ne-demek/) : An individual with no diabetes will have a blood glucose level of 120 mg / dl if he is fasting and not more than 140 mg / dl in toughness (two hours after eating). The blood glucose level measured in fasting or toughness above these values indicates the presence of diabetes.

Data Analysis

-There are 768 instances in my data set.

All persons in the data set are at least 21 years old female patients.

-There are 9 attributes in my data set.

1. Number of times pregnant

2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test

3. Diastolic blood pressure (mm Hg)

4. Triceps skin fold thickness (mm)

5. 2-Hour serum insulin (mu U/ml)

6. Body mass index (weight in kg/(height in m)^2)

7. Diabetes pedigree function

8. Age (years)

9. Class variable (0 or 1)

-Missing Attribute Values: None

-Relabeled values in attribute 'class'

From: 0 To: tested\_negative

From: 1 To: tested\_positive

-Attribute number: Mean: Standard Deviation:

1. 3.8 3.4

2. 120.9 32.0

3. 69.1 19.4

4. 20.5 16.0

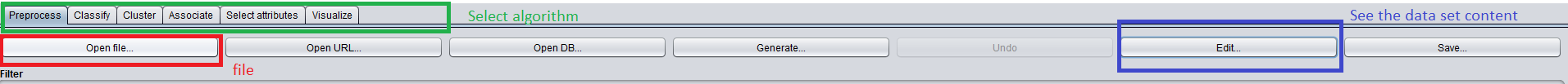
5. 79.8 115.2

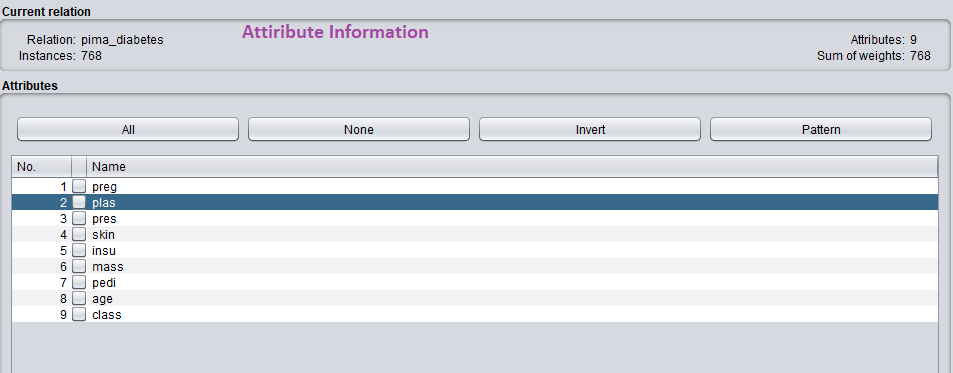
6. 32.0 7.9

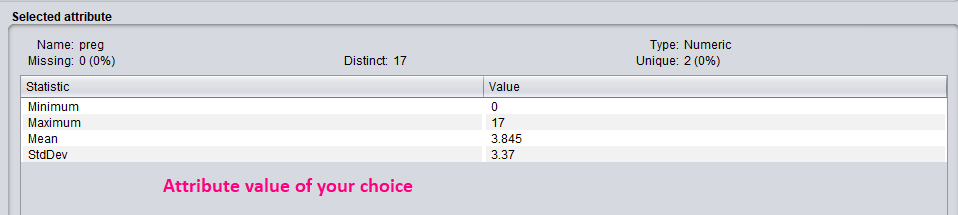
7. 0.5 0.3

8. 33.2 11.8

Introduction to WEKA





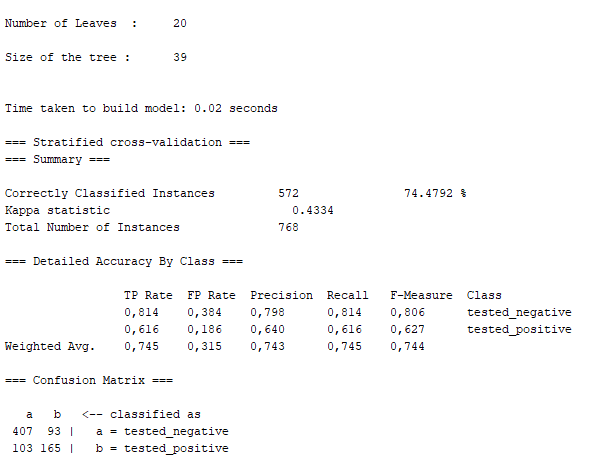


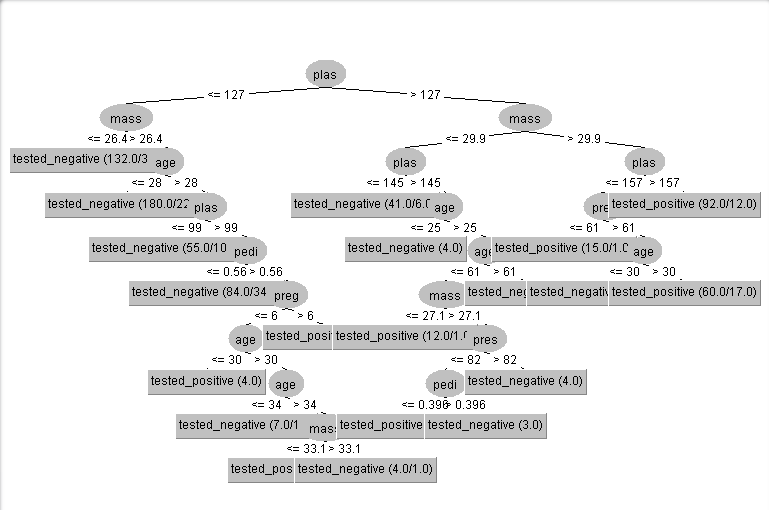
Algorithms to be used and their results

1-Decision Tree (J48)

1.State

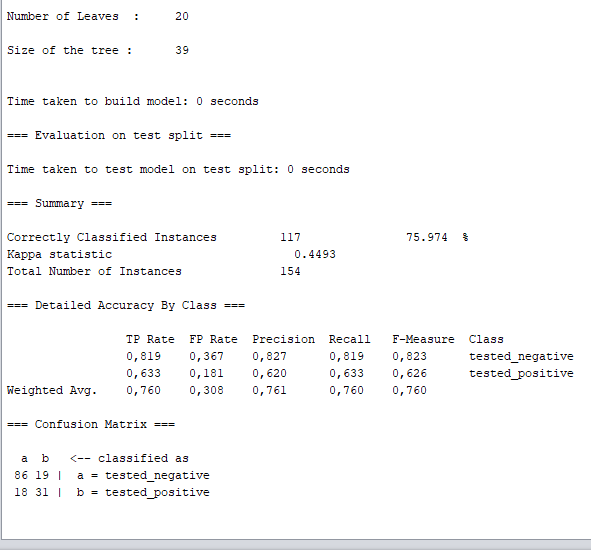
Cross-Validation 8 pieces 96 instances

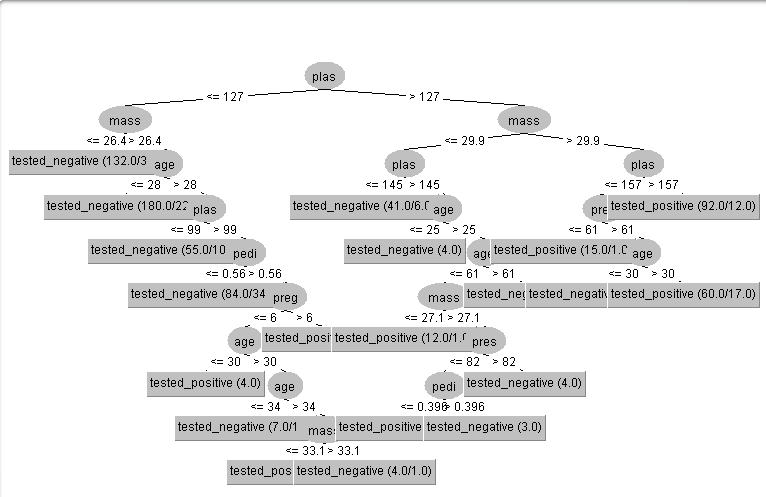




2.State

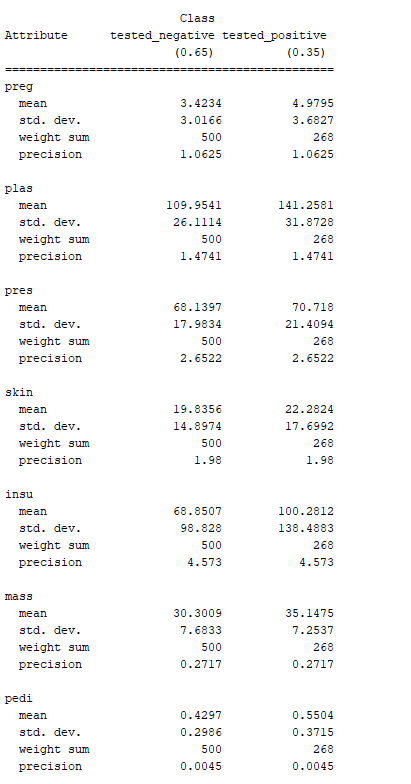
Percentage Split %80 : %20

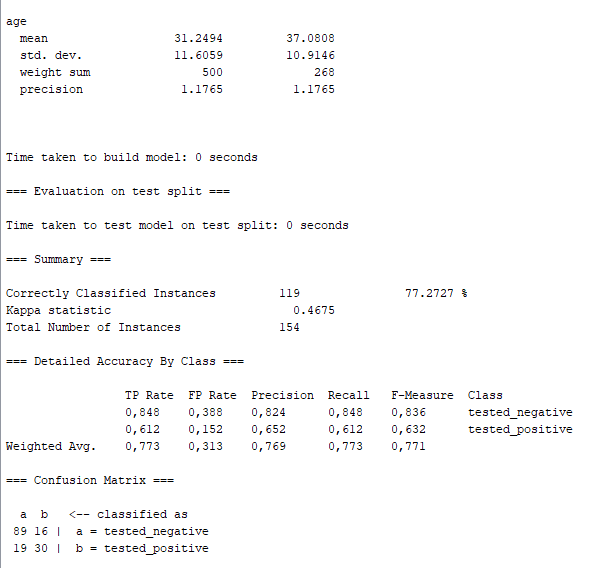




2-Bayes(Naïve Bayes)

Percentage Split %80 : %20





3-Clustering

Simple KMeans

Percentage Split %66 : %34

