**A REPORT ON**

**SUMMER INTERNSHIP**

**JUNE 2018**

**Submitted to: Submitted By:**

**Gaurav Bhatt Sir Ritesh Ranjan**

**Instructor Btech(CSE)**

**Raman Classes GEU**

**ACKNOWLEDGEMENT**

I would like to express my deepest appreciation to all those who provided me the possibility to complete this report.  A special gratitude I give to our project  manager, Dr. Ankush Mittal, whose contribution in stimulating suggestions and encouragement,  helped me to coordinate my project especially in writing this report.

**ABSTRACT**

This report consist of four different projects :

**1. Handwritten Digit Recognition of MNIST Dataset using SVM**

**2. Heart Disease Detection Using Neural Network in Orange**

**3. Breast Cancer Recognition Using Neural Network Using Orange**

**4. Soybean Disease Detection Using J48 Calssifier**

**5. Implementation of KNN (k- Nearest Neighbour) from scratch in python 2.7**

**6. Gender Recogniton From Faces Using Neural Network in Orange**

Primary purpose of first project is to find out the accuracy obtained in recognizing handwritten digits of MNIST Database using SVM (Support Vector Machine) which give a good accuracy.

Primary purpose of second project is to find out the maximum accuracy of the data set provided for Heart Disease Detection Using Neural Network in Orange.

Primary purpose of third project is to find out the maximum accuracy of the data set provided for Breast Cancer Recognition Using Neural Network Using Orange.

Primary purpose of fourth project is to find out the maximum accuracy of the data set provided for Soyabean Diseases Detection. The tool used is Weka 3.8 and method used is Decision Tree (J48) under trees classifier. In this project, maximum possible accuracy is determined by using feature selection on some different sets of attributes.

Primary purpose of fifth project is to implement a well know supervised machine learning tool known as KNN(K Nearest Neighbour) from scratch.

Primary purpose of sixth project is to recognize gender from a well known wiki dataset using neural network.

**Table of Content**

**Content**

* Acknowledgement
* Abstract
* Introduction
* Table of content
* Handwritten Digit Recognition of MNIST Dataset using SVM

1. Dataset and Attribute and Class
2. Techniques used and explanation
3. Result
4. Conclusion and Future Work

* Heart Disease Detection Using Neural Network in Orange

1. Dataset and Attribute and Class
2. Techniques used and explanation
3. Result
4. Conclusion and Future Work

* Breast Cancer Recognition Using Neural Network Using Orange

1. Dataset and Attribute and Class
2. Techniques used and explanation
3. Result
4. Conclusion and Future Work

* Soybean Disease Detection Using J48 Calssifier

1. Dataset and Attribute and Class
2. Techniques used and explanation
3. Result
4. Conclusion and Future Work

* References

**Handwritten Digit Recognition of MNIST Dataset using SVM**

**MNIST Dataset**

The **MNIST database** (Modified [National Institute of Standards and Technology](https://en.wikipedia.org/wiki/National_Institute_of_Standards_and_Technology) database) is a large [database](https://en.wikipedia.org/wiki/Database) of handwritten digits that is commonly used for [training](https://en.wikipedia.org/wiki/Training_set) various [image processing](https://en.wikipedia.org/wiki/Image_processing) systems. The dataset we used is not the original one with 60,000 images but a preprocessed dataset (with 64\*64 size) images built in the sklearn Library with over 1797 entries. We have used 1700 images for testing purpose and remaining (97) for testing purpose.

**Support Vector Machine (SVM)**

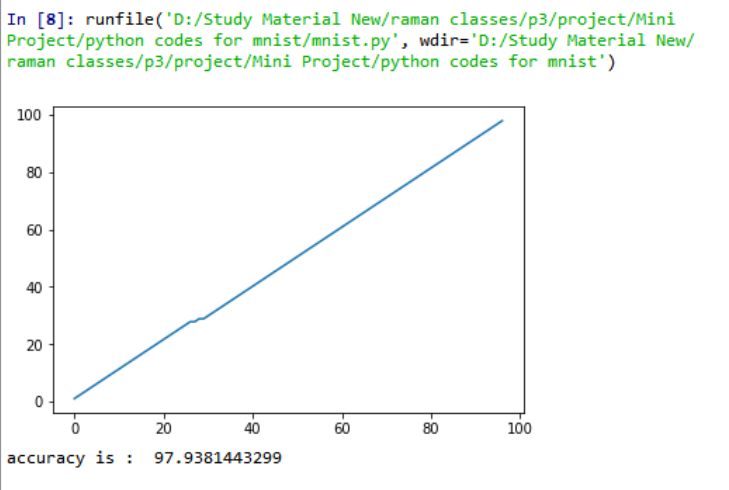
In [machine learning](https://en.wikipedia.org/wiki/Machine_learning), **support vector machines** (**SVMs**, also **support vector networks**) are [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning) models with associated learning [algorithms](https://en.wikipedia.org/wiki/Algorithm) that analyze data used for [classification](https://en.wikipedia.org/wiki/Statistical_classification) and [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis). Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-[probabilistic](https://en.wikipedia.org/wiki/Probabilistic_classification) [binary](https://en.wikipedia.org/wiki/Binary_classifier) [linear classifier](https://en.wikipedia.org/wiki/Linear_classifier) (although methods such as [Platt scaling](https://en.wikipedia.org/wiki/Platt_scaling) exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.



We have used gamma as 0.0001 and C as 100 and RBF (Radial Basis Function) as kernel.

**Result and Accuracy**

With SVM (SVC(gamma = 0.0001, C=100))), we get an acuuracy of **97.9381443299 %** below is an screenshot of result and graph obtained.

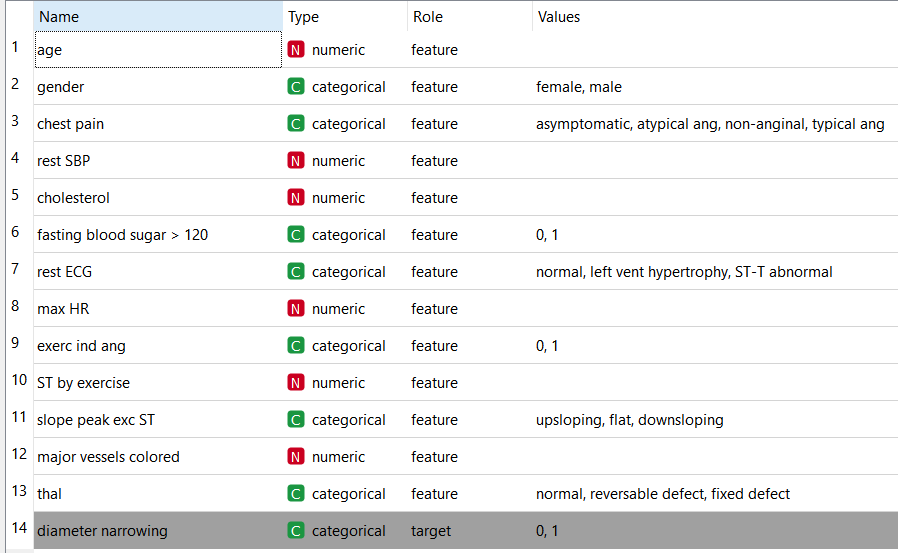


**Conclusion and Future Work :** The above project can be used in many purposes, like recognizing digit in captchas, identifying digits in unreadable texts etc.

**Heart Disease Detection Using Neural Network in Orange**

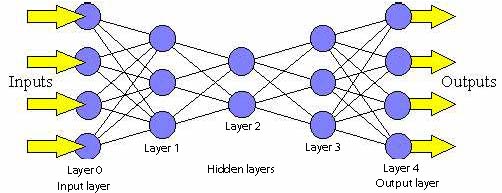
**Dataset**

Coronary artery disease (CAD), also called coronary heart disease (CHD), or simply, heart disease, is a narrowing of the blood vessels (coronary arteries) that supply oxygen and blood to the heart. In this dataset our aim is to detect diameter narrowing as 2 or 4 which will give clear idea about presence of heart disease.

****

**Neural Network**

A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. Neural networks can adapt to changing input so the network generates the best possible result without needing to redesign the output criteria.

****

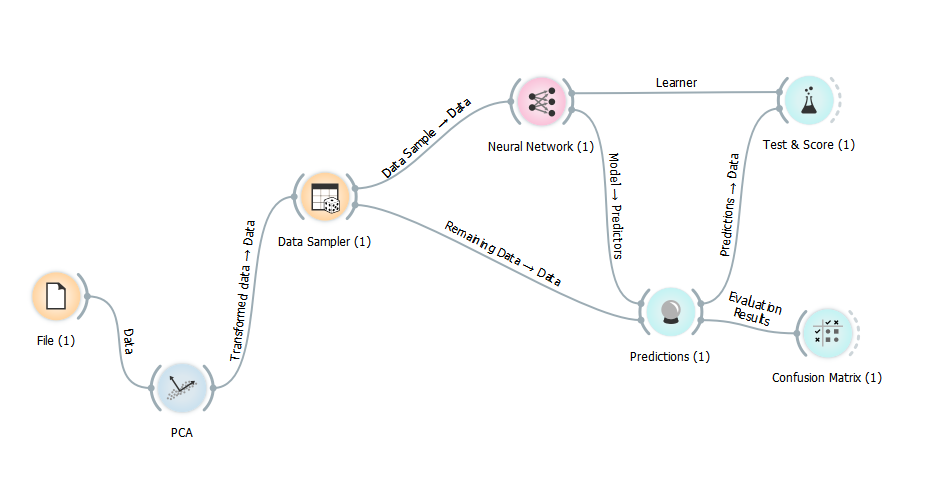
**Neurons per hidden layer : 100**

**Activation Function : Rectified Linear Unit (ReLu)**

**Alpha: 0.00001**

**Number of Iterarions : 50**

**Flow Chart**

****

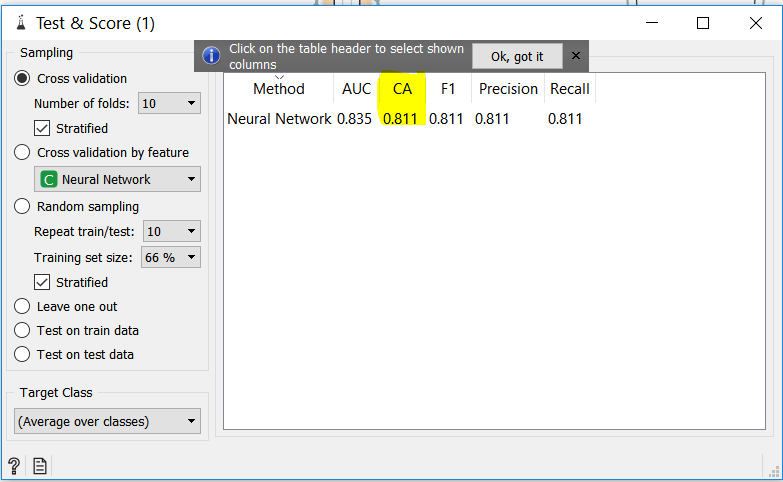
**TOOL USED**

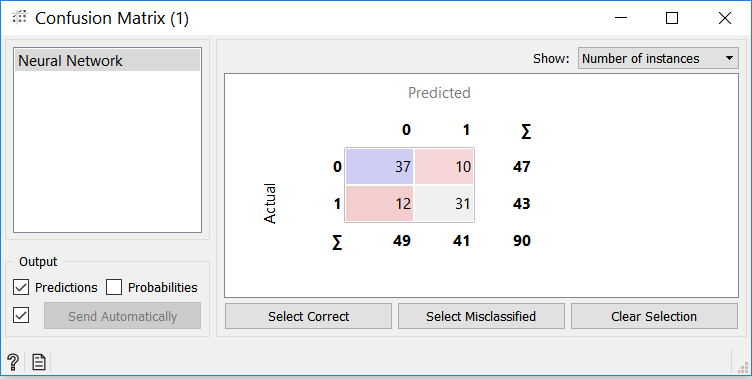
**ORANGE Toolbox** : Open source machine learning and data visualization for novice and expert. Interactive data analysis workflows with a large toolbox.



**Result and Accuracy**

The accuracy obtained with above attributes and parameters is **81.1%.** Below is the screenshot of it.

****

****

**Conclusion and Future Work :** The above project can be used to detect heart disease at an early stage so that one must be careful from starting and can be used to save lives.

**Breast Cancer Recognition Using Neural Network Using Orange**

**Dataset**

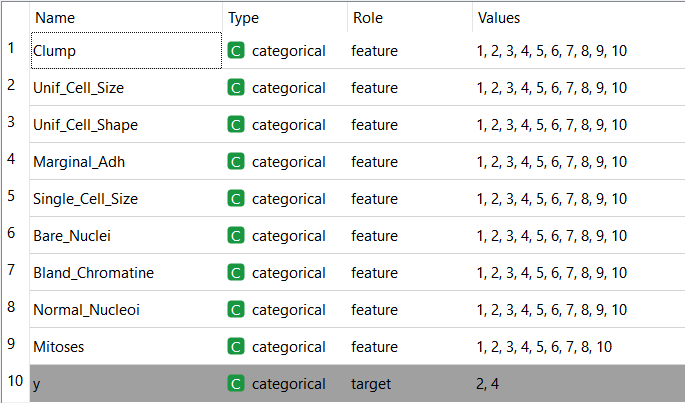
We have used Breast Cancer Wisconsin Data Set (available) on **UCI Machine Learning Repository** to detect breast cancer.

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. A few of the images can be found at <http://pages.cs.wisc.edu/~street/images/>

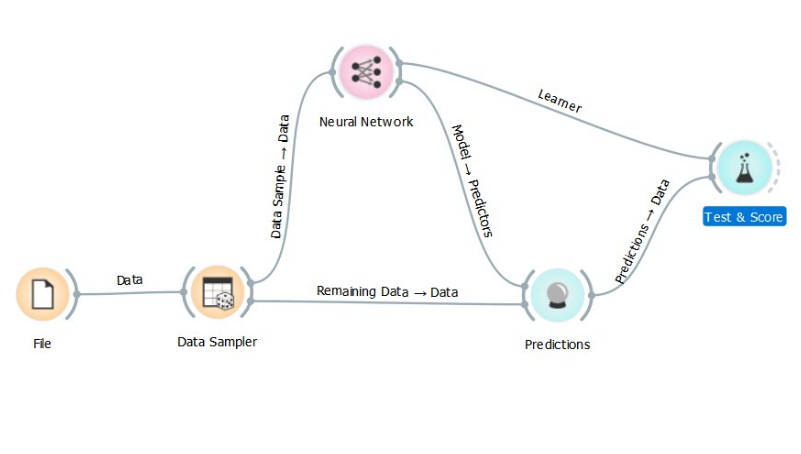
Number of instances: 683

Number of attributes: 9

Attribute information:

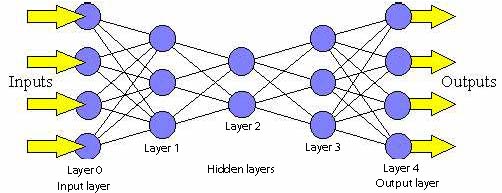


**Flow Chart**



**Neural Network**

A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. Neural networks can adapt to changing input so the network generates the best possible result without needing to redesign the output criteria.



**Neurons per hidden layer : 100**

**Activation Function : Rectified Linear Unit (ReLu)**

**Alpha: 0.00001**

**Number of Iterarions : 50**

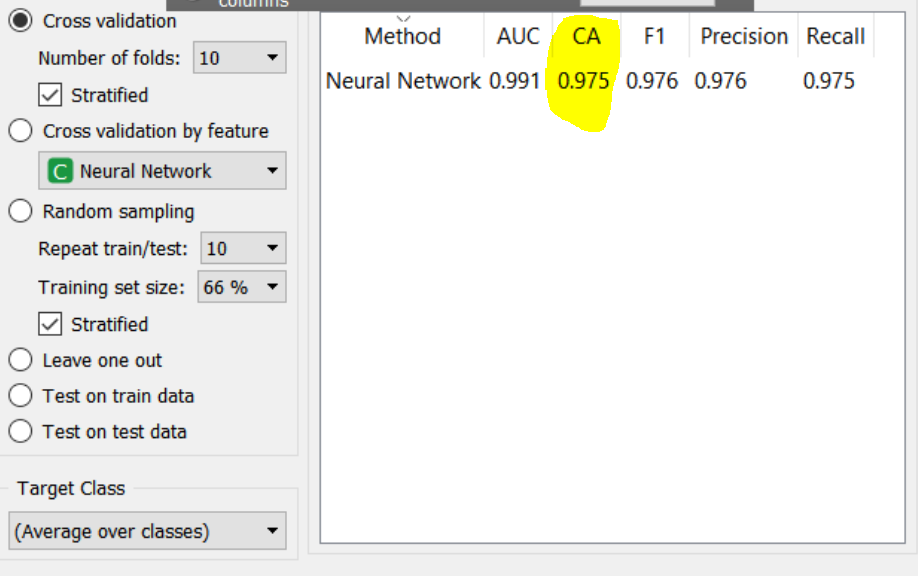
**TOOL USED**

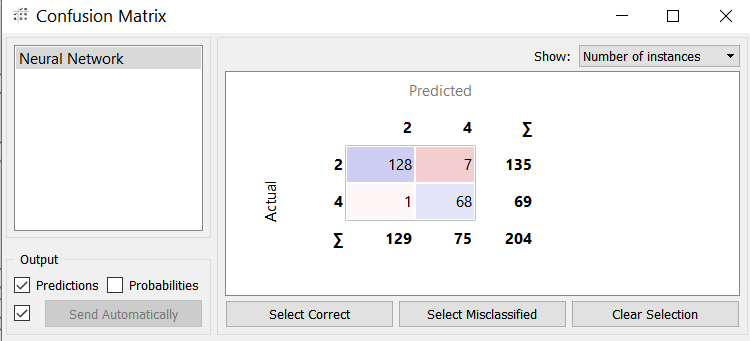
**ORANGE** Toolbox : Open source machine learning and data visualization for novice and expert. Interactive data analysis workflows with a large toolbox.



**Result and Accuracy :**

The accuracy obtained with above attributes and parameters is **97.5%.** Below is the screenshot of it.





**Conclusion and Future Work :** The above project can be used to detect Breast Cancer disease at an early stage so that one must be careful from starting and can be used to save lives.

**Soybean Disease Detection Using J48 Calssifier**

**Dataset:**

19 Classes

( diaporthe stem canker, charcoal rot, rhizoctonia root rot,

phytophthora rot, brown stem rot, powdery mildew,

downy mildew, brown spot, bacterial blight,

bacterial pustule, purple seed stain, anthracnose,

phyllosticta leaf spot, alternarialeaf spot,

frog eye leaf spot, diaporthe pod & stem blight,

cyst nematode, 2 4 d injury, herbicide injury. )

1. date: april, may, june, july, august, september, october.

2. plant stand: normal, lt normal .

3. precip: lt norm, norm, gt norm .

4. temp: lt norm, norm, gt norm .

5. hail: yes, no .

6. crop hist: diff lst year, same lst yr, same lst two yrs,

same lst sev yrs .

7. area damaged: scattered, low areas, upper areas, whole field .

8. severity: minor, pot severe, severe .

9. seed tmt: none, fungicide, other .

10. germination: 90%, 100%, 80 89%, lt 80% .

11. plant growth: norm, abnorm .

12. leaves: norm, abnorm.

13. leafspots halo: absent, yellow halos , no yellow halos .

14. leafspots marg: w s marg, no w s marg, dna .

15. leafspot size: lt 1/8, gt 1/8,dna .

16. leaf shread: absent, present .

17. leaf malf: absent, present .

18. leaf mild: absent, upper surf, lower surf .

19. stem: norm, abnorm .

20. lodging: yes, no .

21. stem cankers: absent, below soil, above soil, above sec nde .

22. canker lesion: dna, brown, dk brown blk,tan .

23. fruiting bodies: absent, present .

24. external decay: absent, firm and dry, watery .

25. mycelium: absent, present .

26. int discolor: none, brown, black .

27. sclerotia: absent, present .

28. fruit pods: norm ,diseased, few present, dna .

29. fruit spots: absent,colored, brown w/blk specks,distort,dna.

30. seed: norm, abnorm .

31. mold growth: absent, present .

32. seed discolor: absent, present .

33. seed size: norm, lt norm .

34. shriveling: absent, present .

35. roots: norm, rotted, galls cysts .

**Class Distribution:**

1. diaporthe stem canker: 10

2. charcoal rot: 10

3. rhizoctonia root rot: 10

4. phytophthora rot: 40

5. brown stem rot: 20

6. powdery mildew: 10

7. downy mildew: 10

8. brown spot: 40

9. bacterial blight: 10

10. bacterial pustule: 10

11. purple seed stain: 10

12. anthracnose: 20

13. phyllosticta leaf spot: 10

14. alternarialeaf spot: 40

15. frog eye leaf spot: 40

16. diaporthe pod & stem blight: 6

17. cyst nematode: 6

18. 2 4 d injury: 1

19. herbicide injury: 4

**Experimental Techniques and Tool Used**

The tool we used is **WEKA 3.8**.

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre processing, classification, regression, clustering, association rules, and visualization. It is also well suited for developing new machine learning schemes.

Experimental Technique used is **J48 –C 0.25 –M 2**.

J48 is an extension of ID3. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. In the WEKA data mining tool, J48 is an open source Java implementation of the **C4.5** **algorithm**.

What is **J48 decision tree**?

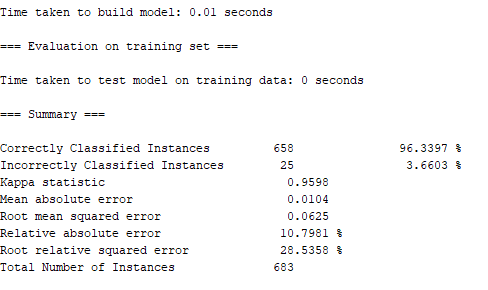
Imagine that you have a dataset with a list of predictors or independent variables and a list of targets or dependent variables. Then, by applying a decision tree like J48 on that dataset would allow you to predict the target variable of a new dataset record.

**Result and Conclusion**

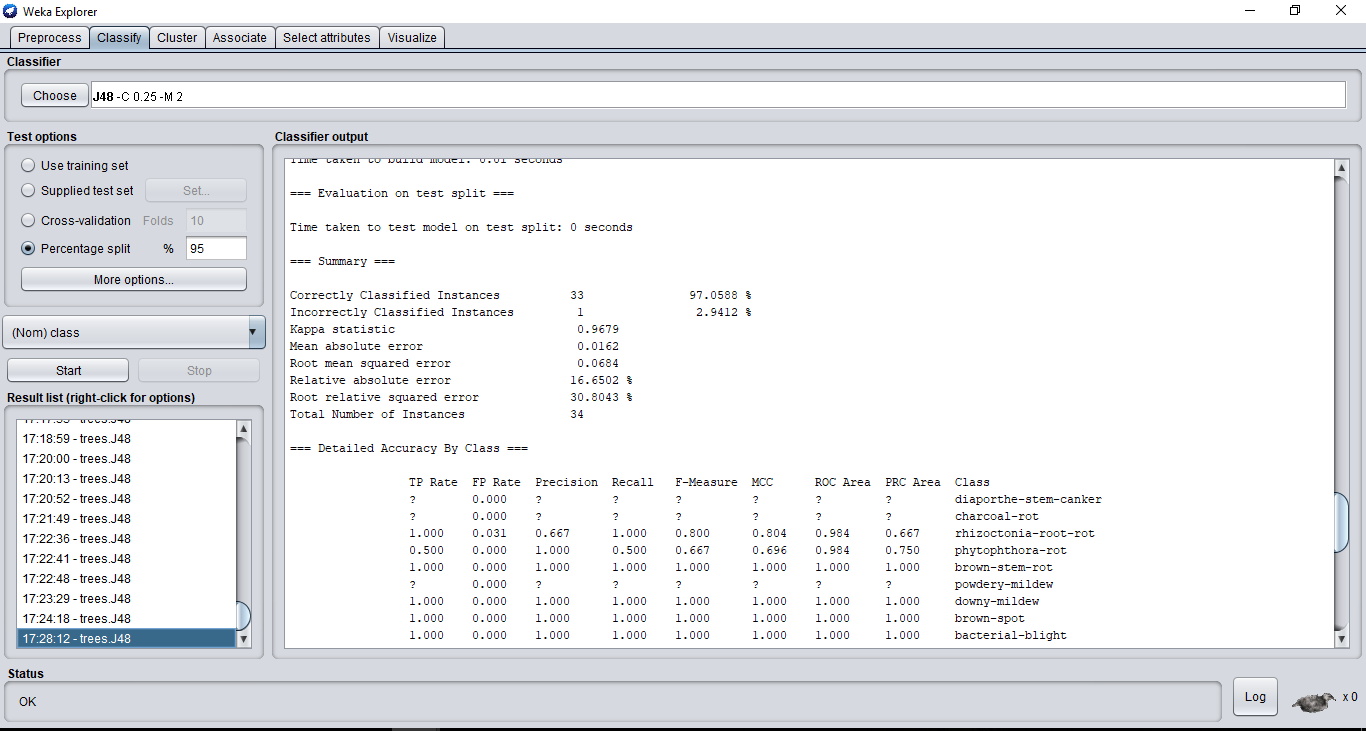
We got some amazing results by changing the training size of datasets and we have also reduced the number of attributes from **35** to **15** most influential ones and increased its accuracy from **96.3397** to **97.0588**, using the information related to soybean diseases.

Followings are the screenshots of the result of our experiments and the previous original result.

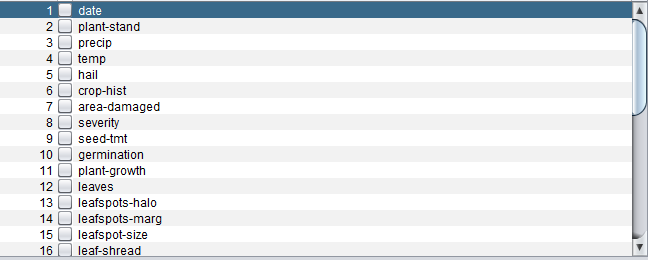
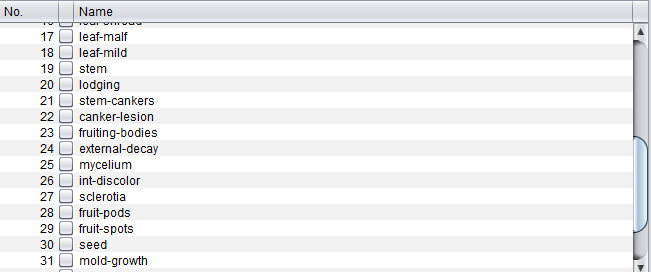
**Previous Results Accuracy**

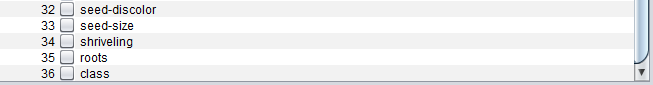
****

**Final Accuracy**

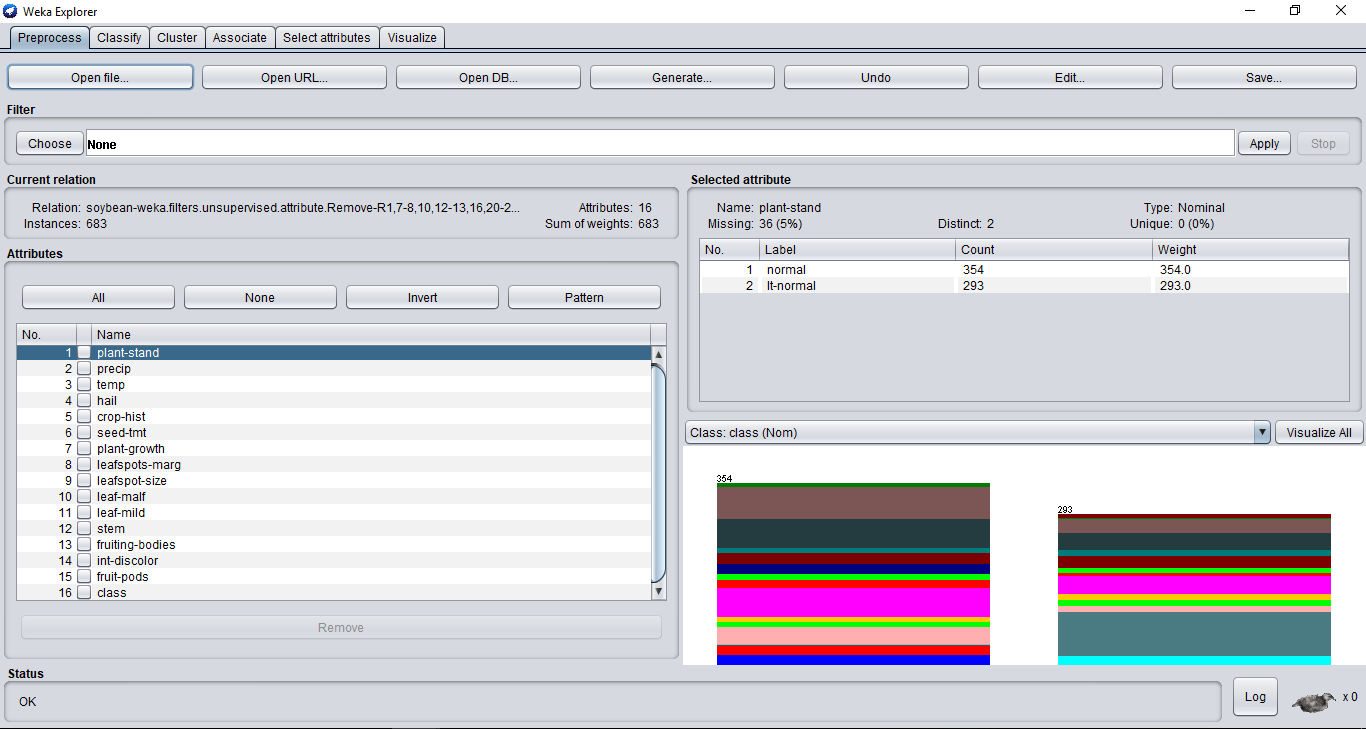
****

**Previous number of Attributes**

****

****

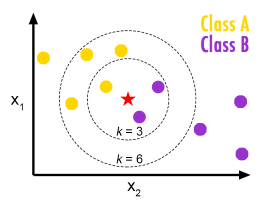
**Final Attributes**

****

**Implementation of KNN (k- Nearest Neighbour) from scratch in python 2.7**

k-nearest neighbors algorithm (k-NN) is a method used for [classification](https://en.wikipedia.org/wiki/Statistical_classification) and [regression](https://en.wikipedia.org/wiki/Regression_analysis).In both cases, the input consists of the k closest training examples in the [feature space](https://en.wikipedia.org/wiki/Feature_space). The output depends on whether k-NN is used for classification or regression:

* In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its  k nearest neighbors (k is a positive [integer](https://en.wikipedia.org/wiki/Integer), typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.
* In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors.

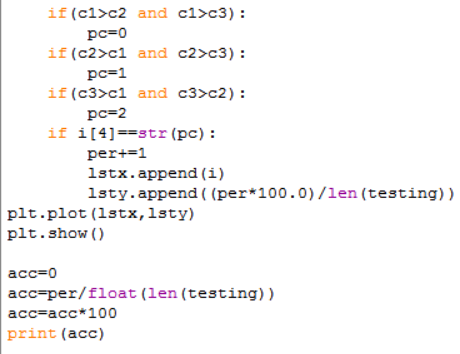


**Dataset**

The data set consists of 50 samples from each of three species of *Iris* ([*Iris setosa*](https://en.wikipedia.org/wiki/Iris_setosa), [*Iris virginica*](https://en.wikipedia.org/wiki/Iris_virginica) and [*Iris versicolor*](https://en.wikipedia.org/wiki/Iris_versicolor)). Four [features](https://en.wikipedia.org/wiki/Features_(pattern_recognition)) were measured from each sample: the length and the width of the [sepals](https://en.wikipedia.org/wiki/Sepal) and [petals](https://en.wikipedia.org/wiki/Petal), in centimeters. Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other.

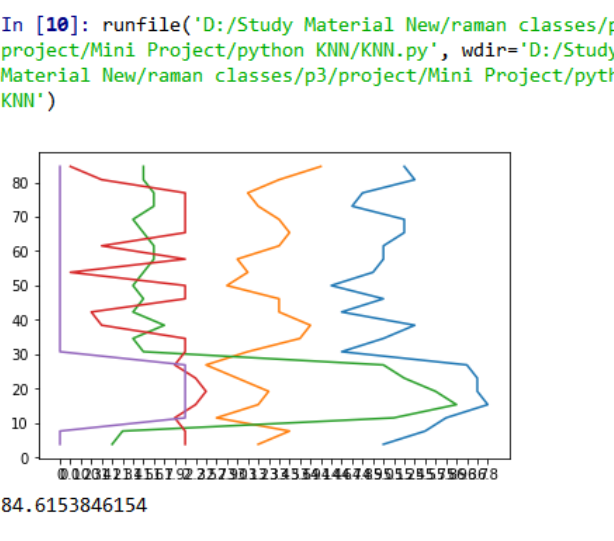
**Python Code:**





**Result and Accuracy**

We have got an accuracy of **84.615384%**



**Gender Recognition Using Neural Network in Orange**

**Dataset :**

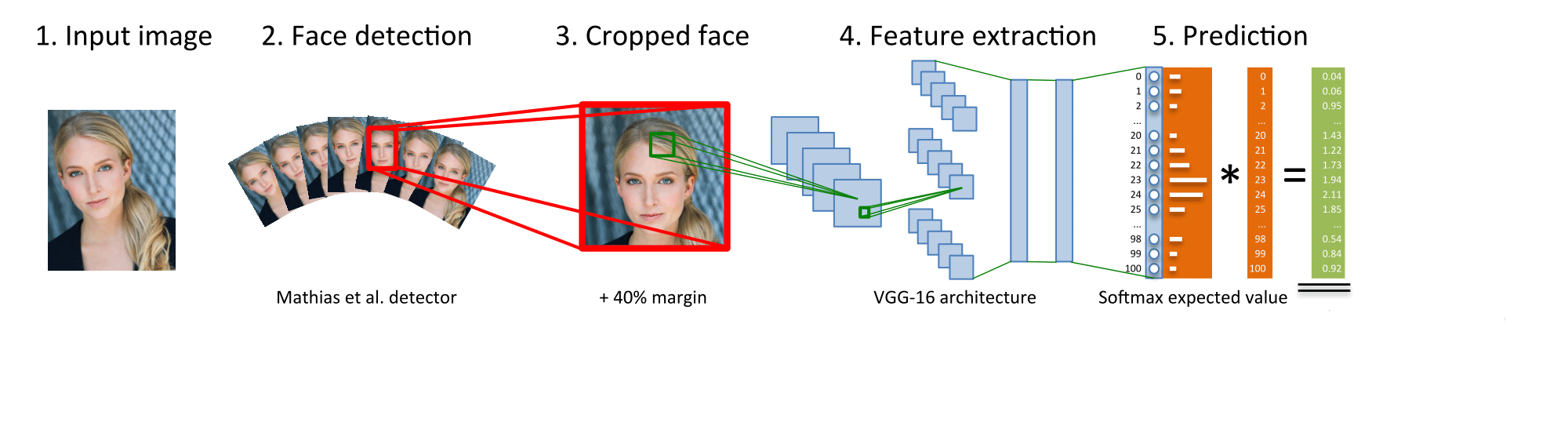
To the best of our knowledge this is the largest publicly available dataset of face images with gender and age labels for training. We provide pretrained models for

Grnder prediction.

Since the publicly available face image datasets are often of small to medium size, rarely exceeding tens of thousands of images, and often without age information we decided to collect a large dataset of celebrities. For this purpose, we took the list of the most popular 100,000 actors as listed on the IMDb website and (automatically) crawled from their profiles date of birth, name, gender and all images related to that person.

**NEURAL NETWORK :**

A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. Neural networks can adapt to changing input so the network generates the best possible result without needing to redesign the output criteria.



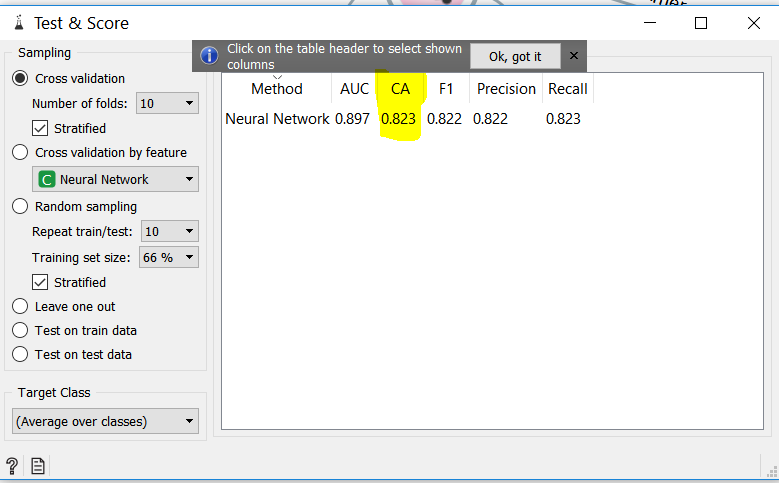
**TOOL USED**

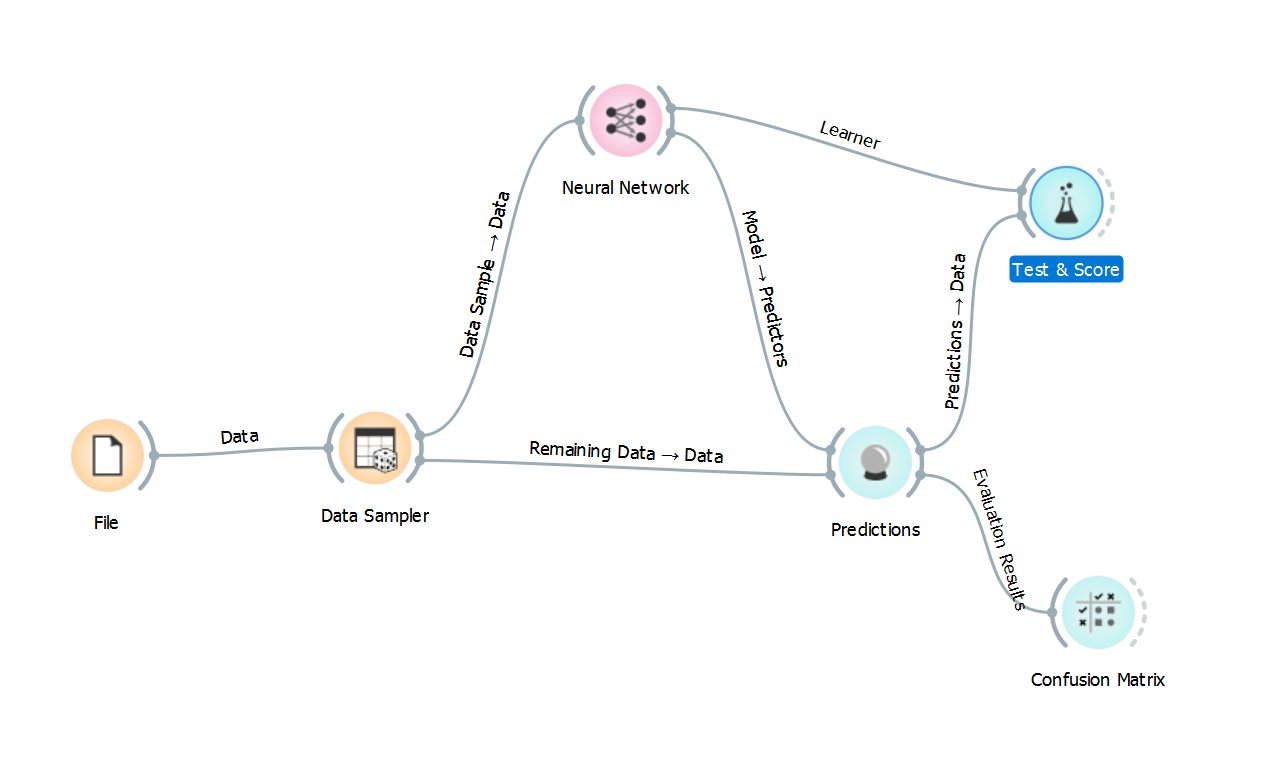
**ORANGE Toolbox** : Open source machine learning and data visualization for novice and expert. Interactive data analysis workflows with a large toolbox.

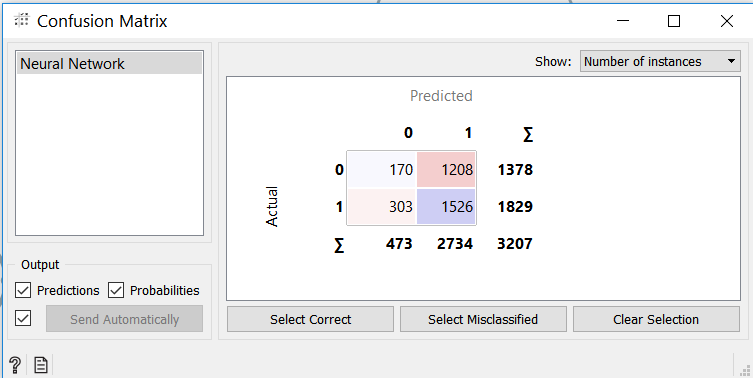


**Result and Accuracy :**

The accuracy obtained with above attributes and parameters is **97.5%.** Below is the screenshot of it.







**Conclusion and Future Work :** The above project can be used to detect gender which is very useful in investigating purposes.

**REFERENCES**

<https://archive.ics.uci.edu/ml/index.php> (UCI Machine Learning Repository)

<http://scikit-learn.org/stable/> (ScikitLearn Library)

<https://www.kaggle.com/>

<https://orange.biolab.si/> (Orange Software)