**PROJECT REPORT ON:**

**DERMATOLOGY**

Submitted By:

ANJANA AJIKUMAR

Reg No:200011024266

**MSC DATA ANALYTICS**

****

**BASELIUS COLLEGE, KOTTAYAM**

**MAHATMA GANDHI UNIVERSITY**

**INTRODUCTION**

The differential diagnosis of erythemato-squamous diseases is a real problem in dermatology. They all share the clinical features of erythema and scaling, with very little differences. The diseases in this group are psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, and pityriasis rubra pilaris. Usually a biopsy is necessary for the diagnosis but unfortunately these diseases share many histopathological features as well. Another difficulty for the differential diagnosis is that a disease may show the features of another disease at the beginning stage and may have the characteristic features at the following stages. Patients were first evaluated clinically with 12 features. Afterwards, skin samples were taken for the evaluation of 22 histopathological features. The values of the histopathological features are determined by an analysis of the samples under a microscope.

In the dataset constructed for this domain, the family history feature has the value 1 if any of these diseases has been observed in the family, and 0 otherwise. The age feature simply represents the age of the patient. Every other feature (clinical and histopathological) was given a degree in the range of 0 to 3. Here, 0 indicates that the feature was not present, 3 indicates the largest amount possible, and 1, 2 indicate the relative intermediate values.

**Attribute Information:**

Clinical Attributes: (take values 0, 1, 2, 3, unless otherwise indicated)

1: erythema

2: scaling

3: definite borders

4: itching

5: koebner phenomenon

6: polygonal papules

7: follicular papules

8: oral mucosal involvement

9: knee and elbow involvement

10: scalp involvement

11: family history, (0 or 1)

34: Age (linear)

Histopathological Attributes: (take values 0, 1, 2, 3)

12: melanin incontinence

13: eosinophils in the infiltrate

14: PNL infiltrate

15: fibrosis of the papillary dermis

16: exocytosis

17: acanthosis

18: hyperkeratosis

19: parakeratosis

20: clubbing of the rete ridges

21: elongation of the rete ridges

22: thinning of the suprapapillary epidermis

23: spongiform pustule

24: munromicroabcess

25: focal hypergranulosis

26: disappearance of the granular layer

27: vacuolisation and damage of basal layer

28: spongiosis

29: saw-tooth appearance of retes

30: follicular horn plug

31: perifollicular parakeratosis

32: inflammatory monoluclearinflitrate

33: band-like infilterate

Class :

* 1: psoriasis
* 2: seboreic dermatitis
* 3: lichen planus
* 4: pityriasis rosea
* 5: cronic dermatitis
* 6: pityriasis rubra pilaris

**This dataset contains instances of dermatology cancer occurrence**

So, for that we can use the Machine Learning Techniques. Machine Learning Services provide an extensible, scalable platform for integrating machine learning tasks and tools with the applications that consume machine learning services. There are many Classification Algorithms that can be used for this purpose. The Machine Learning Algorithms that I used here are:

➢ Logistic Regression

➢ Random Forest Classifier

➢ Support Vector Machine (SVM)

➢ K Nearest Neighbor Classifier (KNN)

➢XGBoost/Gradient Boosting

This Classification project is divided into a few number of stepsThey are:

➢ Loading necessary libraries

➢ Loading Dataset from a CSV file

➢ Exploratory Data Analysis (EDA) used to understand the dataset by using graphs and other

Visualization tools.

➢ Data pre-processing

➢ Parameter selection for the classification models.

➢ Splitting of data into training and testing data set.

➢ Applying different Classification algorithms on the training dataset.

➢ Evaluating the performance and accuracy of the fitted model using evaluation metrics like

confusion matrix, precision, recall, accuracy score

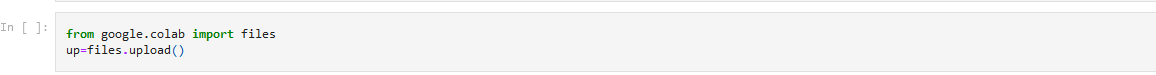
**DATA**

The source of the dataset isDatahub. This dataset consists of integer and float data types. This dataset consists of 35 columns and 366rows of data points. For the classification problem we consider the last column that is “class” as the classification column and the other columns as the independent columns or the columns which supports the classification.

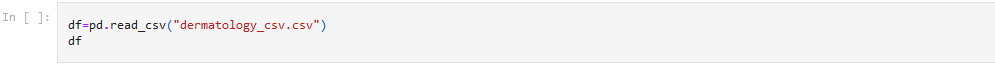
<https://datahub.io/machine-learning/dermatology/r/0.html>

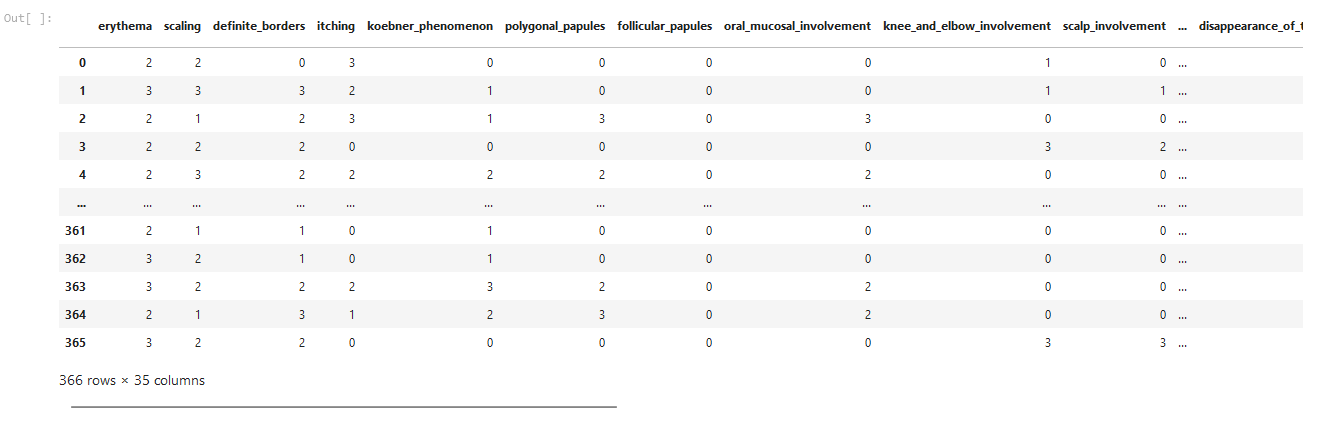
**IMPORTING REQUIRED MODULES**

****

****

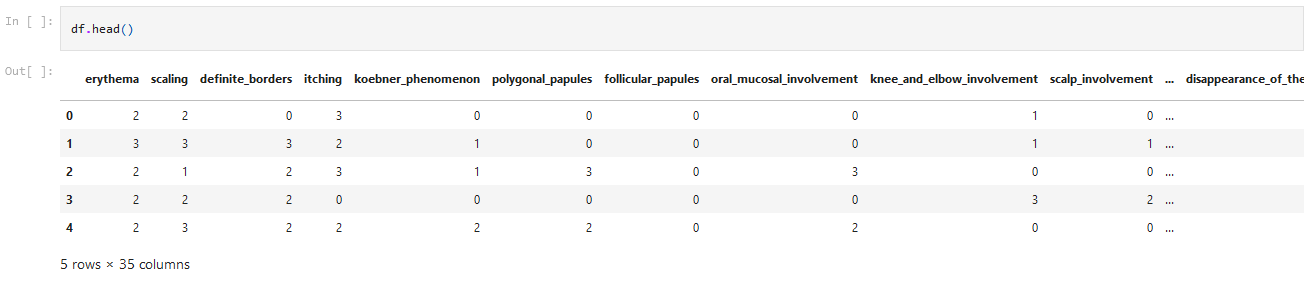
**READING THE DATASET**

****

****

**EDA on the dataset**

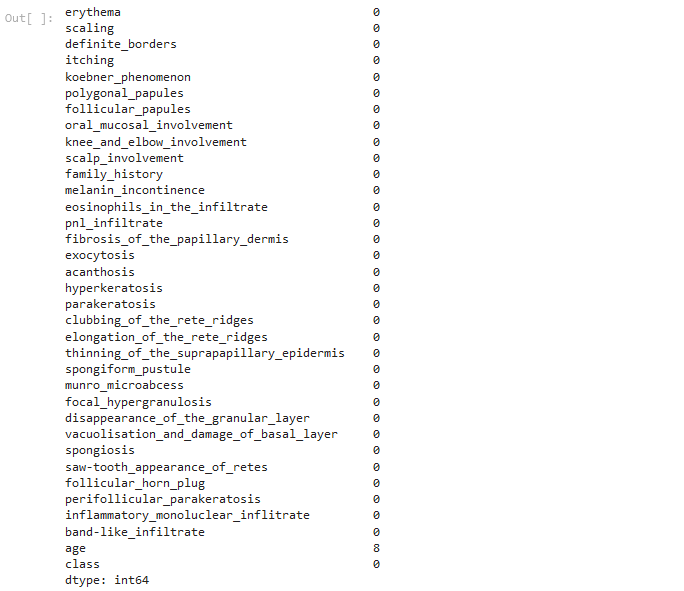
This is a process by which we get a detailed information about the data set. This process includes Descriptive statistical methods like Describe(),info(),Data visualization techniques, correlation etc

****

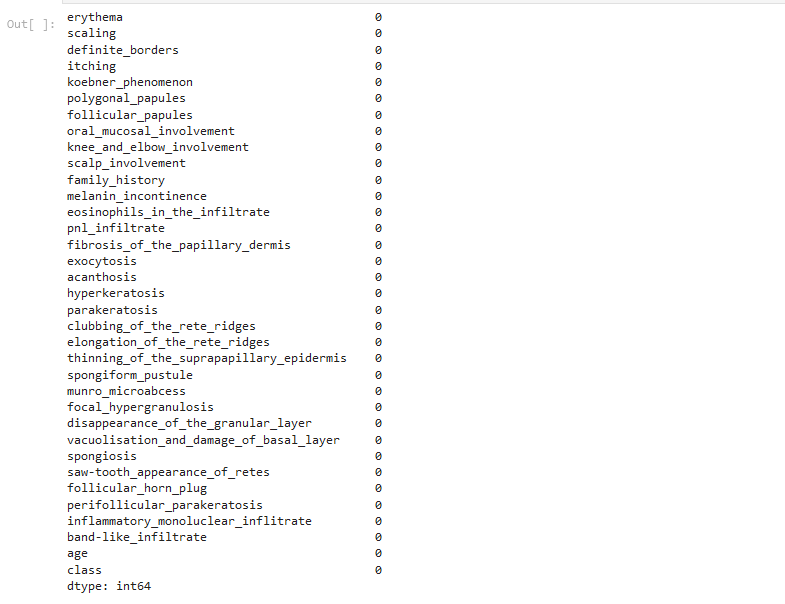
**Checking for null value**

In this step we check whether there is any NULL values present in the data set.

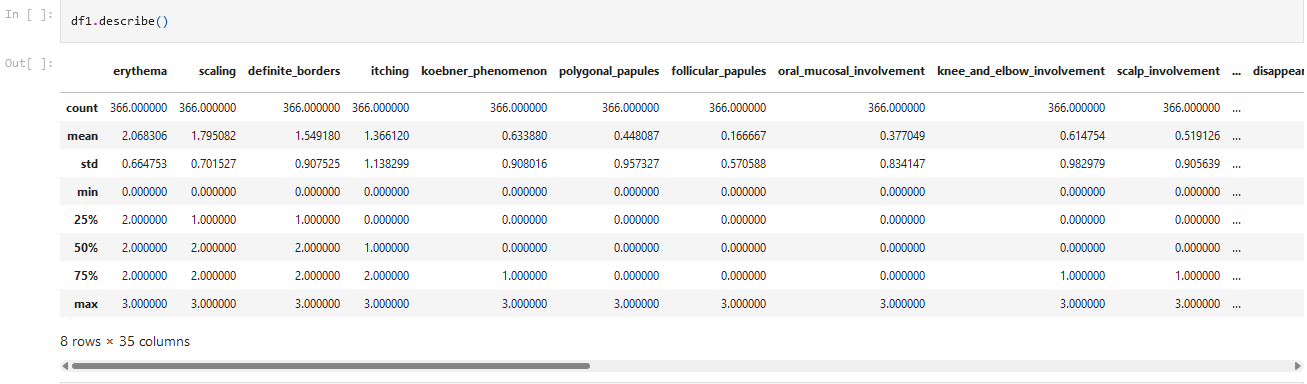
****

****

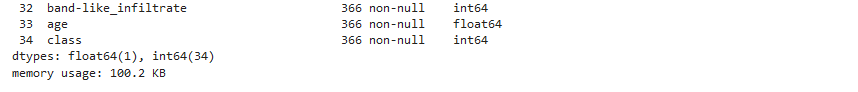
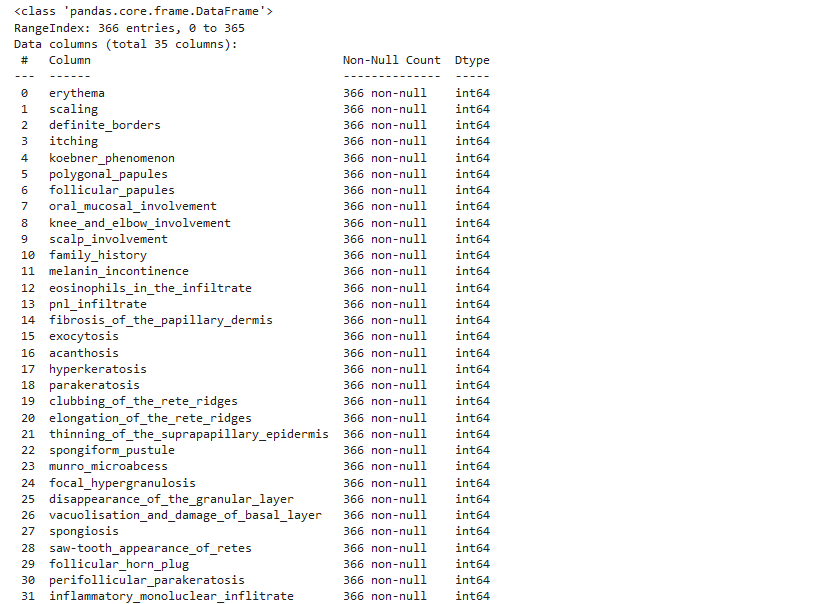
****

****

****

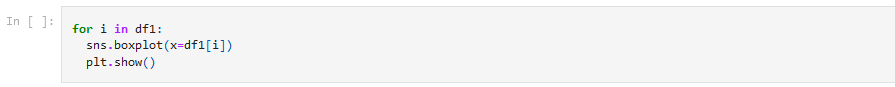
****

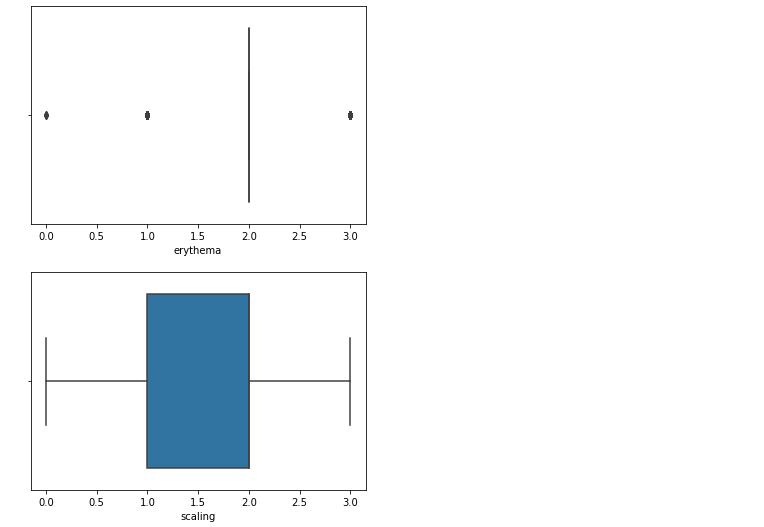
****

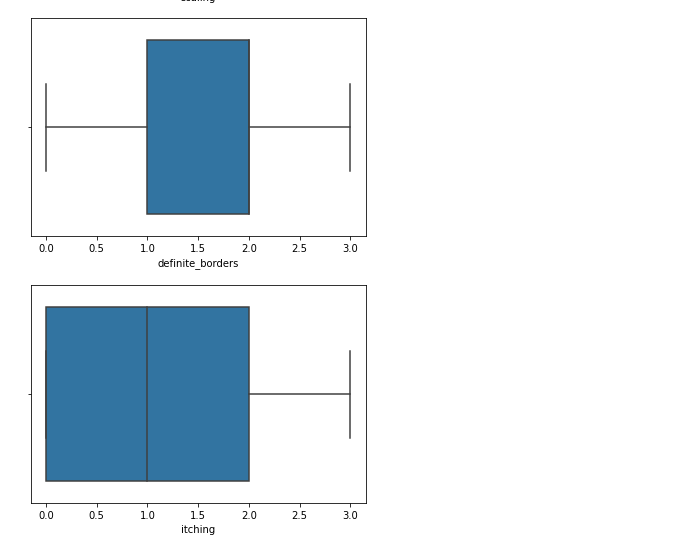
****

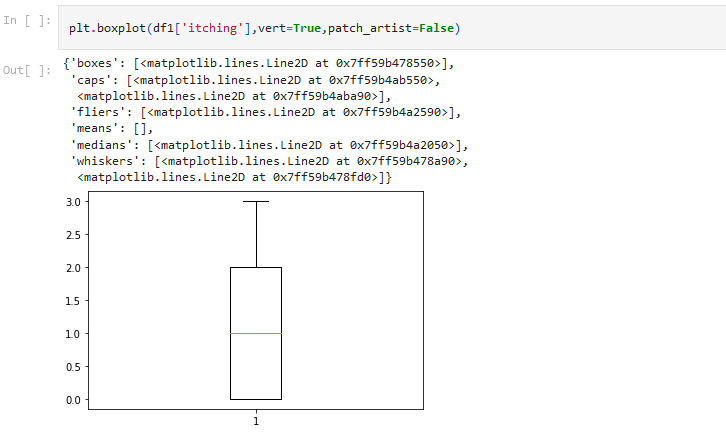
**Using Boxplot DerminingThe Outliers**

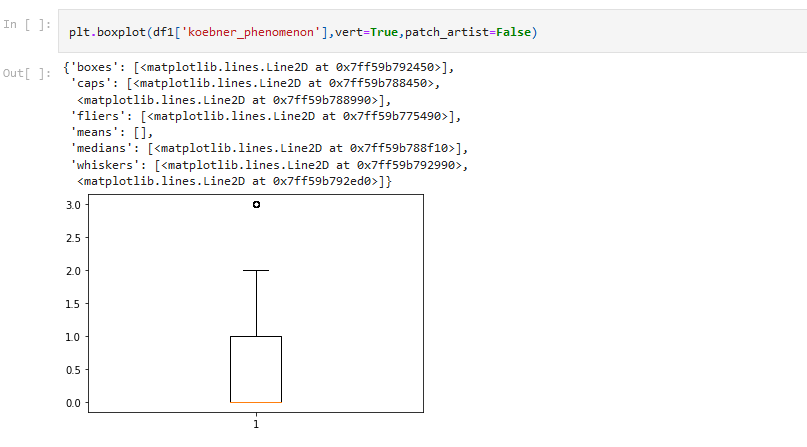
The main features that the boxplot shows the median of the data, which represents where the middle datapoint is. Box plot shows the range of values of different numerical variables. Finally, boxplots also display “outliers” as individual dots that occur outside the upper and lower extremes. With boxplots, you can easily spot outliers.

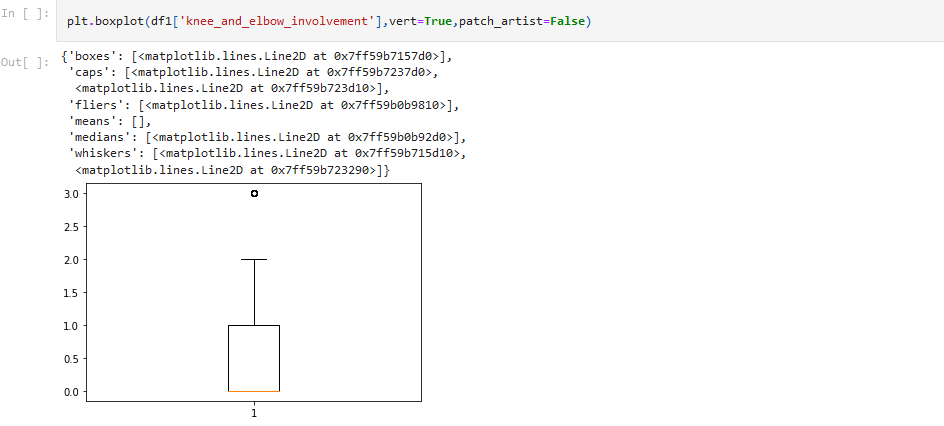
****

****

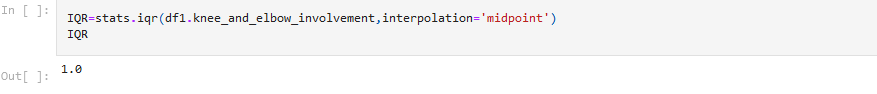
****

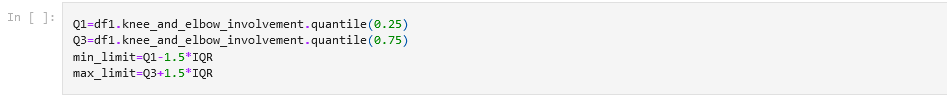
****

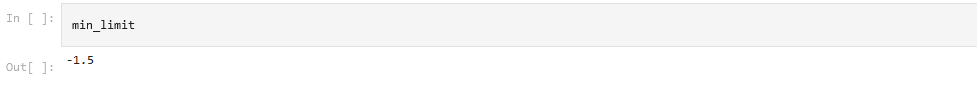
****

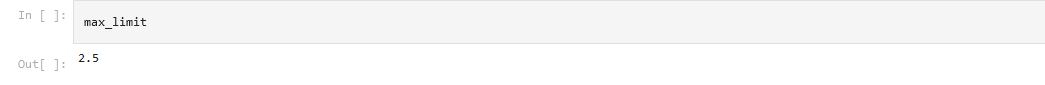
****

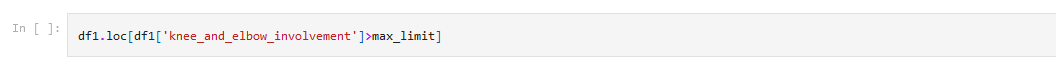
Outlier removal of the knee\_and\_elbow\_involvement column by using the Inter Quartile Range Method.

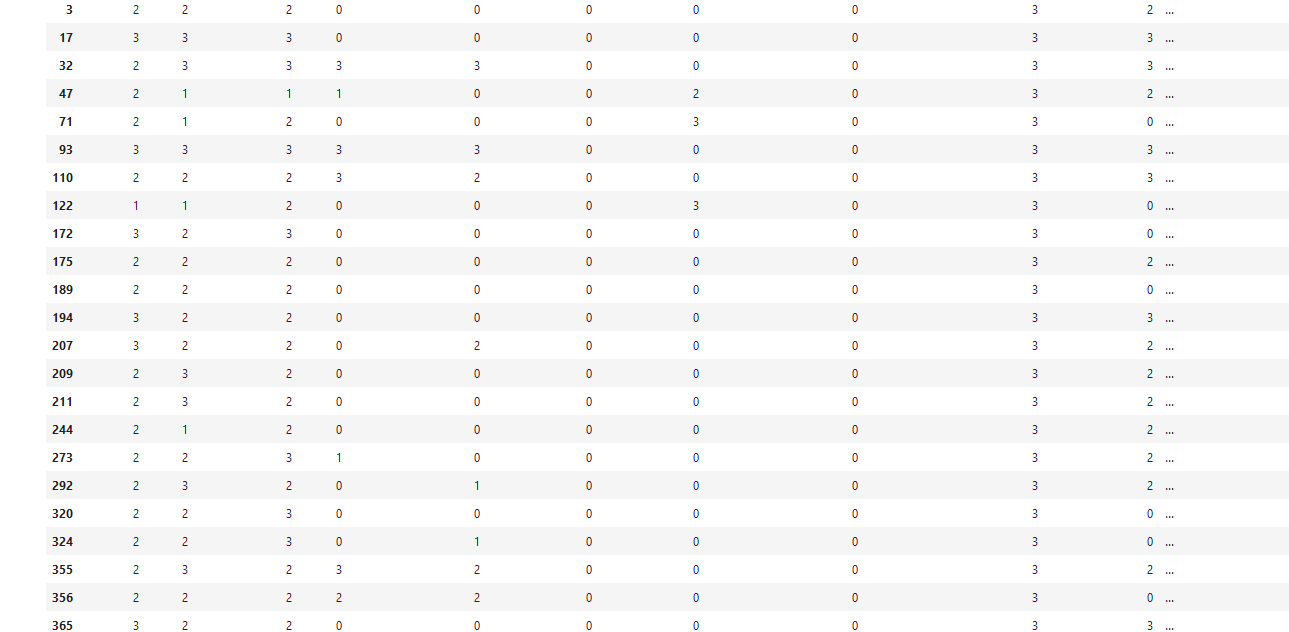


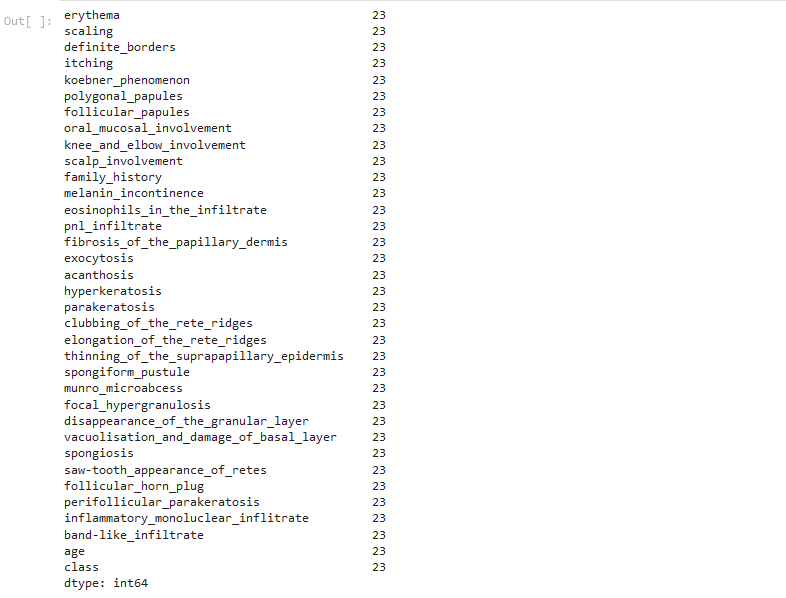






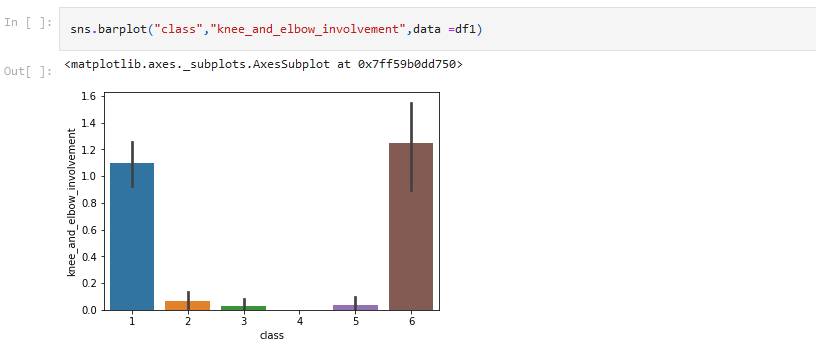




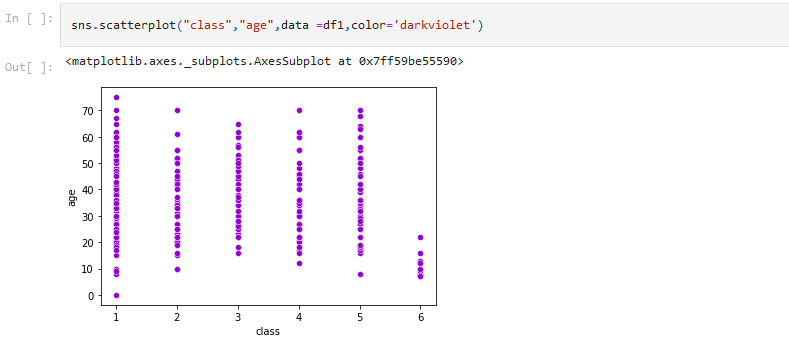




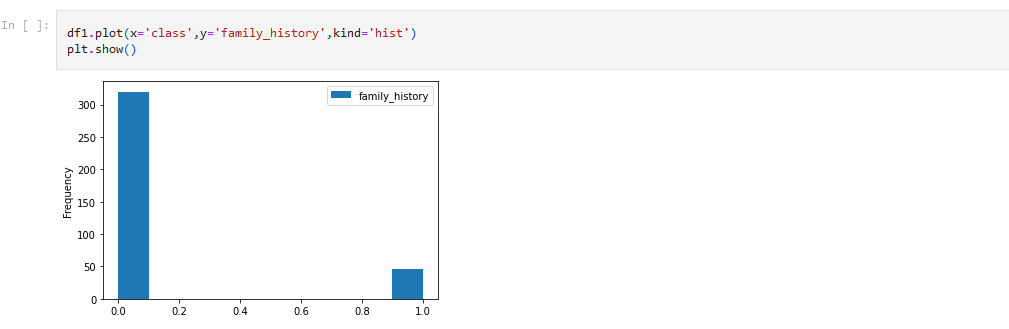
Data Visualization



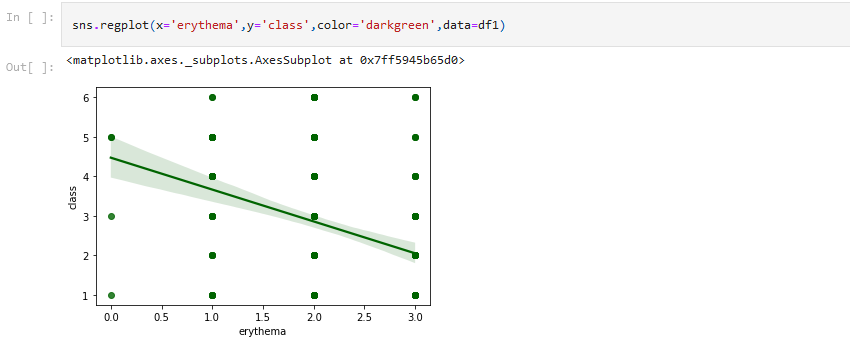
From the above plot we see that knee\_and\_elbow\_involvement has great involment in class 1 and 6,class 4 has no involvement,class 2,3,5 has least involment

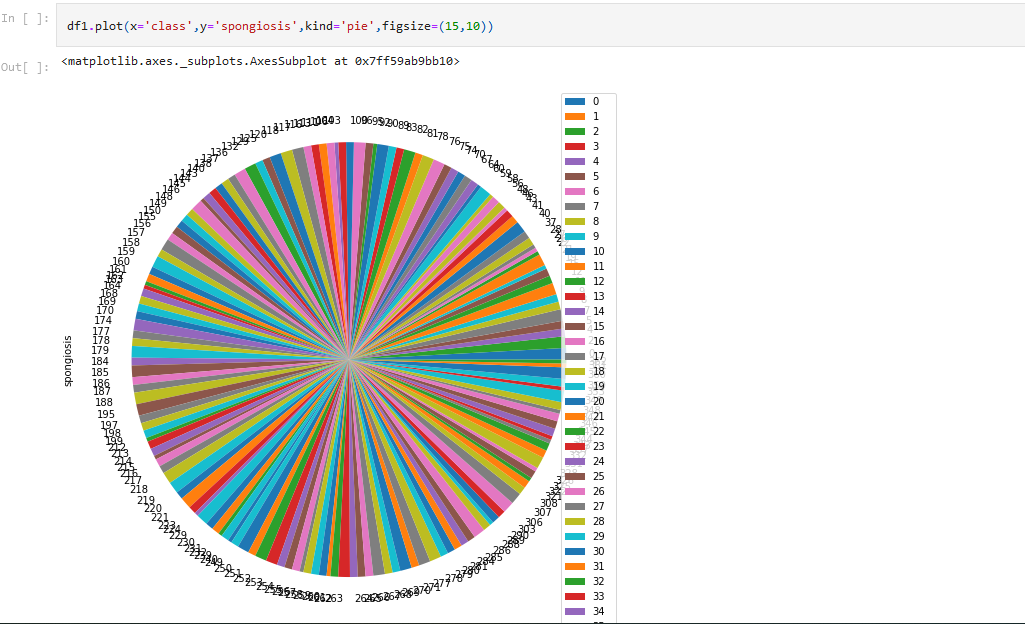


Here all class affect the age symptom,only 10-25 age people can affect the class 6 disease,other classes have great infulence on the age from 10-70



In most of the family we can see that the disease not present,few family has observe the disease present

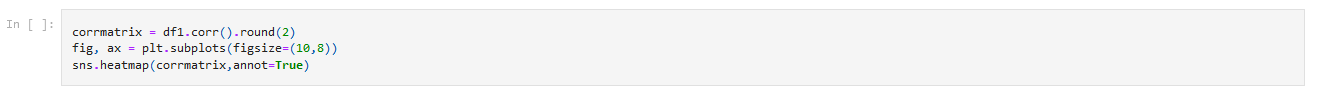


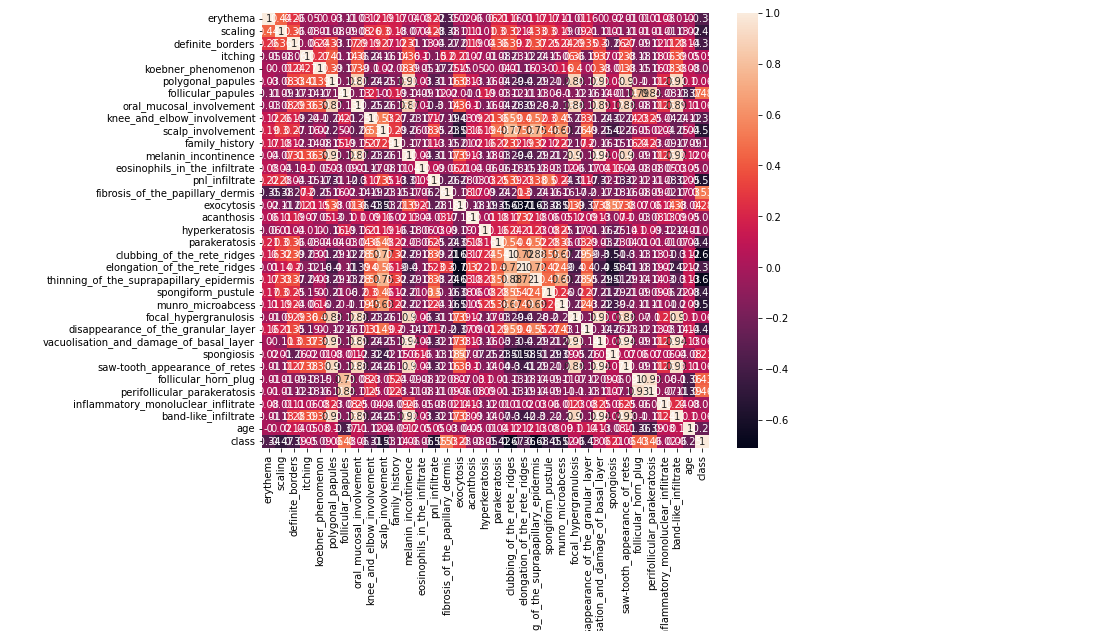


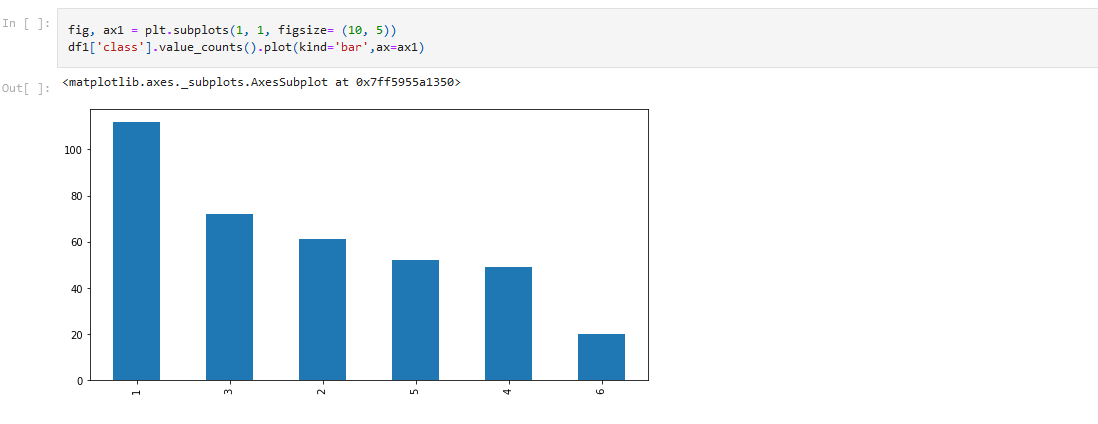


Heat map

A heat map can explain the correlation between the variables with the color intensity. It is a great way to plot the target variable over multiple variables and through this get visual clues of the relationship between variables and the target. As we now have to move into building machine learning models to automate our analysis, feeding the model with variables that meaningfully affect our target variable will improve our model’s prediction performance.

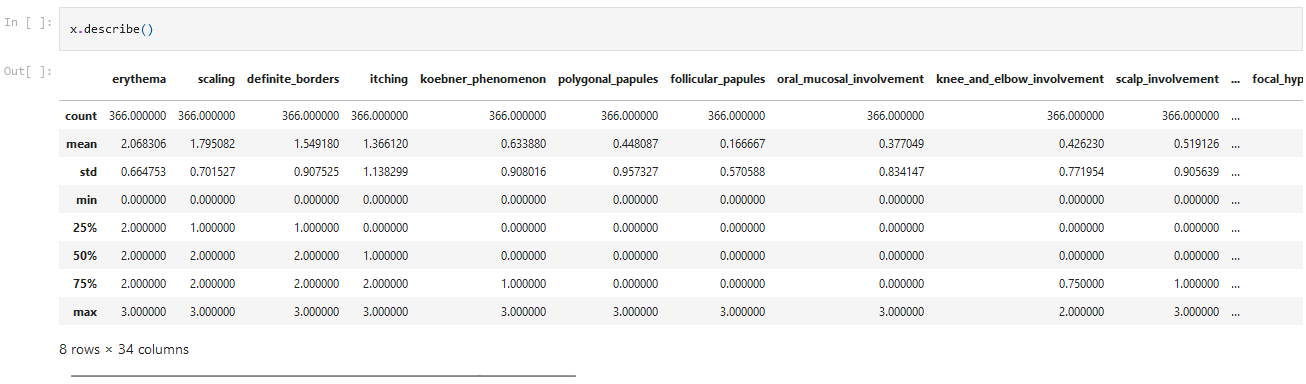


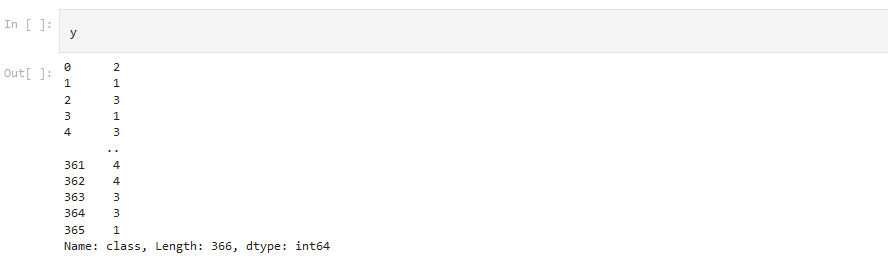




Splittiong to x and y







Splitting the data into training and testing dataset

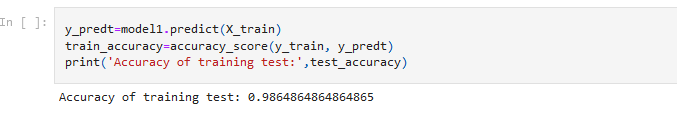


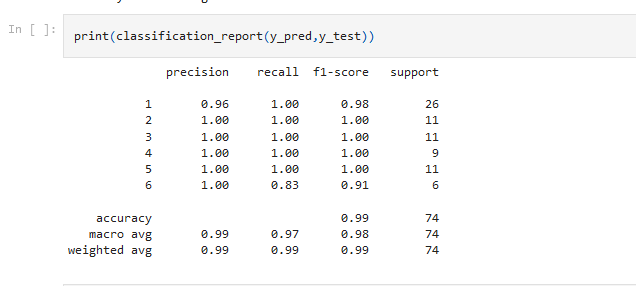
MODEL BUILDING AND PREDICTION

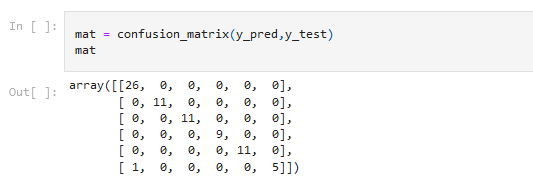
1.SVM Classifier

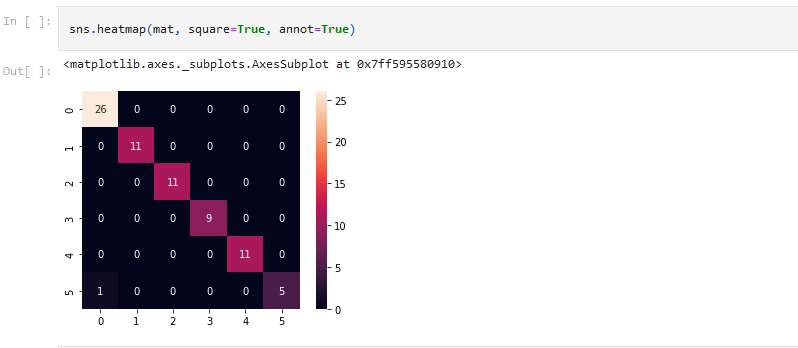
Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection. The advantages of support vector machines includes, Effective in high dimensional spaces. Still effective in cases where number of dimensions is greater than the number of samples. Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.

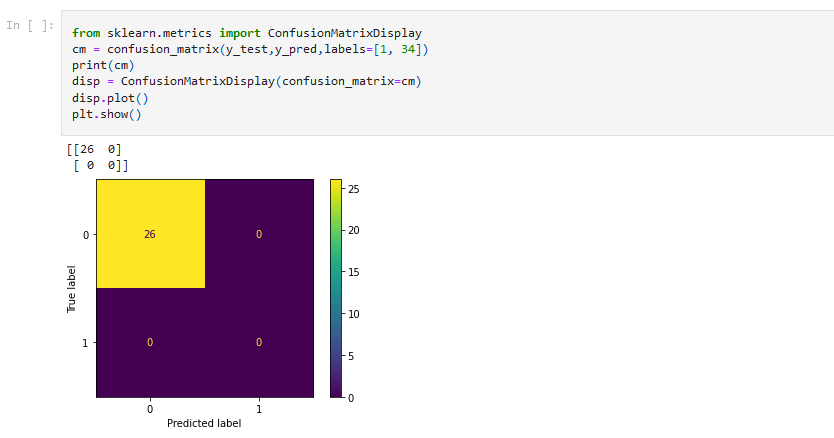






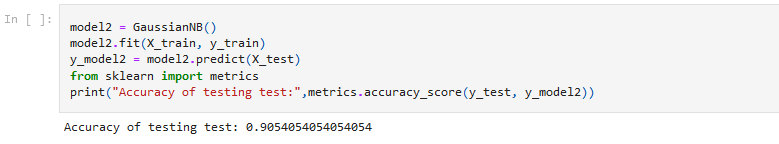


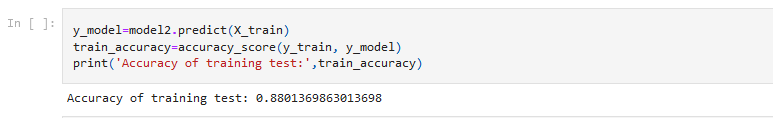


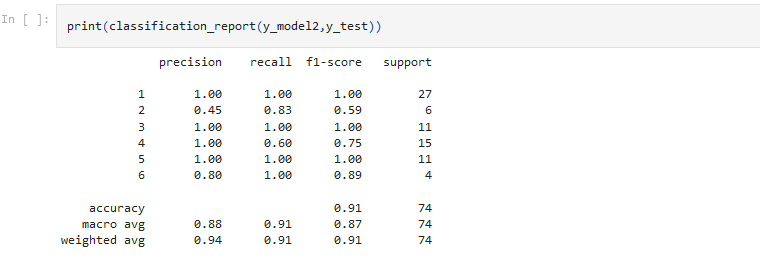


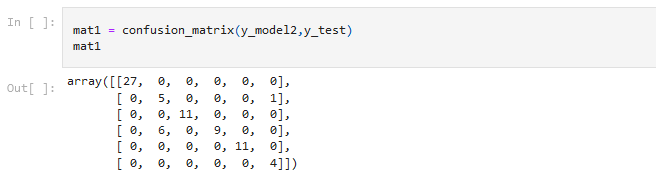
2.GaussianNB

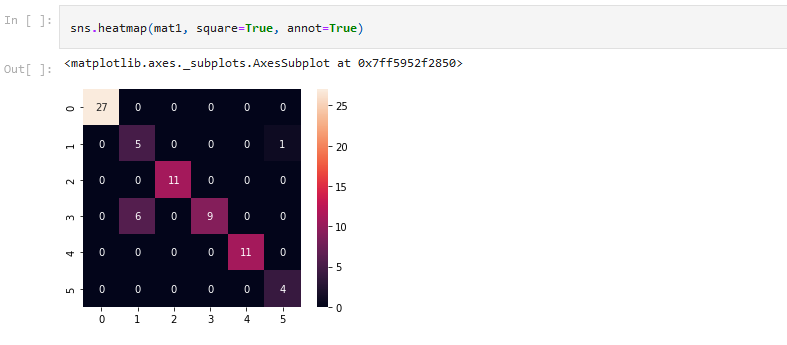
Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

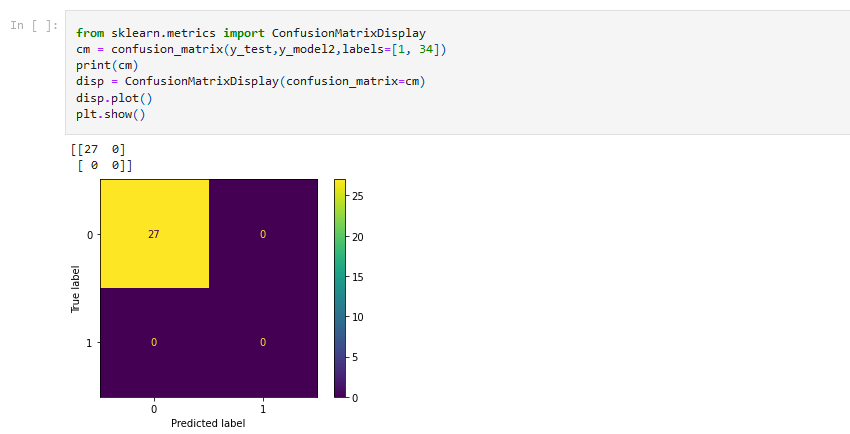






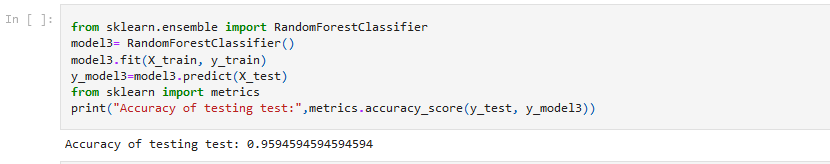


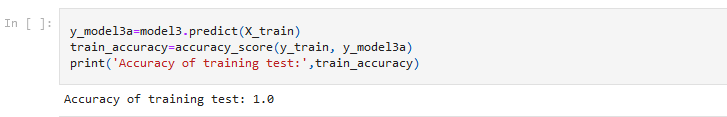


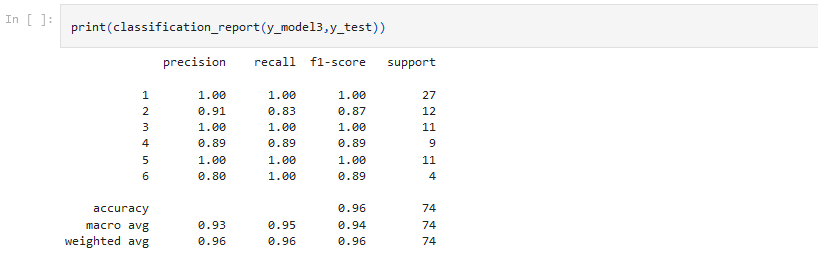


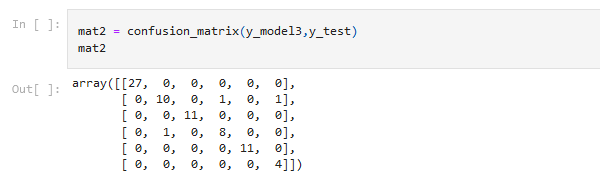
3.Random Forest Classifier

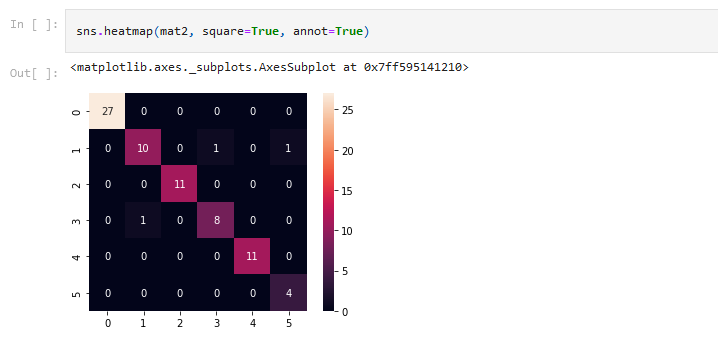
Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.As the name suggests, Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

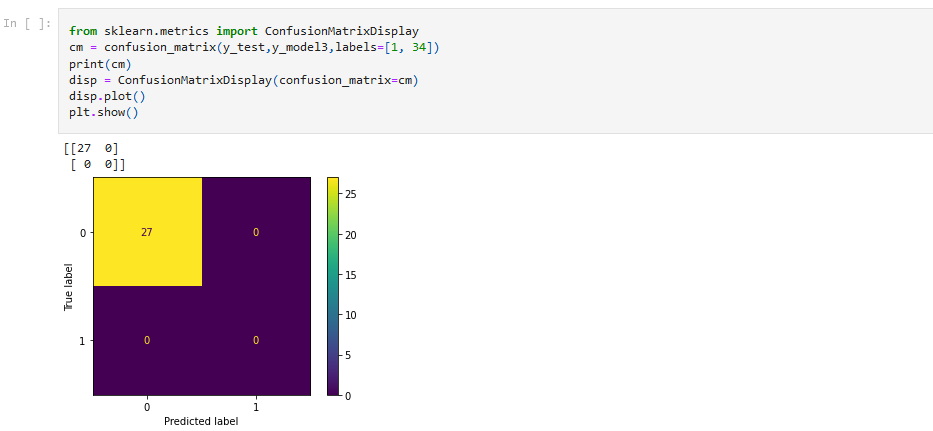






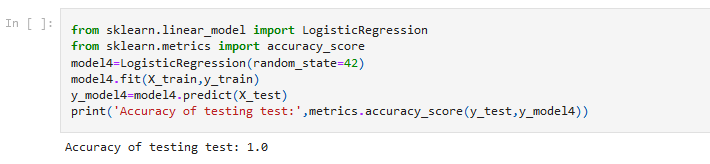


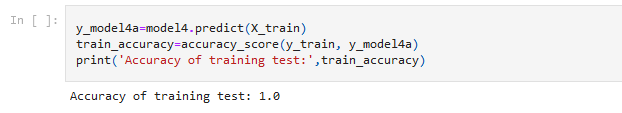


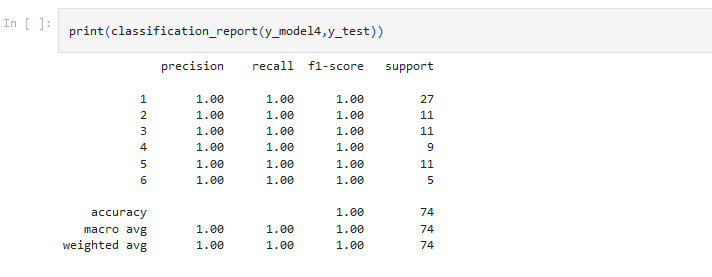


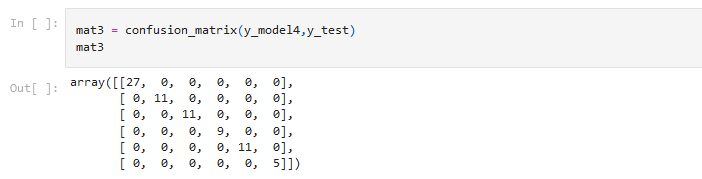
4.LogisticRegression

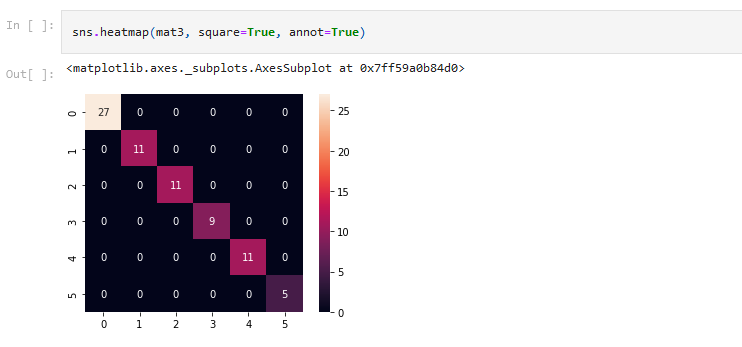
Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on. Multinomial logistic regression can model scenarios where there are more than two possible discrete outcomes. Logistic regression is a useful analysis method for classification problems, where you are trying to determine if a new sample fits best into a category.

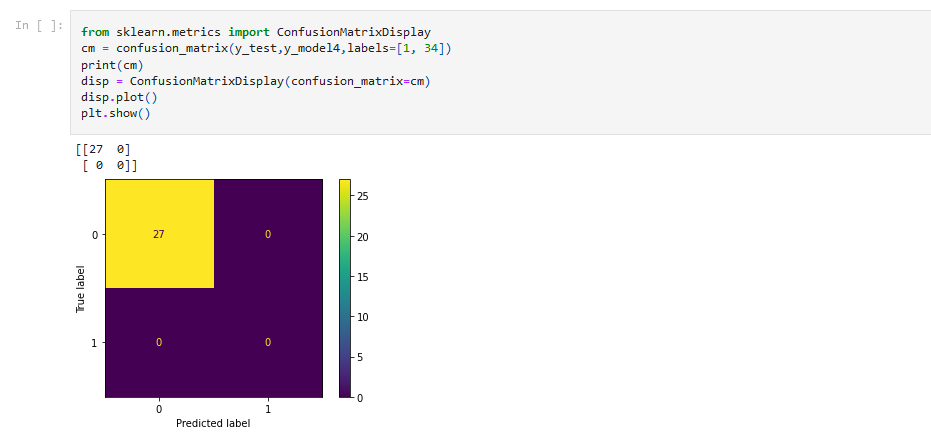






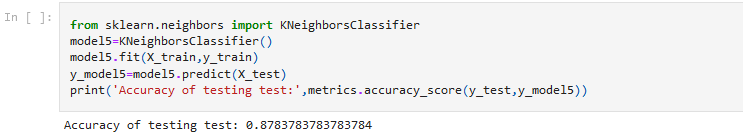


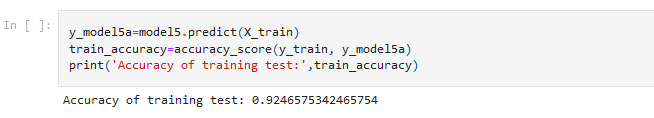


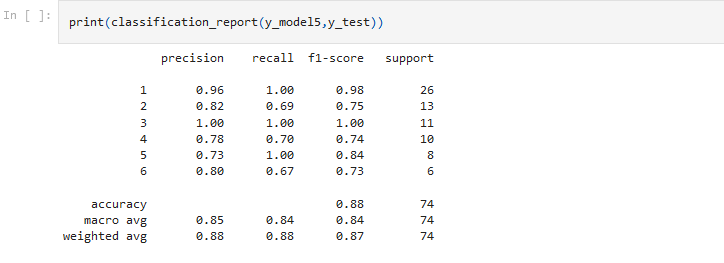


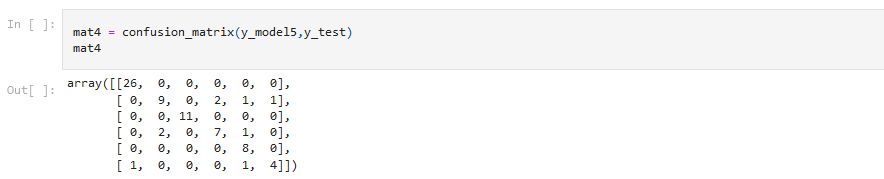
5.KNN Classifier

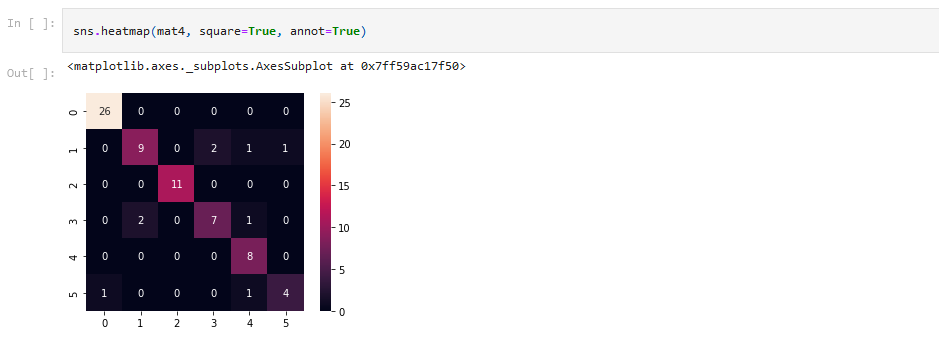
The k-nearest neighbours (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems.It is easy to implement and understand, but has a major drawback of becoming significantly slow as the size of that data in use grows.

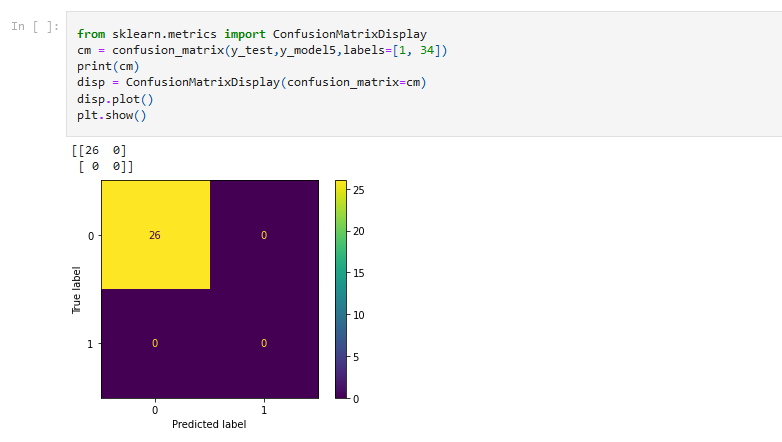






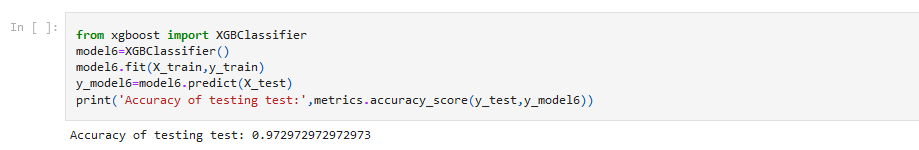


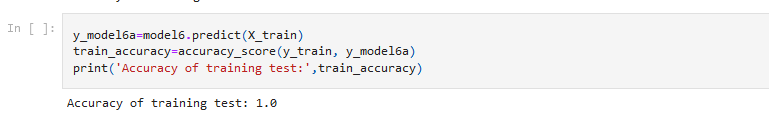


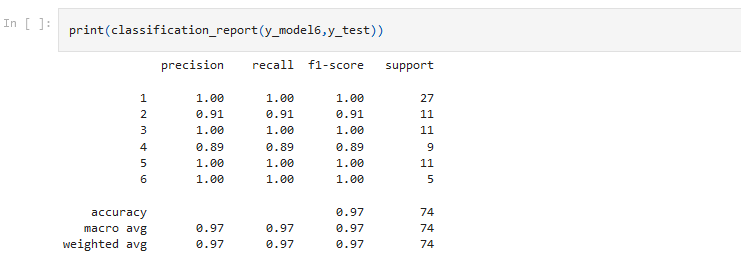


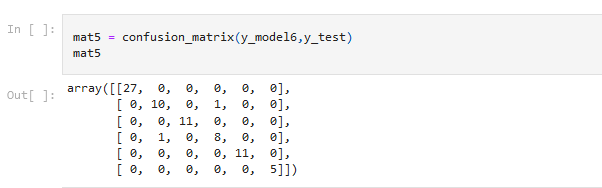
6.XGBClassifier

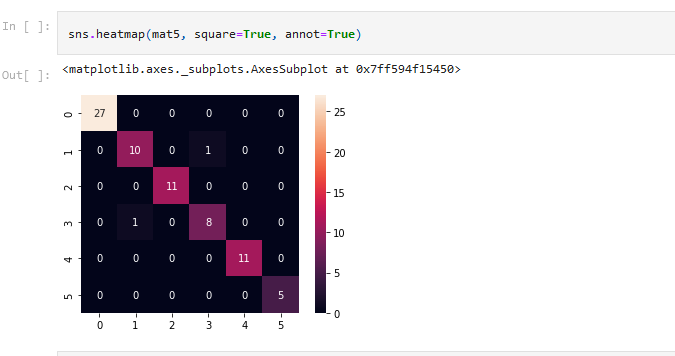
Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees. [1][2] When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest. A gradient-boosted trees model is built in a stage-wise fashion as in other boosting methods, but it generalizes the other methods by allowing optimization of an arbitrary differentiable loss function.

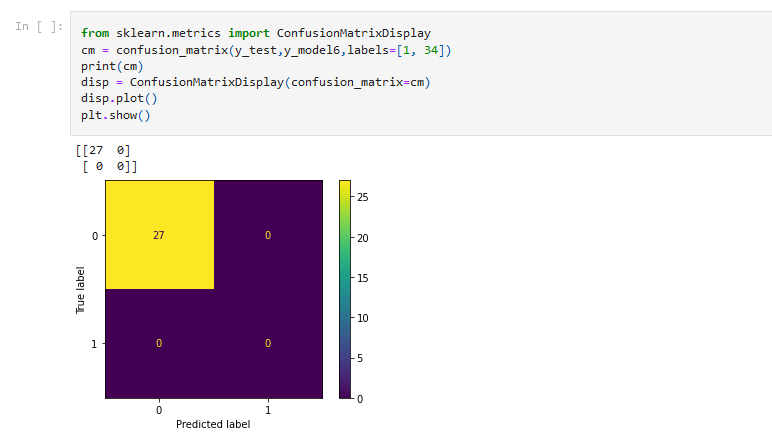








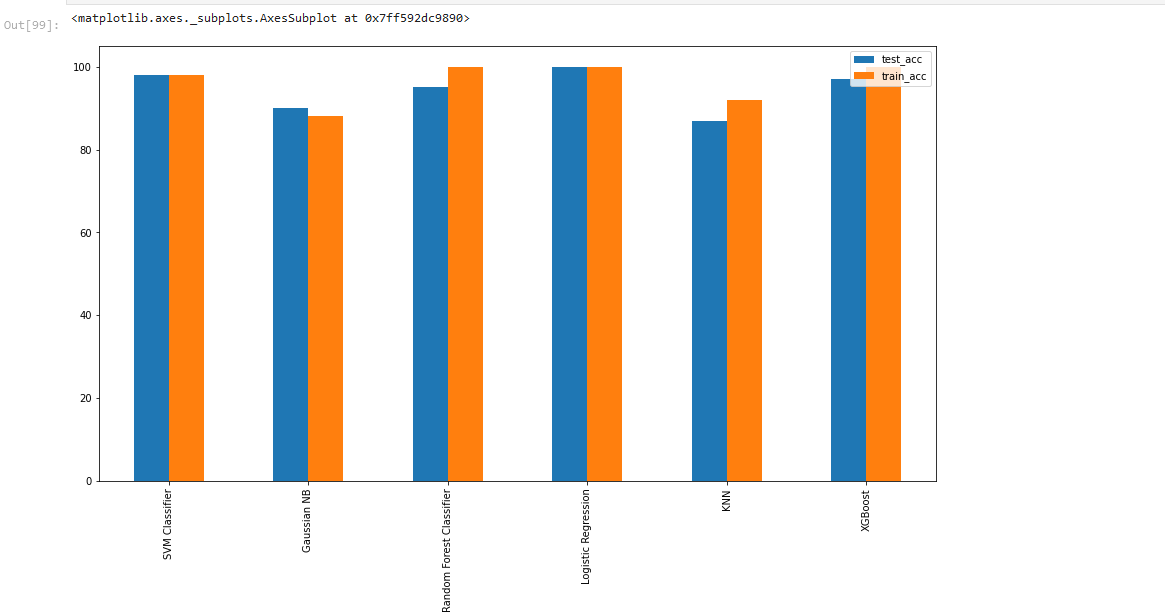




Evaluating the Accuracy of the models







Conclusion:

We have analysed using machine learning to predict class. We have used SVM Classifier,GaussianNB,Random Forest Classifier,LogisticRegression,KNNClassifier,XGBoost Classifier. We trained our models with 34 features. All the models gives the best to the data.

References:

G. Demiroz, H. A. Govenir, and N. Ilter, "Learning Differential Diagnosis of Eryhemato-Squamous Diseases using Voting Feature Intervals", Aritificial Intelligence in Medicine  
[[Web Link]](http://rexa.info/paper/f6790d06b3e02eb2de5f4af243b3a23a934aad9e)

Vassilis Athitsos and Stan Sclaroff. [Boosting Nearest Neighbor Classifiers for Multiclass Recognition](http://rexa.info/paper/692880d7b3356df64bfa0f06a683f89e4ce6955b). Boston University Computer Science Tech. Report No, 2004-006. 2004. [[View Context](https://archive.ics.uci.edu/ml/support/dermatology#692880d7b3356df64bfa0f06a683f89e4ce6955b)].  
  
Gisele L. Pappa and Alex Alves Freitas and Celso A AKaestner. [Attribute Selection with a Multi-objective Genetic Algorithm](http://rexa.info/paper/30b640cb0a4ba73302609a2f1ec74291c70efd85). SBIA. 2002. [[View Context](https://archive.ics.uci.edu/ml/support/dermatology#30b640cb0a4ba73302609a2f1ec74291c70efd85)].  
  
Rafael S. Parpinelli and Heitor S. Lopes and Alex Alves Freitas. [PART FOUR: ANT COLONY OPTIMIZATION AND IMMUNE SYSTEMS Chapter X An Ant Colony Algorithm for Classification Rule Discovery](http://rexa.info/paper/f4405e32dbb5dea3ece303e2a5b3edb6b413271e). CEFET-PR, Curitiba. [[View Context](https://archive.ics.uci.edu/ml/support/dermatology#f4405e32dbb5dea3ece303e2a5b3edb6b413271e)].  
  
Rafael S. Parpinelli and Heitor S. Lopes and Alex Alves Freitas. [An Ant Colony Based System for Data Mining: Applications to Medical Data](http://rexa.info/paper/899bdb470e48c308144216cc22048c88816ee035). CEFET-PR, CPGEI Av. Sete de Setembro, 3165. [[View Context](https://archive.ics.uci.edu/ml/support/dermatology#899bdb470e48c308144216cc22048c88816ee035)].  
  
Gisele L. Pappa and Alex Alves Freitas and Celso A AKaestner. [AMultiobjective Genetic Algorithm for Attribute Selection](http://rexa.info/paper/1a387e912f3bd68e66957cdfdebbb068ffc2149a). Computing Laboratory Pontificia UniversidadeCatolica do Parana University of Kent at Canterbury. [[View Context](https://archive.ics.uci.edu/ml/support/dermatology#1a387e912f3bd68e66957cdfdebbb068ffc2149a)].  
  
Perry Moerland. [Mixtures of latent variable models for density estimation and classification](http://rexa.info/paper/43019420b94637b63ac4fc3f94f64abd11b64f0a). E S E A R C H R E P R O R T I D I A P D a l l e M o l l e I n s t i t u t e f o r Pe r cep t ua l A r t i f i c i a l Intelligence . [[View Context](https://archive.ics.uci.edu/ml/support/dermatology#43019420b94637b63ac4fc3f94f64abd11b64f0a)].  
  
H. Altay Guvenir. [A Classification Learning Algorithm Robust to Irrelevant Features](http://rexa.info/paper/8eb7bb53b63501db4eb1c49eab30d76f7febba8d). Bilkent University, Department of Computer Engineering and Information Science. [[View Context](https://archive.ics.uci.edu/ml/support/dermatology#8eb7bb53b63501db4eb1c49eab30d76f7febba8d)].  
  
M. V. Fidelis and Heitor S. Lopes and Alex Alves Freitas. [Discovering Comprehensible Classification Rules with a Genetic Algorithm](http://rexa.info/paper/25c83dd7b43db4459e6eec12a71a229e10afac90). UEPG, CPD CEFET-PR, CPGEI PUC-PR, PPGIA Praa Santos Andrade, s/n Av. Sete de Setembro. [[View Context](https://archive.ics.uci.edu/ml/support/dermatology#25c83dd7b43db4459e6eec12a71a229e10afac90)].