**Dr.D.Y.Patil Institute of Technology**

Pimpri, Pune-411018

A MINI-PROJECT REPORT

ON

**“****Apply the Principal Component Analysis for Feature Reduction on Wine Company Dataset”**

Submitted by

BCOA41 Gauri Gulwane

BCOA43 Sampada Gaonkar

BCOA46 Devashish Dhande

**Department of Computer Engineering**

2019-2020

**Dr.D.Y.Patil Institute of Technology**

PIMPRI, PUNE-411018

**Department of Computer Engineering**

CERTIFICATE

Certified that the mini-project work entitled **“Apply the Principal Component Analysis for feature reduction on Wine Company Dataset”** is a bonafide workcarried out by

**Gauri Gulwane BCOA41**

**Sampada Gaonkar BCOA43**

**Devashish Dhande BCOA46**

The report has been approved as it satisfies the academic requirements in respect of mini-project work prescribed for the course.

……………...…………………………

**Subject In-charge**

**Abstract**

Large datasets are increasingly common and are often difficult to interpret. Principal component analysis (PCA) is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables that successively maximize variance. Finding such new variables, the principal components, reduces to solving an eigenvalue/eigenvector problem, and the new variables are defined by the dataset at hand, not *a priori*, hence making PCA an adaptive data analysis technique. It is adaptive in another sense too, since variants of the technique have been developed that are tailored to various different data types and structures. This article will begin by introducing the basic ideas of PCA, discussing what it can and cannot do. It will then describe some variants of PCA and their application.

**Content**

|  |  |  |
| --- | --- | --- |
| **Sr. No** | **Name** | **Page No** |
| **1** | **Introduction** | **5** |
| **2** | **Implementation** | **7** |
| **3** | **Result and Analysis** | **9** |
| **4** | **Conclusions** | **10** |
| **5** | **References** | **11** |

**List of Figure**

|  |  |  |
| --- | --- | --- |
| **Sr. No** | **Name** | **Page No** |
| **1** | **Wine Company Dataset** | **7** |
| **2** | **Variable Explorer** | **7** |
| **3** | **Visualization of the Training and Test sets.** | **8** |
| **4** | **Comparison between the Test and Predict sets** | **8** |
| **5** | **Confusion Matrix** | **9** |

**I. Introduction**

Large datasets are increasingly widespread in many disciplines. In order to interpret such datasets, methods are required to drastically reduce their dimensionality in an interpretable way, such that most of the information in the data is preserved. Many techniques have been developed for this purpose, but principal component analysis (PCA) is one of the oldest and most widely used. Its idea is simple—reduce the dimensionality of a dataset, while preserving as much ‘variability’ (i.e. statistical information) as possible.

Although it is used, and has sometimes been reinvented, in many different disciplines it is, at heart, a statistical technique and hence much of its development has been by statisticians.

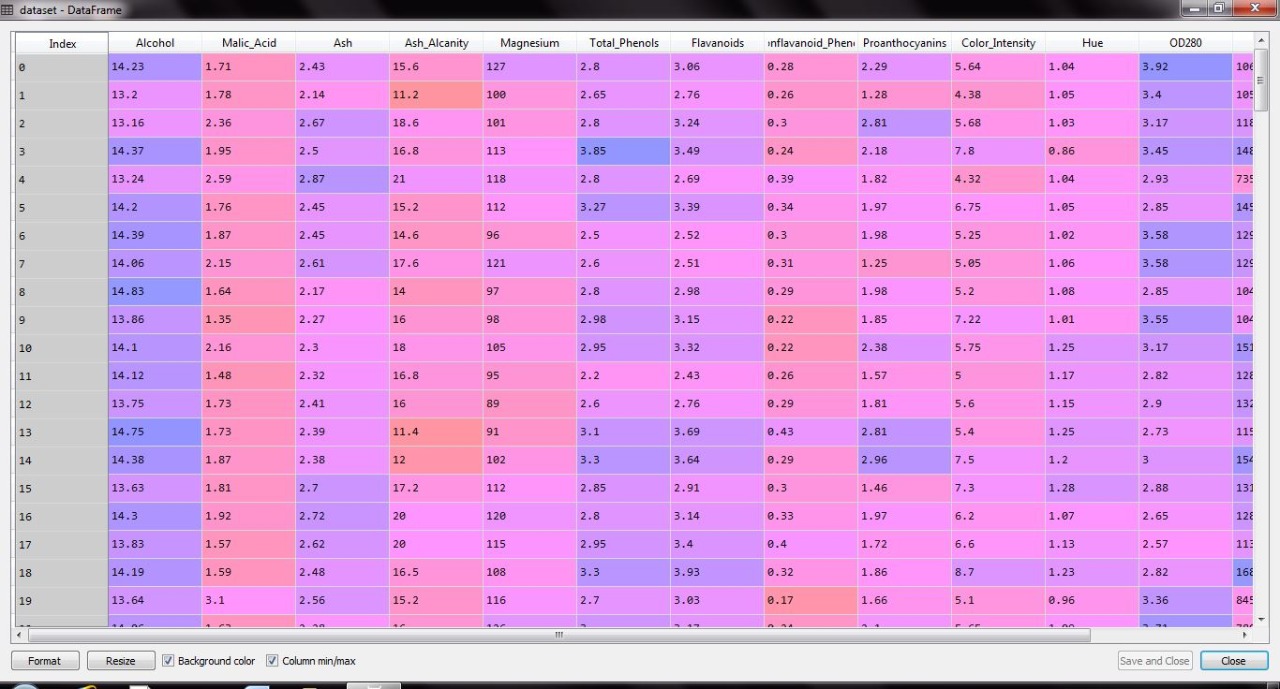
This means that ‘preserving as much variability as possible’ translates into finding new variables that are linear functions of those in the original dataset, that successively maximize variance and that are uncorrelated with each other. Finding such new variables, the principal components (PCs), reduces to solving an eigenvalue/eigenvector problem. The earliest literature on PCA dates from Pearson and Hotelling, but it was not until electronic computers became widely available decades later that it was computationally feasible to use it on datasets that were not trivially small. Since then its use has burgeoned and a large number of variants have been developed in many different disciplines. Substantial books have been written on the subject and there are even whole books on variants of PCA for special types of data. PCA can be based on either the covariance matrix or the correlation matrix. The choice between these analyses will be discussed. In either case, the new variables (the PCs) depend on the dataset, rather than being pre-defined basis functions, and so are adaptive in the broad sense.

Given a collection of points in two, three, or higher dimensional space, a "best fitting" line can be defined as one that minimizes the average squared distance from a point to the line. The next best-fitting line can be similarly chosen from directions perpendicular to the first. Repeating this process yields an orthogonal basis in which different individual dimensions of the data are uncorrelated. These basis vectors are called **principal components**, and several related procedures **principal component analysis** (**PCA**).

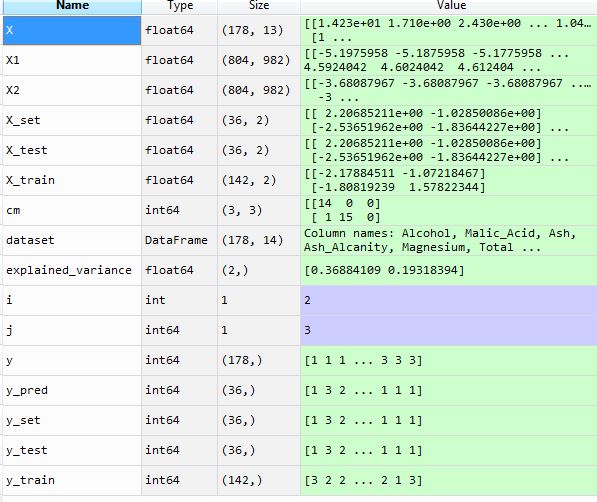
PCA is mostly used as a tool in exploratory data analysis and for making predictive models. It is often used to visualize genetic distance and relatedness between populations. PCA is either done by singular value decomposition of a design matrix or by doing the following 2 steps:

1. calculating the data covariance (or correlation) matrix of the original data
2. performing eigenvalue decomposition on the covariance matrix

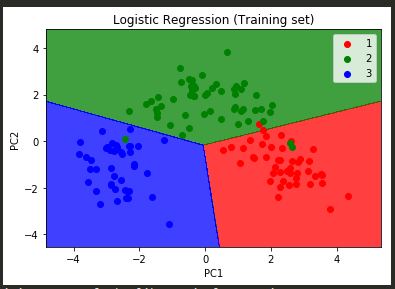
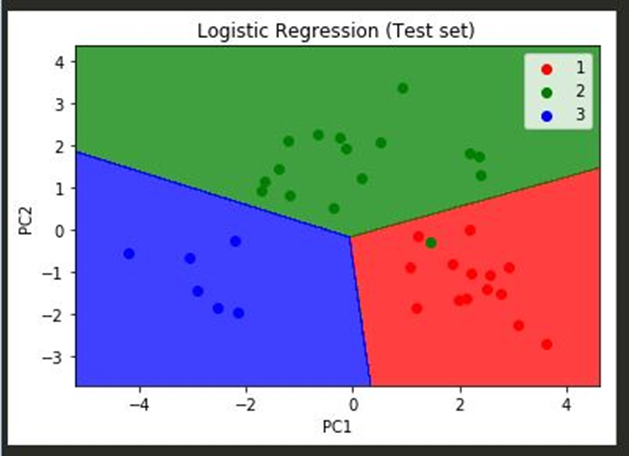
**II. Implementation**



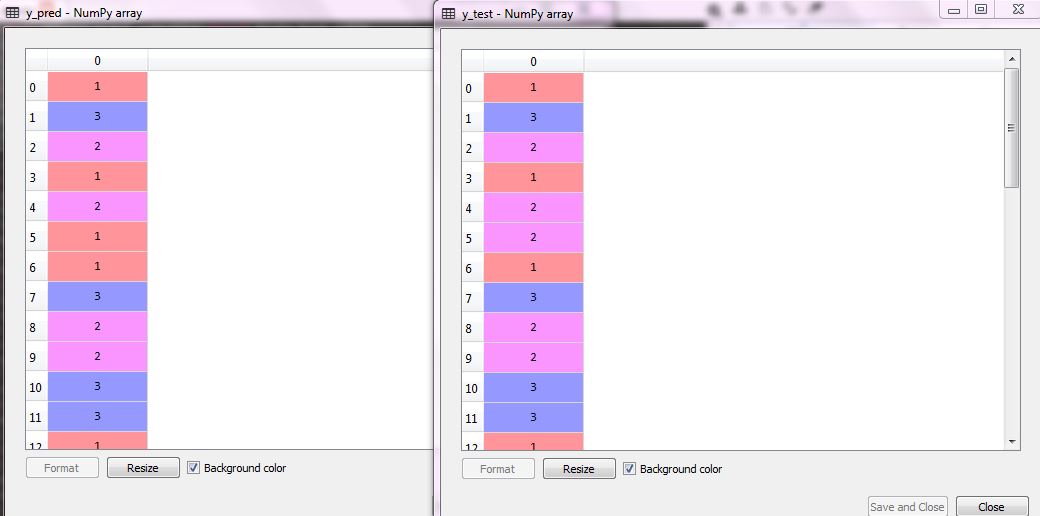
**Fiq 1: Wine Company Dataset**

****

**Fiq 2: Variable Explorer**

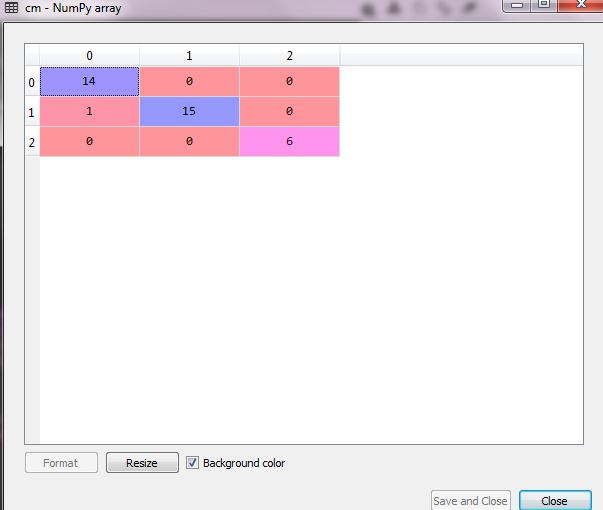


**Fiq 3: Visualization of the Training and Test sets.**



**Fiq 4: Comparison between the Test and Predict sets**

**III. Result and Analysis**

 The analysis is done on the gathered data which is stored in the CSV file. For analysis purposes, we used logistic regression. And successfully predicting the category of wine. Feature reduction is applied to get more important features which can help in predicting the type of wine more precisely. And also removing unnecessary features which can hinder the prediction.

**Fiq 5: Confusion Matrix**

**IV. Conclusions**

Hence, using this application we can successfully predict the category of the wine to which it belongs. It is implemented using Principal Component Analysis (PCA) for Feature reduction .

**V. References**

* Google.com
* [www.stacksoverflow.com](http://www.stacksoverflow.com)
* /scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
* https://en.wikipedia.org/wiki/Principal\_component\_analysis