Combo-17 brighter galaxy anaylysis

Supervised By: Dr. Tahseen A. Jillani.

Ratan Kumar

Musab Bin Tariq

Muhammad Zubair

Kashif Jan

Table of Contents

[1. Introduction 4](#_Toc92635429)

[1.1. Data 4](#_Toc92635430)

[1.2. Data Warehouse 4](#_Toc92635431)

[1.3. Why Do We Use Data Mining 4](#_Toc92635432)

[1.4. How Does Data Mining Work 5](#_Toc92635433)

[2. Literature Review 7](#_Toc92635439)

[2.1. Overview 7](#_Toc92635441)

[2.2. Astronomical Background 7](#_Toc92635442)

[2.3. Understanding the Features 8](#_Toc92635443)

[2.4. Methodology 9](#_Toc92635444)

[3. Data Pre-Processing and Data Mining Techniques. 11](#_Toc92635445)

[3.1. Data Pre-Processing 11](#_Toc92635449)

[3.1.1. Data Integration 11](#_Toc92635454)

[3.1.2. Data Cleaning 11](#_Toc92635455)

[3.1.3. Data Normalization 12](#_Toc92635456)

[3.1.3.1. Min Max Normalization 12](#_Toc92635457)

[3.1.4. Outlier Detection 13](#_Toc92635458)

[3.1.5. Correlation 14](#_Toc92635459)

[3.1.6. Data Reduction 14](#_Toc92635460)

[3.1.6.1. Principle Component Analysis 15](#_Toc92635461)

[3.2. Data Mining Techniques 16](#_Toc92635462)

[3.2.1. K-Mean Clustering Algorithm 16](#_Toc92635468)

[3.2.2. DENSITY BASED CLUSTRING 17](#_Toc92635475)

[3.2.3. SUM OF SQUARED ERROR (SSE) 18](#_Toc92635483)

[3.2.4. Silhouette Score 18](#_Toc92635492)

[3.2.5. Support Vector Machine 19](#_Toc92635493)

[3.2.6. Stochastic Descent Gradient Pipelines Classifier 20](#_Toc92635504)

[3.2.7. Random Forest Classifier 20](#_Toc92635505)

[3.2.8. Confusion Matrix 22](#_Toc92635518)

[4. Results 23](#_Toc92635519)

[5. References 24](#_Toc92635520)

LIST OF FIGURES

[Figure 1. Data Mining Application 5](file:///C:\Users\Ratan%20Kumar\Downloads\Report%20(Data-mining)%20(2).docx#_Toc92634015)

[Figure 2 How Data Mining Works 6](file:///C:\Users\Ratan%20Kumar\Downloads\Report%20(Data-mining)%20(2).docx#_Toc92634016)

[Figure 3. Types of Galaxies 8](file:///C:\Users\Ratan%20Kumar\Downloads\Report%20(Data-mining)%20(2).docx#_Toc92634017)

[Figure 4. Data Preprocessing 11](file:///C:\Users\Ratan%20Kumar\Downloads\Report%20(Data-mining)%20(2).docx#_Toc92634018)

[Figure 5 Outlier Detection using scatter plot by plotting chi square cut 13](file:///C:\Users\Ratan%20Kumar\Downloads\Report%20(Data-mining)%20(2).docx#_Toc92634019)

[Figure 6Outer Detection by Box Plot and Histogram 13](file:///C:\Users\Ratan%20Kumar\Downloads\Report%20(Data-mining)%20(2).docx#_Toc92634020)

[Figure 7. Correlation Matrix 14](#_Toc92634021)

[Figure 8. Scree Plot for Eigen Values (PCA) 15](file:///C:\Users\Ratan%20Kumar\Downloads\Report%20(Data-mining)%20(2).docx#_Toc92634022)

[Figure 9. K-mean Clustering 16](#_Toc92634023)

[Figure 10DB Scan Clustering Algorithm 17](#_Toc92634024)

[Figure 11 SSE PLOT OF K-MEAMNS 18](file:///C:\Users\Ratan%20Kumar\Downloads\Report%20(Data-mining)%20(2).docx#_Toc92634025)

[Figure 12Silhouette score on K-Means 19](file:///C:\Users\Ratan%20Kumar\Downloads\Report%20(Data-mining)%20(2).docx#_Toc92634026)

[Figure 13. Support Vector Machine 19](#_Toc92634027)

[Figure 14Stochastic Descent Gradient Pipelines Classifier 20](file:///C:\Users\Ratan%20Kumar\Downloads\Report%20(Data-mining)%20(2).docx#_Toc92634028)

[Figure 15 Random Forest Classifier 21](#_Toc92634029)

[Figure 16 Confusion Matrix 22](file:///C:\Users\Ratan%20Kumar\Downloads\Report%20(Data-mining)%20(2).docx#_Toc92634030)

Abstract

Data mining is the process of extracting knowledge hidden from large volumes of raw data and represent in visualized form.

Combo-17 Dataset is a repository of [Wolf et al. (2004)](http://arxiv.org/abs/astro-ph/0403666) provided the first public catalog of brightness galaxies measurements in 17 bands in the visible band. In this project we are analyzing brighter galaxies and its properties of given features. We are using different data mining techniques like data pre-processing, data-visualization, clustering and classification.

# Introduction

Data mining defines in many different ways, some of this has to deal with large, ‘global’ buildings, and the purpose is to model the shape, or structural features, of the distribution. The other is about small, ‘local’ buildings, and the purpose is to find out what is confusing and to decide whether they are real events or lucky. However, therefore, further discussion of the initial feature is also required. This provides a clear overview of data mining and your relationship to statistics. Data mining software is one of the data analysis tools. Allows users to analyze from different sizes or angles, classify them, and summarize identified relationships.

* 1. Data

The word Data means “to give”. Data is really given “raw facts”. Sometime data and information are interchangeably used, but the question is that is this interchangeability correct? Some writer use it interchangeably. Which is not considered as standard.

The derivation as explained by some writers as is that Data is considered as Raw facts, where information considered as processed data from which we could be able to extract some meaningful facts and figures about records.

* 1. Data Warehouse

Data Warehouse is a process called as collection facts from multiple sources .Data Warehouse are created by an integrating data, data processing, data cleaning, data transformation and data loading. For example: A Store has multiple branches and every branch has its own database and our task is to collect whole data into one place, for this task we use Data Warehouse application to make a single repository for multiple databases.

* 1. Why Do We Use Data Mining

Data may be one of the most important assets of your organization but only if you know how to disclose important information hidden in raw data. Data mining techniques allow you to dig the data and extract most important information from you Data and make a usable prediction for your organization.

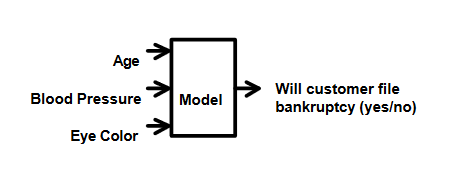


Figure 1. Data Mining Application

* 1. How Does Data Mining Work

Data mining operates in the sense of modelling. Suppose an organization wants to achieve something. By analyzing organization’s dataset, in which result is known. So data mining techniques will be used to build a model software that is already trained by its techniques.

Here’s an overview of data mining model:

2. **Start with historical dataset**

Suppose a company wants to know the best customer service opportunities on a new marketing site. It starts by testing its customers.

1. **Analyze the historical data**

The software scans the collected data using a combination of algorithms from mathematics, machine intelligence and learning, looking at patterns and relationships in data.

1. **Write down the rules**

When patterns and relationships are revealed, the software identifies them as rules. The rule may be that most 51- to 65-year-old customers buy twice a week and fill their baskets with fresh food, while 21- to 50-year-old customers usually buy once a week and buy some packaged food.

1. **Apply Data mining rules**

Here, the rules for data mining model of new marketing database is, If the this company is in food providing package, then it will look for 21-50 years old.

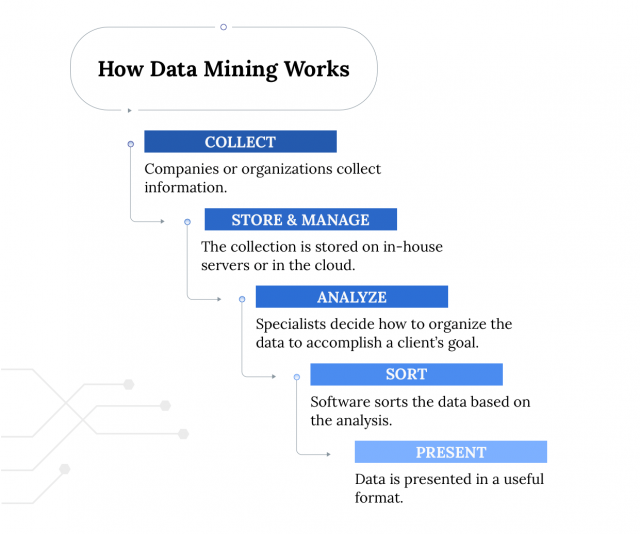


Figure 2 How Data Mining Works

# Literature Review

1. 1. Overview

Combo-17 dataset is a first public large dataset (63,501 objects) provided by [Wolf et al.(2004)](http://arxiv.org/abs/astro-ph/0403666). The dataset contains brightness measurements in 17 bands in the visible bands. Here dataset provides subset of their catalogs with 65 features of information on 3462 galaxies. These are objects from Chandra Deep Field from Wolf and colleagues have classified as galaxies.

* 1. Astronomical Background

We know that the fundamental structure of Universe are galaxies. Like our sun stands in Milky Way Galaxy, in the night we can see visible patchy band of lights of our galaxy across the sky. We can typically define the components of a galaxy by vast number of stars (total weight ~ 106-1011 Mo where Mo is a unit of solar mass), interstellar complex of gas and dust from which stars (usually 1-100% of stellar component mass), one large black hole in the center (usually <1% of stellar component mass), and an obscure component called the Dark Matter weighing ~ 5-10-times all other components are combined. During the 14 billion years since the Big Bang, galaxies have not changed much, and thus the light and color of galaxies change during cosmic times. This phenomenon has several names for the astronomical community: the history of the formation of stars in the universe, the chemical evolution of galaxies, or the mere appearance of the galaxy. Much effort has been made in the last few decades to measure and understand the emergence of the galaxy using telescopes in all wavelengths. A traditional tool for such studies has been optical spectroscopy that easily identifies star signatures in nearby galaxies. However, to study the constellations in the galaxies that have just emerged after the Big Bang, we must examine the weakest galaxies that seem invisible, even using the largest telescopes available. Another possible way to obtain images of obscure galaxies in random places in the sky is by small spectral bands, and then by creating an unmodified spectra. First, mathematical analysis of this multiband photometric information is used to classify galaxies, stars, and quasars. Second, in galaxies, a multivariate retreat was performed to improve the photometric measurements of the redshift, which is a measure of both the distance from us and the age from the Big Bang. Third, one can explore the colors of the galaxy as a redshift function (after various adjustments) to study the appearance of star formation. The current database is taken after the completion of these first two steps.

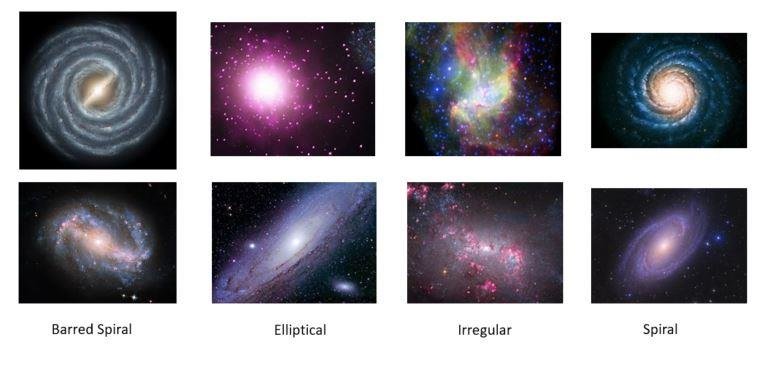


Figure 3. Types of Galaxies

* 1. Understanding the Features

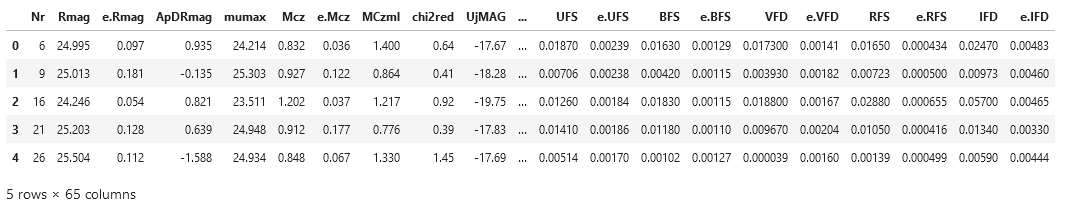


Table 1 Combo-17 Data Features

* + - **Feature 1: Numbers of Observed Object**
    - **Feature 2 -3: Rmag = Red Band Magnitude and it’s error**

Total R (red band) Mag: and error. This was the band on which the basic catalog was built. Mag: is the distorted logarithmic steps of light. The galaxy with R = 21 is 100 times brighter than the one with R = 26. Error common deviations found in detailed information of the measurement process. This database is an excellent example of a star database where each variant is accompanied by heteroscedastic measurement errors for known variations.

* + - **Feature 4-5: ApDRmag = Size of Galaxy, mumax = Core of Galaxy**

ApDRmag difference between total and aperture magnitude in band R. This is a critical measure of the galaxy mag: in the image where ApDRmag = 0 corresponds to the source source. Wrong values mean nothing physically. Max is the center light of an object in the R-band. The difference between Rmag and max should also be an indication of the size of the galaxy.

* + - **Feature 6-9:**

Mcz and MCzml are two redshift values. Mcz is the preferred value. E.Mcz has its own limited error, and chi2red doubled the value of a small chi-square of 17-band square measurement in the spectrum of a very similar galaxy image. A galaxy with a large e.Mcz or chi2red may be left untouched.

* + - **Feature 6-29:**

This gives the abs: mag: (i.e. internal light) of the galaxy by 10 bands, and its measurement errors. They are based on dimensions and redshifts, and represent light within galaxies; the galaxy with M = -15 is 100 times brighter than the one with M = -20. These officers are not all independent of each other, but are important in representing the internal features of galaxies. Below is one of the few redshift-stratified episodes of B-band absolute magnitude (abscissa) against the difference in magnitude (i.e. light intensity) between the ultraviolet band 2800A and blue, which is a critical indicator of star formation. Redshift-based bimodal distribution is visible.

* + - **Feature 30-55**:

Light detected in 13 bands respectively from 420 nm in ultraviolet to 915 nm in very red. This is provided by line flexibility and units of photon flux densities, photons / m2 / s / nm. Also, each measurement is accompanied by a measurement error that can be used to differentiate the measurement and internal dispersion in the distribution.

* + - **Feature 56-65:**

Light detected on 5 wide spectral traditional bands, UBVRI. These are very confusing with 13 bands

## Methodology

The Wolf and its colleagues provided the data of the galaxies and had applied some statistical analysis like, Standard deviation of Galaxy Magnitude is known, and they also calculated the Chi-Square, now we can easily identify how much data is useful by plotting features chi-square cut line with red magnitude, center/core of galaxy and size of galaxy. By these provided values we can easily pre-process data like detecting outliers, and reduction of data. After pre-process it would easier to apply data mining techniques, like dimensionality reduction using PCA, clustering techniques, errors, Classification techniques, Accuracy and confusion matrix etc.

# Data Pre-Processing and Data Mining Techniques.

3. 1. Data Pre-Processing

Data Pre-processing is the process of transformation into meaningful data. Data Quality should be checked before applying data mining algorisms. It is mainly to check the following data quality Accuracy: To check Correct or not.

Completeness: To Check Data recorded or not available.

Consistency: To check same data is stored in all the same or different locations.

Timeliness: Data updated correctly.

Believability: Data should be trustable or not.

Interpretability: Data easily understandable or not.

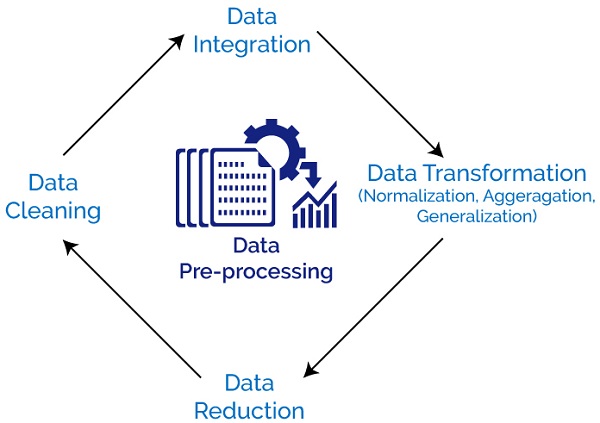


Figure 4. Data Preprocessing



### Data Integration

Data integration includes data from multiple sources to related data stores, such as a data repository or data cube. Data conversion as normal can be used. There are many problems to consider during data consolidation, schema integration can be difficult.

Redundancy is an important issue. The attribute may not work if it cannot be "found" in another table. Other job losses can be detected by a combination analysis.

### Data Cleaning

The process of removing invalid data, incomplete data and incorrect data from data sets, and changing missing values is called Data cleaning

There are some techniques for cleaning the data.

Detect or remove outliers, remove noisy data, and resolve inconsistencies.

**Before Data Shape:**

(3438, 65)

Table 2 Remove outliers.

|  |  |
| --- | --- |
| CODE | df = df[df.chi2red<5]  df = df[df.ApDRmag>=0]  df = df[df.mumax>=20]  df.shape |

**After Data Shape Output:**

(1170, 35)

### Data Normalization

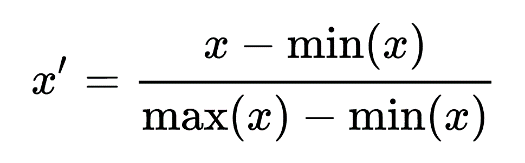
Some data mining methods are usually those based on calculating the distance between points in an n dimensional space e.g. [-], [] or [0, 1]. If the values are unusual the distance estimates will be more than the weight of those features with medium, large values. There are many techniques of normalizing the data. . Here are three easy ways to get used to it: -Standard deviation normalization

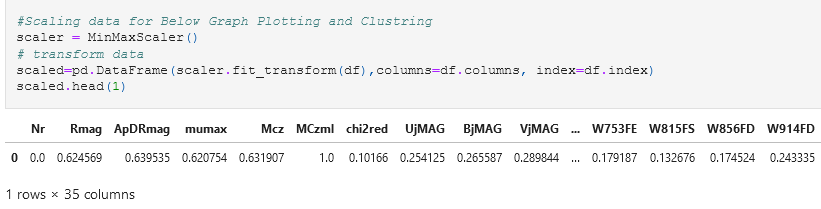
-Smooth data

-Differences and dimensions

-Min max normalization

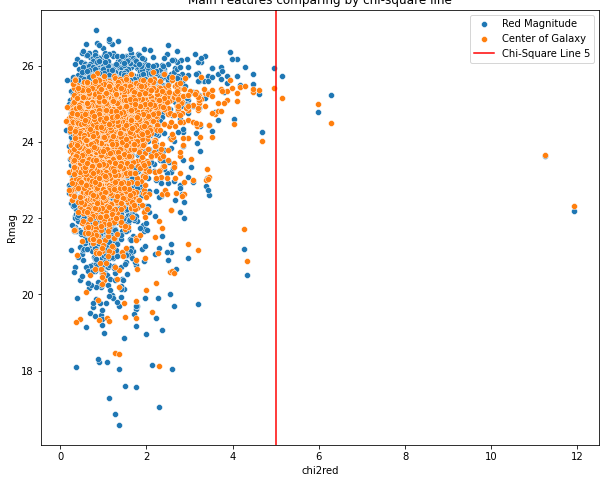
### Min Max Normalization

Minimum normalization methods such as Z-score, decimal decimation and standard deviation for standard deviations. It is to make the data normal. It will measure data between 0 and 1. Simplification helps us to understand data easily. In this process, transformation is performed on the original data. The min, max values collect from the datasets and each value is changed according to the following formula.



### Outlier Detection

In large data sets there are samples that do not confirm to the normal behavior of the data model. Such samples are very different or inconsistent with the remaining set of data called outliers. External factors may be caused by measurement error or may be the result of data variability. If e.g. while receiving fraudulent credit card purchases from a bank, external providers are typical examples that can illustrate fraudulent activity and the entire data mining process focused on their acquisition. But for some data mining applications, especially if they are supported by large data sets, the external objects are less efficient and are the result of errors in data collection than the data set features.

External detection and possible removal of a data set can be defined as the process of selecting k and n samples that are very different, distinct or inconsistent with respect to the remaining data. The problem of defining outliers is no small feat, especially for various samples. Data display methods that are useful for finding out of one to three dimensions are weak in multidimensional data due to the lack of sufficient statistical samples and external visual detection.

By Plotting Red Magnitude and Center of Galaxy vs Chi square Values, we found that we can draw chi cut line at where x is 5, so now can detect the outliers of these features.

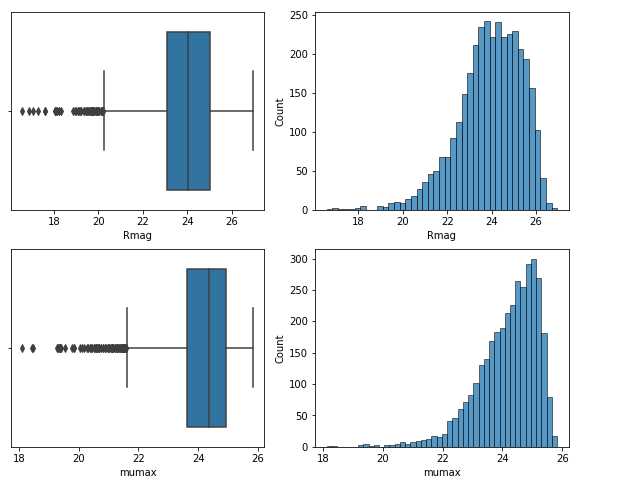


Figure 5 Outlier Detection using scatter plot by plotting chi square cut

Here we detect outliers of Red Magnitude and Center of Galaxy with the help of Box Plot and Histogram. And we found that Values of Rmag <20 are outliers. But as per our dataset these values are important for us so we cannot consider these values as outliers.

Figure 6Outer Detection by Box Plot and Histogram

### Correlation

A broad-term relationship is a measure of the relationship between flexibility. Typically, the term affiliation is used in the context of the linear relationship between 2 continuous variants and is expressed as a temporary affiliate product of Pearson. The Pearson coefficient of integration is usually used for shared data (data following normal bi variate distribution). With continuous data that is no longer distributed in the normal way, of ordinal data, or data with qualified outsiders. Both correlation coefficients are measured in terms of 1 to + 1, where 0 indicates no line or monotonic correlation, and the relationship is strong and reaches a straight line (Pearson connection) or an ever-increasing curve or decreasing. As the coefficient approaches the total value of 1.

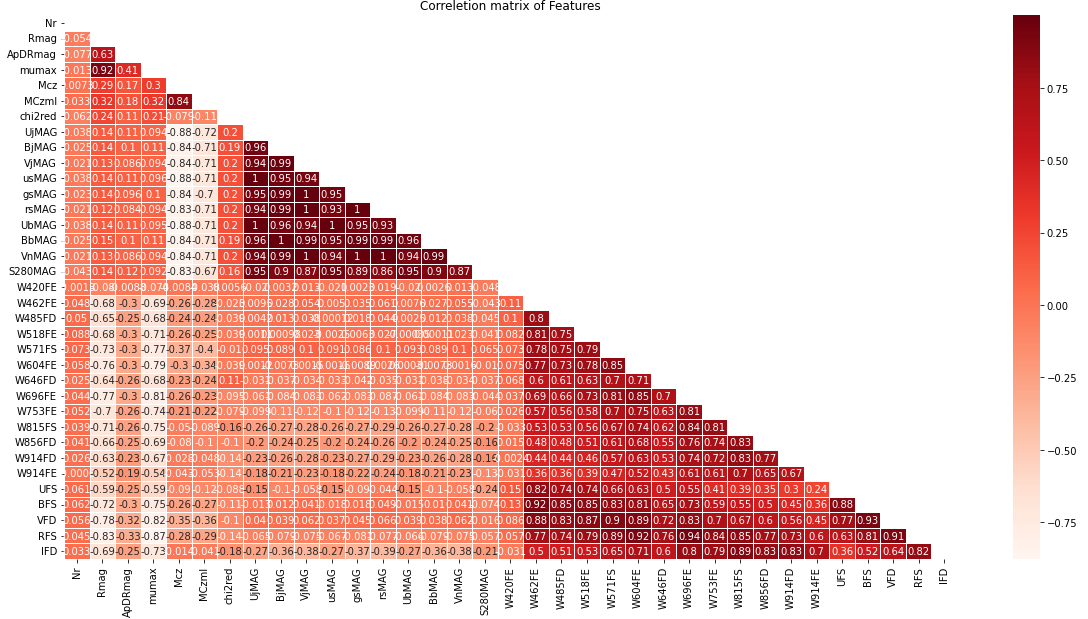


Figure 7. Correlation Matrix

Here by applying correlation on dataset we found our main correlated features which are highly co-related to each other, there **are RMag, ApdRmag, mumax and Values that contains Magnitude.**

### Data Reduction

Data reduction techniques are used to obtain a reduced representation of a very small data set in volume but which closely maintains the integrity of the original data. That is to say, mining in a reduced data set should work very well but work in the same way or almost the same as the results of the analysis.

### Principle Component Analysis

Big data sets are becoming more common and often difficult to do interpret. Key Component Analysis (PCA) is a method reducing the size of such data sets, increases, interpretation but at the same time reduces the loss of information. It does so by creating new unrelated dynamics in sequence increase diversity. Discovering such new flexibility, principal components, reducing eigenvalue / eigenvector resolution problem, and new variables are defined by existing databases, not a priori, which is why PCA analyzes dynamic data technology. Analysis of the main part of the data matrix is ​​extracted dominant patterns in the matrix according to the corresponding set of points and loading sites. It is the responsibility of the data analyst to make an existing science issue about PC speculation, PLS retreats, etc.

The results of the Principle analysis depend on the measurement of matrix, to be clarified. A different measure, in each case variability is measured so that the unit variance, can be recommended normal use, as long as there are still unchanged variables unrated. Combining different types of variable warranties block measurement.

Example: eigenvalue of Combo-17 Dataset

|  |  |  |  |
| --- | --- | --- | --- |
| 1.39292685e+01 | 1.11414716e+01 | 2.11094633e+00 | 1.22167798e+00 |
| 1.00382040e+00 | 9.55928964e-01 | 9.41419208e-01 | 4.71139128e-01 |
| 3.76207254e-01 | 3.46442994e-01 | 3.15333554e-01 | 2.79382861e-01 |
| 2.49460955e-01 | 2.28580805e-01 | 2.16686736e-01 | 2.09305239e-01 |
| 1.71258484e-01 | 1.50462554e-01 | 1.35875919e-01 | 1.26779632e-01 |
| 9.96222132e-02 | 7.43785446e-02 | 5.52091899e-02 | 5.25385951e-02 |
| 4.14647116e-02 | 3.21297852e-02 | 2.59103828e-02 | 1.93461429e-02 |
| 1.64712189e-02 | 1.30330261e-03 | 1.26998339e-04 | 4.10624122e-05 |
| 3.46525191e-06 | 2.84614235e-06 | 2.38926782e-06 |  |

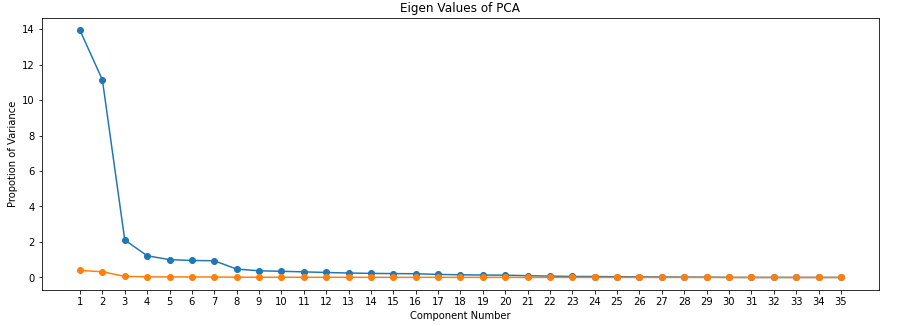
 Table 3 Find eigenvalues of Datasets

Figure 8. Scree Plot for Eigen Values (PCA)

* 1. Data Mining Techniques

Bulk of data are generated daily on most organizations. To pull out hidden predictive information from that bulk/large data volumes, data mining methods (DM) are required. Organizations are beginning to see the importance of data mining in their strategic plans and the effective use of DM strategies can be of great benefit to organizations. This paper discusses the needs and challenges of DM, and describes the main DM strategies such as mathematics, practical wisdom, decision-making method, genetic algorithm, and visualization.



### K-Mean Clustering Algorithm

K-means combining the most popular and unattended machine learning algorithm.

Typically, unsupervised algorithms perform predictions from data sets using input vectors without reference to known, or labeled results. Collection refers to a collection of data points that are grouped together for specific reasons. We needed the k, centroid number on the website. Representing collection center of real place is a centroid. . All data points are assigned to each collection by reducing the total number of squares in the cluster. The K-means algorithm detects the k number in inches, and then delivers each data point to the nearest collection. ‘Methods’ in K-methods mean data rate; that is, to find a centroid.



Figure 9. K-mean Clustering

**By Applying K-Mean We Found Below Clusters:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Clusters 1** | **Clusters 2** | **Clusters 3** | **Clusters 4** |
| -0.34205321 | 0.06839859 | 0.27148294 | -0.06956034 |
| 1.05374332 | 0.37537188 | 0.0986479 | -0.00426436 |
| -0.13789675 | -0.2660048 | -0.01482924 | , 0.08575043 |
| -0.44064723 | 0.05104455 | -0.21551508 | -0.02555367 |
| 1.29058758 | -0.0416233 | -0.25966151, | -0.05548346 |
| -0.11916294 | -0.17109548 | -0.35355228 | 0.01485758 |
| 0.19999844 | 2.20884842 | -0.00762482 | , 0.87655921 |
| -0.11744255 | -0.26050434 | 0.32466503 | , 0.01837362 |
| -0.13981448 | 0.3471124 | -0.15774641 | , -0.1207312 |
| 0.39739687 | 0.54025413 | 0.08652383 | -0.07759182 |
| 1.53506752 | -0.22671654 | 0.20076071 | 0.03288343 |
| -0.47211538 | 0.73017693 | 0.1657628 | , 0.08254285 |

Table 4 Applying K-mean Algorithm find out clusters



### DENSITY BASED CLUSTRING

DBSCAN is a compacting algorithm based on congestion that assumes that clusters are dense regions of space divided by low density regions. It combines data points ‘compactly consolidated’ into a single collection. It can detect clusters in large data databases by checking the local density of data points. Epsilon is a circular radius that should be created near each data point to check congestion and minutes Points the minimum number of data points needed within that circle so that that data point is classified as an important point.

At higher intensity the circle becomes a hyper sphere, the epsilon becomes the radius of that hyper sphere, and min Points the minimum number of data points required within that hyper sphere.

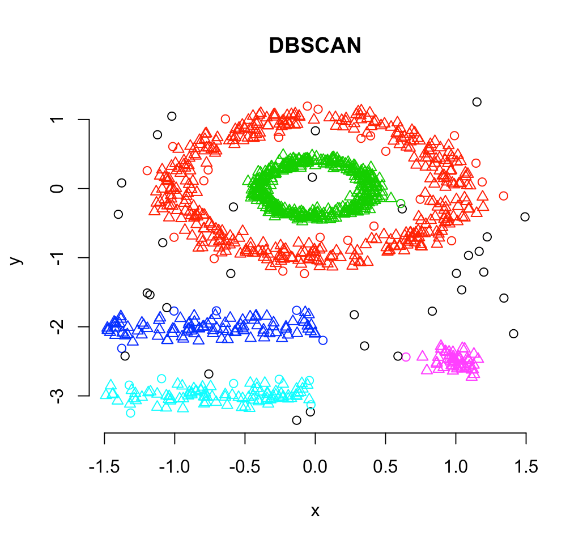


Figure 10DB Scan Clustering Algorithm

In our DBSCAN result we found noise, epsilons and mid-point.

|  |  |  |
| --- | --- | --- |
| -0.1 | -0.3539114299191538 | 9 |
| -0.2 | 0.4463249210248632 | 9 |
| -0.3 | 0.48729031565098124 | 9 |
| -0.4 | 0.588990323411522 | 9 |
| -0.5 | 0.644614294577061 | 9 |
| -0.6 | 0.644614294577061, | 9 |
| -0.7 | 0.67140874334388 | 9 |
| -0.8 | 0.6895696359800023 | 9 |
| -0.9 | 0.7373941889146873 | 9 |

Table 5 Applying DBSCAN Algorithm and find out noise, epsilons and mid-point



### SUM OF SQUARED ERROR (SSE)

SSE is total value of the square error between each observation and your team description. It can be used as a measure of diversity within a collection.

We found the clusters of sum of squared error

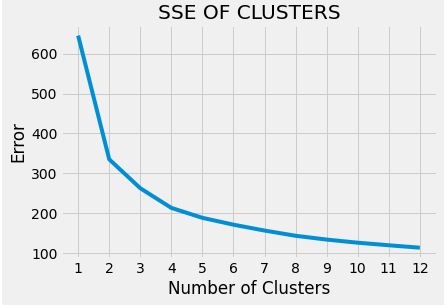


Figure 11 SSE PLOT OF K-MEAMNS

|  |
| --- |
| 645.1405624207305 |
| 335.09918043155517 |
| 262.64673108524096 |
| 213.32982824470855 |
| 188.4204850052321 |
| 171.26649414711633 |
| 156.54979905394973 |
| 143.4560309323208 |
| 133.95273083085638 |
| 126.08284092188995 |
| 119.66912186609275 |
| 113.57505616032562 |



### Silhouette Score

The silhouette plot is used to calculate goodness of clustering technique. Its display to measure of how close each point of one cluster into point the neighbor cluster and its range is -1 to +1.

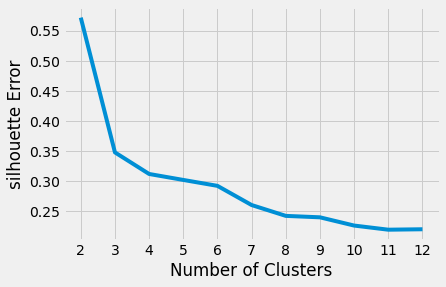
**We find out Silhouette score**

Figure 12Silhouette score on K-Means

|  |
| --- |
| 0.5718575504256521 |
| 0.34782245829984765 |
| 0.3119732553827454 |
| 0.30205607502469584 |
| 0.29221420197454934 |
| 0.2603393016055867 |
| 0.24231111328942415 |
| 0.23992229350458197 |
| 0.22622415319293399 |
| 0.21934583941458807 |
| 0.22007537191936313 |
| 0.5718575504256521 |

### Support Vector Machine

SVM (Support vector machine) are a set of integrated methods of supervised learning, which work on both classification and retraining problems.The hyperplane solution parameters were taken from the quadratic system development problem.



Figure 13. Support Vector Machine

After applying support vector machine algorithm we find out score of Regularization and Gamma.

|  |  |  |  |
| --- | --- | --- | --- |
| **Regularization** | **Parameters** | **Gamma** | **Parameters** |
| 1 | 0.9829059829059829 | 1 | 0.9786324786324786 |
| 2 | 0.9871794871794872 | 2 | 0.9786324786324786 |
| 3 | 0.9786324786324786 | 3 | 0.9743589743589743 |
| 4 | 0.9829059829059829 | 4 | 0.9658119658119658 |
| 5 | 0.9829059829059829 | 5 | 0.9572649572649573 |
| 6 | 0.9829059829059829 | 6 | 0.9444444444444444 |
| 7 | 0.9786324786324786 | 7 | 0.9444444444444444 |



### Stochastic Descent Gradient Pipelines Classifier

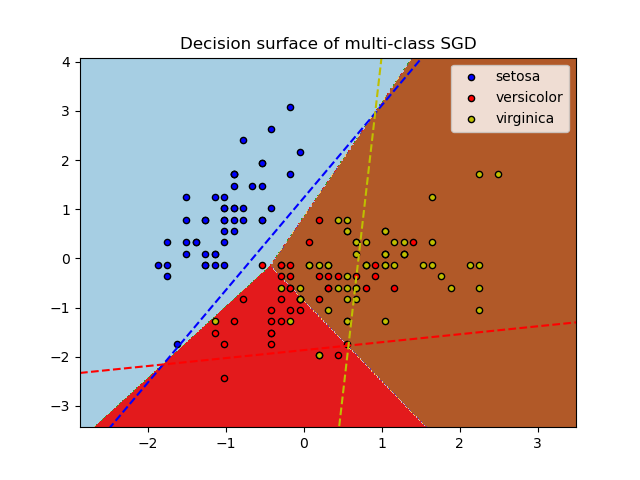
The Stochastic Gradient Descent (SGD) has gained a reputation for solving large-scale machine learning problems. It provides a quick way to reduce the number of losses and is effective in Vector Support (SVM) and Logistic efficiency. It is generally found to provide significant reductions in training time without sacrificing precision. (SGD) can be used for normal convex loss activities with routine goals that have a positive effect. This has become a widely used method and is used to divide a set of training into non-compliant pieces that include a new training set and a validation set. Then for each training passing the new training sets one test set of the validation set and that number of duplicate training data that is found to be valid is recorded.

Figure 14Stochastic Descent Gradient Pipelines Classifier

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **precision** | **recall** | **f1-score** | **support** |
| 0 | 0.88 | 0.91 | 0.90 | 47 |
| 1 | .97 | 0.97 | 0.97 | 30 |
| 2 | 0.97 | 0.96 | 0.97 | 157 |
| **accuracy** |  |  | 0.95 | 234 |
| **macro average** | 0.94 | 0.95 | 0.94 | 234 |
| **weighted average** | 0.95 | 0.95 | 0.95 | 234 |

**Confusion Matrix**

|  |  |  |
| --- | --- | --- |
| 41 | 1 | 3 |
| 0 | 29 | 1 |
| 6 | 0 | 151 |

### Random Forest Classifier

Random Forests is a mathematical algorithm and machine learning tool for predicting forest trees. Random forests is a combination of tree predictions so that each tree depends on the randomly calculated vegan values and the uniform distribution of all the trees in the forest. A common forest error involves the number of trees in a forest becoming larger. Using random selection of features to separate each node produces error rates that are well compared to Ada-boost. It measures error, power, and correlation and this is used to identify features to increase the amount of feedback used in classification.

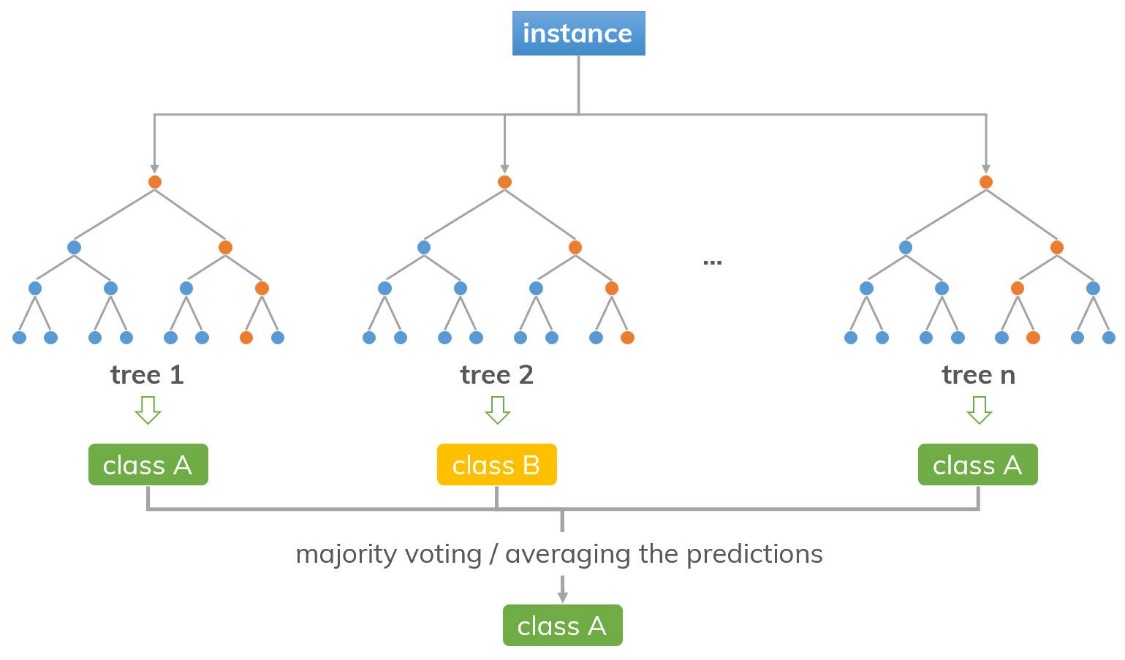


Figure 15 Random Forest Classifier

We applying Random forest algorithm we find accuracy of our model.

Score of Random Forest: 0.9658119658119658

Accuracy for Random Forest: 0.9658119658119658

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **precision** | **recall** | **f1-score** | **support** |
| 0 | 0.91 | 0.95 | 0.93 | 73 |
| 1 | 1.00 | 0.94 | 0.97 | 51 |
| 2 | 0.98 | 0.98 | 0.98 | 227 |
| **accuracy** |  |  | 0.97 | 351 |
| **macro average** | 0.96 | 0.95 | 0.96 | 351 |
| **weighted average** | 0.97 | 0.97 | 0.97 | 351 |

Table 9 Apply Random forest algorithm and find Accuracy



### Confusion Matrix

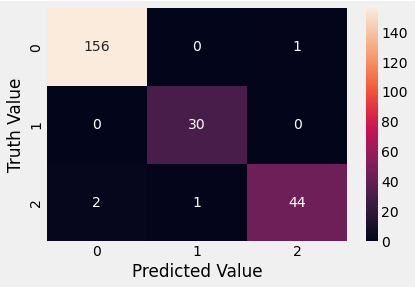
A comprehensive method of machine detection test is to classify predictable and actual decision-making categories into a confusion matrix, also called an error matrix. A classification tool that does not consider distribution parameters but only information contained Predicted and true class membership values can be categorized and calculated by a confusion matrix. If the success of a class divider is measured in error rate, confusion matrices can be used to analyze and compare dividers.

Figure 16 Confusion Matrix

# Results

We Analyzed this large data set (contains on 63,501) objects. Of brightness measure in 17 band. First we had to find out the outliers and clean the data so we applied Data pre-processing. We know that in this dataset clustering is already applied, so tried clustering by using Principle Component Analysis (PCA),

By applying PCA we reduced dimension from 34, to 4. On 5 PCA.

We applied classical Clustering technique like K-mean and modern DB scan, by applying kneed on K-mean we found that 4 Clusters would be enough for further calculations, so by using guess of 4 clusters we, applied further techniques.

We applied SVM and on 4 clusters we found the result of regulation and gamma with the help of SVM (support vector machine algorithm). Find out the accuracy of regulation is 98.7% on point 2 and gamma accuracy is 97.8% on point 2.

We also applied The Stochastic Gradient Descent (SGD) and found accuracy of 94% on 4 classes.

We finally applied Random Forest algorithm is found 96% accuracy.

# References

(Barnett & Lewis, T., 1994)

(N. Ghadiri,, M. Ghaffari,, & M. A. Nikbakht,, 2007)

(Bazaraa M. S., Sherali H. D.,, & Shetty C. M., 1993)

(D.P., 1995)

(Chang C.-C. & Lin C.-J., 2001)

(Chew H. G., Lim C. C,, & Bogner R. E., 2003)