Project Report

mcs final 2021

**Course Supervisor: Dr Tehseen Ahmed Jilani**

1. **means Clustering and PCA**

**Data Warehousing and Data Mining (CS-626)**

UBIT KARACHI UNIVERSITY

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | Group Members 1. (P19101039) Muhammad Rizwan  2. (P1810028) Muhammad Humayun  3. (P1510010) Aasadullah Siddiqui  4. (P19101030) Khuda bakhsh | mrizwan555666777@gmail.com |
|  |  |  |

Table of Contents

1. **Table of contents**
2. **Abstract**
3. **Data Mining**

* **3.1 Introduction**
* **3.2 Knowledge Discovery and Data Mining**
* **3.3 Kind of Data that be mined**

1. **Supervised Data and Unsupervised Data**

* **4.1 What is supervised data in data mining?**
* **4.2 What is unsupervised data ?**
* **4.3 Difference between Supervised and Unsupervised.**

1. **Principal Component Analysis (PCA)**

* **5.1 Working Principle for the PCA**

1. **K-Mean Clustering Analysis**

* **K-Mean Working Principal**

1. **Data Detail.**
2. **Data Mining.**
3. **Result and Conclusion.**

# 2 Abstract

This report is all about the details of some selected methods like PCA and K-mean some selected data in the data science. Although data science is combination of different fields computer science and statistics which are used to discover different kinds of patterns from the information collection. Today’s real world there is large amount available but we have to find some useful information to check about the behavior like customers for achieving the particular task into an understandable structure for future use. There are many methods and techniques but here only two methods used to carry out data mining successfully. Main objective of data mining is that finding useful information or knowledge having significant role in human activities by processing of large amount of data from different perspective. Data mining is essential part of human life because it extracts important information or knowledge from large bank of data from daily life or real world and summarize it into useful patterns showing behavior of anything predicting whole picture about the given data. There is a significant advancements in Statistics, Machine Learning, Artificial Intelligence, Pattern Recognition and computation capabilities have evolved the present day data mining application and use of this almost in every field of life like business, education, medical and scientific showing best information for the future use.

# Data Mining Project Report

**3.1 Introduction to Data Mining**

Data mining is a process of finding information or knowledge from large volume of data for useful purpose. Data mining is evolved according to the technologies and it is combination of Artificial Intelligence, Machine Learning and computation capabilities. In data mining these are following things to be discover and change it to useful information like noisy data, missing values, dynamic data, and relationship between the data, spare and dynamic data. When there is noise and missing values in the data then it is imprecise and the data structure is complex. Now a days there is popular and well-known tool for analyzing the large datasets. Data mining is simple method of solutions when there is statistical techniques becomes fail and it extracts effective and efficient of particular information for important assets for management.

3.2  **Knowledge Discovery and Data Mining**

A generic KDD process typically consists of six different stages Domain Understanding, Data Understanding, Data Preparation, Data Mining, Evaluation and Knowledge Consolidation. For each of these, a general description will be given Domain Understandingconcerns familiarizing oneself with the domain at hand. This entails a sufficient comprehension of the current state of affairs, on the problems therein, and on the goals that have to be reached. Data Understanding concerns the data analyzing, thereby inspecting its quality. This involves the identification of anomalies, such as noise, outliers, and missing values, as well as the selection of interesting subsets or features. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision making. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

1. Data cleaning: It is also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.

2. Data integration: In this stage, multiple data sources, often heterogeneous, may be combined in a common source.

3. Data selection: At this step, the data relevant to the analysis is decided on and retrieved from the data collection.

4. Data transformation: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.

**3.3 Kind of data that can be mined.**

**Flat files**: Flat files are actually the most common data source for data mining algorithms, especially at the research level. Flat files are simple data files in text or binary format with a structure known by the data mining algorithm to be applied. The data in these files can be transactions, time-series data, scientific measurements, etc.

**Relational Databases:** a relational database consists of a set of tables containing either values of entity attributes, or values of attributes from entity relationships. Tables have columns and rows, where columns represent attributes and rows represent tuples. A tuple in a relational table corresponds to either an object or a relationship between objects and is identified by a set of attribute values representing a unique key.

**Transaction Databases:** A transaction database is a set of records representing transactions, each with a time stamp, an identifier and a set of items. Associated with the transaction files could also be descriptive data for the items.

**Multimedia Databases:** Multimedia databases include video, images, and audio and text media. They can be stored on extended object-relational or object-oriented databases, or simply on a file system. Multimedia is characterized by its high dimensionality, which makes data mining even more challenging.

**Spatial Databases:** Spatial databases are databases that, in addition to usual data, store geographical information like maps, and global or regional positioning. Such spatial databases present new challenges to data mining algorithms.

**Time-Series Databases:** Time-series databases contain time related data such stock market data or logged activities. These databases usually have a continuous flow of new data coming in, which sometimes causes the need for a challenging real time analysis. Data mining in such databases commonly includes the study of trends and correlations between evolutions of different variables, as well as the prediction of trends and movements of the variables in time.

**World Wide Web:** The World Wide Web is the most heterogeneous and dynamic repository available. In WWW there is organized data available through inter connected documents.

**4. Supervised and unsupervised data in data mining**

Data mining makes use of a plethora of computational methods and algorithms to work on knowledge extraction. Classification is perhaps the most basic form of data analysis. A common task in data mining is to examine data where the classification is unknown or will occur in the future, with the goal to predict what that classification is or will be. Similarly, data where the classification is known are used to develop rules, which are then applied to the data where the classification is unknown. That being said, the techniques of data mining come in two main forms: supervised and unsupervised. Supervised is a predictive technique whereas unsupervised is a descriptive technique. Although both the algorithms are widely used to accomplish different data mining tasks, it is important to understand the difference between the tow.

4.1  **What is Supervised Data Mining?**

Supervised data mining, as the name suggests, refers to learning algorithms that are used in classification and prediction. Supervised algorithm learns from the training data which is labeled and the task is controlled by the knowledge engineer and system designer. With supervised data, we must have known inputs corresponding to known outputs, as determined by domain experts. The data mining task is often referred to as supervised learning because the classes are determined before examining the data. This technique uses an objective function and a set of data elements which are independent variables. Supervised technique attempts to identify relationships between dependent and independent variables, identify the degree of correlation for each set of variables, and build a model showing the web of dependencies. The model is then applied to the data for which the target value is unknown.

4.2  **What is Unsupervised Data Mining?**

Unlike supervised technique, unsupervised data mining does not have a predetermined objective function, nor does it predict a target value. Unsupervised techniques are those where there is no outcome variable to predict or classify. Hence, there is no learning from cases where such an outcome variable is known. The algorithm requires the user to specify the number of intervals and/or how many data points should be included in any given interval. It helps you identify all kinds of unknown patterns in data. Unsupervised model is also called descriptive model because it looks for unknown patterns in a data set with no predetermined labels and with no or minimal human supervision. Unsupervised learning methods include clustering, association, and extraction methods. This type of learning technique is used when a specific goal is not available or when the user seeks to find hidden relationships in data.  
Supervised learning is the data mining task of using algorithms to develop a model on known input and output data, meaning the algorithm learns from data which is labeled in order to predict the outcome from the input data. Supervised technique is simply learning from the training data set. Unsupervised learning, on the other hand, is the technique of using algorithms where there is no outcome variable to predict or classify, meaning there is no learning from cases where such an outcome variable is known.

4.3  **Difference between Supervised and Unsupervised data in data mining?**

1. The main difference between the supervised and unsupervised is the use of labeled data sets. In simple way supervised data mining includes labeled data for input and output but unsupervised does not use labeled data for input and data.

2. Supervised data mining also known as predictive or directed and unsupervised data mining can be briefed as descriptive or undirected.

3. When there is specific target value for prediction about any data set then we use supervised data mining techniques. When there is not predetermined attributes not any predict a target value then we use unsupervised data mining techniques.

4. In supervised data mining techniques there are following some specific demonstrations including Classification, Regression, and Anomaly Detection while the Unsupervised having Clustering which is most open ended technique, Association and Feature Extraction.

**5. Principal Component Analysis (PCA)**

Principal Component Analysis (PCA) is a feature extraction method for detection of the variance of the data. It generates projections that are scaled with the data variance. PCA is in simple word that it is statistical approach that changes a set of observations of possibly correlated variables into a set of values of linearly independent variables called Principal Components. Normally PCA is used to simplify the data, reduce noise, and find unmeasured latent variables. It standardizes range of continuous initial variables. It identifies correlations in computing the covariance matrix. It computes the eigenvectors and eigenvalues of the covariance matrix to identify the principal components. There are following steps we use in PCA given below:

1. Standardization
2. Covariance matrix computation
3. Compute the eigenvectors and eigenvalues of covariance matrix to identify PCA
4. Vector features
5. Recasting of the given data along with the PCA

Principal Component Analysis summarizes the information content in large data tables by means of a smaller set of summary indices that can be easily visualized and analyzed. The processed data can describe production samples, chemical compounds or reactions, process, batches from a batch process, biological individuals or trials. The basic idea behind this technique is that it combines highly correlated variables together to form a smaller number of an artificial set of variables which is called principal components that account for most variance in the data. Now this method we applied the selected data and we have taken the results which are attached below which shows how we process data and segregate according to requirements and conclude results. Next we are discussing another similar method of unsupervised data mining method which is K-mean method.

**6. K-Mean Clustering Method**

K-mean clustering is the simplest and well known method of unsupervised machine learning algorithms. Typically, in unsupervised algorithms having inferences from datasets using only input vectors referring to popular or labelled, outcomes. A cluster is a collection of data points aggregated together because there is lot of specific similarities in datasets. I’ll define a certain number k, which shows the number of centroids required in the dataset where centroid is the imaginary or real location showing the center of the cluster. Each data point is assigned to each of the clusters through reducing the in-cluster sum of squares. In simple words, the K-mean algorithm identifies k number of centroids, then it assigns each data point to the nearest cluster but centroids should be small as possible. The word mean in the method of K-mean is showing the averaging of the dataset which finds the centroids.

6.1 **Working Principal of K-mean Algorithm**  K-mean algorithms works on the dataset centroids which are selected randomly selected from the dataset for first centroids as a reference for the rest of the dataset clustering. This first centroid which is taken from the randomly selected group of the dataset is the beginning points for every cluster, and then performing the iterative steps or repetitive calculations to optimize the positions of the centroids. This process will be stopped if there are two following conditions meets give below

1. If centroids have stabilized because there is no change in their values due to clustering has been successful.
2. In the K-mean method there is defined number of iterations when this number achieved then stops creation and optimization of clusters.

**7. Data Detail**

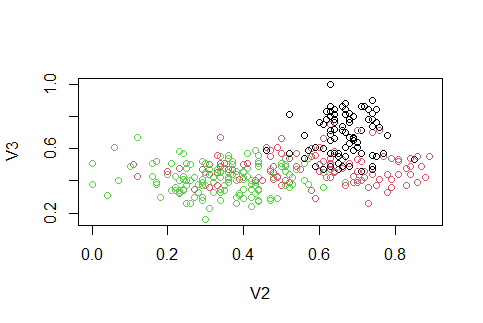
Dataset consist of 336 samples from each of 8 species (cp , im , imL , imS , imU , om , omL , pp)

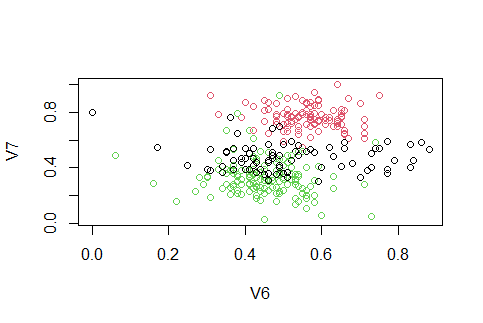
Seven features were measured from each sample

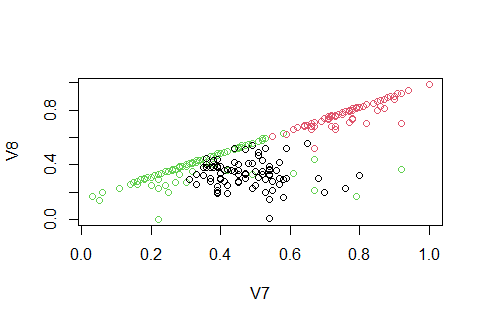
(V2 , V3 , V4 , V5 , V6 , V7 , V8) and based on the combination of these features

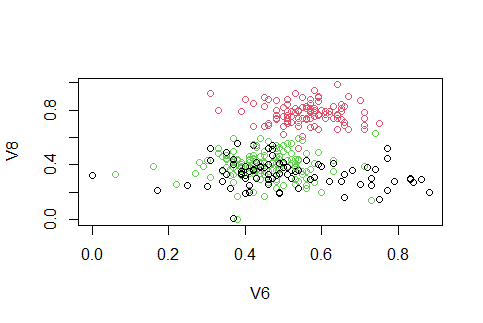
**8. Data Mining / Visualization / Analytics**

**K-means Clustering :**

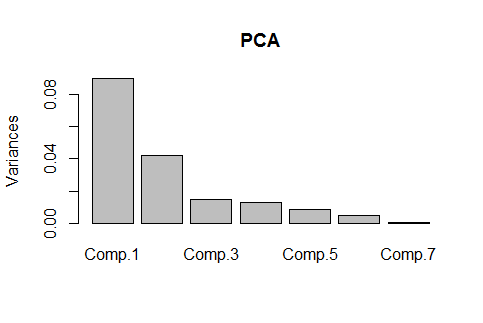
****

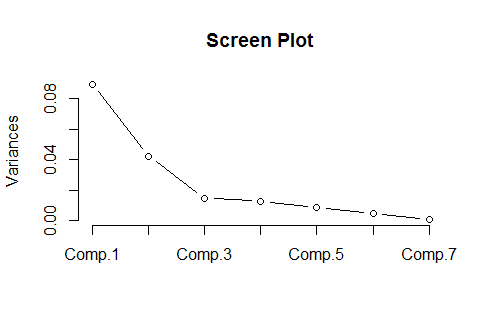
****

****

****

**PCA :**

****

****

**9. Result and Conclusion**

In this project, we have made a survey on work carried out by different researchers using K-means clustering approach. We also discussed the evolution, limitations and applications of K-means clustering algorithm. It is observed that a lot of improvement has been made to the working of K-means algorithm in the past years. Maximum work carried out on the improvement of efficiency and accuracy of the clusters. This field is always open for improvements. Setting appropriate initial number of clusters is always a challenging task. At the end it is concluded that although there has been made plenty of work on K-means clustering approach, there is a scope for future enhancement.