**Instructor: DR TEHSEEN JILANI**

GROUP MEMBER:

1. Syed Osama Ali -Ep:20101054
2. Muhammad Janshair Khan - Ep:20101035
3. Hassan Moeed -Ep:20101017
4. Tabish Idrees -Ep:20101057

GitHub Links:

Repository: https://github.com/SyedOsamaAli007/Data-Mining-Project

CODE: https://github.com/SyedOsamaAli007/Data-Mining-Project/blob/main/Data\_Mining\_Project.ipynb

Data: https://github.com/SyedOsamaAli007/Data-Mining-Project/blob/main/diabetes.csv

Table of Contents

[1. Introduction](#_26in1rg)

[1.1. Data](#_gjdgxs)

[1.2. Data Warehouse](#_30j0zll)

[1.3. Why Do We Use Data Mining](#_1fob9te)

[1.4. Steps Involved In Data Minning](#_lnxbz9)

[2. Literature Review](#_3znysh7)

[2.1. Overview](#_2et92p0)

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

[3.1. Data Pre-Processing](#_35nkun2)

[3.1.1. Data Integration](#_3dy6vkm)

[3.1.2. Null Testing](#_1ksv4uv)

[3.1.3. Min Max Normalization](#_44sinio)

[3.1.4. GROUP BY](#_4d34og8)

[3.1.5. DATA REDUCTION](#_2jxsxqh)

[3.1.5. PAIR PLOT](#_z337ya)

[3.2. Data Mining Techniques](#_17dp8vu)

[3.2.1. K-Mean Clustering Algorithm](#_3rdcrjn)

[3.2.2.](#_3j2qqm3) **Error! Bookmark not defined.**

[3.2.3. LOGISTIC REGRESSION](#_1y810tw)

[4. Results](#_4i7ojhp)

List Of Figure

[**Figure1: Step involved in data mining**](about:blank)

[**Figure2: Step involved in data preprocessing**](about:blank)

[**Figure3: Null testing**](about:blank)

[**Figure4: Min-Max Normalization**](about:blank)

[**Figure5: Group By**](about:blank)

[**Figure6: Pair plot**](about:blank)

[Figure7:K-mean classification](#_2xcytpi)

[**Figure8: Decision tree**](about:blank)

[**Figure9: Logistic regression**](#_1ci93xb)

***DIABETES PREDICTION***

**1. INTRODUCTION:**

In this project we are going to check out the impact of different features on human’s diabetes. Diabetes is increasing at an alarming rate and has been recognized as one of the leading causes of death and disabilities worldwide. Mostly people thought that age is the main factor which causes diabetes diseases, but in the real world age is not the only factor which cause such a type of disease, although there are various factors that causes this type of diseases.

Recent estimate indicate that in the year of 2017, more than 400 million people are living with this disease and this figure was predicted to increase by approximately 200 million in 2035. However in 2017, 3.2 million deaths are attributable to diabetes in people aged between 60 to 99 years. These dataset raise an urgent need to further assess the impact of diabetes on the lives of patients.

So, In this project we are going to consider various factors for testing diabetes such as blood pressure test, heart rate, gender, chest pain, cholesterol, diabetes, smoking, obesity, stress, depression, age etc. in our project, we have taken the data of heart disease prediction from the Kaggle. In this data we consider the following features, blood pressure, chest pain, heart rate, gender, fat levels, angina, old level and also the report of ECG. Our dataset contains both numeric and alpha-numeric data. We made our project by using python Jupyter Notebook. In our project we perform clustering, principal component analysis, Min-Max normalization Algorithm, logistic regression and decision tree. On the basis of these techniques we predict results.

“Sometimes people think that data can answer everything and sometimes we have to acknowledge that is simply untrue.”

* 1. *Data*

We are very much familiar the word Data which means “to give”. Data is really give “raw facts”. This data can be includes (observations, pictures, videos, measurements, numbers, genders, words and much more).

Sometime data and information are interchangeably used, but the question is that is this interchangeability correct? Some writers use it interchangeably. Which is not considered as standard.

*Data-Analysis:*

Data analysis is a

1. Collection
2. Transformation
3. Organization

Here we arrange data in order for conclusions then predict and drive informed decision-making. And it does not stop there:

Data evolve over time, which mean this analysis or analytics as we collect can give us new information.

* 1. **Data Warehouse**

As we know that Data is a collection of rows facts and figure:

After this the data goes to database which help to generate Data warehouse.

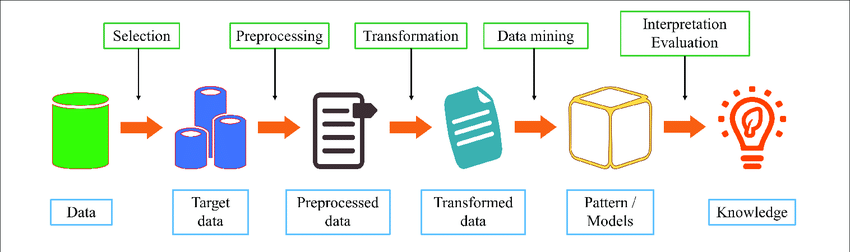
Data-Warehouse basically “Collection of different kind of database into particular format OR style.” Data warehouse is the science of the Data.

As we know that Data-warehouse is:

1. Subject Oriented 2) Integrated
2. Time Variants 4) Non-Volatile
   1. **Why Do We Use Data Mining?**

Data is everywhere we use and create data every day. 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.

* 1. **Steps Involved In Data Mining**

**Figure1: Step involved in data mining**

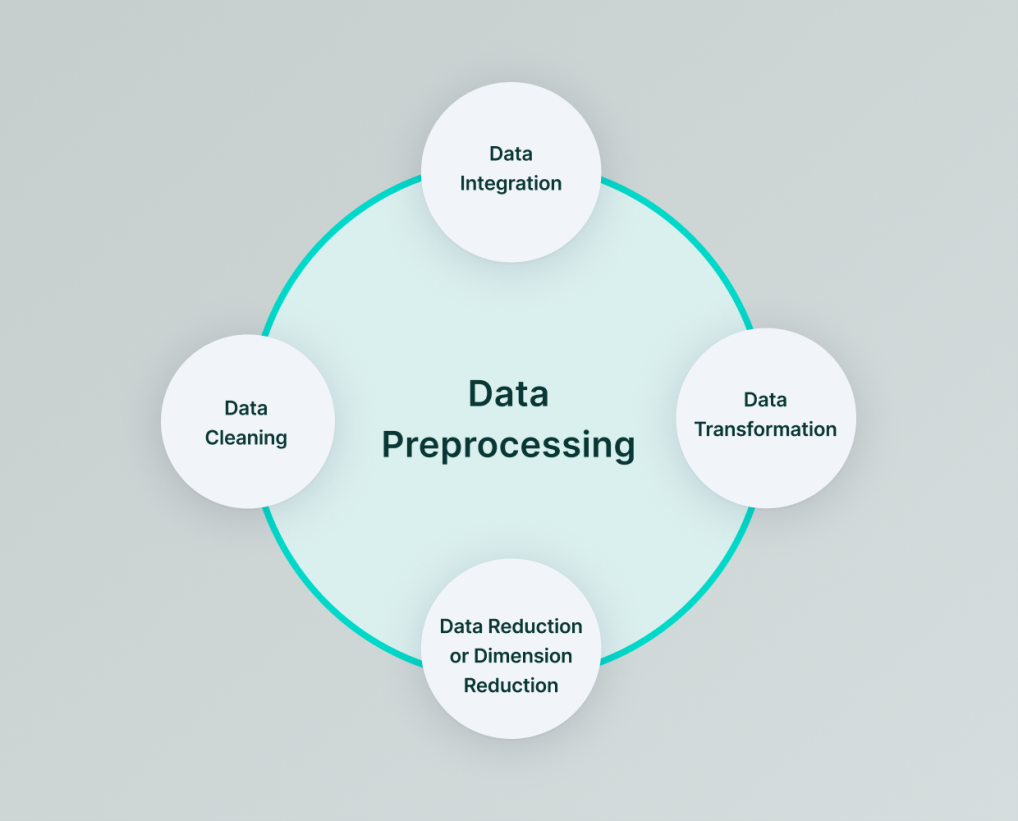
# Literature Review

* 1. Overview

In this Project, we will be elaborating the given dataset that whether the patient has diabetes or not on the basis of the Machine Learning Algorithm and for that we used the [**Pima Indians Diabetes Database**](https://www.kaggle.com/uciml/pima-indians-diabetes-database)for our data mining project.

# 3.1 Data Pre-Processing and Data Mining Techniques.

# 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.

 **Figure2: Step involved in data preprocessing**

### Data Integration

As we know Data Integration is a data preprocessing technique that will be combining data from multiple heterogeneous and data sources into a coherent data store and provide a unified view of the dataset.

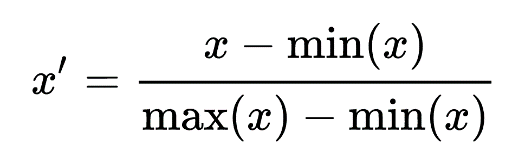
**NULL TESTING:**

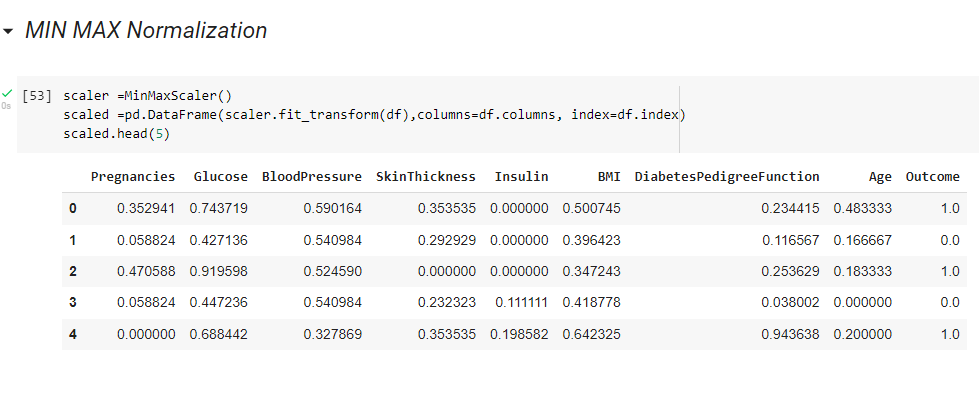
We perform null testing on our dataset so that we can assure there are no null values in any feature ocolumn of the data. You can see below results of our test that no feature column have any single null value. It can be very odd for any data mining algorithm if data set contains any null value in any feature so it’s better to check the data is not nullified from any portion.

 **Figure3: Null testing**

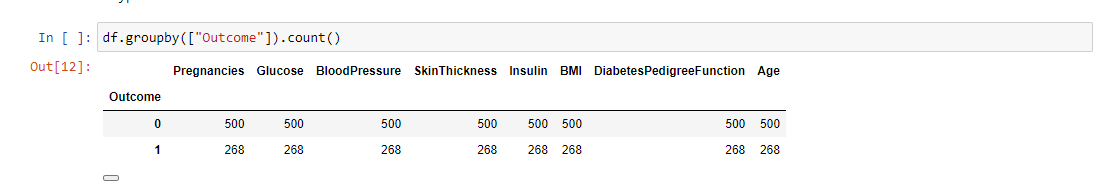
### Min Max Normalization

Min-max normalization is one amongst the foremost common ways to normalize data. it's accustomed scale the information of an attribute in order that it falls when in a smaller range, like -1.0 to 1.0 or 0.0 to 1.0.  
When multiple attributes are having values on different scales, this could result in poor data models for data processing operations. So, they're normalized to bring all the attributes on same scale.



**Figure4: Min-Max Normalization**

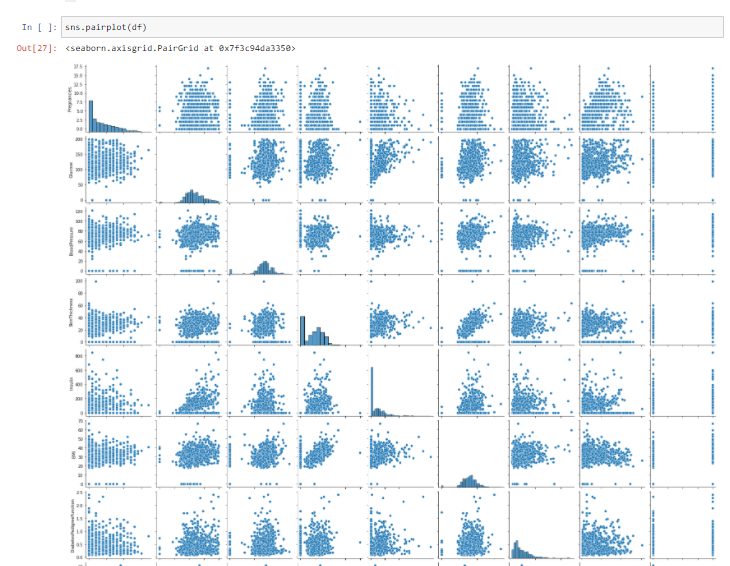
**GROUP BY**

** Figure5: Group By**

### DATA REDUCTION

Data reduction techniques are accustomed obtain a reduced representation of a really small data set in volume but which closely maintains the integrity of the first data. that's to mention, mining in a very reduced data set should work all right but add the identical way or almost the identical because the results of the analysis.

**PAIR PLOT**

A pair-plot plot a pairwise relationship in a very dataset. The pair-plot creates a grid of Axes such that each variable in data will share the y-axis across one row and share the x-axis across one column. This makes plots as shown below. Related course: Matplotlib **Figure6: Pair plot**

* 1. **Data Mining Techniques**

Data mining is an iterative process. OR

It is the effort of humans and system like Computer and etc.

Basically there are two main goals of data mining which are following.

1. **Predictive Data mining:**  This produces the models of the system described by the given dataset.
2. **Descriptive Data Mining:** Describes your data with produces new nontrivial information based on the available dataset.

### K-Mean Clustering Algorithm

K-Mean is an Unsupervised Machine leaning algorithm. 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. Aims of this algorithm to partition n observation into K cluster.

K-Means clustering is one of the hierarchical clustering method which identifies clusters on basis of centroids or we can say it is a distance based algorithm. The clustering is done by minimizing sum of squares of distance between data and the corresponding cluster centroid.

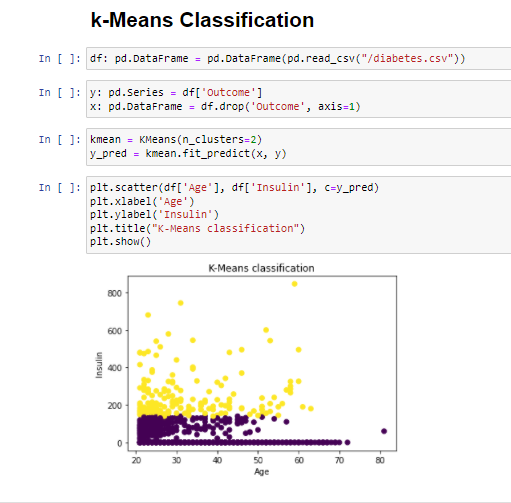
Hierarchical clustering determines cluster assignments by building a hierarchy. This is implemented by either a top-down or a bottom-up like approach. This methods produce a tree-based hierarchy of points called as a Dendrogram. Similar to partition clustering, in hierarchical clustering the number of clusters (k) is often pre-determined by the user. Clusters are assigned by cutting the Dendrogram at a specified depth that results in k groups of smaller Dendrograms.

The strengths of hierarchical clustering methods include the following:

* They often reveal the finer details about the relationships between data objects.
* They provide an interpretable Dendrogram.

The weaknesses of hierarchical clustering methods include the following:

* They’re computationally expensive with respect to algorithm complexity.
* They’re sensitive to noise and outliers.



**Figure7: K-MEAN CLASSIFICATION**

### Decision tree

A machine researcher named J.Ross Quinlan in 1989 developed decision tree algorithm referred to as ID3 (Iterative Dichotomiser). Later, he presented C4.5, which was the successor of ID3 and C4.5 adopt a greedy approach.

A Decision tree could be a structure that includes a root node, braches and leaf nodes. Each internal node denotes a test on an attributes, each branch denotes the outcome of a test, and each of the leaf nodes hold a class labels.

Decision tree uses a tree like structure and their possible combinations to solve a particular problem. It lies in the class of supervised learning algorithms and it can be used for both regression and classification purposes.

A decision tree is a tree data structure which have a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, and each of the branch denotes the outcome of a test, finally each leaf node holds a class label which predicts the class. The first and the topmost node of the tree is the root node.

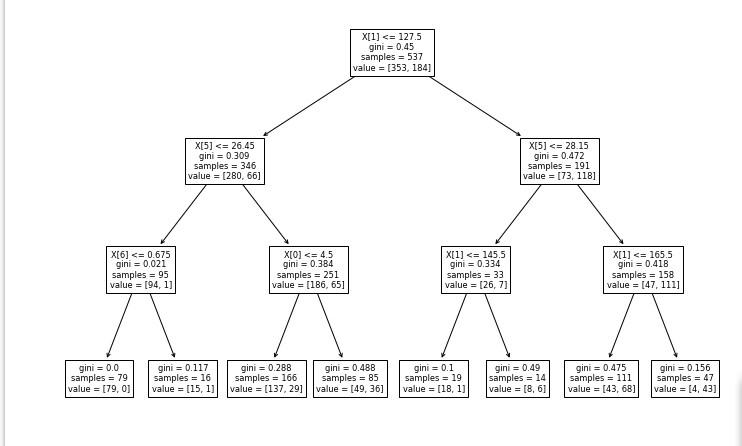
**BENEFITS OF DECESION TREE:**

The benefits of getting decision trees are the following:

1. It doesn’t need any domain knowledge.
2. The learning and classification steps of a decision tree are simplest and fast.
3. It is simple to grasp.
4. It helps us to settle on the several course of action.
5. Decision tree help to lay out the problem.

While some of the disadvantages are:

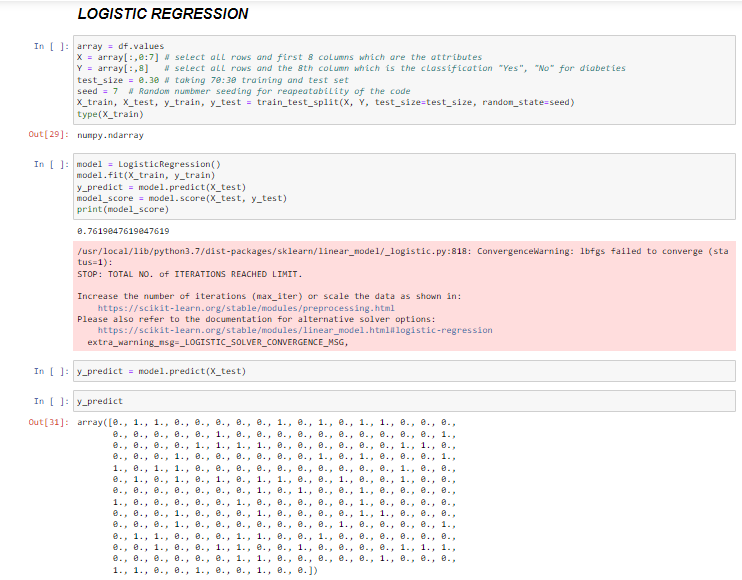
* It sometimes take huge amount of time to train a decision tree model.
* Calculations can get sometimes very complex and thus it can be expensive on training.
* It can only be used for classification purpose.

 **Figure8: DECESION TREE**

**LOGISTIC REGRESSION**

Logistic regression such a type of statistical analysis method which help to predict a Boolean outcome, like 0 or 1 But it is much different from Fuzzy logic algorithm. It supported prior observations of a knowledge set.  
  
A logistic regression model is help us to predict a dependent data variable by analyzing the link between one or more existing independent variables.

As an example, a logistic regression are accustomed predict whether It will be rain in Karachi or not or whether the price of petrol will be decreases per liter in this month. These binary Result allow straightforward decisions between two alternatives.

 **Figure9: LOGISTIC REGRESSION**

**RESULT:**

First we had to define the introduction of data mining, data warehouse then and data analysis then we show our data then we applied Data pre-processing. In the preprocessing section we applied Min-Max normalization (use for normalizing the data, provide the scalability to the data but do not handle the outliers), Null Testing, Pair plot (provide the information for the relationship between two variables and design most isolated cluster), Group By and etc.

We applied classical Clustering technique like K-mean classification, Decisions tree and logistic regression to elaborate our data more.