**University Of Karachi**

**Department of Computer Science(UBIT)**

**MCS Final Year (Morning Program)**

**Course: Data Mining & Data Warehousing**

**Project Title:** **Web Traffic Time Series Forecasting**

**Supervisor: Dr Tahseen Jilani**

**Group Members:**

**Rida Khan (P19101056)**

**Ruqia Rasheed (P19101058)**

**Iqra Zaheer (P19101027)**

**Abdul Basit (P19101001)**

Table of Contents

[What is Data Mining? 3](#_Toc92742410)

[Data Mining Techniques: 3](#_Toc92742411)

[Data cleansing and preparation: 3](#_Toc92742412)

[Association rule learning: 3](#_Toc92742413)

[Clustering: 3](#_Toc92742414)

[Classification and prediction: 3](#_Toc92742415)

[Data analysis: 4](#_Toc92742416)

[Data warehousing: 4](#_Toc92742417)

[Artificial Intelligence (AI): 4](#_Toc92742418)

[How does AI work? 4](#_Toc92742419)

[Machine Learning: 4](#_Toc92742420)

[Unsupervised Machine Learning 5](#_Toc92742421)

[Supervised Machine Learning 5](#_Toc92742422)

[Regression: 5](#_Toc92742423)

[Abstract: 5](#_Toc92742424)

[Introduction: 6](#_Toc92742425)

[Web Traffic Time Series Dataset: 6](#_Toc92742426)

[Include the procedures listed below: 6](#_Toc92742427)

[Principal Component Analysis(PCA): 6](#_Toc92742428)

[African Country Recession Dataset (2000 to 2017) 7](#_Toc92742429)

[Cluster Analysis: 7](#_Toc92742430)

[**APPENDIX** 8](#_Toc92742431)

[PCA Code: 8](#_Toc92742432)

[Cluster Analysis: 8](#_Toc92742433)

# What is Data Mining?

One of the first data mining developments took place in 1936, when Alan Turing introduced the concept of a universal machine capable of performing calculations similar to modern computers. Analytical data is used with tools and techniques that rely heavily on mathematical methods to inform a business problem (or any other problem from which data can provide a solution).

Statistics mining is the process of analyzing huge amounts of statistics to determine business intelligence. information mining is a manner utilized by organizations to transform uncooked statistics into beneficial statistics. by the usage of software to look for styles in huge statistics sets, organizations can research extra about their customers in order to develop greater effective marketing strategies, growth sales and reduce fees. Data mining is based on active data collection inventory Processing.

Data mining is the process of gathering information about customers and using tools and strategies to inform business policy or marketing strategies.

# Data Mining Techniques:

Obtaining the excellent consequences from facts mining requires a spread of gear and techniques. some of the most broadly used sports encompass:

## Data cleansing and preparation:

The step at which information is transformed into the proper form for further analysis and processing, which include identity and debugging and missing information.

## Association rule learning:

Those equipment, additionally known as marketplace basket evaluation, look for relationships between one of a kind objects within the database, together with determining which products are commonly purchased together.

## Clustering:

The process of separating databases into a set of logical sub-classes, called collections, to help users understand natural collections or data structures.

Clustering is the assignment of dividing the populace or facts points into a number of groups such that statistics points in the equal corporations are more just like different records factors within the equal group and varied to the data factors in different corporations. It is largely a set of gadgets on the premise of similarity and dissimilarity among them.

## Classification and prediction:

This process assigns objects to the database into specific categories or classes with the goal of accurately predicting the target class in each data setting. Type is the system of designing a set of fashions to predict the class of gadgets whose class label is unknown. The derived version can be represented in diverse bureaucracy, such as if-then regulations, selection trees, or mathematical formulation.

A selection tree is a flow-chart-like tree structure where each node denotes a check on an attribute cost, every branch represents an final results of the check, and every tree leaf represents a class or magnificence distribution. decision bushes may be converted to category guidelines. Category may be used for predicting the class label of records gadgets. Prediction encompasses the identification of distribution tendencies primarily based on the available facts.

## Data analysis:

The process of testing digital information into a useful business intelligence.

# Data warehousing:

A large collection of business facts used to help an agency make selections. it's miles an critical a part of many records mining efforts.

statistics warehousing is the at ease digital garage of information with the useful resource of a business or distinct business enterprise. The motive of information warehousing is to create a trove of ancient facts that may be retrieved and analyzed to offer beneficial notion into the corporation's operations.

# Artificial Intelligence (AI):

### 

Those programs carry out analytical duties related to human intelligence alongside making plans, mastering, wondering, and trouble solving.

artificial Intelligence is the simulation of human intelligence strategies with the useful resource of way of machines, in particular computer structures. specific applications of AI encompass expert systems, natural language processing, speech recognition and device vision.

**How does AI work?**

because of the fact the hype spherical AI has prolonged, vendors were scrambling to sell how their products and services use AI. regularly what they speak to as AI is in truth one component of AI, inclusive of machine analyzing. AI requires a foundation of specialized hardware and software program for writing and training tool reading algorithms.

programming language is synonymous with AI, however a few, such as Python, R and Java, are famous.

In desired, AI systems paintings with the aid of the use of eating huge quantities of classified training facts, reading the information for correlations and styles, and using those styles to make predictions approximately future states. in this way, a chatbot this is fed examples of textual content chats can learn to produce existence like exchanges with humans, or an photo recognition tool can discover ways to discover and describe objects in images with the aid of reviewing lots and lots of examples.

# Machine Learning:

System mastering is method which develops complex algorithm for processing big information and provides effects to its users. Its uses complex software that may learn via enjoy and make predictions. The algorithms are improved by using itself via normal enter of schooling data. The aim of device getting to know is to recognize records and construct models from statistics that may be understood and utilized by human.

A laptop programming device that makes use of mathematical possibilities to give computer systems the capacity to “study” with out specific programming.

There are two Types of Machine Learning

1.Unsupervised Machine Learning

2. Supervised Machine Learning

## Unsupervised Machine Learning

Unsupervised getting to know does no longer depend on educated facts units to expect the consequences however it uses direct strategies such as clustering and affiliation with the intention to predict effects. educated statistics units suggest the enter for which the output is known.

## Supervised Machine Learning

Supervised gaining knowledge of is like instructor-scholar learning. The relation between the enter and the output variable is thought. The device mastering algorithms will anticipate the final results at the enter records at the way to be in contrast with the anticipated outcome.The mistake could be corrected and this step can be finished iteratively until a appropriate degree of usual overall performance is finished.

Regression:

A way used to are expecting the variety of numerical values, consisting of income, temperatures, or inventory fees, primarily based on a hard and fast of particular facts. Regression evaluation creates models that designate dependent variables through the assessment of impartial variables. as an example, the prediction for a product’s earnings performance can be created thru correlating the product fee and the average purchaser income diploma.

# 

# Abstract:

Net traffic forecasting is a big issue in recent times, as it may disrupt the operations of critical web sites. Time-collection forecasting is a outstanding issue in academia. one of the maximum difficult obligations in the industry is predicting future time series values. From inference and analysis to forecasting and categorization, the time collection discipline covers a extensive range of subjects. The maximum green way to present the information would be to forecast community traffic and display it in a dashboard that adjustments in real-time. using a dashboard to display and analyses real-time facts would be beneficial. we're overly reliant on Google servers nowadays, but if we desired to host a server for a massive variety of people, we'd have projected the quantity of users from previous years to avoid server failure. multiple domains rely on time collection forecasting. we have tested the accuracy of most of the vintage timeline series records with the simple reality statistics acquired from Google's Kaggle net prediction competition. a new technique to the seasonal, trend and cycle sample is used for a particular time collection of each day facts. we have proposed the use of a mixture of four conventional techniques to lessen RMSE and accordingly obtain better forecasting accuracy. The consequences confirmed the error rate reduced by 10 to 20 percent. After analyzing the capabilities of the internet visitors time series, we brought the Generative antagonistic model (GAN) with long-quick term reminiscence (LSTM) as a generator and a deep Multi-Layer Perceptron (MLP) as a predictor of internet site visitors time series. Predictability overall performance is as compared among traditional mathematical methods and a deep efficient competitive network. We concluded with the aid of testing that there is no sizable difference in this series of precision time collection the usage of these two types of strategies.

# Introduction:

As extra humans advantage get entry to to the internet around the sector, the boom in site visitors to nearly all web sites has emerge as unavoidable. The boom in website site visitors should bring a slew of troubles, and the company that is capable of cope with the variations in site visitors the maximum efficaciously will win. As the general public have skilled a crashed web page or a completely slow loading time for a internet site when there are plenty of people using it, including whilst numerous buying websites may additionally crash just before fairs as greater people try to log into the website than it become at first capable of, inflicting a number of inconveniences for the customers and, as a result, reducing the user's scores of the site and instead the usage of every other website online, lowering their business. As a end result, a site visitors control approach or plan ought to be carried out to restriction the threat of such disasters, that can jeopardize the employer's existence. till currently, there has been no want for such tools because most servers ought to manage the extended visitors.

# Web Traffic Time Series Dataset:

Internet visitors is essentially the quantity of classes in a given time, and varies substantially in phrases of time of day, what day of the week, and so on., and what kind of web platform traffic. Can withstand as much as the dimensions of the platform-helping servers. If the visitors is past what the servers can handle, the website may display this 404 blunders, which we do no longer need to take place. it'll make the guests pass. One way to this hassle is to increase the range of servers. but, the disadvantage of treatment for the reason may also boom, which is likewise ugly. So, what's the solution?

You may switch multiple servers based on historical traveler volume facts or based totally on internet visitors history records. And that brings us to the science of statistics, which predicts net traffic or some instances primarily based on historic information.

Internet visitors Time collection Forecasting train dataset is the facts used for this challenge to calculate validation score.

## Include the procedures listed below:

7-49 days on average

With the holidays, the median 7 to 49 days

(holiday /holiday & log/ yearly & log)

# Principal Component Analysis(PCA):

Principal Component Analysis (PCA) is a feature extraction technique that use orthogonal linear projections to capture the data's underlying variance.

Here we are applying PCA **Using R** on Web Traffic Time Series Dataset**.**

# [African Country Recession Dataset (2000 to 2017)](https://www.kaggle.com/mollywiener/cluster-analysis-african-recession-data/data)

There are 49 feature variables in the dataset and one goal variable (the 'growthbucket' variable). The dataset contains 486 samples in total. The "0," or "No Recession" class accounts for 92.81 percent of the samples. Furthermore, 7.82 percent of the samples fall into the "1" or "Recession" category. In other words, there is a class imbalance in the dataset. It's useful for learning approaches like Cost-Sensitive Classification, Oversampling, and Undersampling for dealing with class imbalance.

The data set spans the years 2000 to 2017 & 27 African countries covered by the dataset.

# Cluster Analysis:

Cluster Analysis in Data Mining refers to the process of identifying a group of objects that are similar to one another but distinct from those in other groups.

Here we are applying cluster analysis **Using R** on [African Country Recession Dataset](https://www.kaggle.com/mollywiener/cluster-analysis-african-recession-data/data) .

Clustering is a method for locating subgroups of observations within a facts set.Even as we are doing clustering, we need observations within the equal group with comparable styles and observations in specific companies to be various. If there is no response variable, then suitable for an unmonitored method, which means that that it seeks to discover relationships a number of the n observations without being knowledgeable with the useful resource of a reaction variable.

Clustering allows us to grow to be aware of homogenous organizations and categorize them from the dataset. One of the excellent clustering’s is k-method, the maximum usually used clustering method for splitting a dataset into a set of n companies.

If datasets include no response variable and with many variables then it comes below an unmonitored approach.Cluster analysis is an unmanaged technique and set for segmenting markets into agencies of similar customers or patterns.

# 

# **APPENDIX**

## PCA Code:

library(factoextra)

library(fastDummies)

library(psych)

library(GPArotation)

library(datasets)

library(readr)

file <- read.csv('C:/Users/user/Downloads/validation\_score.csv')

data <- na.omit(file)

data <- data[-c(18)]

data <- scale(data)

KMO(data)

cortest.bartlett(data, n=NULL, diag=TRUE)

fa.parallel(data, fm="pa", fa="both", n.iter=100)

fit <- principal(data, 3, rotate="varimax")

print(fit$loadings, cutoff=.3)

factor.plot(fit)

fa.diagram(fit)

fit <- fa(data, 3, rotate="promax", fm="pa")

print(fit$loadings, cutoff=.3)

factor.plot(fit)

fa.diagram(fit)

fviz\_nbclust(data, kmeans, method = "wss")

cluster <- kmeans(data, 3, nstart = 24)

fviz\_cluster(cluster, data = data, ellipse.type = "euclid", star.plot = TRUE, repel = TRUE, ggtheme = theme\_minimal())

## Cluster Analysis:

library(tidyverse)

library(cluster)

library(haven)

library(ggdendro)

library(NbClust)

library(factoextra)

library(klaR)

library(data.table)

library(rlang)

library(dplyr)

library(NbClust)

library(ggpubr)

library(readr)

theme\_set(theme\_pubr())

library(corrplot)

data <- read\_csv("C:/Users/user/Downloads/africa\_recession.csv")

head(data)

data <- na.omit(data)

head(data)

data <- data[-c(50)]

summary(data)

data <- scale(data)

head(data)

fviz\_nbclust(data, kmeans, method = "wss") +

geom\_vline(xintercept = 4, linetype = 2)+

labs(subtitle = "Elbow method")

fviz\_nbclust(data, kmeans, method = c("silhouette", "wss", "gap\_stat"))

set.seed(123)

fviz\_nbclust(data, kmeans, nstart = 25, method = "gap\_stat", nboot = 50)+

labs(subtitle = "Gap statistic method")

dist <- dist(data, method = "euclidean")

fit <- hclust(dist, method = "ward.D2")

plot(fit)

groups <- cutree(fit, k=5)

rect.hclust(fit, k=5, border="red")

fit <- kmeans(data, 5)

aggregate(data,by=list(fit$cluster),FUN=mean)

data <- data.frame(data, fit$cluster)

databind <- cbind(data, Cluster = fit$cluster)

head(databind)

library(ggplot2)

library(ggpubr)

theme\_set(theme\_pubr())

ggplot(databind, aes(Cluster)) +

geom\_bar(fill = "#0073C2FF") +

theme\_pubclean()

library(cluster)

clusplot(databind, fit$cluster, color=TRUE, shade=TRUE,

labels=2, lines=0)

library(fpc)

plotcluster(databind, fit$cluster)

library("mclust")

fit <- Mclust(data)

plot(fit)

summary(fit)

library("mclust")

fit <- Mclust(data)

plot(fit)

summary(fit)