**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 (PCA)**

**Group Members:**

**Rida Khan (P19101056)**

**Ruqia Rasheed (P19101058)**

**Iqra Zaheer (P19101027)**

**Abdul Basit (P19101001)**

**Table of Content**

1.What is data mining?

2.Data Mining Techniques.

i: Data cleansing and preparation

ii: Artificial Intelligence (AI)

iii: Association rule learning

iv: Clustering

v: Classification

vi: Data analysis

3. Data warehousing

i: Machine Learning:

ii: Regression

4.Introduction

5.Abstract

6.web Traffic Time Series Dataset

7.Principal Components Analysis(PCA)

8. Cluster Analysis

9.APPENDIX

**What is Data Mining?**

Data mining is not a fictional thing that comes with the digital age. The concept has been around for over a century, but it began to focus more on society in the 1930's. 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). Data mining is the process of analyzing large amounts of data to determine business intelligence. Data mining is a process used by companies to convert raw data into useful information. By using software to look for patterns in large data sets, businesses can learn more about their customers in order to develop more effective marketing strategies, increase sales and reduce costs. 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 best results from data mining requires a variety of tools and techniques. Some of the most widely used activities include:

**Data cleansing and preparation:**

The step at which data is converted into an appropriate form for further analysis and processing, such as identification and debugging and missing data.

**Artificial Intelligence (AI):**

These programs perform analytical tasks associated with human intelligence such as planning, learning, thinking, and problem solving.

**Association rule learning:**

These tools, also known as market basket analysis, search for relationships between different items in the database, such as determining which products are usually 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.

**Classification:**

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.

**Data analysis:**

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

**Data warehousing:**

A large collection of business data used to help an organization make decisions. It is an integral part of many data mining efforts.

**Machine Learning:**

A computer programming system that uses mathematical possibilities to give computers the ability to “read” without explicit programming.

**Regression:** A method used to predict the range of numerical values, such as sales, temperatures, or stock prices, based on a set of specific data.

**Introduction:**

As more people gain access to the internet around the world, the increase in traffic to practically all websites has become unavoidable. The increase in website traffic could bring a slew of issues, and the company that is able to deal with the variations in traffic the most effectively will win. As most people have experienced a crashed site or a very slow loading time for a website when there are a lot of people using it, such as when various shopping websites may crash just before festivals as more people try to log into the website than it was originally capable of, causing a lot of inconveniences for the users and, as a result, decreasing the user's ratings of the site and instead using another site, reducing their business. As a result, a traffic management approach or plan should be implemented to limit the danger of such disasters, which could jeopardize the company's existence. Until recently, there was no need for such tools because most servers could handle the increased traffic.

However, the smartphone era has increased demand to such a high level for some websites that businesses have been unable to respond quickly enough to maintain their normal customer service levels. Many approaches for projecting web traffic have been proposed in the literature. Based on the models examined, they can be divided into two categories: nonlinear prediction and linear prediction. The most popular models The HoltWinters Algorithm, the AR Model, and the MA Model are all linear forecast models. Nonlinear prediction frequently employs forecasts based on repeating neural networks. The discrete wavelet transform (DWT) separates data into linear and non-linear components, which aids forecast accuracy. ES-RNN improves performance by training the dataset with GPU computation.

**Abstract:**

Web traffic forecasting is a big issue nowadays, as it can disrupt the operations of important websites. Time-series forecasting is a prominent issue in academia. One of the most difficult tasks in the industry is predicting future time series values. From inference and analysis to forecasting and categorization, the time series discipline covers a wide range of topics. The most efficient way to present the information would be to forecast network traffic and display it in a dashboard that changes in real-time. Using a dashboard to monitor and analyses real-time data would be beneficial. We are overly reliant on Google servers these days, but if we wanted to host a server for a huge number of people, we might have projected the amount of users from prior years to avoid server failure. Multiple domains rely on time series forecasting. We have tested the accuracy of a few of the old timeline series statistics with the basic truth data obtained from Google's Kaggle web prediction competition. A new approach to the seasonal, trend and cycle pattern is used for a specific time series of daily data. We have proposed the use of a combination of four traditional methods to reduce RMSE and thus achieve better forecasting accuracy. The results showed the error rate reduced by 10 to 20 percent. After studying the features of the web traffic time series, we introduced the Generative Adversarial Model (GAN) with Long-Short Term Memory (LSTM) as a generator and a deep Multi-Layer Perceptron (MLP) as a predictor of web traffic time series. Predictability performance is compared between traditional mathematical methods and a deep productive competitive network. We concluded by testing that there is no significant difference in this series of precision time series using these two types of methods.

**Web Traffic Time Series Dataset:**

Web traffic is basically the number of sessions in a given time, and varies greatly in terms of time of day, what day of the week, etc., and how much web platform traffic. can withstand up to the size of the platform-supporting servers. If the traffic is beyond what the servers can handle, the website may display this 404 error, which we do not want to happen. It will make the guests go. One solution to this problem is to increase the number of servers. However, the downside of remedy for the cause may increase, which is also unpleasant. So, what is the solution?

You can switch multiple servers based on historical visitor volume data or based on web traffic history data. And that brings us to the science of data, which predicts web traffic or a few times based on historical data.

Web Traffic Time Series Forecasting train dataset is the data used for this project 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) .

**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)