**Route Optimization**

An analysis of the transport system and solution to traffic congestion

**Data Mining And Predicative Analysis Project-2017**

**Abstract**

Bangalore, the IT capital of India, is growing at a faster rate than any other city in the country because of the technological advancement over the last decade. The city consists of tech parks that consist of several multi-national companies, each having thousands of employees. In a developing city like Bangalore, there’s bound to be plenty of urban problems. One of the major problems in this city is in the field of transportation. Traffic congestion on roads occurs as use increases, and is characterized by slower speeds, longer trip times, and increased vehicular queueing.  During the day, especially at peak hours, it takes a very long time to commute within the city due to heavy traffic and lack of efficient public transport. Employees of most companies across the city, especially those who have strict working hours find it extremely hard to reach on time.

This project is based on a survey conducted at the Manyata Tech Park, Bengaluru. It aims to help reduce the traffic in the city by optimizing routes and using company buses to its maximum capacity. This document gives an overview of the results after the required predictive analysis and a summary of the different methodologies and functionalities implemented in the course of the project.

**Problem statement**

To reduce number of vehicles on the road and thus save time and money, company buses have been deployed to pick up and drop employees from this tech park. However, to use minimum number of buses and fuel and to accommodate as many employees until it reaches its maximum capacity, a routing algorithm is essential. Employees occasionally have to travel by different routes or from different locations but it is difficult to predict how long it will take to reach the office. The project has been developed to overcome these issues and to improve the overall commuting challenges faced by the public.

**Objective**

The primary objective is to decrease the overloaded transport networks and understand the traffic condition in the city. This project aims to use different clustering algorithms to predict the best routes, that is the fastest and the most efficient in terms of fuel and capacity and least number of buses to commute employees that belong to the tech park to and from home every day. Additionally, it also analyses the time taken based on your exact location to travel to work considering all factors such as traffic and time of the day.

**Introduction**

Traffic congestion has become the primary concern for the city of Bangalore and to reduce this, different companies have come together along with the government of Karnataka and the Bangalore Metropolitan Transport Corporation. One such company known as Mapunity, conducted a survey on the employees at the tech park to get a better understanding of the life of a daily commuter in the city. Important information like time taken for travel, mode of transportation and approximate monthly cost was noted. This project aims to analyze this survey and propose optimum and feasible solutions as well as a clear picture of the traffic scenario in Bangalore. The different tools used for this project are Rstudio for R programming, Rapid Miner, CartoDB and extensive use of Excel for database.

**Design and database**

The survey conducted is the database for this project and it is on the basis of the responses of all these employees that further predictions and analysis will take place. The employees had to give their address, or atleast the location of the building they live in for the survey. This was later converted in the form of latitude and longitude to get the exact location of the point from which they commute to and fro daily. The time taken to go to work in the in the morning and back from work at night, that is, usually during peak working hours was taken into account. The mode of transport used by these employees and approximate monthly expenditure on travelling was also considered.

Number of tuples in the dataset: 297

Schema:

|  |  |
| --- | --- |
| **Attributes** | **Datatype** |
| Id | Integer |
| Latitude | Float |
| Longitude | Float |
| Time taken to commute to work | String\* |
| Time taken to commute from work | String\* |
| Mode of Transport | String |
| Monthly expenditure on transport | String\* |

\*Tuples were a range, for eg. “Between 30-45 minutes”

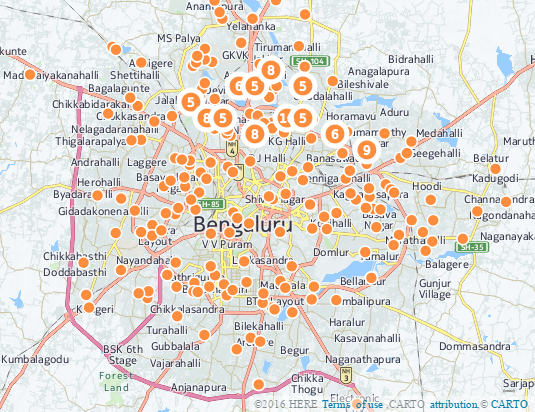
**Methodology**

**Understand the dataset**

With the help of various bar graphs and pie charts created using R and Excel, both extremely powerful tools to plot graphs, the data in hand appears more structured and the most frequent values for each attribute are highlighted.

Transport-demand mapping is essentially done to understand when certain parts of an area require an increase in the transport facilities provided to them and what means of transport is best suitable for that area.

In this project it has been implemented on the city map of Bangalore to give a clear picture of the exact location of these employees so that clustering can be understood better.



**Data Cleansing and Pre-processing**

To ensure structured data with no missing values to get accurate predictions is one of the most important aspects of data analytics. In this project, R programming has been used to make the data uniform by using different functionalities like sapply, gsub, sample, levels and ggplot. The code snippet is provided below:

manyata.R

library(ggplot2)

complete<-read.csv("manyata.csv")

dim(complete)

names(complete)

summary(complete)

miss\_val <- sapply(complete, function(x) sum(is.na(x)))

sapply(complete, class)

complete$to.work <- gsub("Less than 30 minutes","1",complete[[3]])

**Data Analysis and Prediction:**

Two most note-worthy algorithms K-means and K-medoids have been used for the purpose of clustering. These have been used keeping in mind that the actual number of employees in the tech park are a much larger number. Each cluster created is allocated to one bus. These clusters have been formed by using the location of the employees as input. So the number of items in a cluster depict the total number of people that belong to that region which has been clustered.

For the sake of accurate comparison and analysis, the following assumptions have been made:

1. The tech park is only financially capable of buying 10 buses
2. Each bus has the capacity of 30 people to fit the dataset of approximately 300 people.

Using Rapid Minor and keeping all factors constant the following clusters are made when k=10:

**K-medoids**:

Cluster 0: 39 items

Cluster 1: 18 items

Cluster 2: 12 items

Cluster 3: 31 items

Cluster 4: 31 items

Cluster 5: 24 items

Cluster 6: 28 items

Cluster 7: 16 items

Cluster 8: 49 items

Cluster 9: 58 items

Total number of items: 297

**K-means**:

Cluster 0: 30 items

Cluster 1: 18 items

Cluster 2: 36 items

Cluster 3: 44 items

Cluster 4: 26 items

Cluster 5: 19 items

Cluster 6: 26 items

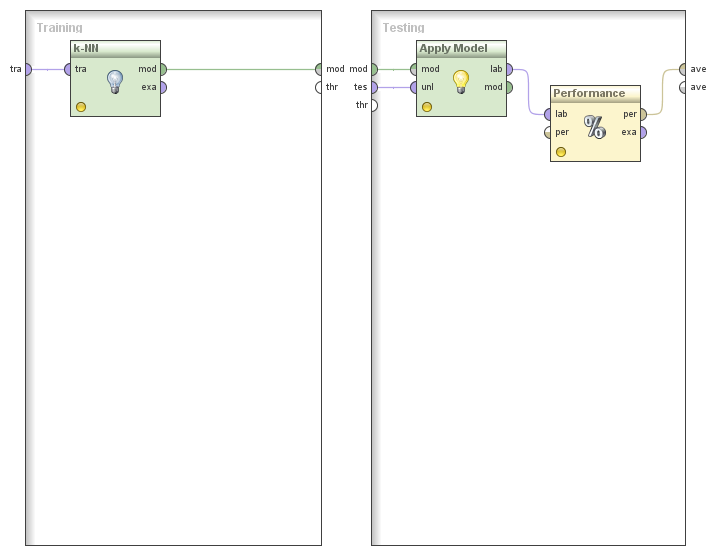
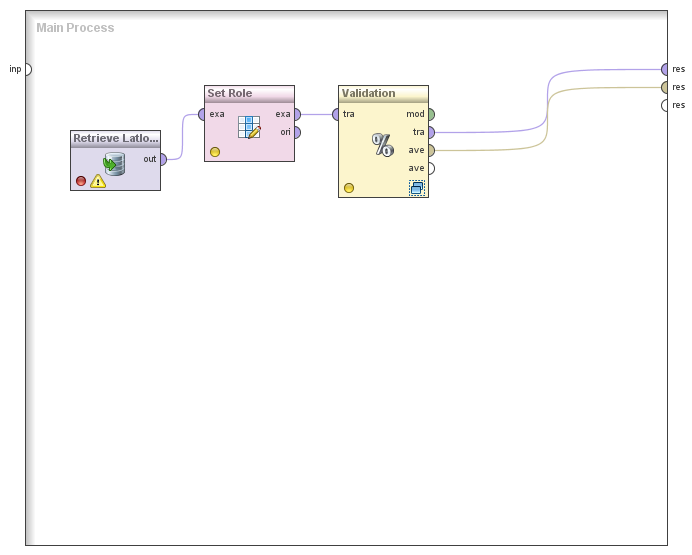
Cluster 7: 23 items

Cluster 8: 28 items

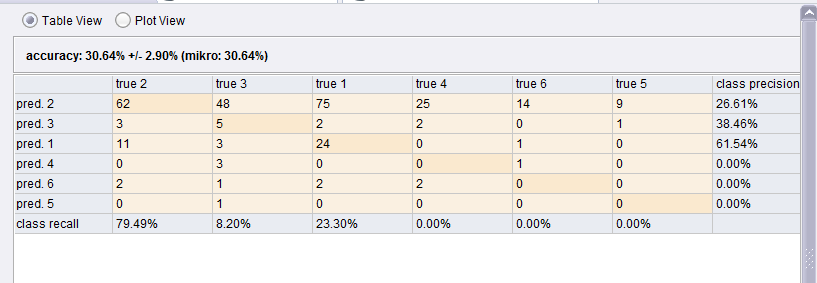
Cluster 9: 38 items

Total number of items: 297

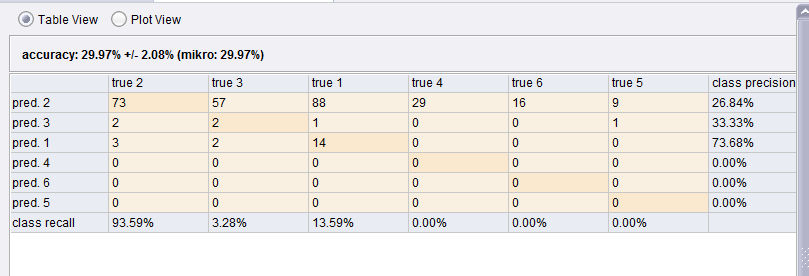
On further calculation (by comparing excess number of people in some number of buses) it can be said that the clustering done by K-means is far more accurate and it would utilize the resources of the tech park more efficiently.



With k=1



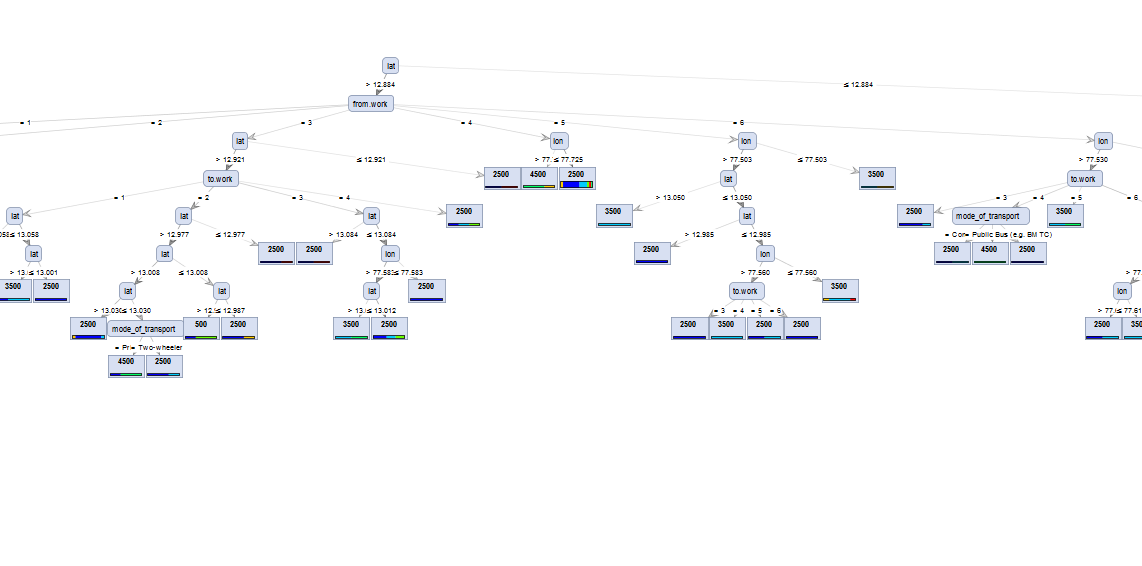
With k=5



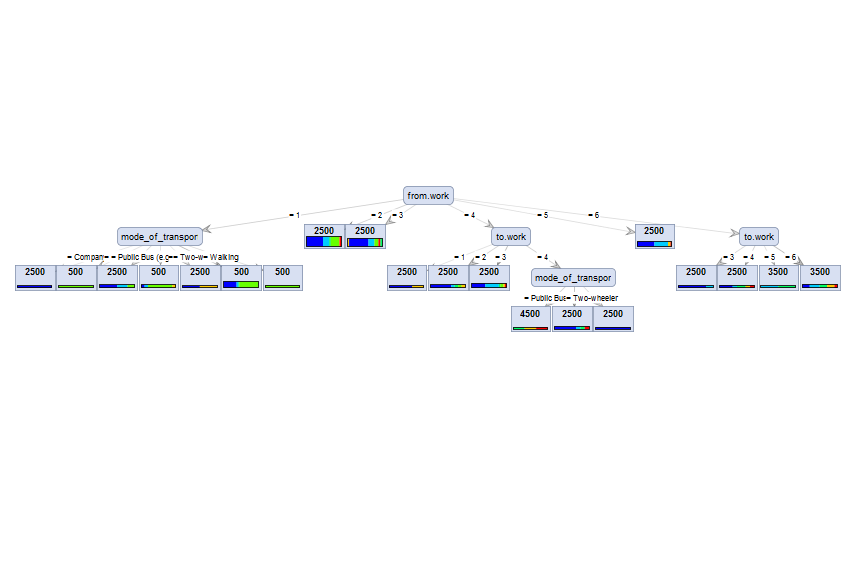
Cross-validation and use of K-NN to predict values of time with the location of the employee as input is used. Thus with a higher value of k the precision is better in predicting the time taken to go to work in the morning hours. This can be done even in case of time taken from work during the evening hours.

**Decision tree:**

Using the data that was acquired after the cleaning, it was applied to form a decision tree using all valid parameters and part of the decision tree is displayed below. But this was found to be contradictory to the fields denoted according to the actual time that the person in question required to commute to and fro from their place of work. This classification can be said to be the more logical inferences from the data that was used but is invalidated due to the complex inner working of the traffic and cost based on location.

****

Hence another decision tree was created using only time and mode of transportation as parameters that signifies the reality of the situation and how even though the location can be said to be a factor, when it comes to breaking it down to their smaller clusters, the internal factors such as traffic and the local road system nullifies the argument that based only on location that the time taken is dependent upon. The secondary decision tree is displayed below.

****

**Conclusion:**

To improve the situation of extreme traffic congestion, it is important to first reduce the number of private vehicles and instead increase public buses or company buses. If such a survey is conducted in all tech parks around the city and using the k-means algorithm, the people are clustered based on their location, fewer company buses will be required to pick and drop these employees. Employees using private vehicles or cabs will instead begin to use these company buses, thus reducing the over population of vehicles on Bangalore roads. Using the maps, it has also been noted that the monthly cost for existing company buses are already lower than other means of transportation. If each bus is allocated to one cluster, the time saved is more and the cost will vary based on distance of the cluster from the destination. Thus, with a combination of data mining and analysis, traffic congestion in Bangalore can be easily curbed.

**Reference:**

<https://www.mybmtc.com/>

[www.carto.com](http://www.carto.com)

[www.tutorialspoint.com](http://www.tutorialspoint.com)

[www.stackoverflow.com](http://www.stackoverflow.com)

**Submitted By:**

Akriti Nagar -140953024