**DEMOGRAPHICS**

**Introduction:**

Demographics is an important field of study with important implications for planning and resource management. Changing lifestyle patterns and human migration can change the infrastructure needs of communities, but infrastructure projects by their nature require large investments in both time and capital. We applied data mining techniques to the socio-economic data from the '90 Census, law enforcement data from the 1990 Law Enforcement Management and Admin Stats survey. We predicted that there is a strong correlation between population and mass transit usage. We also predicted that there is a strong correlation between income and population. In other words, income and mass transit usage are both higher in heavily populated areas.

**Dataset:**

For finding an appropriate dataset we went through several websites and finally found this dataset on the link. [http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized#](http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized).

This dataset is a collection of several attributes based on the US census which predict many interesting things. This is a very large dataset with 147 attributes and 2215 instances. The creator of this dataset is Michael Redmond; Computer Science; La Salle University; Philadelphia, PA, 19141, USA.

**Research questions:**

Is there a definable population density threshold which sees a large increase in commuting times and mass transit usage?

Does the data not conform well to predictable patterns based on human planning and decision making?

Can we predict where mass transit infrastructure will be most and least needed based on demographic trends?

What is the relationship between commuting times and mass transit usage? Does mass transit decrease commuting times (it is clear that it has a positive ecological impact)?

**Dataset with arff format:**

At this step our dataset is ready to be given to data mining software for which we chose Weka.

**Attributes and Relations:** We identified thatsome further pre-processing of the dataset must be done for better understanding of the correlation between population, mass transit usage and income. Using Weka's Unsupervised Attribute Remove Filter we removed the unwanted attributes. So the source dataset consists of 16 attributes. Many variables are included so that algorithms that select or learn weights for attributes could be tested. We have to identify the attributes that are not predictive, and would get in the way of some algorithms. However, clearly unrelated attributes were not included. Attributes were picked if there was any plausible connection between them.

**Project Management:**

| **Team member** | **Roles and skills** | **Contributions** |
| --- | --- | --- |
| Bethi Bharath Reddy | code development | Updated the project proposal, analyzed the available traffic data and working on coding part. |
| Venkateswara rao alluri | Works on testing | Cleaning of data and eliminated some of the bad data |
| Roja suryuadhevara | Collected data | Progressing demographics |
| Deepthi kota | Data analysis | Analysing data required, reporting accurate data |
| Priyanka pappuri | Installation and setup | Finished installing and setting up the environment required to test our project outcomes. |

**PROJECT CHECKPOINTS:**

| **Checkpoint date** | **Expected deliverables** | **Responsible Team members** | **Checkpoint results** |
| --- | --- | --- | --- |
| 02/16/2016 | Exploring more data | Deepthi, Venkatesh | Collected the data |
| 03/15/2016 | Coding and execution in weka | Bharath, Roja | Successfully executed |
| 04/02/2016 | Debugging | Priyanka, Venkatesh | Data was being successfully executed and tested by all team members |
| 04/10/2016 | Test the project again | Bharath | Project is ready for final report |
| 04/12/2016 | Preparing final report | All team members | All the sections are filled |