Project for Data Analysis and Visualization

Project Title: District Level Health Care Planning

Member's Names: Shubham Arora, Akash Likhar, Rhythm Gupta, Vaibhav K Verma

Q1. Business/Research/Social Objective of the project:

Our aim is to analyse the data and rank the districts (state wise) according to below points, which help make observations regarding the need of organising **health and child welfare programmes**.

1. Household characteristics
2. Wealth index
3. Sex ratio
4. Literacy rate
5. School & work status
6. Age at time of marriage
7. Injury, disability, illness, personal habits
8. Fertility, abortion
9. Family planning practices
10. Child health after birth
11. Awareness on HIV/AIDS and some other diseases
12. Mortality

Also, given the outcome of pregnancy, predict the outcome of pregnancy for future cases (live birth, still birth, abortion).

Q2. Sources of data, method of data collection and estimate of data cleaning, transformation effort.

<Even Kaggle datasets have a source. If possible, students should get some more recent data for testing and validation of their models>

Source: [Health Management Information System - Ministry of Health & Family Welfare]

In surveys, a lot of questions are optional and thus remain unfilled. This causes a lot of **NA** values in the dataset. Also, there are 201 columns in the dataset and not all are required to analyse data and predict the objective. So, there must be some detailed analysis to find which columns are required and which can be dropped. Thus there is a lot of effort in data cleaning and transformation.

Q3. Proposed techniques of data analysis, modeling and model validation/testing.

<Please state references, also state if some technique would be new for you and would like the instructor to cover in the class>

We will use unsupervised learning to cluster the districts according to their need for health and child welfare programmes. So, for data analysis and modeling , we will use:

1. Neural Network (For prediction of outcome of pregnancy)
2. Dimensionality Reduction (KNN or PCA)
3. Clustering

For testing/validation: Hand labeling some results based on government records can be used to test/validate the model.

Q4. How do you intend to visualize your raw data, intermediate and final results?

Plots that are mainly used in the analysis of survey datasets include,

1. Pie charts
2. District level Maps (Mainly to show final results)
3. Informative tables
4. Stacked Bar Graphs
5. Regression plots
6. Bubble Charts
7. Contour Plots
8. Some standard plots like histograms, scatter plots, box plots etc.

Thus, tools used to visualise are **matplotlib**, **seaborn**, **Tableau**, and some survey data visualisation tools like **Chartio**, **Qualtrics**.

Q5. Duties of team members - Not yet decided. But, will be mentioned clearly in the final report. For now, “Yes” is written in every cell.

Team Coordinator: Shubham Arora

Member Name	Data Collect/Clean	Data Model/Analysis	Data Viz./Report
Shubham Arora	Yes	Yes	Yes
Akash Likhar	Yes	Yes	Yes
Rhythm Gupta	Yes	Yes	Yes
Vaibhav K Verma	Yes	Yes	Yes

Any additional comments: