Project Plan : Genetic Correspondence and Comparative Spatial Analysis with Age and Gender specification of ICMR data concerning Cancer Incidence in India

Group : The Outlaws

Group Members :

Swapnonil Mondal, Sarbojit Das, Sayanta Biswas, Sujash Krishna Basak, Sameer Verma

Introduction

Cancer is a high concern diseases in the world due to its massive suffering and incurability in many cases. It sometimes due to Genes and sometimes it causes due to the various surrounding components. There are many types of cancer, some are curable at some extent and some are not. Patient goes through many tough procedures to get cured.

This project is all about the cancer scenario in India as per spatial and genetic viewpoint by using statistical concepts concerning some specific recent years and various types of genes that may contribute effects causing cancer. Here we are focusing mainly on India population. Moreover we are also interested in the gender specific cancer types and the age specific cancer incidence in India.

For the completion of the project in the following we are proposing a tentative project plan that we are trying to execute in this course project under your guidance.

Dataset

Here we are dealing with many datasets that is based on the cancer patient and the nature of the cancer in human bodies, both male and female.

Here is the link of the datasets :

1. Caner Gene Dataset : The input dataset contains 802 samples for the corresponding 802 people who have been detected with different types of cancer. Each sample contains expression values of more than 20K genes. Samples have one of the types of tumours: BRCA, KIRC, COAD, LUAD, and PRAD.

2. Caner Label Dataset : Sample contains different types of tumours: BRCA, KIRC, COAD, LUAD, and PRAD.

3. Estimated state wise Cancer Incidence : The table provide estimated incidence of cancer cases in India by States and Union Territory wise for All sites and for both sexes.
4. Gender wise different sites of Cancer : The table presents the estimated cancer incidence, number of cases, crude rate and cumulative risk by sex and anatomical sites in India for the year 2022.
5. Age wise Cancer Incidence : The table provides gender disaggregated, estimated top five leading sites of cancer (%) in India by age group (0-14, 15-39, 40-64, 65+ age group) for the year 2022.

Purpose

The purpose of this project is mainly divided in some parts;

1. Determining those genes which are mainly causing cancer in the human body in India.
2. Analysis if there is any dependency of those genes in effecting different types of cancer.
3. Comparing the estimated cancer incidence through out India according to the states on some coherent years and trying to infer about the region wise dependencies diagrammatically.
4. Trying to fit a model to age wise cancer incidence in India.
5. Checking various components of the distribution.
6. Trying to observe how the different sites of cancer are distributed among different sexes.

Methodology

According to the fulfillment of the purpose we are dividing the work into some subparts :

1. To determine the genes which are mainly causing cancer in the human body of India, we will use **Principle Component Analysis** to find those Genes which captures maximum amount of variability and then we can ignore those which are reluctant in that sense. For a high dimensional data is will be a very useful analysis.
2. to analysis if there is any dependency of those genes in effecting different cancer, we will use **Logistic Regression** and regress different type of cancer incidence probability by it
3. Comparing the estimated cancer incidence through out India according to the states on some coherent years and trying to infer about the region wise dependencies diagrammatically by using `ggmap` or `leaflet` packages, constructing the map and highlighting the areas by different colours with colour gradient according to the number of cancer patients at that region.

4. Trying to fit a **Probability Model** to age wise cancer incidence in India diagrammatically and try to observe its nature.

5. To check various components of the distribution of age wise cancer incidence in India, we can use various **Non-parametric Tests** and infer about there nature.

6. Which cancer caused according to the site of the cancer for two different Genders, diagrammatically we will try to observe, which cancer is mostly caused to the males and the females.