CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)

Kaushik Shivakumar

Project Number

S0830

Project Title

A Machine-Learning Approach to Correlate Environmental and Demographic Factors to Cancer Incidences across US Counties

Abstract

Objectives/Goals

About 80% of all cancer incidences are sporadic and largely caused by environmental factors. This implies that cancer incidence varies widely based on geography, and my preliminary findings for lung cancer confirm this by showing elevated levels in the eastern U.S. Thus, it is hypothesized that cancer incidence for a county is linearly correlated with factors relating to the local demography and environment. This study focuses on identifying these underlying causes of the majority of cancer incidences.

Methods/Materials

Two types of data are used to analyze the cancer (lung, colorectal, pancreatic, and overall) incidences: 69 demographic factors and, based on EPA toxic emissions data, 274 chemicals.

For demographic data, the features with absolute value of Pearson correlation coefficient>0.3 are selected and normalized. The linear regression model is built and features with $p<0.05$ are chosen as statistically significant. The coefficients are compared to see which factors have the greatest impact on cancer incidence.

Chemical data, however, is sparse, making it difficult to perform linear regression on. For each chemical, counties are marked as having levels signifying contamination (>3 times average) or not. Then the compound is checked for whether it shows a statistically significant increase in cancer incidences when present in increased levels.

Results

The number of demographic factors (eg: ethnicity, diabetes) that are statistically significant range from 9 to 15, depending on the cancer type. Also, predicting using the machine learning model trained on 3/4th of the data yields accurate predictions of cancer rates for the remaining counties (~10% error), confirming that the factors in the model strongly relate to incidences.

The number of chemicals that are statistically significant, showing increased cancer rates when present in elevated levels (eg: Methyl Isocyanate; 2,4-Dinitrotoluene), range from 5 (pancreatic) to 18 (all cancer).

Conclusions/Discussion

Looking at data by county offers a means for identifying sources for increased cancer incidence rates and can also be applied to other diseases. This type of analysis enables each county to identify and fix its specific problems, for instance, by improving living conditions or regulating emissions of certain chemicals. Overall, analyzing data available by county is very powerful and can lead to a major step forward in preventative medicine.

Summary Statement

This project focuses on using machine-learning and statistical methods to identify environmental and demographic factors that are responsible for the majority of cancer incidences.

Help Received

Dr. Anu Aiyer and Dr. Srikant Ramakrishnan advised me on the statistical analysis and machine learning techniques. My father guided me on the use of SQL for data cleaning and Ms. Anu Datar was the mentor for the project.