# CMSC 33750:
# Machine Learning in Cancer

## Assignment 5

## November 19, 2020

This lab uses county level cancer incidence rate data from the State Cancer Profiles resource jointly developed by the NCI and the CDC.   The resource can be found here:

https://www.statecancerprofiles.cancer.gov/incidencerates/index.php

I have downloaded and reformatted a dataset of the incidence rate of lung and bronchus for the most recent five years of the available data that you can find on the course website.  The name of the file is incd-v3.txt and it is reformatted a bit to make it easier to read into R.   It is a tab separated variable (TSV) file, in which missing numbers and low counts are represented with an asterisk “*” and comments have been removed.  It was created by filling in the form’s fields on the website with: “US by County”, “Colon & Rectum (All Stages)”, “All Races”, “Both Sexes” and “All Ages”.

**Enriching the model**

Please also download the following county level data:

median income level for all counties from:

https://statecancerprofiles.cancer.gov/demographics/index.php
Please select: “US by county”, “Income”, “Median family income”, “All races”

**Modeling the relationship of the level of lung cancer and income level**

The goal of this assignment is to build a very simple model to model the relationship, if any, between the incidence level of colon and rectum cancer for each county and the income level for that county. We don’t expect the model to be very good, but simply to gain some insight into whether household income may potentially have a relationship with the incidence level of colon and rectum cancer.

Please carefully write up your assignment, explaining:

1. How you prepared the data for modeling
2. What features you used and how they were computed.

3. The model that you built to model the relationship, if any.
4. How you evaluated the model.

Please prepare:

1. a written report addressing questions 1-4 above.
2. The code you wrote to clean the data, build the features, build the model, and evaluate the performance of the model.

Remember, we don't expect the model to have much explanatary power.

The assignment must be turned in by midnight on **December 3, 2020.**

**Please work individually.**

**Extra Credit**

For extra credit, please:

- Develop a more complex model that includes additional social economic factors in your model
- Look at which cancers have incidence levels that seem to be related to income level and other social economic factors.

**Notes**

1. The note below describes how the age-adjusted incidence rate is computed:

   Incidence rates (cases per 100,000 population per year) are age-adjusted to the 2000 US standard population [http://www.seer.cancer.gov/stdpopulations/stdpop.19ages.html] (19 age groups: <1, 1-4, 5-9, ... , 80-84, 85+). Rates are for invasive cancer only (except for bladder cancer which is invasive and in situ) or unless otherwise specified. Rates calculated using SEER*Stat. Population counts for denominators are based on Census populations as modified [https://seer.cancer.gov/popdata/] by NCI. The 1969-2015 US Population Data File [https://seer.cancer.gov/popdata/] is used for SEER and NPCR incidence rates.

   Source: https://www.statecancerprofiles.cancer.gov/incidencerates/index.php

2. The cancer incidence data, income data, and sunlight data can all be joined by the FIPS code using the R merge function. The FIPS code is an integer that uniquely identifies US counties and is included in each of the datasets.