**Machine Learning Final project**

**Title: Predicting cancer mortality rate in US counties.**

**Objective:**

The main aim of the project is to predict the cancer in before and to mitigate the risk in the people who belong to same county and to decrease the chances of mitigating the cancer effect in the people.The dataset supplied contains information gathered from different counties in the United States. Given some training data, the regression aim for this evaluation is to predict cancer death rates in "unseen" US counties. The training data ('Training data.csv') includes a variety of features/predictors relating to socioeconomic variables, as well as other forms of information for individual counties throughout the country.

The following are the various columns present in this dataset.

1) avgAnnCount: Mean number of reported cases of cancer diagnosed annually

2)avgDeathsPerYear: Mean number of reported mortalities due to cancer

3)incidenceRate: Mean per capita (100,000) cancer diagoses

4)medianIncome: Median income per county

5)popEst2015: Population of county

6)povertyPercent: Percent of populace in poverty

7)MedianAge: Median age of county residents

8)MedianAgeMale: Median age of male county residents

9)MedianAgeFemale: Median age of female county residents

10)AvgHouseholdSize: Mean household size of county

11) PercentMarried: Percent of county residents who are married

12) PctNoHS18\_24: Percent of county residents ages 18-24 highest education attained: less than high school

13) PctHS18\_24: Percent of county residents ages 18-24 highest education attained: high school diploma

14) PctSomeCol18\_24: Percent of county residents ages 18-24 highest education attained: some college

15) PctBachDeg18\_24: Percent of county residents ages 18-24 highest education attained: bachelor's degree

16) PctHS25\_Over: Percent of county residents ages 25 and over highest education attained: high school diploma

17)PctBachDeg25\_Over: Percent of county residents ages 25 and over highest education attained: bachelor's degree

18)PctEmployed16\_Over: Percent of county residents ages 16 and over employed

19)PctUnemployed16\_Over: Percent of county residents ages 16 and over unemployed

20) PctPrivateCoverage: Percent of county residents with private health coverage

21)PctPrivateCoverageAlone: Percent of county residents with private health coverage alone (no public assistance)

22)PctEmpPrivCoverage: Percent of county residents with employee-provided private health coverage

23)PctPublicCoverage: Percent of county residents with government-provided health coverage

24)PctPubliceCoverageAlone: Percent of county residents with government-provided health coverage alone

25)PctWhite: Percent of county residents who identify as White

26)PctBlack: Percent of county residents who identify as Black

27)PctAsian: Percent of county residents who identify as Asian

28)PctOtherRace: Percent of county residents who identify in a category which is not White, Black, or Asian

29)PctMarriedHouseholds: Percent of married households

30)BirthRate: Number of live births relative to number of women in county

31) Target\_DeathRate: Predicted death rate value

**Dataset Link:**

<https://www.kaggle.com/code/meghnumnirwal/cancer-predictions/notebook>

**Supervised Learning method which is going to be implemented in this project:**

Linear Regression and Random forest

**Target Variable used in this dataset:**

Target\_deathrate is the value where we take it as a target variable in our dataset. Which could be useful for training and testing purposes.