Health Status Indicators

Project Introduction

1. Goal of The Project

1. Analysing health status indicators and understand how behavioural factors such as obesity, tobacco use, diet, physical activity, drugs, alcohol usage and others contribute and relate to leading causes of death like obesity, heart disease, cancer etc and determine what affects them at the county level.

2. Using Predictive Modeling for predicting and understanding the leading causes of death like Lung Cancer, Breast Cancer, Colon Cancer, Poverty, Heart D iseases etc using various feature selection techniques.

2. Expected Outcome

- 1.After exploratory data analysis, a one-time report to be prepared providing insights and measures to be taken to improve healthcare across the country.
- 2.Predictive Analytics to be done using Machine Learning based models to provide and assist with evidence based decisions for better healthcare on Leading Ca uses of Death like Diabetes, Cancer, Heart Diseases etc.
- 3. Spread awareness about the issues affecting public health and a tool to local public health agencies for improving their community's health and provide the m with insights to assist in the development of public policies, health programs and prioritise funding in the most effective pathway.

3. About the Data

```
2.DEMOGRAPHICS.csv (Demographics indicator domain)

3.LEADING_CAUSES_OF_DEATH.csv (Leading Causes of Death indicator domain)

4.SUMMARY_MEASURES_OF_HEALTH.csv (Summary Measures of Health indicator domain)

5.MEASURES_OF_BIRTH_AND_DEATH.csv (Measures of Birth and Death indicator domain)

6.RELATIVE_HEALTH_IMPORTANCE. (Relative Health Importance indicator domain)

7.VULNERABLE POPS AND ENV HEALTH.csv (Vulnerable Populations and Environmental Health)

3141 rows, 44 columns
3000+ rows, 235 columns
```

1.HEALTHY_PEOPLE_2010.csv (Healthy People 2010 Targets and the U.S. Percentages or Rates)

8.PREVENTIVE_SERVICES_USE. (Preventive Services indicator domain)

9.RISK_FACTORS_AND_ACCESS_TO_CARE.csv (Risk Factors and Access to Care indicator domain)

3141 rows, 43 columns

4. Goals Updated

Earlier Goal: Broad goals of predicting heart disease, obesity and cancer.

Refined Goal: Other leading causes of death like average life expetancy, diabetes, different types of cancers added(Please see dependent variables below)

- We strive to provide as much useful insights as possible from the dataset to spread awareness about general issues pertaining to health and give correlations of each of the above attribute with other attributes in the dataset and find actionable insights.
- We intend to determine the important factors which are responsible or have correlation for higher values for each of the attributes in leading causes of deat h.

5. About Data Cleaning

A data dump was picked up from the link below in a zip file consisting of 10 csvs. The data dump consisted of the data dictionary as well as the description of the default values therein.

```
1. All missing values and default values were replaced to nan or handled appropriately for plotting purposes.
```

Default Values= [-9999,-2222,-2222.2,-2,-1111.1,-1111,-1,-9998.9]

2.During Modeling Phase missing values in all numerical attributes were replaced by the mean of the column

In [4]: import os from os import listdir from os.path import isfile, join codepath=r'C:\Users\Varun\Desktop\IDS Project\Codes' datapath=r'C:\Users\Varun\Desktop\IDS Project\Dataset' os.chdir(datapath) onlyfiles = [f for f in listdir(datapath) if isfile(join(datapath, f))] os.chdir(codepath) %run LibrariesImport.py # Loading Libraries %run DD.py# Loading datadictionary

All Libraries loaded

6. Data Dictionary (Description About All CSVs and Columns)

In [6]: DD.head(5) #manually change the integer to display more rows

Out[6]:

	PAGE_NAME	COLUMN_NAME	DESCRIPTION	IS_PERCENT_DATA
0	Demographics	State_FIPS_Code	Two-digit state identifier, developed by the N	N
1	Demographics	County_FIPS_Code	Three-digit county identifier, developed by th	N
2	Demographics	CHSI_County_Name	Name of county	N
3	Demographics	CHSI_State_Name	Name of State or District of Columbia	N
4	Demographics	CHSI State Abbr	Two-character postal abbreviation for state name	N

7. Exploratory Data Analysis

We have studied every attribute in all the files, merged all of them on the basis of the primary keys: 'State_FIPS_Code', 'County_FIPS_Code', 'CHSI_County_Name', 'CHSI_State_Name', 'CHSI_State_Abbr', 'Strata_ID_Number'.

- 1. The Data Granularity is of a county level.
- 2. Each row represents various values in percentages/or numeric of a particular county of a particular state.
- 3. The scope of the dataset is entire population of United States of America.
- 4. One Time Survey Data 1993-2003
- 5.Merged CSV of 9CSVs (Check About the Data Heading)

Independent Attributes: 'No_Exercise', 'Few_Fruit_Veg', 'Obesity', 'High_Blood_Pres', 'Smoker', 'Uninsured', 'Elderly_Medicare', 'Disabled_Medicare', 'Prim_Care_Phys_Rate', 'Dentist_Rate', 'FluB_Rpt', 'HepA_Rpt', 'HepB_Rpt', 'Meas_Rpt', 'Pert_Rpt', 'CRS_Rpt', 'Syphilis_Rpt', 'FluB_Rpt', 'HepB_Rpt', 'HepB_Rpt', 'Meas_Rpt', 'Pert_Rpt', 'CRS_Rpt', 'Syphilis_Rpt', 'Pap_Smear', 'Mammogram', 'Proctoscopy', 'Pneumo_Vax', 'Flu_Vac', 'Pap_Smear', 'Mammogram', 'Proctoscopy', 'Pneumo_Vax'', 'Flu_Vac', 'Population_Size', 'Population_Density', 'Poverty', 'Age_19_Under', 'Age_19_64', 'Age_65_84', 'Age_85_and_Over', 'White', 'Black', 'Native_American', 'Asian', 'Hispanic', 'No_HS_Diploma', 'No_HS_Diploma', 'Unemployed', 'Unemployed', 'Sev_Work_Disabled', 'Sev_Work_Disabled', 'Major_Depression', 'Major_Depression', 'Recent_Drug_Use', 'Recent_Drug_Use', 'Ecol_Rpt', 'Salm_Rpt', 'Shig_Rpt', 'Toxic_Chem', 'All_Death', 'Health_Status', 'Unhealthy_Days', 'LBW', 'VLBW', 'Premature', 'Under_18', 'Total_Births', 'Total_Births', 'Total_Beaths', 'Total_Deaths', 'Total_Deaths', 'Over_40', 'Unmarried', 'Late_Care', 'Infant_Mortality', 'IM_Neonatal', 'IM_Postneonatal' 'Homicide', 'Homicide'', 'Dependent Attributes(Leading Causes of Death): 'ALE', 'Diabetes', 'Lung_Cancer', 'Brst_Cancer', 'Col_Cancer', 'MVA', 'Stroke', 'Suicide', 'Injury', 'CHD'

After Careful EDA of all attributes we found many Observations, some of the intuitive ones have been listed below:

In [20]: PSU_Demo_VPEH_SMOH_RFAC_df.head()# merged, collated, cleaned ready for analysis dataset

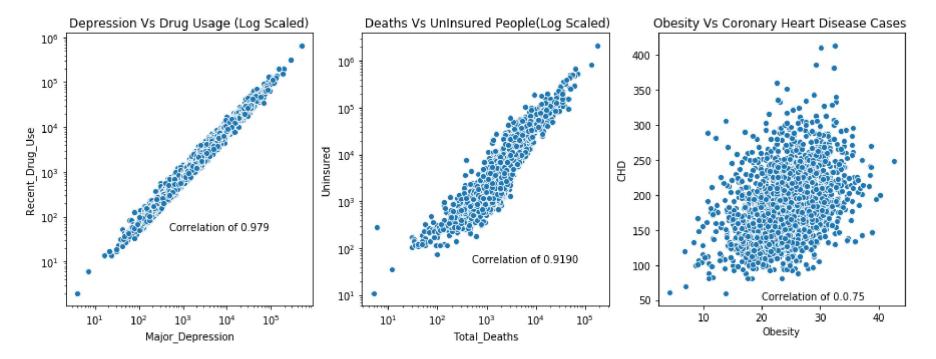
Out[20]:

	State_FIPS_Code	County_FIPS_Code	CHSI_County_Name	CHSI_State_Name	CHSI_State_Abbr	Strata_ID_Number	No_Exercise	Few_Fruit_Veg	Ob
0	1	1	Autauga	Alabama	AL	29	27.8	78.6	
1	1	3	Baldwin	Alabama	AL	16	27.2	76.2	
2	1	5	Barbour	Alabama	AL	51	NaN	NaN	
3	1	7	Bibb	Alabama	AL	42	NaN	86.6	
4	1	9	Blount	Alabama	AL	28	33.5	74.6	

5 rows × 71 columns

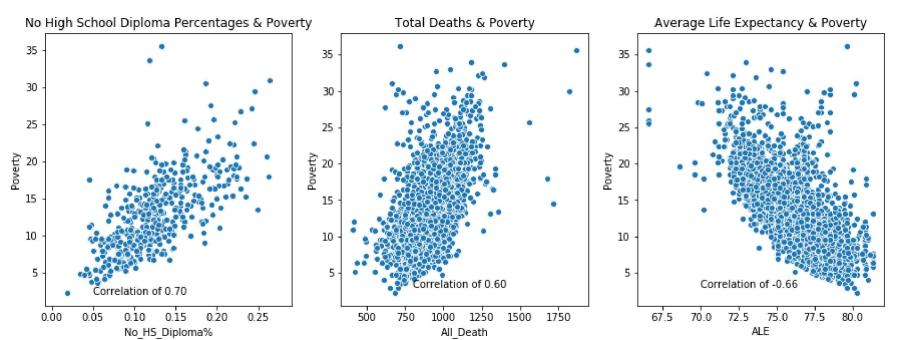
In [8]: os.chdir(codepath)
%run EDA.py
os.chdir(codepath)
%run plotter1.py

- 1. Higher cases of depression in a county are correlated with higher drug usage
- 2. Counties with more insured people, have higher death rates
- 3. Higher Coronary Heart Disease Cases in a county seen in counties with more obesity levels

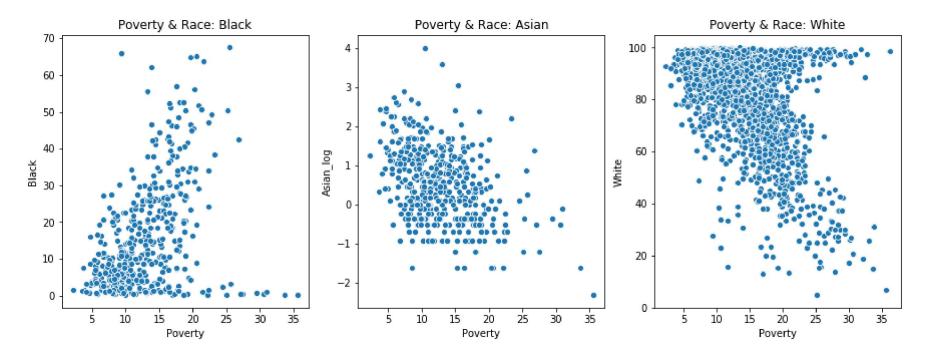


In [9]: os.chdir(codepath) %run plotter2.py

- 4 . Counties with Higher Povery Level have more people who are less educated
- 5 . Counties with Higher Povery Level have more death rates
- 6 . Counties with Higher Povery Level have more lower average life expectancy levels



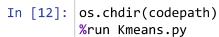
Poverty Ridden Counties are the ones which have higher population of Blacks

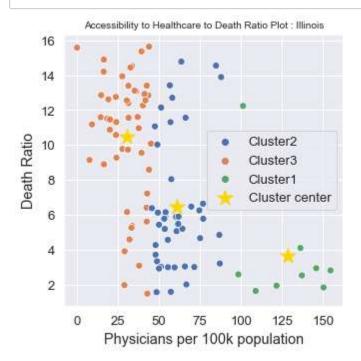


In [23]: os.chdir(codepath)
%run plotter4.py

<u>'</u>		1,7																								
Correlation Values of Various Attributes																										
No_Exercise	1.	0.4	0.51	0.30	0.55	0.40	-0.043	-0.51	-0.021	H0691	-0.013	-0.007	0.0009	-0.12	-0.11	0.40	0.54	0.14	0.21	0.50	0.41	V.J2	0.42	0.21	0.019	-0.04
Few_Fruit_Veg	0.4	1	0.38	0.18	0.27	0.22	-0.17	-0.41	-0.14	0.29	-0.17	-0.21	-0.062	-0.24	-0.23	0.21	0.12	0.082	0.21	0.3	0.24	0.21	0.37	0.24	0.1	-0.34
Obesity	0.57	0.38	1	0.52	0.41		-0.084	-0.26	-0.092	0.41	-0.069	-0.11	-0.051	-0.15	-0.14	0.37	0.38	0.1	0.22	0.39	0.38	0.38	0.3	0.26	-0.0094	-0.52
High_Blood_Pres	0.56	0.18	0.52	1	0.42	0.58	-0.08	-0.15	-0.065	0.39	-0.055	-0.098	-0.029	-0.12	-0.11	0.35	0.33	0.11	0.21	0.42	0.27		0.23	0.23	0.1	-0.52
Smoker	0.53	0.27	0.41	0.42	1	0.28	-0.097	-0.2	-0.056	0.31	-0.077	-0.1	-0.045	-0.12	-0.12	0.29	0.29	0.004	0.18	0.4	0.17		0.21	0.18	0.15	-0.49
Diabetes	0.48	0.22	0.52		0.28	1	-0.054	-0.17	-0.034	0.41	-0.037	-0.074	-0.014	-0.097	-0.09	0.33	0.37	0.086	0.11	0.31	0.3	0.33	0.25	0.19	0.053	-0.45
Uninsured	-0.043	-0.17	-0.084	-0.08	-0.097	-0.054	1	0.17	0.31	0.017	0.99	0.96	0.46	0.95	0.96	0.024	0.064	-0.024	-0.09	-0.02	-0.0017	-0.077	-0.19	-0.086	-0.12	0.057
Prim_Care_Phys_Rate	-0.31	-0.41	-0.26	-0.15	-0.2	-0.17	0.17	1	0.24	-0.14	0.17	0.23	0.087	0.26	0.23	-0.071	0.0087	-0.0017	-0.1	-0.21	-0.1	-0.11	-0.41	-0.085	-0.17	0.18
Population_Density	-0.021	-0.14	-0.092	-0.065	-0.056	-0.034	0.31	0.24	1	0.016	0.35	0.34	0.14	0.35	0.32	0.018	0.058	0.021	-0.023	0.015	0.028	-0.046	-0.17	-0.1	-0.11	0.026
Poverty	0.57	0.29	0.41	0.39	0.31	0.41	0.017	-0.14	0.016	1	0.02	-0.033	0.018	-0.067	-0.053	0.55	0.64	0.09	0.11	0.37	0.64	0.35	0.44	0.25	0.18	-0.66
No_HS_Diploma	-0.019	-0.17	-0.069	-0.055	-0.077	-0.037	0.99	0.17	0.35	0.02	1	0.96	0.49	0.94	0.96	0.034	0.076	-0.02	-0.081	-0.001	0.0039	-0.06	-0.2	-0.089	-0.13	0.04
Unemployed	-0.087	-0.21	-0.11	-0.098	-0.1	-0.074	0.96	0.23	0.34	-0.033	0.96	1	0.41	0.98	0.97	0.018	0.06	-0.019	-0.1	-0.041	-0.032	-0.076	-0.26	-0.1	-0.16	0.075
Sev_Work_Disabled	0.0069	-0.062	-0.051	-0.029	-0.045	-0.014	0.46	0.087	0.14	0.018	0.49	0.41	1	0.43	0.43	0.0079	0.036	-0.02	-0.038	0.0021	0.00086	-0.045	-0.087	-0.057	-0.062	0.029
Major_Depression	-0.12	-0.24	-0.15	-0.12	-0.12	-0.097	0.95	0.26	0.35	-0.067	0.94	0.98	0.43	1	0.98	0.0075	0.038	-0.028	-0.12	-0.069	-0.071	-0.089	-0.29	-0.12	-0.17	0.11
Recent_Drug_Use	-0.11	-0.23	-0.14	-0.11	-0.12	-0.09	0.96	0.23	0.32	-0.053	0.96	0.97	0.43	0.98	1	-0.0073	0.033	-0.03	-0.11	-0.061	-0.066	-0.089	-0.26	-0.11	-0.16	0.1
Premature	0.46	0.21	0.37	0.35	0.29	0.33	0.024	-0.071	0.018	0.55	0.034	0.018	0.0079	0.0075	-0.0073	1	0.6	0.13	0.088	0.31	0.55	0.37	0.22	0.28	-0.035	-0.6
Unmarried	0.34	0.12	0.38	0.33	0.29	0.37	0.064	0.0087	0.058	0.64	0.076	0.06	0.036	0.038	0.033	0.6	1	0.13	0.16	0.27	0.67	0.34	0.2	0.25	0.079	-0.6
Brst_Cancer	0.14	0.082	0.1	0.11	0.004	0.086	-0.024	-0.0017	0.021	0.09	-0.02	-0.019	-0.02	-0.028	-0.03	0.13	0.13	1	0.22	0.12	0.26	0.13	0.098	0.1	-0.015	-0.18
Col_Cancer	0.27	0.21	0.22	0.21	0.18	0.11	-0.09	-0.1	-0.023	0.11	-0.081	-0.1	-0.038	-0.12	-0.11	0.088	0.16	0.22	1	0.23	0.3	0.25	0.15	0.13	0.087	-0.23
CHD	0.56	0.3	0.39	0.42	0.4	0.31	-0.02	-0.21	0.015	0.37	-0.001	-0.041	0.0021	-0.069	-0.061	0.31	0.27	0.12	0.23	1	0.29	0.46	0.25	0.21	0.037	-0.52
Homicide	0.41	0.24	0.38	0.27	0.17	0.3	-0.0017	-0.1	0.028	0.64	0.0039	-0.032	0.00086	-0.071	-0.066	0.55	0.67	0.26	0.3	0.29	1	0.21	0.44	0.27	0.2	-0.64
Lung_Cancer	0.52	0.21	0.38	0.44	0.59	0.33	-0.077	-0.11	-0.046	0.35	-0.06	-0.076	-0.045	-0.089	-0.089	0.37	0.34	0.13	0.25	0.46	0.21	1	0.18	0.3	0.087	-0.65
MVA	0.42	0.37	0.3	0.23	0.21	0.25	-0.19	-0.41	-0.17	0.44	-0.2	-0.26	-0.087	-0.29	-0.26	0.22	0.2	0.098	0.15	0.25	0.44	0.18	1	0.15	0.37	-0.42
Stroke	0.27	0.24	0.26	0.23	0.18	0.19	-0.086	-0.085	-0.1	0.25	-0.089	-0.1	-0.057	-0.12	-0.11	0.28	0.25	0.1	0.13	0.21	0.27	0.3	0.15	1	0.022	-0.42
Suicide	0.079	0.1	-0.0094	0.1	0.15	0.053	-0.12	-0.17	-0.11	0.18	-0.13	-0.16	-0.062	-0.17	-0.16	-0.035	0.079	-0.015	0.087	0.037	0.2	0.087	0.37	0.022	1	-0.12
ALE	0.64	0.34	0.52	0.52	0.40	0.45	0.067	0.18	0.026	0.66	0.04	0.075	0.020	0.11	0.1	0.6	0.6	0.18	0.23	0.52	0.64	0.65	0.42	0.42	0.12	1
	Exercise	Veg	Obesity	Pres	Smoker	Diabetes	Uninsured	Rate	ensity	Poverty	oloma	loyed	abled	ssion	_Use	ature	Unmarried	ancer	ancer	유	Homicide	ancer	MVA	Stroke	Suicide	ALE
	No_Ex	Few_Fruit_Veg	٥	High_Blood_Pres	ି ଓ	Dia	Z	Phys	on_D	ď	No HS Diploma	Unemployed	A Dis	Depre	Drug	Prematur	Unm	3rst_Cancel	Col_Cancer		Hon	Lung_Cance		93	Ø	
	z	Few		High				im_Care_Phys_	Population_Density		N N)	Sev_Work_Disabled	Major_Depression	Recent_Drug_Use							3				
								Prim	Ā				8	2	œ											

Few Observations:
No High School Diploma, Has high correlation with drug use, depression and unemployment.
Average life expectancy is negatively correlated with obesity and high blood pressure.
No Excercise Correlated with Obesity and high Blood Pressure, Heart Disease.
Lung Cancer & Smoker are highly correlated



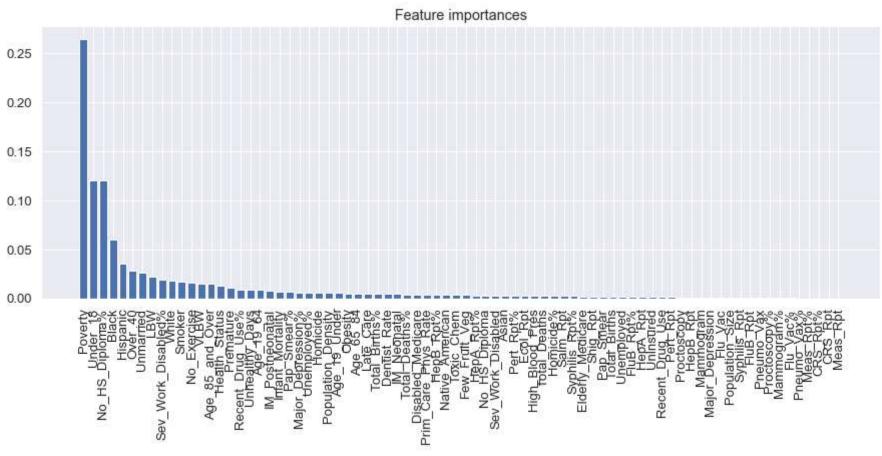


Ratio of Deaths Vs Physician Rate have been clustered by Kmeans, an inversely proportional relationship shows that staffing of medical depts is one of the top concerns.

8. Modelling & Inferencing

- # choose one of the leading causes from the list above eg. ALE(Average Life Expectancy)
- # the baseline model predicts the median always

```
In [14]: os.chdir(codepath)
          %run Modeling.py
          ToBePredicted=['ALE','Diabetes','Lung_Cancer','Brst_Cancer','Col_Cancer','Brst_Cancer%','Col_Cancer%','Lung_Cancer%','MVA
          strval='ALE' # choose one of the leading causes from the list above eg. ALE(Average Life Expectancy)
          colname=ToBePredicted[ToBePredicted.index(strval)]
          modelrun(colname,mlmodel,X1)
         ['ALE', 'Diabetes', 'Lung_Cancer', 'Brst_Cancer', 'Col_Cancer', 'Brst_Cancer%', 'Col_Cancer%', 'Lung_Cancer%', 'MVA', 'M VA%', 'Stroke', 'Stroke%', 'Suicide', 'Suicide%', 'Injury', 'Injury%', 'CHD', 'CHD%']
          ------Data Specs-----
          Amount of Training Data 2197
          Amount of Training Labels Data 2197
          Amount of Testing Data 942
          Amount of Testing Labels Data 942
          Baseline : Train Root Mean Squared Error: 2.003518752569599
          RandomForest: Train Root Mean Squared Error: 0.3702769004339355
          Baseline : Test Root Mean Squared Error: 2.0085195401339235
          Random Forest: Test Root Mean Squared Error: 0.9241368699409664
          Feature ranking:
          Attribute Predicted
                                   ALE
          Predictor Columns
          36
                         Poverty
          65
                      Under_18
          47
                No_HS_Diploma%
          42
                          Black
          45
                      Hispanic
          Name: cols, dtype: object
```



- 1. What is hardest part of the project that you've encountered so far?
 - Defining a concrete problem definition since the dataset consists of wide variety of health related data.
 - Collating all datasets/csvs and merging them into one dataframe.
 - Imputing Missing and Default Values
 - The incomplete data dictionary and missing reference links.
- 2. What are your initial insights?
 - Understanding and exploration of the data (of all attributes) in all csvs.
 - We have shortlisted the attributes (by calculating feature importance using random forest)

from the entire dataset which are correlated or are responsible for their 'effects' with the leading causes of death attributes

- Useful Correlations achieved during Exploratory Data Analysis.
- Modelling phase summarises all the model outputs with useful statistical information

Observations: Exploratory Data Analysis:

- 1. There is a positive correlation of poverty and unemployment.
- 2. Negative correlation of poverty and population density.
- 3. No relationship between poverty and depression observed.
- 4. Strong positive correlation between population density and depression.
- 5. Positive correlation between poverty and No High School Diploma Percentages.
- 6. Population size and E.Colli, Salmonella and Shigella Correlated (Hygiene Related Diseases)
- 7. Depression & Drug Use Positive Correlation
- 8. Povery & Number of Deaths Positive Correlation
- 9. Poverty & Average Life Expectancy Negative Correlation
- 10. UnInsured People Vs Number of Deaths Positive Correlation
- 11. Depression Vs Suicide Rate Positive Correlation
- 12. Heart Disease Vs Obesity Positive Correlation

Results: Random Forest Regressor(9 Models) for each leading cause of death & Variable Importance (determining attribute for that cause):

```
1.Average Life Expectancy Train MSE=0.361304375017007
                                                        Test MSE=0.977328605007327
1.Poverty, 2.No_HS_Diploma, 3.Under_18, 4.Black, 5.Over_40
Diabetes
                          Train MSE=0.814102003689945
                                                       Test MSE=2.29140400291951
1.Obesity, 2.No_HS_Diploma, 3.Unmarried, 4.Poverty, 5.Recent_Drug_Use%
                          Train MSE=3.35822064794277 Test MSE=8.40868361996625
1. Smoker, 2. NO_HS_Diploma, 3. Poverty, 4. Major_Depression, 5. Sev_Work_Disabled%
4.Breast Cancer:
                                                      Test MSE=5.09185013278892
                          Train MSE=1.96858600852315
1.Major_Depression, 2.Recent_Drug_Use%, 3.Black, 4.Unemployed%, 5.Population_Size
5.Colon_Cancer
                          Train MSE=1.40279823772743
                                                       Test MSE=3.99720364350038
1.Unmarried, 2.Hispanic, 3.No_Exercise, 4.Smoker, 5.Major_Depression
6.Motor Vehicle Injuries: Train MSE=2.2402199328038
                                                      Test MSE=5.96104038324399
```

7.Heart Stroke Train MSE=4.58634474894955 Test MSE=13.4173059174712

1.Under_18, 2.Population_Density, 3.Recent_Drug_Use%, 4.Asian, 5.No_HS_Diploma

- 1.Black, 2.High_Blood_Pres, 3.No_HS_Diploma, 4.Premature, 5.Mammogram
- 8. Suicide: Train MSE=1.15818004036267 Test MSE=3.12497294467384 1.Population_Density, 2.Unemployment, 3.Recent_Drug_Use%, 4.Major_Depression, 5.Under_18
- 9. Coronary Heart Disease: Train MSE=12.4806249718433 Test MSE=34.9709746818494
- 1.No_Excercise, 2.No_HS_Diploma, 3.Smoker, 4.Unemployment, 5.Major_Depression
- 3. Are there any concrete results you can show at this point? If not, why not?
- Set of features for every leading cause of death has been listed above. (Goal of the project)
- We intend to increase our accuracy of our models(9) using grid search technique.
- Making a generic framework to understand the models better which takes custom input of data (work in progress)
- 4. Going forward, what are the current biggest problems you're facing
- Current problems include the deployment of the models as a web application (work in progress).
- Increasing the accuracy of the models (work in progress)
- 5.Do you think you are on track with your project? If not, what parts do you need to dedicate more time to?
 - Moving forward, We intend to improve all the models using grid search, generalise them and extract more insights.
- 6. Given your initial exploration of the data, is it worth proceeding with your project, why? If not, how are you going to change yo ur project and why do you think it's better than your current results?

The dataset is quite promising because of so many interesting and useful features.

10. Next Steps

- Project is 90% Complete.
- We still intend to improve the accuracy of all the models using grid search.
- A web-page with django framework to be deployed, with backend in python for the project. (Not yet decided)
- Generalising the code for custom inputs for the model to get better insight into the data.