NORTHWESTERN DATA SCIENCE BOOT CAMP

Impact of Environment Quality on Lung Cancer Mortality

Amelia Horton, Jenna Jorstad, Nina Padgett, Indranil Roy, Ying Sun

Project Quad Chart

What?

Present study aims to determine whether ambient air pollutant exposures and various environment quality indexes are associated with lung cancer deaths.

Why?

Exposure to ambient air pollutants has been associated with increased lung cancer incidence and mortality. However, little is known about the impacts of poor quality land and water environments alongside the air pollution exposures on lung cancer-related deaths.

How?

Using state-of-the-art supervised learning methods–i) linear regression, ii) random forest regression, and iii) support vector regression–and ETL.

Future Scope...

Global dataset, use of forecasting model...
Building an app that can work on real-time inputs, giving lung cancer risk outputs...

Dataset Snapshot & Glossary of Column Names

Cleaned dataset was prepared by dropping unnecessary columns, deleting rows with Null values, and deleting rows containing asterisk symbols.

FIPS_code	Lung_Cancer	PM2.5	Land_EQI	Sociod_EQI	Built_EQI	CLU50_1	PM10	SO2	NO2	O3	CO	CN	Disel	CS2	Air_EQI	Water_EQI	EQI	LCI	UCI
1001	73.9	12.06	-0.7065906	0.6704357	-0.4973013	1	15.07	10.66109	123.6576	522.38	4.463225	0.054815	0.388556	0.00808	0.9553846	-1.109728	0	64.3	84.6
1003	68.4	11.12	-1.084299	0.5530728	0.4015849	2	19.99	17.14685	247.7423	540.79	12.87583	0.021069	0.428278	0.00109	0.7179643	-0.5659107	0.2	63.9	73.1
1005	76.1	12.36	-1.28147	-1.236294	0.0488544	3	15.77	23.25712	183.1936	896.42	19.62054	0.014027	0.199725	0.000513	0.1310074	-0.9780902	-0.95	63.3	90.9
1007	86.4	12.24	-0.8274103	-0.6000178	-1.290857	4	14.92	7.630953	127.7799	563.48	2.951976	0.009613	0.211741	0.000225	0.065289	-0.9681726	-1.09	71.2	104.1
1009	73.1	12.97	-0.6229339	0.2965088	-1.26274	5	17.9	8.913795	95.19809	561.94	9.362216	0.022128	0.3001	0.000429	0.4021944	-0.7186447	-0.51	64.5	82.6
1011	72.4	12.16	-1.258014	-1.82397	-1.795921	6	15.95	21.0864	172.55	652.28	15.3636	0.004206	0.16053	0.000728	-0.309186	-1.451335	-2.08	52.8	97.3

- FIPS_code - Federal Information Processing System codes are used to define US states and counties.
- Lung_cancer - lung cancer fatalities per 100,000 people
- EQI - Environmental Quality Index; different types deal with different aspects of environment quality: land, air water, built (indoor), and sociod (socioeconomic)
- CLU50_1 - a type of air pollutant particles
- LCI - lower confidence interval
- UCI - upper confidence interval

- PM2.5 - tiny particles or droplets in the air, < 2.5 microns width
- PM10 - course pollutants such as dust, 10 microns width
- SO2 - sulfur dioxide
- NO2 - nitrogen dioxide
- O3 - ozone
- CO - carbon monoxide
- CN - Cyano Compound
- Disel - diesel (misspelled in dataset)/Hydrocarbon
- CS2 - carbon disulfide

Data Handling

Split cleaned dataset into features and target arrays

FIPS_code	Lung_Cancer	PM2.5	Land_EQI	Sociod_EQI	Built_EQI	CLU50_1	PM10	SO2	NO2	O3	CO	CN	Disel	CS2	Air_EQI	Water_EQI	EQI	LCI	UCI
1001	73.9	12.06	-0.7065906	0.6704357	-0.4973013	1	15.07	10.66109	123.6576	522.38	4.463225	0.054815	0.388556	0.00808	0.9553846	-1.109728	0	64.3	84.6
1003	68.4	11.12	-1.084299	0.5530728	0.4015849	2	19.99	17.14685	247.7423	540.79	12.87583	0.021069	0.428278	0.00109	0.7179643	-0.5659107	0.2	63.9	73.1
1005	76.1	12.36	-1.28147	-1.236294	0.0488544	3	15.77	23.25712	183.1936	896.42	19.62054	0.014027	0.199725	0.000513	0.1310074	-0.9780902	-0.95	63.3	90.9
1007	86.4	12.24	-0.8274103	-0.6000178	-1.290857	4	14.92	7.630953	127.7799	563.48	2.951976	0.009613	0.211741	0.000225	0.065289	-0.9681726	-1.09	71.2	104.1
1009	73.1	12.97	-0.6229339	0.2965088	-1.26274	5	17.9	8.913795	95.19809	561.94	9.362216	0.022128	0.3001	0.000429	0.4021944	-0.7186447	-0.51	64.5	82.6
1011	72.4	12.16	-1.258014	-1.82397	-1.795921	6	15.95	21.0864	172.55	652.28	15.3636	0.004206	0.16053	0.000728	-0.309186	-1.451335	-2.08	52.8	97.3
1013	58.5	11.66	-1.678537	-1.295534	-0.0580627	7	13.18	13.8314	134.137	578.7	8.018968	0.018163	0.206896	0.000311	-0.070452	-1.528175	-1.3	46.5	72.9
1015	79.6	13.15	-0.0666635	-0.3698209	0.4047821	8	16.17	22.16249	176.6921	531.53	25.99952	0.038176	0.499753	0.00101	1.027328	-0.9281401	0.17	73.1	86.5
1017	72.5	12.31	-0.2369899	-0.699318	-0.4082792	9	13.97	39.79517	198.9622	634.03	22.57713	0.029555	0.318671	0.00133	0.6116033	-1.004405	-0.46	61.7	84.7
1019	85.9	13.44	-0.1945019	-0.6738133	-0.6189157	10	16.78	34.67909	221.7159	838.14	80.6963	0.01446	0.276152	0.000334	0.1769833	-0.9704146	-0.66	73.3	100.5
1021	70.5	12.28	-0.6654943	-0.141334	-0.447298	11	15.96	8.044921	124.5018	534.52	3.216023	0.01391	0.236606	0.000259	0.1707421	-1.059535	-0.54	60.6	81.5
1023	50.2	11.79	-1.716982	-1.242582	-0.7427115	12	8.64	21.07971	134.7725	641.07	18.18652	0.004407	0.126131	0.000118	-0.362408	-1.24212	-1.58	37.3	66.7
1025	67.4	11.74	-1.877054	-0.9321283	-0.2272212	12	10.23	17.97536	137.608	592.35	10.71781	0.005216	0.12008	0.000275	-0.130679	-1.122216	-1.22	55.5	81.3
1027	87.5	13.29	-0.189493	-0.7478549	-0.8776811	10	16.1	20.85784	164.3854	485.97	12.27522	0.008423	0.224393	0.000122	-0.172236	-0.9876922	-0.9	69.6	109.1
1029	85.8	13.25	-0.2598838	-0.4756131	-1.233658	5	14.36	37.77223	256.7616	592.32	43.44772	0.010388	0.272729	0.000166	0.0225133	-0.8583899	-0.86	68	107.1
1031	72.1	11.78	-1.371213	-0.1633784	0.0374083	11	15.7	14.78714	170.562	789.66	16.28554	0.025552	0.229195	0.000822	0.4223093	-1.086293	-0.49	62.9	82.4
1033	78.3	11.65	-0.2068101	0.0195102	0.3970989	13	8.01	2.053658	100.4087	1178.5	81.24455	0.041337	0.455037	0.00111	1.04411	-0.9147808	0.29	69.5	88.1
1035	74.7	11.63	-2.215354	-1.386962	-1.882343	6	14.72	18.87658	134.8703	551.7	11.60842	0.004849	0.142302	0.000109	-0.458279	-0.9339556	-2.17	57.9	95.4
1037	66.6	12.2	-0.3489624	-0.7830176	-2.316545	14	16.73	12.70444	145.7348	534.47	5.380743	0.009228	0.209585	0.000155	-0.108911	-1.515493	-1.56	50	87.9

y

X

... train-test Split… followed by implementing different supervised regression models…

Linear Regression Model

- Split cleaned dataset into features and target arraysOur goal was to find a correlation between different environmental factors and the correlation they have to lung cancer, so we chose to test a linear regression model, as it finds the best linear line and the best values of intercept and coefficients to reduce error
- Once we cleaned up the dataset and removed unnecessary columns we could define our x and y values to test a linear regression model
- We used train test split to create training and testing data, and from there we were able to create the linear regression model using sklearn
- We used the multiple linear regression model in order to apply multiple independent variables

Results from Linear Regression Model

Training Score: 0.999206338853848

Testing Score: 0.9991387624768184

What does this mean?

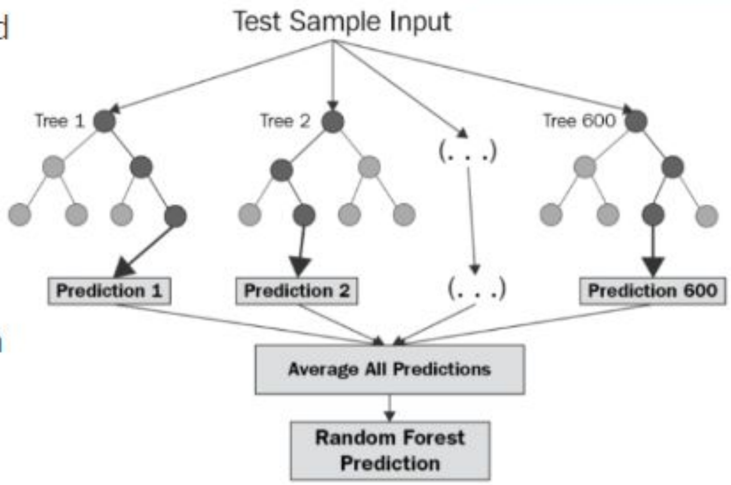
- This r-squared value tells us the strength of the relationship between our model and the dependent variable is very good.
- The regression model fits our observations almost to 100%

Random Forest Regression Model: Quick Revisit

- (i) A supervised learning algorithm that uses **ensemble learning** method for regression.

- (ii) Ensemble learning combines predictions from multiple machine learning algorithms to make a more accurate prediction.

- (iii) A Random Forest operates by constructing several parallel decision trees during training and outputting the mean of the classes as the prediction of all the trees.



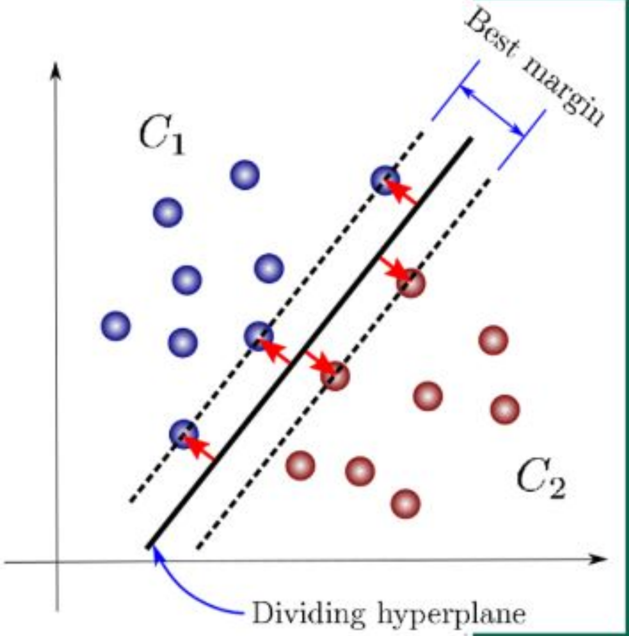
https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

Performance of Random Forest Regression Model on Cleaned Dataset

- Quantifying RandomForestRegressor model with mean squared error (MSE) and r2 Score as they are commonly used metrics
- Mean squared error (MSE): 1.0901631707664408
- R-squared (r2): 0.9961216810575816
- A good MSE score should be close to zero, while a good r2 score should be close to 1
- The obtained value suggest that RandomForestRegressor works efficiently on the dataset
- Five-fold cross validation (r2) : 0.99646495, 0.93781531, 0.99576459, 0.99580746, 0.99470601
- These results suggest RandomForestRegressor works efficiently on the dataset

Support Vector Regression (SVR) Model: Quick Revisit

- The objective of Support Vector is to find a hyperplane in N dimensional space which can classify data-points.
- The data points lying on the margin and nearest to Hyperplane are called Support Vectors.
- Now when most of the data lies mostly within the best margin towards each side of hyperplane then SVR or Support Vector Regression can be used to identify and predict the dependent data.



https://www.numpyninja.com/post/a-dive-into-support-vector-regression-with-python

Performance of SVR Model on Cleaned Dataset

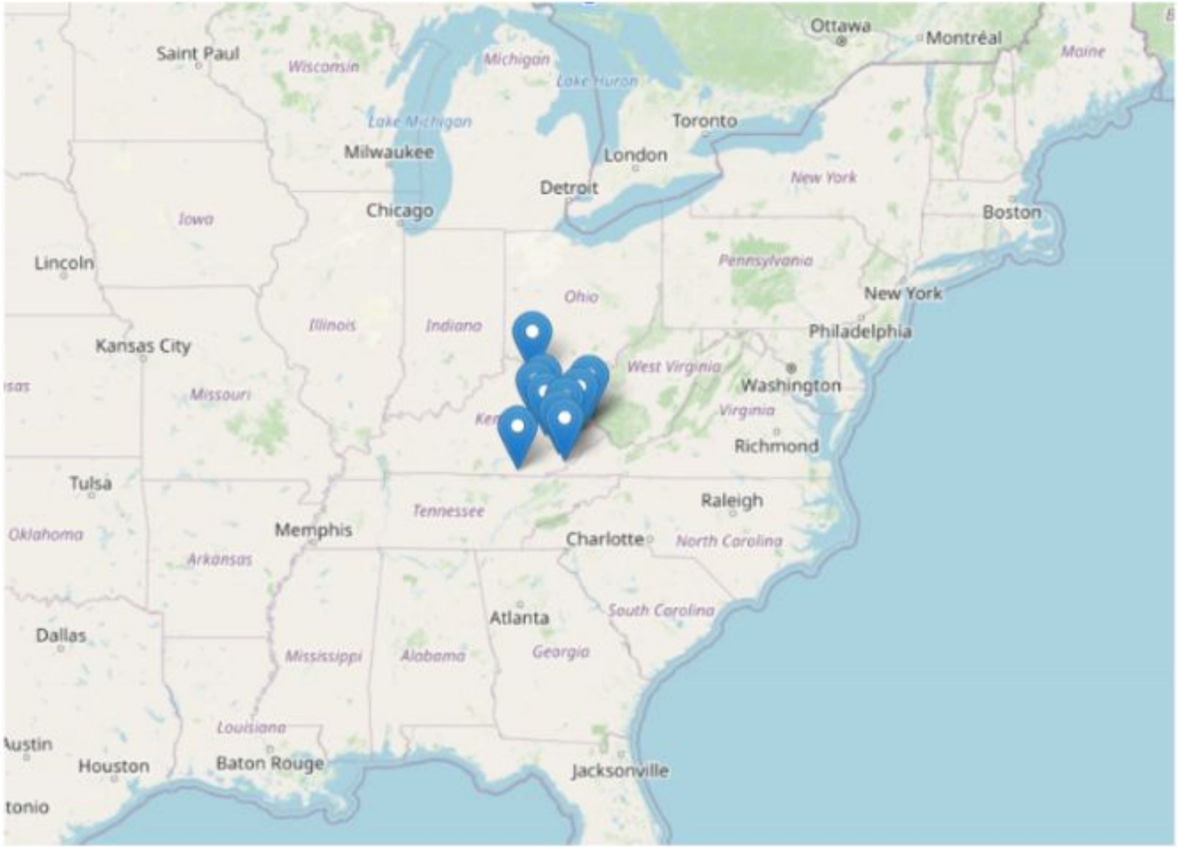
- Mean squared error (MSE): 3.286872270493343
- R-squared (r2): 0.9647445579626635
- The obtained value suggest that SVR works well on the dataset but it is less efficient than RandomForestRegressor
- Five-fold cross validation (r2) : 0.91820326, 0.81007791, 0.96688675, 0.86465052, 0.86909039

Different Model's Performances & Final Analyses

Model Name	mean squared error (MSE)	R-squared (r2)
LinearRegression	0.5137256106065202	0.9991387624768184
RandomForestRegressor	1.0901631707664408	0.9961216810575816
SVR	3.286872270493343	0.9647445579626635

- LinearRegression model works a touch better than RandomForestRegressor model.
- Both LinearRegression and RandomForestRegressor works much better than SVR model.
- SVR model can also be used for prediction with reasonable accuracy for this particular dataset
- Overall, our findings can help understanding the use of various supervised learning algorithms for predicting lung cancer mortality based on various air pollutants features as well as land and water quality indexes.
- The scope of this work is limited. The results solely depend on the dataset, data range used for training, and the features considered for training.

Kentucky Findings



All of the top 10 highest rates of lung cancer in our dataset occur in eastern Kentucky.