Life Expectancy Predictions Based on the on the EPA's EJScreen Dataset

Levi Schult [*1] **Mary Stirling Brown** [*2]

Abstract

This paper utilizes machine learning techniques in order to predict the life expectancy of a county based on its environmental conditions sourced from the EPA's EJScreen dataset. We compared predictions using only environmental features versus those combined with demographic information. Based on the results from the regression models, environmental features from the EPA's EJScreen dataset are not well-suited to perform accurate predictions for the life expectancy when compared to the utilizing the demographic features as well.

1. Introduction

Environmental data science is a field that seeks to leverage newly-available, massive datasets to understand the effects of the ever-evolving environment on mental and physical health. These studies can focus on generating higher resolution measurements for pollution levels (Lary et al., 2015) to understanding how air pollution conditions can have detrimental mental health impacts on urban dwellers (Saxena & Dodell-Feder, 2022). Frequently, the topic of environmental racism, or how pollution and other forms of environmental degradation most often affects minoritized and or low-income communities is investigated (Patnaik et al., 2020). Environmental racism can take the form of increased air pollution from traffic or chemical manufacutring plants or frequent placement of hazardous waste sites near Black, Indigenous, and People of Color (BIPOC) neighborhoods, among other forms (White, 2018; Patnaik et al., 2020). Anthropogenic climatic changes are also already affecting minoritized communities disproportionately due to droughts, intense heat waves, and severe weather events (Berberian et al., 2022). With an established body of research that assuredly only scratches the surface of the size and scope of harmful environmental impacts facing BIPOC communities across the United States, it is abundantly clear that community members and policymakers must know how to prioritize communities for solutions to these conditions. The method frequently used is a mapping system for screening what pollution challenges face a community. These screening tools have been used in California as well as Maryland on a state level as well as the Environmental Protection Agency's (EPA) EJScreen mapping tool (English et al., 2013; Williams et al., 2022; EPA, 2022). EJScreen is composed of various environmental factors that can potentially harm human life as well as demographic information on a Census block group level. Some of the environmental factors are that of proximity to pollution sources like traffic or EPA Risk Mitigation Protocol (RMP) sites. Others are pollution measurements like ozone levels, particulate matter, or prevalence of homes with lead paint. Demographic features include income, education, unemployment, and racial breakdowns of the Census block group (EPA, 2022). There have been previous studies using these datasets in large urban areas such as Detroit and Atlanta, however, this study is the first of its kind in its national scope (Shkembi et al., 2022; Heidger et al., 2021). Mullen et al. (2023) have highlighted serious oversights in the creation of the EJD, mainly that it does not include some environmental impacts relating to mining, land degradation, energy extraction, and indigenous sovereignty. Further, the EJD also has its own caveats related to how measurements of risks were determined for each census tract. These shortcomings are described in Section 2. The EPA states, however, that the EJScreen Dataset (EJD) is a screening tool, highlighting areas that may be facing serious environmental justice issues and may need follow up studies performed (EPA, 2022). We additionally use the U.S. Small-area Life Expectancy Estimates Project (USALEEP) dataset from the National Center for Health Statistics for our target to be predicted by features in the EJD (Arias et al., 2018). Our study is an exploration of the capabilities of the EJD in predicting life expectancy across the United States. We acknowledge, however, that life expectancy is closely tied to other demographic factors such as income, race, and education and that the causes for this trend are difficult to disentangle due to the layers of economic, health, and racial oppression that frequently overlap for communities facing cases of environmental injustice.

2. Methods

2.1. Caveats

The EJD is a substantial dataset, however it is not without its drawbacks. Due to its national-scale Census block group level of resolution, the EPA was limited by what datasets were appropriate for inclusion. Thus, other environmental impacts that are localized to a particular census block, a constituent member of a block group similar to how block groups combine to create a tract, are not included whatsoever in the EJD. Additionally, many of the environmental features are proxies for potential impact rather than direct measurement. This is especially true when considering the proximity-based variables included in the EJD, such as traffic proximity, Superfund proximity, or hazardous waste proximity. For these measures, there is no determination of true exposure since the level of emissions from these sources can vary drastically from source to source, and it is unknown how prevalent the pollutant is in the air, water, or soil. Finally, the particulate matter (PM), ozone, and air toxics features were determined on a census tract level, so such measurements were applied to all block groups within them (EPA, 2022). These caveats are important to remember, as there are significant levels of uncertainty, but the EJD does capture a rough outline of the environmental threats facing communities across the United States.

2.2. Pre-Processing

These datasets required several steps of pre-processing before being suitable for machine learning techniques. The largest challenge was the difference in naming conventions across the USALEEP and the EJD. We initially wanted to do a block group level analysis, however, USALEEP used block group numbers from the 2015 census while EJD used numbers from the 2020 census (EPA, 2022; Arias et al., 2018). Due to limits on the population size of a census block group, a collection of census blocks, the number denoting a census block can change, as the block will split or merge (U.S. Census Bureau, 2022). This complicated the comparison between the 2010 and the 2020 census, as the changes were irregular and difficult to predict. Thus, we decided to do a county-level analysis. This still required some editing, as USALEEP included words like ”County, Borough, Parish, Census Area” in their county names, unlike the EJD. Using some basic regular expressions, we were able to remove these titles to get a list of 1,802 counties in common. For each county, the mean and median life expectancy was calculated, as well as the mean, median, and standard deviation of all environmental and demographic features included in the EJD. NaNs were removed before usage in the various machine learning methods as described below, leaving 1,543 counties in our dataset. Initially we did not separate the features in the dataset into environmental versus demographic ones, but in our analyses including demographic variables, they outperformed all environmental ones in predicting life expectancy. This is due to relationships within USALEEP, where according to Arias et al. (2018), ”Among census tracts that belong to the lowest quartile of life expectancy at birth (56.3-75.7), more than one-half ... have predominantly non-Hispanic black populations (51.0%), and consist of populations with low educational attainment (56.7%) and low median income (60.9%).” The corrollary of these trends is also seen in USALEEP, where census tracts with the highest life expectancy were more frequently those that had higher education levels and high median income. Realizing this, we decided to keep these results, but focus on the environmental indicators by not including demographic features in our machine learning models.

2.3. Feature Selection

Because the EJD had over 200 features, we tested two different methods to perform feature extraction. First, we performed Principal Component Analysis (PCA) to reduce the dimensionality of our dataset. We were able to capture 95% of the variance in the EJD with 55 principal components, which can be visualized in Figure 1. When removing the demographic features from the EJD, we achieve 95% variance with only 39 principal components. When we specify using the PCA reduced versions of our datasets or principal components in various models, we are using the number of principal components that captures 95% of the variance.

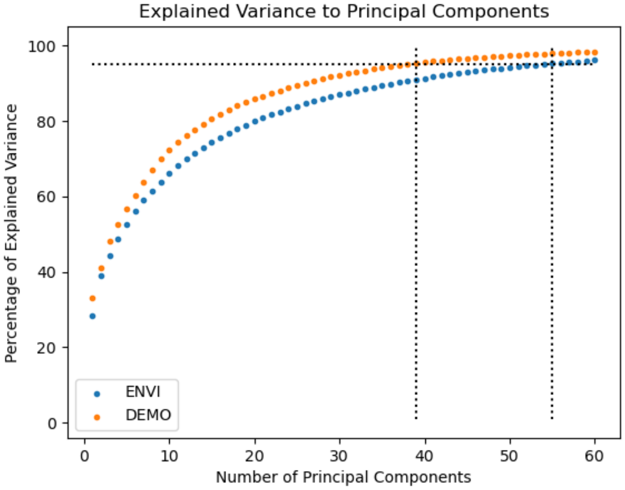


Figure 1. Principal Components needed to achieve 95% variance with 55 principal components for all features in the dataset (DEMO) and 39 for only environmental (ENVI).

We also used a filter method to perform feature selection. Forward sequential search and backwards sequential search were not implemented because they are model-dependent, and this was an inefficient process due to their large computational cost, especially when testing multiple models. The metric used for the filter method was Mutual Information (MI). MI measures the mutual dependence between two variables. In other words, it is able to represent how much an independent feature is able to predict the target variable. A high MI means that that there is greater certainty that the feature is able to predict the target variable, whereas a very low MI means that there is weak correlation between these variables.

This approach for feature selection demonstrated to us that certain demographic features were outweighing environmental factors. This can be seen in the Figure 2, which graphs the 10 features with the highest MI value for the EJD with and without demographic features. From Figure 2(a) with the entire dataset, it can be observed that demographic features, such as education and income statistics, overshadow environmental ones. By taking these demographic features out, we could focus on the effect of environmental features such as those seen in Figure 2(b). This MI metric allowed us to rank features in the dataset and explore how many to include from this ranking for the best performance of whatever model was being tested.

2.4. Regression Methods

Various regression methods were used in order to predict the life expectancy mean or median from the EJD. These were tested first on our principal components and then various numbers of features as ranked by MI. All regression methods were coded in Python with the SciKit-Learn libraries.

All experiments were tested using k-fold cross validation to compute the model performance. Because our dataset has over 1,500 samples, we chose a k = 5 and averaged the model performance across all 5 folds. The metric used to measure performance is the Mean Squared Error (MSE) to compute the amount of error from our predicted life expectancy values to the true value. The equation is:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y_i})^2 \tag{1}$$

where m is the number of samples in the dataset, y_i is the actual value of the ith data point, and $\hat{y_i}$ is the predicted value of the ith data point. The overall measure of the model performance, P_j, is defined as the average of the MSE for each k-fold test represented by the index j:

$$P_j = \frac{1}{k}\sum_{j=1}^{k}(MSE_j) \tag{2}$$

2.4.1. Multiple Linear Regression

Multiple linear regression works by fitting a linear model to observed data with multiple n features. For our case, the model would be in the equation form:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n \tag{3}$$

where θ_0 is the intercept term, $\theta_1, \theta_2, \cdots, \theta_n$ are the regression coefficients, $\hat{y}$ is the predicted life expectancy, and n is the number of features or dimensions.

It is able to calculate this linear model by minimizing the loss of the sum-of-squared errors cost equation:

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(y^{(i)} - \theta^T x^{(i)})^2 \tag{4}$$

where m is the number of samples, $y^{(i)}$ is the actual life expectancy value for the ith sample, $x^{(i)}$ is the vector of predictor variables for the ith sample, and θ is the vector of coefficients for the model. This loss is minimized via gradient descent.

In addition to multiple linear regression, polynomial regression was also tested with degrees of 2, 3, and 4. However, due to the amount of features being used, it proved to be computationally demanding, and the MSE varied from 10 to over 1,000, demonstrating a poor fit to the data.

2.4.2. Locally-Weighted Linear Regression

Locally-Weighted Linear Regression (LWLR) is non-parametric, meaning that the parameters θ's that are seen in linear regression are computed for each query point x_q. A weight is computed for each sample in comparison to a given x_q. Higher weights indicate that the sample is closer to x_q compared to samples with lower weights. The weight for each sample is calculated as:

$$w^{(i)} = \exp\left(-\frac{(x^{(i)} - x_q)^2}{2\tau^2}\right) \tag{5}$$

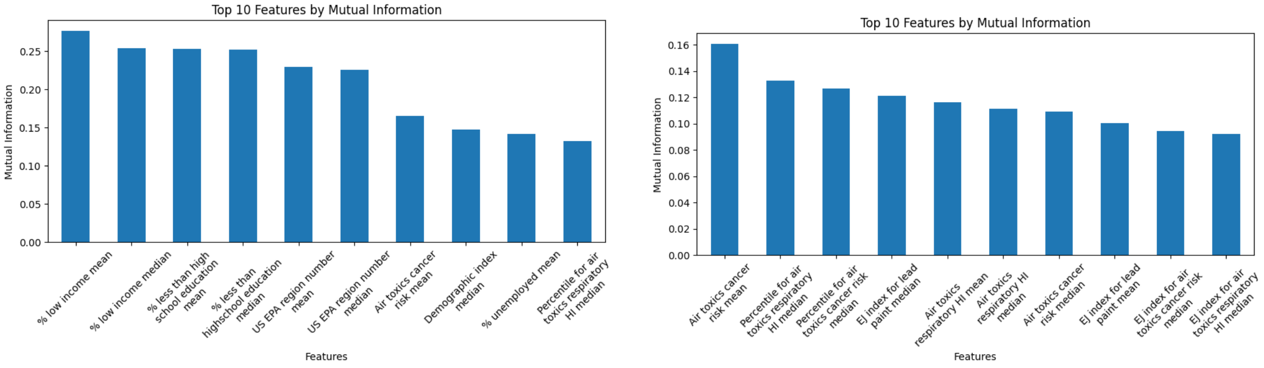
where $w^{(i)}$ is the weight of each sample point and τ is the bandwidth, controlling how much weight each sample has on predicting the life expectancy at x_q. A larger τ gives a smoother fit, whereas a smaller τ produces a more localized fit that can be vulnerable to overfitting. For our experiments, different τ's were tested using the wide range of values of 0.01, 0.05, 0.1, 0.5, 1, 5, and 10.

The optimal weights and parameters are found by minimizing the cost function, Eq. 6, through gradient descent.

$$J(\theta) = \sum_{i=1}^{m} w^{(i)}(y^{(i)} - \theta^T x^{(i)})^2 \tag{6}$$

2.4.3. K-Nearest Neighbor

While also used for classification problems, K-Nearest Neighbor (KNN) can be used for regression by designating k number of neighbors for each testing sample. The algorithm is as follows:



(a) Selected features for DEMO dataset

(b) Selected features for ENVI dataset

Figure 2. Comparison of features extracted with and without demographic features present in the EJD.

1. Choose a value of k.

2. For each new test sample, the distance between it and each training point is calculated.

3. The closest k data points are selected for this new training set.

4. The average of these k points are the the predicted life expectancy mean or median.

For our experiments, we ran KNN by selecting 1 to 25 neighbors and measuring what is the optimal number of neighbors, achieving the lowest MSE.

2.4.4. Support Vector Regression

While using many of these same principles as a Support Vector Machine (SVM), the purpose of Support Vector Regression (SVR) is to compute a hyperplane that best fits the data rather than separating the data for classification when performing SVM. This hyperplane is to maximize the margin between predicted values and actual values of the training data. The margin is referring to the two parallel lines that pass through the support vectors, which are the data points that are closest to the hyperplane.

For our experiments, we tested our given dataset with four different kernels (linear, polynomial, Radial Basis Function (RBF), and sigmoid) and compared the performance between them. The linear kernel is the simplest as it models a linear relationship, but it is not suitable for complex data. The polynomial kernel models captures nonlinear relationships. The RBF kernel captures more complex nonlinear problems. The sigmoid kernel is typically used for binary classification problems, as it proved to not perform well for our purposes.

3. Results

Our regression models were used to predict the target life expectancy median and life expectancy mean. The models tended to behave similarly when predicting the mean versus median life expectancy, thus, some charts in this section will only feature life expectancy mean as the target. Our analysis code is included to verify the similarity in model performance across the two targets in Appendix A. The results shown in this section are from our dataset with all features and with demographic features removed to truly see the impact that environmental conditions have on the life expectancy predictions. These will be called the DEMO dataset and ENVI dataset, respectively.

3.1. Linear Regression

The linear regression model trained on principal components and the features according to MI did not perform as well as other models discussed later. For the principal components, the MSE for both trained on the life expectancy mean and median are shown in Table 1. Our ENVI dataset does not perform as well as the DEMO dataset, so this model is not good for predicting the life expectancy based with only environmental features with solely principal components.

	DEMO	**ENVI**
life expectancy mean	1.95	2.94
life expectancy median	2.09	3.03

Table 1. Comparison of linear regression performance with both dataset versions using principal components.

For the top 50 features as ranked by MI, the results of the linear regression model with both dataset versions are shown in Figure 3. As with principal components, our model performed better with features from the DEMO dataset rather than just with features from the ENVI dataset. The best results when using MI-ranked features for linear regression are achieved when there is about 30 to 40 features. Overall, our linear regression model performed the best with principal components from the entire dataset, showing that this model with environmental features is not as capable at predicting life expectancy.

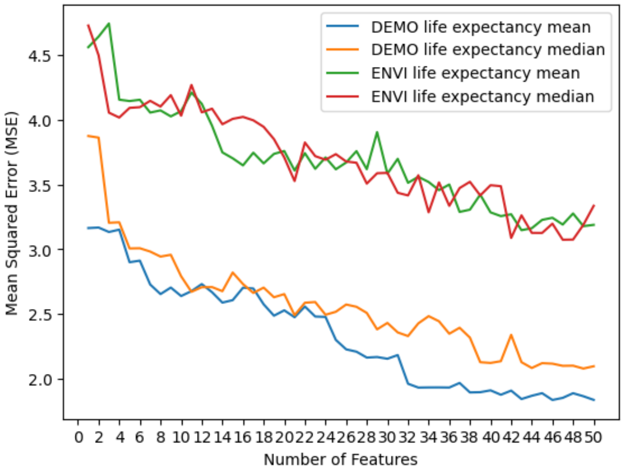


Figure 3. Comparison of linear regression performance for up to 50 features selected with both versions of dataset.

3.2. Locally-Weighted Linear Regression

For LWLR, the model performed slightly better than or similar to linear regression for both principal components and MI-ranked features for a varying number of features and τ's. In Figure 4, MSE values are only slightly lower for LWLR compared to linear regression with PCA for the DEMO dataset. Similar results occur for the ENVI dataset where most MSE measurements are below 3 for the life expectancy mean, providing to be a better slightly better model that linear regression for the ENVI dataset.

For MI-ranked features, LWLR was tested for 5, 10, 15, 20, and 25 features, as seen in Figure 5. The top features in the DEMO dataset in Figure 5(a) outperformed those of the ENVI dataset in Figure 5(b). The ENVI dataset performed best, across linear regression techniques, when using a LWLR model with 5 features and $\tau = 0.05$ resulting in a $MSE = 2.75$. Overall, LWLR is a good model in that it slightly outperforms linear regression when using principal components or MI-ranked features.

3.3. K-Nearest Neighbor

KNN for regression has a worse performance than linear and LWLR models for most values of k when using principal components for both datasets. In Figure 6, the MSE values are much higher than those of linear and LWLR, making it the worst model overall for both versions of the dataset when using principal components.

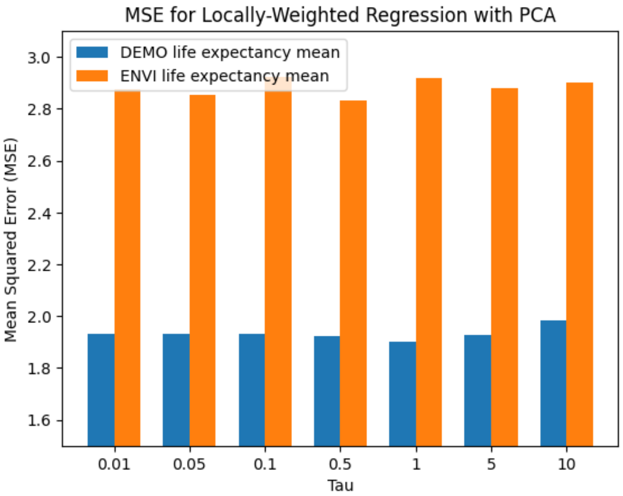


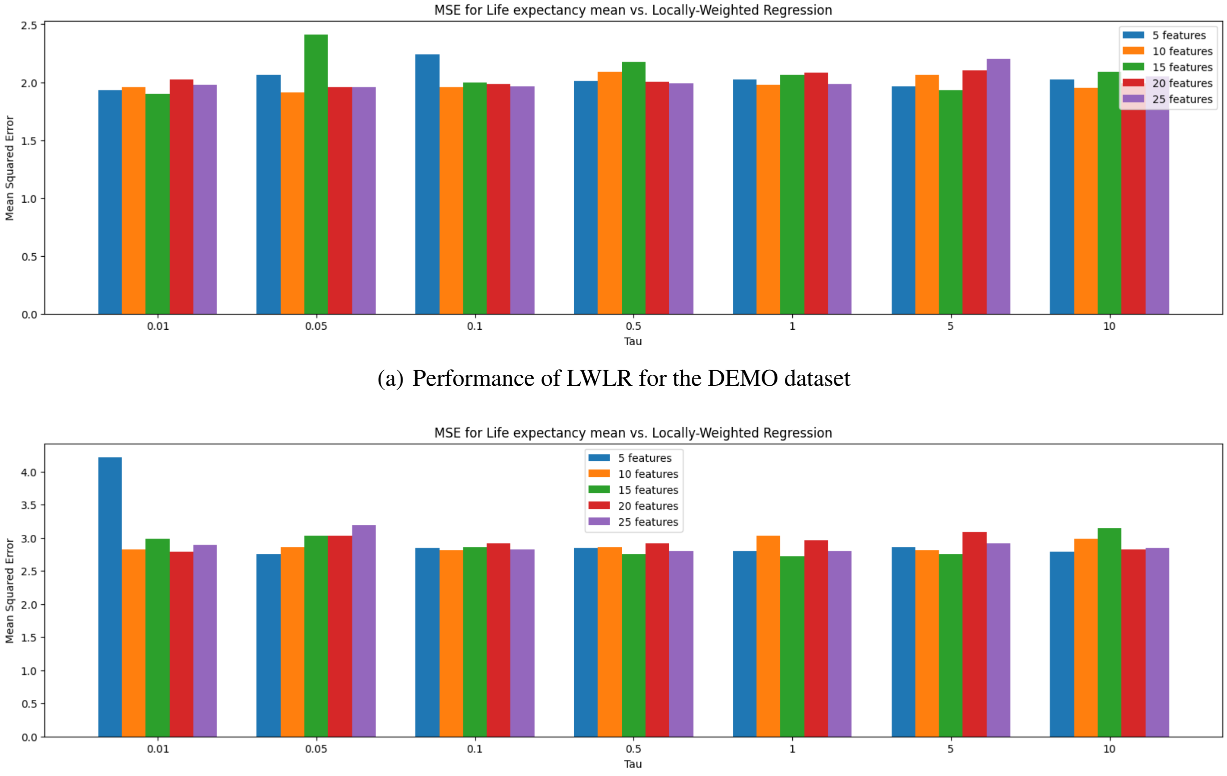
Figure 4. Comparison of locally-weighted linear regression performance for PCA and selected τ with both versions of dataset.

The MSE of the KNN model was significantly lower for MI-ranked features for both the DEMO and ENVI datasets. We searched over the number of neighbors, k, from 1 to 25 for both versions of datasets to determine the best model parameters. Determining the optimal number of MI-ranked features, 5 and 10 features were outperformed by 15, 20, and 25 features in Figure 7. Since 15, 20, and 25 features had extremely similar performance measures for the life expectancy mean, the best number of features to choose would be 15 because it gives approximately the same performance as 20 and 25 features but is less computationally expensive and a simpler model. In terms of choosing the number of k neighbors for this model, all feature numbers hit a minimum MSE value at around 5 neighbors after a steep drop and then either stay at a steady or slightly increasing MSE. Thus, the most optimal model performance model so far is with 5 neighbors and 15 features.

The performance of the MI-ranked features for the DEMO dataset in Figure 7(a) has slightly lower MSE values than those of the ENVI dataset in Figure 7(b). Thus, KNN proves to be more successful than linear and LWLR for both datasets due to its lower MSE values, with the DEMO dataset slightly outperforming the ENVI dataset.

3.4. Support Vector Regression

As mentioned previously, experiments were done on the Support Vector Regression (SVR) model with four different kernels: linear, polynomial, RBF, and sigmoid. In Figure 8, these four different kernel functions are compared when implemented in the SVR model using the principal components of both datasets. Overall, SVR did not perform well when using principal components as compared to other



(a) Performance of LWLR for the DEMO dataset

(b) Peformance of LWLR for the ENVI dataset.

Figure 5. Comparison of locally-weighted linear regression for features selected by MI and selected τ with both versions of dataset.

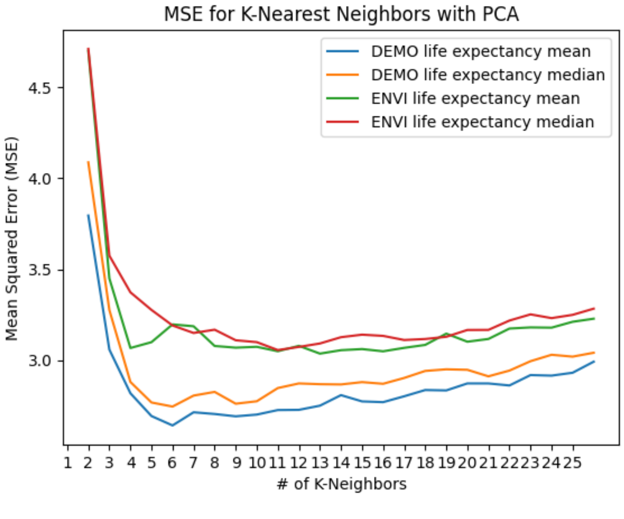


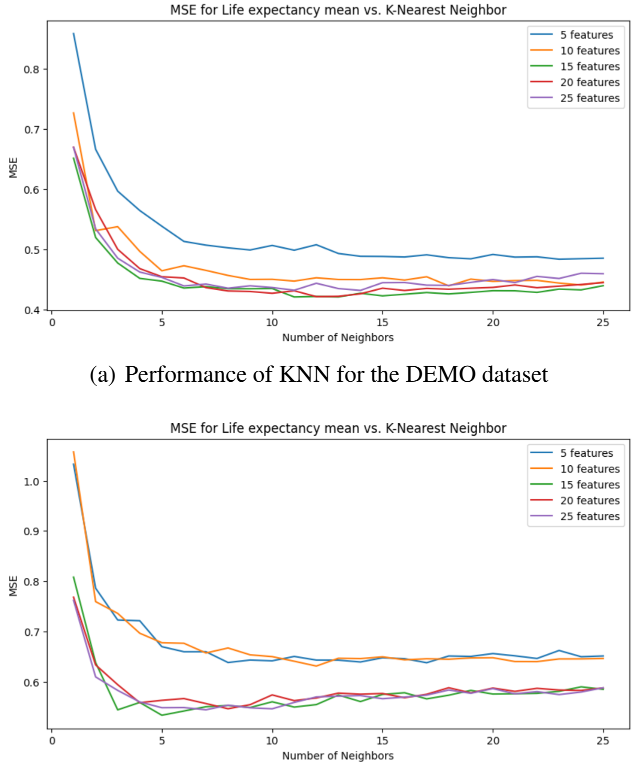
Figure 6. Comparison of KNN model performance for PCA of up to 25 neighbors with both versions of dataset.

models. Polynomial and sigmoid kernels proved to be the least effective such inputs. The linear and RBF performed the best overall with comparative MSE values for both. As with other models using principal components, in Figure 8 the DEMO dataset outperformed the ENVI dataset.

Despite not working for our dataset with PCA due to an unknown error, the best model performance for all regression models tested was with SVR on MI-ranked features. The sigmoid and polynomial kernel results are not plotted for both subplots in Figure 9 to better compare the behavior of the RBF and linear kernels. The model was tested against the polynomial and sigmoidal kernels, but the unplotted MSE values were all greater than 10, skewing the graph to make the better results un-viewable. The RBF kernel proved to be the most successful SVR for both versions of the dataset. It has the best performance with 25 features resulting in a MSE of 0.37 and 0.45 for the DEMO and ENVI dataset, respectively. However, as can see from the graph, 15, 20, and 25 features were all around the same MSE. Therefore, it may be less computationally expensive to get the most optimal performance with 15 features using the RBF kernel.

4. Discussions and Conclusions

From the regression methods tested, it became evident that the dataset with all features present compared to the dataset with only environmental features had higher performance in



(a) Performance of KNN for the DEMO dataset

(b) Peformance of KNN for the ENVI dataset.

Figure 7. Comparison of KNN model for selected features of up to 25 neighbors with both versions of dataset.

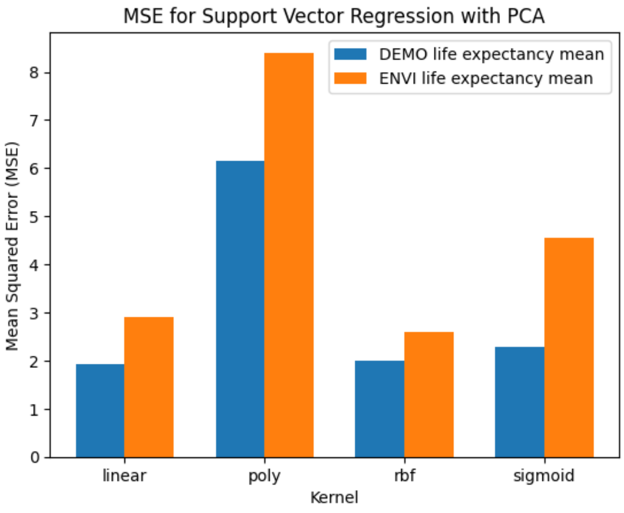
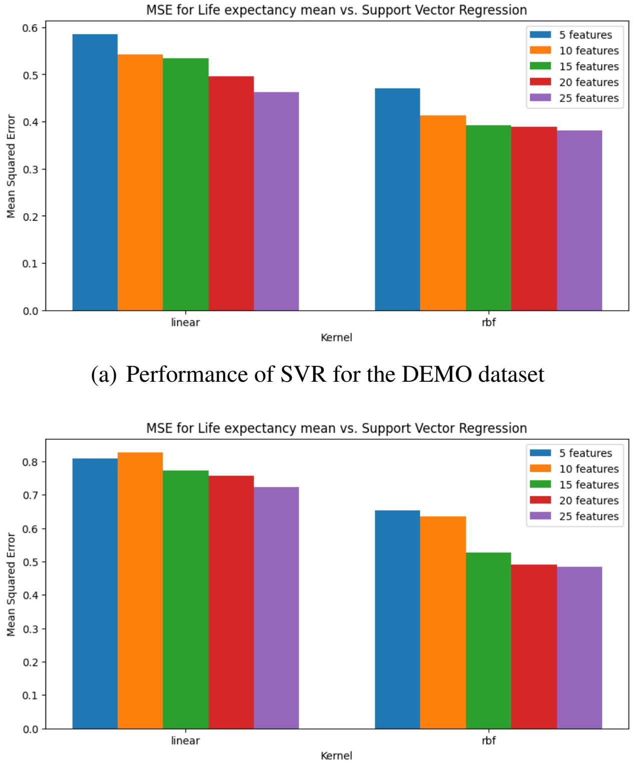


Figure 8. Comparison of SVR model performance for PCA for four different kernels with both versions of dataset.

all experiments. This concludes that environmental factors in the EJScreen dataset may not be suitable to predict the life expectancy on its own without the addition of some of the demographic features. This is consistent with the caveats



(a) Performance of SVR for the DEMO dataset

(b) Peformance of SVR for the ENVI dataset.

Figure 9. Comparison of SVR model for selected features for different kernels with both versions of dataset.

of the EJScreen dataset as well as its spatial and exposure uncertainties.

Furthermore, examining the MSE values of the datasets trained on the life expectancy mean and median, training with the mean seemed to create models that would outperform the median-based models.

Overall, to effectively predict the life expectancy using the environmental features, the SVR model with a RBF kernel give the lowest MSE value using around 15 to 25 features chosen by MI. PCA did not prove to be successful for the models with environmental features. K-nearest neighbor is also a suitable model to use for the top environmental MI-ranked features using around 5 neighbors and 15 to 25 features as well.

5. Acknowledgements

Teamwork was divided equally in this project. We focused on different tasks, while continually communicating with one another to make sure everything was on the right track. Levi primarily worked on pre-processing the data, conducting a literature review, and building the neural network. Mary Stirling focused on doing feature selection and test-

ing the various regression methods for PCA and a different number of features. We met whenever we felt needed to go over our work. We both contributed to this report equally.

A. Code

In our zip file, the following python notebooks are present:

- all_data_regression.ipynb - This notebook provides the PCA method, feature selection, results, and graphs for evaluation of different models based on the dataset with all features present in the EJScreen dataset.

- nondemographic_regression.ipynb - This notebook provides the same methodology from all_data_regression.ipynb, but it is based on the dataset with only the non-demographic features present in the EJScreen dataset.

- mutual_information_Graphs.ipynb - This notebook provides the code for producing the mutual information graphs in Figure 2 for the EJScreen dataset versions.

- combination_plots.ipynb - This notebook plots the DEMO and ENVI dataset together for certain regression models.

- EJScreen_PCA_finalsubmit.ipynb - This notebook contains PCA reduction of ENVI and DEMO datasets as well as the plotting for 1.

B. Resources

- Towards Data Science: Feature Selection in Python Using Filter Method

- Analytics Vidhya: KNN algorithm: Introduction to K-Nearest Neighbors Algorithm for Regression

- Towards Data Science: Locally-Weighted Lineaer Regression in Python

- Analytics Vidhya: Support Vector Regression Tutorial for Machine Learning

References

Arias, E., Escobedo, L. A., Kennedy, J., Fu, C., and Cisewski, J. A. Us small-area life expectancy estimates project: Methodology and results summary. 2018.

Berberian, A. G., Gonzalez, D. J., and Cushing, L. J. Racial disparities in climate change-related health effects in the united states. *Current environmental health reports*, 9(3): 451–464, 2022.

English, P., Richardson, M., Morello-Frosh, R., Pastor, M., Sadd, J., King, G., Jesdale, W., and Jerrett, M. Racial and income disparities in relation to a proposed climate change vulnerability screening method for california. *The International Journal of Climate Change: Impacts and Responses*, 4(2):1–18, 2013.

EPA. Ejscreen technical documentation. *United States*, 2022.

Heidger, L., Leconte, F., Quirino, R., and Petrissans, M. Environmental justice in climate change adaptation context: a case study in atlanta, georgia, using local climate zones and sociodemographic indicators. In *AGU Fall Meeting Abstracts*, volume 2021, pp. GC25G–0721, 2021.

Lary, D., Lary, T., and Sattler, B. Using machine learning to estimate global pm2. 5 for environmental health studies. *Environmental health insights*, 9:EHI–S15664, 2015.

Mullen, H., Whyte, K., and Holifield, R. Indigenous peoples and the justice40 screening tool: Lessons from ejscreen. *Environmental Justice*, 2023.

Patnaik, A., Son, J., Feng, A., and Ade, C. Racial disparities and climate change, 2020. URL https://psci.princeton.edu/tips/2020/8/15/racial-disparities-and-climate-change.

Saxena, A. and Dodell-Feder, D. Explaining the association between urbanicity and psychotic-like experiences in pre-adolescence: The indirect effect of urban exposures. *Frontiers in Psychiatry*, pp. 190, 2022.

Shkembi, A., Smith, L., and Neitzel, R. A ranking of environmental indicators among historically redlined neighborhoods in detroit, michigan, usa. In *ISEE Conference Abstracts*, volume 2022, 2022.

U.S. Census Bureau. Glossary, 2022. URL https://www.census.gov/programs-surveys/geography/about/glossary.html.

White, R. Life at the fenceline: Understanding cumulative health hazards in environmental justice communities. *Coming Clean, The Environmental Justice Health Alliance for Chemical Policy Reform and The Campaign for Healthier Solutions*, 2018.

Williams, E., Polsky, D., Archer, J.-M. J., Rodriguez, A., Han, R., Stewart, K., and Wilson, S. Md ejscreen v2. 0: Visualizing overburdening of environmental justice issues using the updated maryland environmental justice screening tool. *Environmental Justice*, 15(6):385–401, 2022.