A Machine Learning Approach to Predict the Total Age-Adjusted Mortality Rate Using Air Pollution Data and Socioeconomic Status in the United States

Oon Yu Yang[1]

School of Physical and Mathematical Sciences, NTU (Singapore).

Abstract

In this study, we apply a machine learning (ML) approach to predict the total age-adjusted mortality rate (AAMR) using the air pollution data and the population's socioeconomic status across sixty metropolitan areas of the United States. We use three different ML regression techniques, namely linear regression (LR), k-Nearest-Neighbours (kNN), and random forests (RF) models to achieve our aim. The result of this study suggests that the RF model has the best predicting performance (RMSE=24.59). We report that the most significant factor which affects the AAMR is the percentage of the non-white population in the studied areas. We also claim that Support Vector Machine (SVM) with sigmoid kernel achieves the highest accuracy compared with other kernels and the multiclass logistic regression when classifying mortality risk factors using one-vs-one classification. These knowledge and approach are useful for air quality management authority, city development board, and for other demographic, sociologic, and humanitarian study purposes.

Keywords: Machine learning, mortality, air pollution, particulate matter, black community.

1. Introduction

Air pollution is a major environmental and ecological challenge, resulting in potential loss of human lives, disruption of biological diversity, and creating huge economical losses. Since the beginning of the industrial revolution, air pollution has been continually threatening human populations. In recent years, it is well-known that the poorest communities embrace the largest impact as they do not have enough access to medical resources and financial aid[1, 2]. In this study, we determine which factors affect the age-adjusted mortality rate (AAMR) the most. Either they are environmental (e.g., particulate matter combined with temperature and rain precipitation), socioeconomical (e.g., employment rate in white-collar occupations, median school years completed, salary), or both. By comparing and selecting the best model out of three different ML techniques, namely k-Nearest Neighbours (kNN), linear regression (LR), and random forests (RF), we want to predict an urbanised area's AAMR using regression method since mortality rate is a continuous, numerical variable. While for the classification of risk (low, medium, or high), we compare multiclass logistic regression and SVM.

2. Data Source

The dataset used is from McDonald and Schwing (1973), which is now available online on Statlib (http://lib.stat.cmu.edu/datasets/)[3]. This dataset (60 × 17) consists of 16 independent variables are used to train our ML models, and the outcome variable, `age`, is the total AAMR on 60 U.S. metropolitan areas in 1959-1961.

`age`	Total Age-Adjusted Mortality Rate (AAMR).
`prep`	Mean annual precipitation in inches.
`jan.temp`	Mean January temperature in degrees Fahrenheit.
`jul.temp`	Mean July temperature in degrees Fahrenheit.
`older.65`	Percent of 1960 SMSA population that is 65 years of age or over.
`ppl.household`	Population per household, 1960 SMSA.
`school.year`	Median school years completed for those over 25 in 1960 SMSA.
`housing.unit`	Percent of housing units that are found with facilities.
`ppl.sqmile`	Population per square mile in urbanized area in 1960.
`ppl.nonwhite`	Percent of 1960 urbanized area population that is non-white
`white.collar`	Percent employment in white-collar occupations in 1960.
`income`	Percent of families with income under 3,000 in 1960 urbanized area.
`hc`	Relative pollution potential of hydrocarbons (HC).
`nox`	Relative pollution potential of oxides of nitrogen (NO_x).
`so2`	Relative pollution potential of sulfur dioxide (SO_2).
`rel.humidity`	Percent relative humidity, annual average at 1 p.m.
`risk`	Considered "high" if the mortality rate is higher than the 3rd quartile (Q3) of the recorded rates; considered "low" if it is lower than the 1st quartile (Q1); otherwise, it is considered medium.

Table 1. *Variables and their attributes in the dataset. Except for* `risk`*, which is a categorical variable; the rest are numerical (continuous) variables. Refer to Figure 8A and Figure 8B (Appendix) for the descriptive summary on the dataset.*

[1] Email address: P190001@e.ntu.edu.sg. Matriculation number: U1940100K.

The age-adjusted rate (AAR) is a method to make fair comparisons between groups with different age distributions[4]. For instance, a state with more elderly people tends to have higher rate of death or hospitalisation than a state with younger population, merely because the elderly is more likely to die or to be hospitalised. Age adjustment can make the different groups more comparable. A "standard" population is used to adjust the mortality rates in this context, which are the rates that would have existed if the population under study had the same age distribution as the "standard" population[5]. The National Center for Health Statistics of the United States recommends that the U.S. 1940, 1970 or 2000 standard population should be used when calculating age-adjusted rates. Thus, the U.S. 1940 standard population was used in this dataset.

The mortality rate of each U.S. metropolitan area was computed for each age group by dividing the number of mortalities in that age group by the estimated population of the same age group in that area and then multiplying by a constant of 100,000. This results in an age-specific mortality rate (ASMR) per 100,000 population for each age group. Each ASMR is then multiplied by the proportion of the standard population that same age group. The age-specific results are summed to get the AAMR for each state. The formula is:

$$\text{AAMR} = \sum (\text{ASMR} \times \text{standard population})$$

Using the dataset, we hypothesize that urban areas with more non-white, less highly educated population, and worse environmental pollutions tend to have higher AAMR because the community had poor health condition in general and did not have enough access to better medical facilities or higher living standards[6, 7].

3. Preliminary Analysis

First, we analyse the distribution of the total age-adjusted mortality rates across all 60 areas (Figure 1A). It is found that the histogram peaks at 940-960. Second, we analyse the correlations between all variables in the dataset, excluding the categorial variable `risk` (Figure 1B). In the correlation plot, it is found that variables `school.year`, `housing.unit`, and `white.collar` have the most negative correlations with the mortality rate, which are all socioeconomical factors. This implies that population which had better education, higher-paid jobs, and better living conditions had lower mortality rate, which agrees with our hypothesis. On the other hand, `ppl.nonwhite`, `prep`, and `so2` have the most positive correlations with the mortality rate. It implies that areas with higher non-white population had higher mortality rate in general. While for the environmental factors `prep` and `so2`, it implies that urban areas with higher acid rain precipitation tend to have population with higher mortality rate, which again, agrees with our hypothesis.

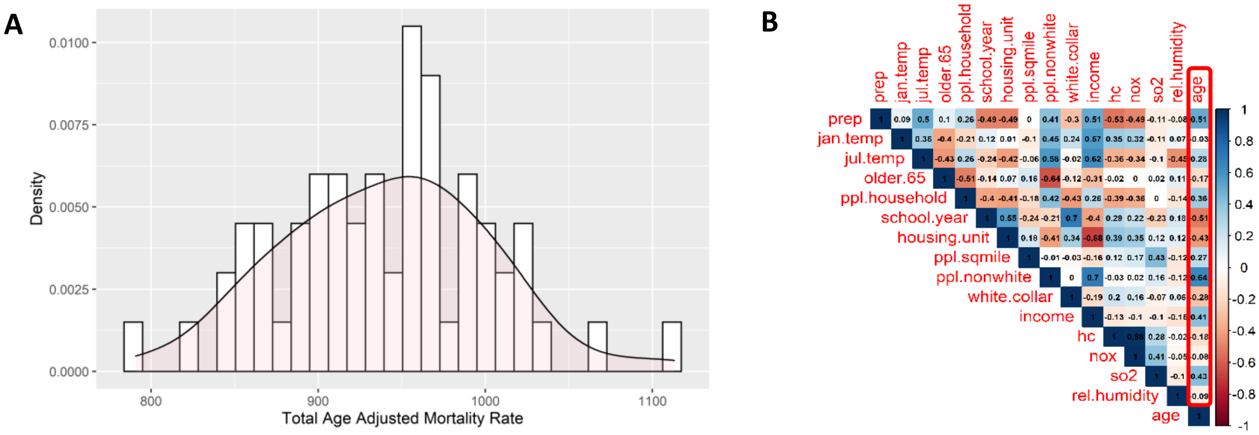


Figure 1. (A) *The histogram for the total-age adjusted mortality rate. In general, it appears symmetric and meets normality assumptions.* **(B)** *The correlation plot showing the correlations between all numerical (continuous) variables in the dataset.*

4. Predictions using ML Models and their Interpretations

In the dataset, there are no NA values. We transform the `risk` variable from low, medium, and high to the factors 1, 2, and 3, respectively. Then, we split the dataset (60 entries) into the training set (48 entries) and the testing set (12 entries) according to the 80:20 rule. To predict the AAMR, which is the variable `age`, we also include other environmental factors (e.g., rain precipitation and temperature, etc.) because study reveals the interactions between particulate matter in the air and temperature and rain have huge impact on human's mortality[8].

4.1 k-Nearest-Neighbour (kNN) Model

We perform the prediction by training the kNN model twice. First, we use all numerical variables in linear order except for the variable `age`, which is our outcome variable. Second, all numerical variables except for the variable `age`, but now we raise the variables which are highly correlated ($|r| \geq 0.4$) with the total AAMR, e.g., `school.year`, `housing.unit`, `income`, `ppl.nonwhite`, `prep`, and `so2` to the second order to enhance the performance of the model. To find the best `k` parameter, we design a for-loop to tune this parameter. The best parameter found is `k=5` (Figure 4, Appendix) since it reports the lowest root-mean-squared error (RMSE). Moreover, we set the following training parameters to improve the kNN performance:

`trControl`	`"cv"`, with `number = 3`
`preProcess`	`"center"` and `"scale"`
`tuneLength`	`10`

4.2 Linear Regression (LR) Model

Next, we train a LR model using the method identical to the training of the kNN model. After first training, we remove outliers and retrain the model, the new adjusted R-squared value is approximately 0.54, which means that the LR model can explain 54% of variations in the AAMR. From the summary details of the retrained LR model (Figure 5A, Appendix), it shows that the non-white population variable has the most significant impact on the AAMR. We present the residual plot of the better LR model (Figure 4B, Appendix) to check the linearity. The pattern shows a roughly symmetrical distributed shape, which means that the LR model is ideal.

4.3 Random Forests (RF) Model

Lastly, we train a RF model using the method identical to the training of the kNN model. The summary details of the RF model shows that the best tuned RF model has 500 trees and 8 variables tried at each split, which can explain approximately 55% of variations in the AAMR (Figure 5B, Appendix).

4.4 Prediction Performance and Results

After having trained all the model, we evaluate the performance of all three models using the testing set. The RMSE of each model, using different training methods, are recorded in Table 2 and their best prediction performance are shown in Figure 2. The best result (lowest RMSE) is 24.5921 from the RF model. In addition, the performance of all models increases except for the kNN model. It is because kNN is very sensitive to the scale on which predictor variable measurements are made. Nevertheless, the performance of kNN is better than the LR model because the relationship between the outcome and predictors is nonlinear. Thus, we conclude that the variables: `school.year`, `housing.unit`, `income`, `ppl.nonwhite`, `prep`, and `so2` are really the important factors to determine the AAMR while the most significant variable among them is the `ppl.nonwhite`, which is the percentage of non-white population in the urbanised area.

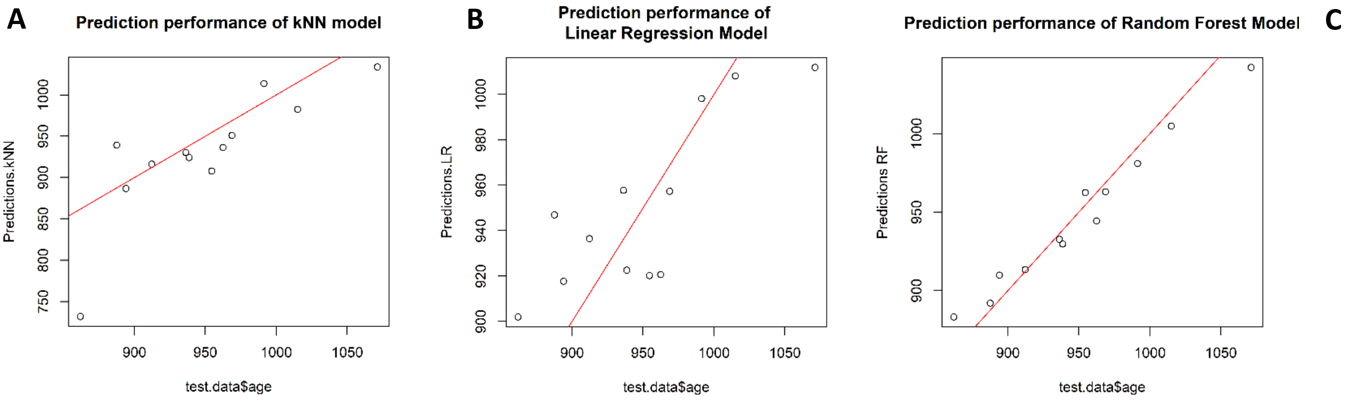


Figure 2. *The best prediction performance of the* **(A)** *kNN,* **(B)** *linear regression (LR), and* **(C)** *random forest (RF) models.*

ML (Regression) Models	RMSE (All variables in first order)	RMSE (Significant variables in second order; otherwise, in first order)
k-Nearest-Neighbour (kNN)	32.3143	33.7568
Linear Regression (LR)	47.5076	46.4968
Random Forest (RF)	25.8105	24.5921

Table 2. *Comparison of RMSE of different regression models using different choices of predictor variables.*

5. Risk Classification

In this section, we aim to classify the (categorical) variable, `risk` factor based on all other variables in the studied areas. Since there are more than two classes in the `risk` factor: low, medium, and high, we consider two multiclass classification methods: multiclass logistic regression and Support Vector Machines (SVM), based on one-versus-one (OvO for short) technique.

OvO is an approach for using binary classification algorithms for multiclass classification. In this method, a multiclass dataset is split into binary classification tasks. Unlike one-versus-rest (OvR) that splits it into one binary dataset for each class, the OvO approach splits the dataset into one dataset for each class versus every other class. In our dataset, we have the following binary classification datasets: low vs medium; low vs high; medium vs high. Each binary classification model predicts one class label and the model with the most votes is predicted by the OvO strategy. Classically, this approach is used for kernel-based algorithms such as SVM[9]. However, we wish to compare its performance with multiclass logistic regression for rigorous studies.

5.1 Multiclass logistic regression

We use the `multinom` function from the R package ***nnet***. Using the same training set from the regression tasks, except for now, our target is the categorial variable `risk`. We train the multiclass logistic regression model twice. First, we use all the predictor variables in linear order. While in the second run, we use all the predictor variables too but now the significant variables found in the last section are raised to second order to examine the model's performance. The results of the classification of testing set are summarised in Table 3.

Risk factor		*Predicted Labels*					
		All predictor variables in linear order			Significant predictor variables in second order; otherwise, linear		
		High	Medium	Low	High	Medium	Low
True Labels	High	3	1	0	3	0	2
	Medium	0	0	3	0	1	0
	Low	0	2	3	0	2	4
Accuracy		50.00%			66.67%		

Table 3. *Comparison of the accuracies of different training methods using the multiclass logistic regression.*

The multiclass logistic regression has the following logistic model equations:

$$\ln\left(\frac{\Pr(\text{risk} = \text{high})}{\Pr(\text{risk} = \text{low})}\right) = b_{1,0} + b_{1,1}X_{\text{prep}} + b_{1,2}X_{\text{jan.temp}} + \cdots + b_{1,15}X_{\text{so2}}$$

$$\ln\left(\frac{\Pr(\text{risk} = \text{high})}{\Pr(\text{risk} = \text{middle})}\right) = b_{2,0} + b_{2,1}X_{\text{prep}} + b_{2,2}X_{\text{jan.temp}} + \cdots + b_{2,15}X_{\text{so2}}$$

From the first trained logistic model summary, we can exponentiate the model coefficients to get the following odds table:

	prep	school.year	housing.unit	ppl.nonwhite	income	so2
low	0.233	0.646	0.481	0.144	0.722	0.702
middle	0.855	1.864	0.947	1.920	1.148	0.982

Table 4. *The odds table of the significant predictor variables from the first multiclass logistic classification model.*

From the table above, we know that, for example, for every one-unit increase in the variable `ppl.nonwhite`, the odds of being high risk (versus middle risk) increase by 1.920 times. Similarly, with one unit increase in the variable `so2`, the odds of being high risk (versus low risk) increase by 0.702 times.

Similarly, we have the following odds table from the second model, where the significant predictors are raised to the second order:

	I(prep^2)	I(school.year^2)	I(housing.unit^2)	I(ppl.nonwhite^2)	I(income^2)	I(so2^2)
low	0.763	0.683	1.056	0.719	1.076	1.000
middle	0.802	0.778	0.970	0.934	0.943	0.997

Table 5. *The odds table of the significant predictor variables from the second multiclass logistic classification model.*

5.2 Support Vector Machine (SVM)

Next, we use the SVM methods from the R package ***e1071*** with three different kernel types: linear, radial, and sigmoid. Similarly, we use all the predictor variables in linear order in the first trail while the significant variables are raised to the second order in the second trail. The results of the classification of testing set are shown below:

Risk factor		*Predicted Labels (All predictor variables in linear order)*								
		Linear Kernel			Radial Kernel			Sigmoid Kernel		
		High	Medium	Low	High	Medium	Low	High	Medium	Low
True Labels	High	3	1	0	1	0	0	1	0	0
	Medium	0	0	1	0	0	0	0	0	0
	Low	0	2	5	2	3	6	2	3	6
Accuracy		66.67%			58.33%			58.33%		

Table 6. *Comparison of the accuracies of different kernels of the multiclass SVM based on OvO. In this first trail, all the predictor variables are in linear order.*

Risk factor		*Predicted Labels (Significant predictor variables in second order; otherwise, linear)*								
		Linear Kernel			Radial Kernel			Sigmoid Kernel		
		High	Medium	Low	High	Medium	Low	High	Medium	Low
True Labels	High	3	1	1	2	0	0	2	0	0
	Medium	0	0	1	0	0	0	0	1	0
	Low	0	2	4	1	3	6	1	2	6
Accuracy		58.33%			66.67%			75.00%		

Table 7. *Comparison of the accuracies of different kernels of the multiclass SVM based on OvO. Now, we raise the significant predictor variables to the second order, while the others are kept at the first order.*

From the results shown, the highest accuracy achieved is 75.00% from the SVM (sigmoid kernel) with significant predictor variables raised to the second order. In general, when we do so, the accuracy of each model increases, except for the SVM linear kernel. This, again, reinforces our claim that `school.year`, `housing.unit`, `income`, `ppl.nonwhite`, `prep`, and `so2` are indeed significant predictor variables.

6. Conclusion

In this study, we conclude that the best ML regression model for predicting the AAMR is the random forest (RF) model, achieving RMSE=24.5921, compared with the kNN (RMSE=33.7568) and the linear regression (LR) (RMSE=46.4968) models. We verify our hypothesis that the significant predictor variables are `school.year`, `housing.unit`, `income`, `ppl.nonwhite`, `prep`, and `so2`. The first four factors are socioeconomical while the last two are environmental. Among these significant variables, `ppl.nonwhite`, which is the percentage of the non-white population has the most profound effect on the models while predicting the AAMR. It implies that in the studied 60 urbanised areas of the United States, mortality rate of the population still largely determined by their socioeconomic status, rather than the air pollution or other environmental factors, which agrees with other similar cases studied before[10-12]. Consequently, training the RF model using the significant predictor variables raised to the second power produces the best prediction. Finally, from the classification tasks using multiclass logistic regression and SVMs, we conclude that the SVM method with sigmoid kernel is the best method to predict the `risk` factor, achieving 75.00% in accuracy.

However, there are still some limitations in this study. First, the dataset used in this study is outdated. We use the 1959-1961 total age-adjusted mortality rate based on the U.S. 1940 standard population. Therefore, future research studies shall focus on the most recent data. Second, the data size is limited in scope. In this study, only 60 metropolitan areas from the U.S. are considered. Hence, more urban areas across the U.S. should be studied in the future. Third, there are rooms for improvement on the performance of the models since not all training parameters are considered and finetuned systematically in this study. Future trainings should apply regularisation technique to prevent overfitting of the model if applicable. Moreover, higher orders for the significant predictor variables should also be investigated while training both classification and regression models. Last but not least, future study shall focus on using the `tuneGrid` method from the ***Caret*** package to search the algorithm parameters by specifying a tune grid manually[13]. In the grid, each algorithm parameter can be specified as a vector of possible values. These vectors combine to define all possible combinations to try.

Data Availability Statement and Supplementary Material

All datasets, R code file, R library packages and their respective versions presented and applied in this study are available online at https://github.com/Isomorpfishm/NTU-PS0002-AY2021-MLProject.

Acknowledgement

The author would like to express his sincere gratitude and appreciation to the course instructor Assoc Prof Xiang Liming for her support and instructions while conducting this study and Mr Tiang Bi Hong from the School of Physical and Mathematical Sciences for his review, comments, and suggestions on this paper.

Appendix

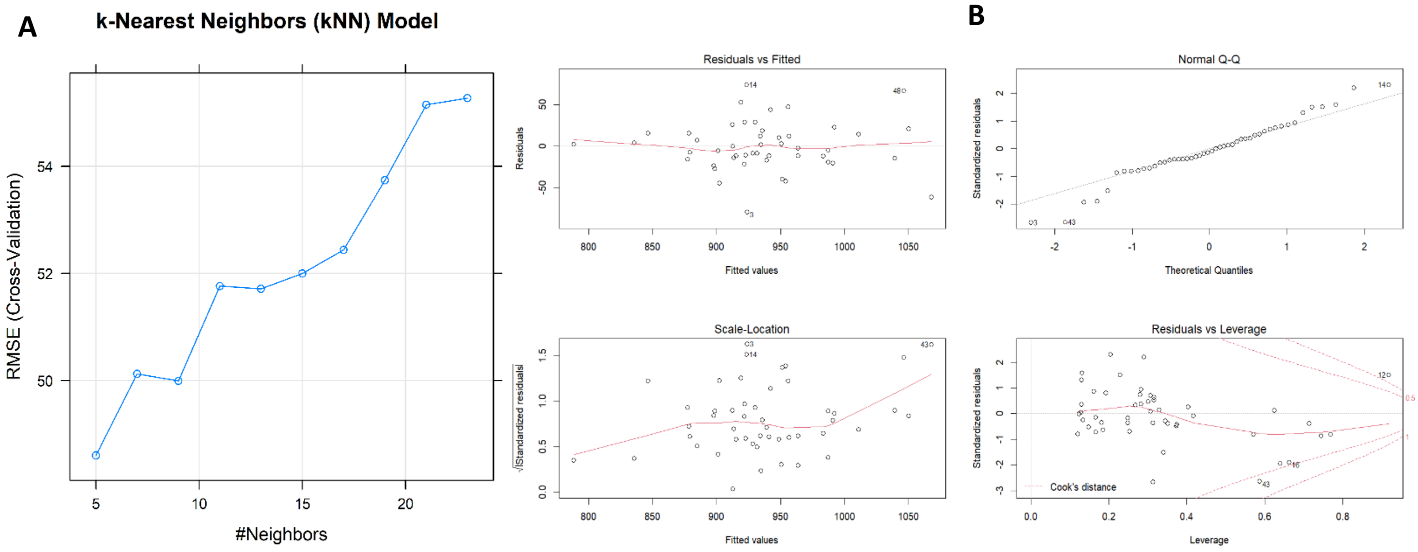
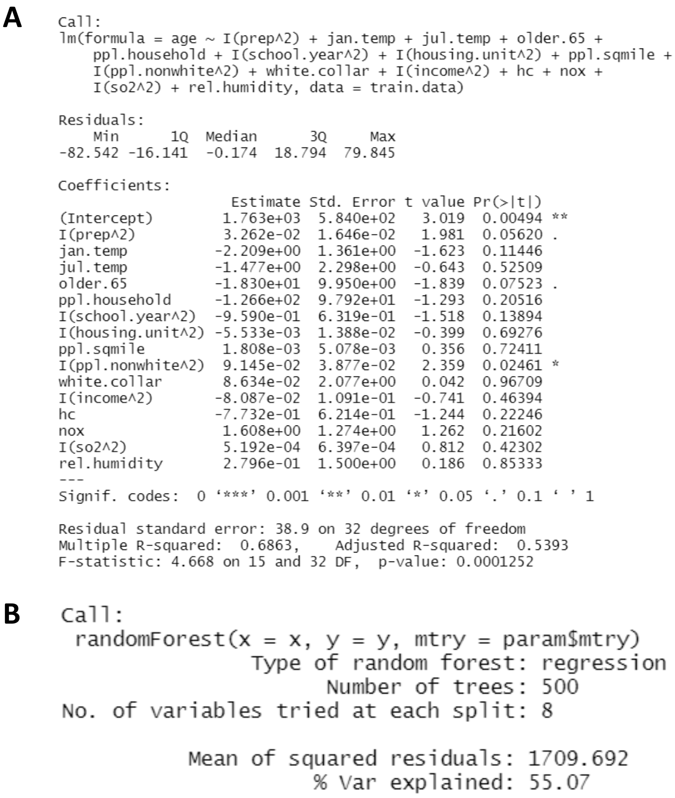


Figure 4. **(A)***The optimal value for the parameter k in the kNN model, and* **(B)** *the residual plot for the linear regression (LR) model.*



```
Call:
lm(formula = age ~ I(prep^2) + jan.temp + jul.temp + older.65 +
    ppl.household + I(school.year^2) + I(housing.unit^2) + ppl.sqmile +
    I(ppl.nonwhite^2) + white.collar + I(income^2) + hc + nox +
    I(so2^2) + rel.humidity, data = train.data)

Residuals:
    Min      1Q  Median      3Q     Max
-82.542 -16.141  -0.174  18.794  79.845

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.763e+03  5.840e+02   3.019  0.00494 **
I(prep^2)          3.262e-02  1.646e-02   1.981  0.05620 .
jan.temp          -2.209e+00  1.361e+00  -1.623  0.11446
jul.temp          -1.477e+00  2.298e+00  -0.643  0.52509
older.65          -1.830e+01  9.950e+00  -1.839  0.07523 .
ppl.household     -1.266e+02  9.792e+01  -1.293  0.20516
I(school.year^2)  -9.590e-01  6.319e-01  -1.518  0.13894
I(housing.unit^2) -5.533e-03  1.388e-02  -0.399  0.69276
ppl.sqmile         1.808e-03  5.078e-03   0.356  0.72411
I(ppl.nonwhite^2)  9.145e-02  3.877e-02   2.359  0.02461 *
white.collar       8.634e-02  2.077e+00   0.042  0.96709
I(income^2)       -8.087e-02  1.091e-01  -0.741  0.46394
hc                -7.732e-01  6.214e-01  -1.244  0.22246
nox                1.608e+00  1.274e+00   1.262  0.21602
I(so2^2)           5.192e-04  6.397e-04   0.812  0.42302
rel.humidity       2.796e-01  1.500e+00   0.186  0.85333
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38.9 on 32 degrees of freedom
Multiple R-squared:  0.6863,    Adjusted R-squared:  0.5393
F-statistic: 4.668 on 15 and 32 DF,  p-value: 0.0001252
```

```
Call:
 randomForest(x = x, y = y, mtry = param$mtry)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 8

          Mean of squared residuals: 1709.692
                    % Var explained: 55.07
```

Figure 5. *The summary details of the results for* **(A)** *the linear regression model. The result suggests that among all the significant factors that we consider, the percentage of the non-white population has the most profound influence on the prediction outcome.* **(B)** *The random forest (RF) model for the regression task. The optimal number of trees is 500 while in each split, 8 variables are tried.*

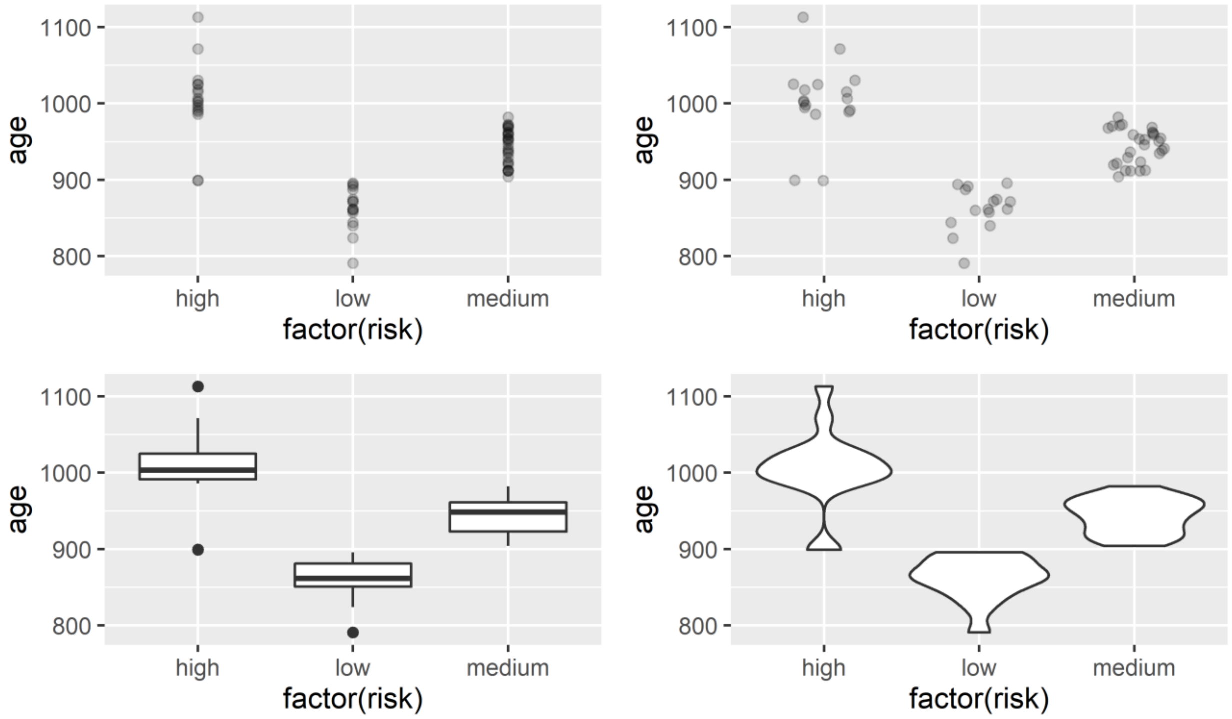


Figure 6. (A) *Strip plots,* **(B)** *jitter plots,* **(C)** *boxplots, and* **(D)** *violin plots of the numerical (continuous) variable* `age` *(total age-adjusted mortality rate, AAMR) against the categorical variable* `risk` *of mortality. A few outliers are observed.*

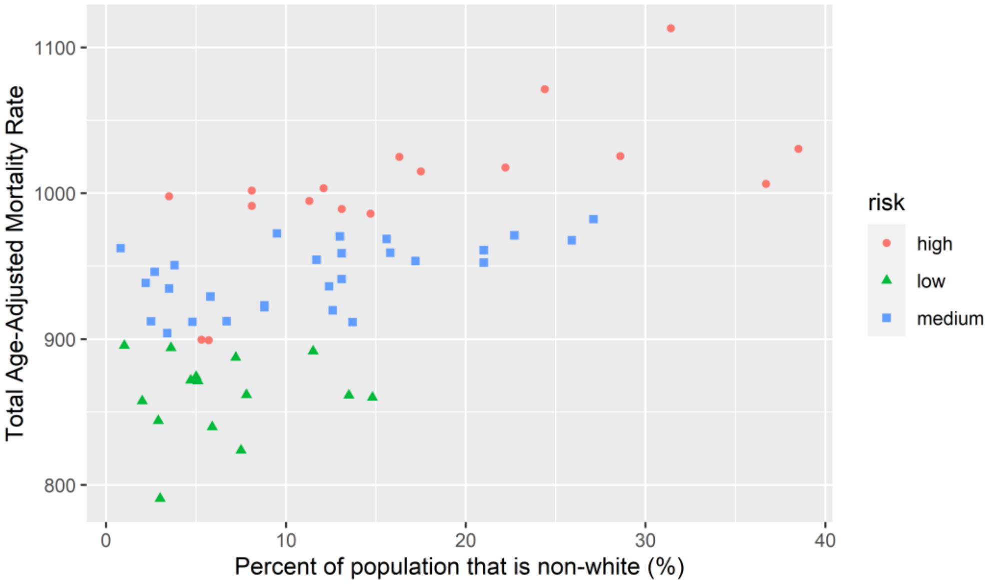
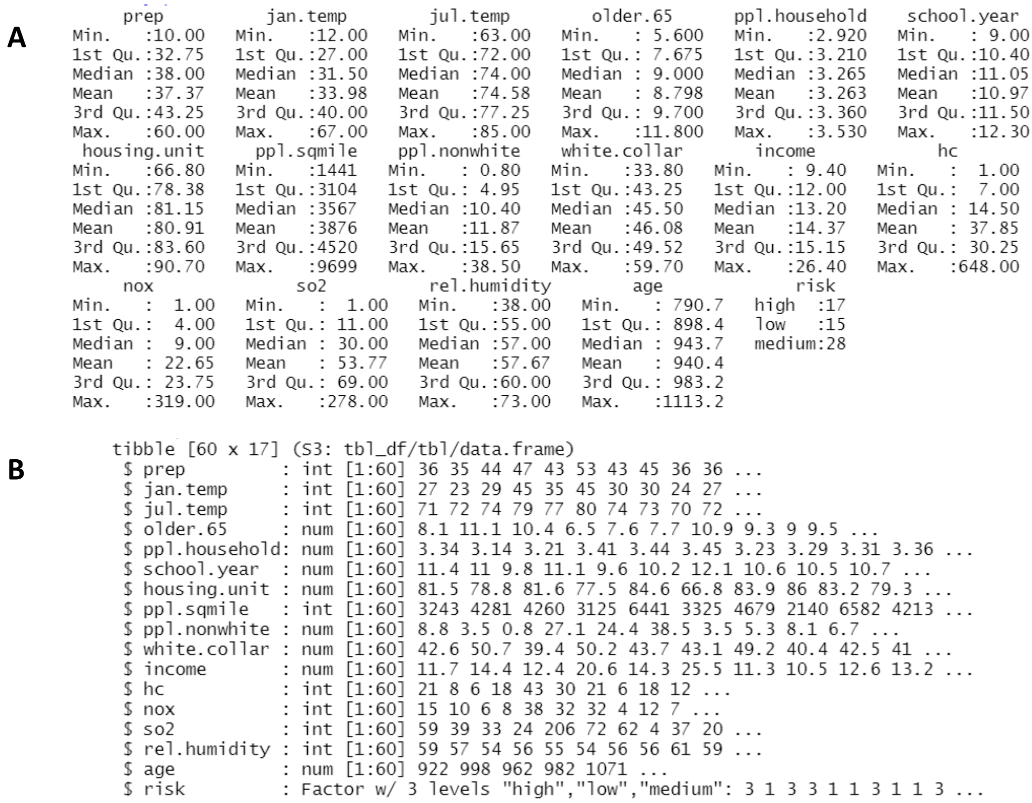


Figure 7. *The scatter plot showing the distribution of the total age-adjusted mortality rate (AAMR) against the most significant predictor variable found in the study,* `ppl.nonwhite`, *categorised using the* `risk` *factor.*



```
     prep          jan.temp        jul.temp        older.65      ppl.household    school.year   
 Min.   :10.00   Min.   :12.00   Min.   :63.00   Min.   : 5.600   Min.   :2.920   Min.   : 9.00  
 1st Qu.:32.75   1st Qu.:27.00   1st Qu.:72.00   1st Qu.: 7.675   1st Qu.:3.210   1st Qu.:10.40  
 Median :38.00   Median :31.50   Median :74.00   Median : 9.000   Median :3.265   Median :11.05  
 Mean   :37.37   Mean   :33.98   Mean   :74.58   Mean   : 8.798   Mean   :3.263   Mean   :10.97  
 3rd Qu.:43.25   3rd Qu.:40.00   3rd Qu.:77.25   3rd Qu.: 9.700   3rd Qu.:3.360   3rd Qu.:11.50  
 Max.   :60.00   Max.   :67.00   Max.   :85.00   Max.   :11.800   Max.   :3.530   Max.   :12.30  
  housing.unit     ppl.sqmile    ppl.nonwhite    white.collar       income            hc        
 Min.   :66.80   Min.   :1441   Min.   : 0.80   Min.   :33.80   Min.   : 9.40   Min.   :  1.00  
 1st Qu.:78.38   1st Qu.:3104   1st Qu.: 4.95   1st Qu.:43.25   1st Qu.:12.00   1st Qu.:  7.00  
 Median :81.15   Median :3567   Median :10.40   Median :45.50   Median :13.20   Median : 14.50  
 Mean   :80.91   Mean   :3876   Mean   :11.87   Mean   :46.08   Mean   :14.37   Mean   : 37.85  
 3rd Qu.:83.60   3rd Qu.:4520   3rd Qu.:15.65   3rd Qu.:49.52   3rd Qu.:15.15   3rd Qu.: 30.25  
 Max.   :90.70   Max.   :9699   Max.   :38.50   Max.   :59.70   Max.   :26.40   Max.   :648.00  
      nox              so2          rel.humidity        age             risk   
 Min.   :  1.00   Min.   :  1.00   Min.   :38.00   Min.   : 790.7   high  :17  
 1st Qu.:  4.00   1st Qu.: 11.00   1st Qu.:55.00   1st Qu.: 898.4   low   :15  
 Median :  9.00   Median : 30.00   Median :57.00   Median : 943.7   medium:28  
 Mean   : 22.65   Mean   : 53.77   Mean   :57.67   Mean   : 940.4              
 3rd Qu.: 23.75   3rd Qu.: 69.00   3rd Qu.:60.00   3rd Qu.: 983.2              
 Max.   :319.00   Max.   :278.00   Max.   :73.00   Max.   :1113.2              
```

```
tibble [60 x 17] (S3: tbl_df/tbl/data.frame)
 $ prep         : int [1:60] 36 35 44 47 43 53 43 45 36 36 ...
 $ jan.temp     : int [1:60] 27 23 29 45 35 45 30 30 24 27 ...
 $ jul.temp     : int [1:60] 71 72 74 79 77 80 74 73 70 72 ...
 $ older.65     : num [1:60] 8.1 11.1 10.4 6.5 7.6 7.7 10.9 9.3 9 9.5 ...
 $ ppl.household: num [1:60] 3.34 3.14 3.21 3.41 3.44 3.45 3.23 3.29 3.31 3.36 ...
 $ school.year  : num [1:60] 11.4 11 9.8 11.1 9.6 10.2 12.1 10.6 10.5 10.7 ...
 $ housing.unit : num [1:60] 81.5 78.8 81.6 77.5 84.6 66.8 83.9 86 83.2 79.3 ...
 $ ppl.sqmile   : int [1:60] 3243 4281 4260 3125 6441 3325 4679 2140 6582 4213 ...
 $ ppl.nonwhite : num [1:60] 8.8 3.5 0.8 27.1 24.4 38.5 3.5 5.3 8.1 6.7 ...
 $ white.collar : num [1:60] 42.6 50.7 39.4 50.2 43.7 43.1 49.2 40.4 42.5 41 ...
 $ income       : num [1:60] 11.7 14.4 12.4 20.6 14.3 25.5 11.3 10.5 12.6 13.2 ...
 $ hc           : int [1:60] 21 8 6 18 43 30 21 6 18 12 ...
 $ nox          : int [1:60] 15 10 6 8 38 32 32 4 12 7 ...
 $ so2          : int [1:60] 59 39 33 24 206 72 62 4 37 20 ...
 $ rel.humidity : int [1:60] 59 57 54 56 55 54 56 56 61 59 ...
 $ age          : num [1:60] 922 998 962 982 1071 ...
 $ risk         : Factor w/ 3 levels "high","low","medium": 3 1 3 3 1 1 3 1 1 3 ...
```

Figure 8. *Descriptions of the preliminary exploratory analysis.* ***(A)*** *The descriptive (summary) statistics of the dataset, and* ***(B)*** *the structure of the dataset.*

References

[1] A. P. Polednak, "Black-white differences in infant mortality in 38 standard metropolitan statistical areas," (in en), *Am J Public Health,* vol. 81, no. 11, pp. 1480-1482, 1991/11// 1991, doi: 10.2105/AJPH.81.11.1480.

[2] M. A. Winkleby and C. Cubbin, "Influence of individual and neighbourhood socioeconomic status on mortality among black, Mexican-American, and white women and men in the United States," (in en), *Journal of Epidemiology & Community Health,* vol. 57, no. 6, pp. 444-452, 2003/06/01/ 2003, doi: 10.1136/jech.57.6.444.

[3] G. C. McDonald and R. C. Schwing, "Instabilities of Regression Estimates Relating Air Pollution to Mortality," *Technometrics,* vol. 15, no. 3, pp. 463-481, 1973/08/01 1973, doi: 10.1080/00401706.1973.10489073.

[4] D. F. Austin and S. B. Werner, *Epidemiology for the health sciences: a primer on epidemiologic concepts and their uses*. Springfield: Thomas (in eng), 1974, p. 75.

[5] N. N. Naing, "Easy way to learn standardization : direct and indirect methods," (in eng), *Malays J Med Sci,* vol. 7, no. 1, pp. 10-15, 2000. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/22844209.

[6] R. T. Anderson, P. Sorlie, E. Backlund, N. Johnson, and G. A. Kaplan, "Mortality Effects of Community Socioeconomic Status," *Epidemiology,* vol. 8, no. 1, pp. 42-47, 1997 1997. [Online]. Available: https://www.jstor.org/stable/3702421.

[7] R. A. Hummer, "Black-White Differences in Health and Mortality: A Review And Conceptual Model," (in en), *The Sociological Quarterly,* vol. 37, no. 1, pp. 105-125, 1996/01// 1996, doi: 10.1111/j.1533-8525.1996.tb02333.x.

[8] S. Roberts, "Interactions between particulate air pollution and temperature in air pollution mortality time series studies," (in en), *Environmental Research,* vol. 96, no. 3, pp. 328-337, 2004/11// 2004, doi: 10.1016/j.envres.2004.01.015.

[9] J. Brownlee, "One-vs-Rest and One-vs-One for Multi-Class Classification," in *Machine Learning Mastery*, ed, 2020.

[10] V. Griskevicius, A. W. Delton, T. E. Robertson, and J. M. Tybur, "Environmental contingency in life history strategies: the influence of mortality and socioeconomic status on reproductive timing," *Journal of personality and social psychology,* vol. 100, no. 2, p. 241, 2011.

[11] O. Laurent, D. Bard, L. Filleul, and C. Segala, "Effect of socioeconomic status on the relationship between atmospheric pollution and mortality," (in en), *Journal of Epidemiology & Community Health,* vol. 61, no. 8, pp. 665-675, 2007/08/01/ 2007, doi: 10.1136/jech.2006.053611.

[12] G. D. Smith, M. J. Shipley, and G. Rose, "Magnitude and causes of socioeconomic differentials in mortality: further evidence from the Whitehall Study," (in en), *Journal of Epidemiology & Community Health,* vol. 44, no. 4, pp. 265-270, 1990/12/01/ 1990, doi: 10.1136/jech.44.4.265.

[13] *caret: Classification and Regression Training*. (2020). Accessed: 2021/04/15/05:59:38. [Online]. Available: https://CRAN.R-project.org/package=caret