Names: Tung Kieu (2044173), Juan Manzano Vasquez (1774201), Brandon Miramontes (1987701), Andrew J. Ballance (2064104), Kevin S. Kappadakkamadathil (1765095)
Course: MATH-4323
Professor's name: Dr. Wenshuang Wang

Introduction

We examined the relationship between CO2 emissions and the use of technology (note that we've constrained our population to countries in Europe and Central Asia), and to say the least, World Bank's database of World Developments Indicators seemed like a comprehensive source for this purpose. As for why we want to examine that particular relationship, we note that climate change and its impact on the environment have become pressing global concerns; thus, by identifying the relationship between CO2 emissions and technology metrics, we can contribute to the understanding of how technology usage affects greenhouse gas emissions and, so, provide the basis for informed decisions to mitigate the technology impact on environment or vice versa. Additionally, by narrowing our focus to Europe and Central Asia, we will provide region-specific insights and recommendations. The dataset has 37 variables (not including the country, country code, series name, and series code), and given that our observations are countries over time (e.g. Albania in 2001, Albania in 2002, etc.), the number of observations is 1334.

We want to use this data to determine if CO2 emissions can be predicted based on technology use and energy consumption. By analyzing the relationship between these variables, statistical models can be developed to estimate CO2 emissions levels. Thus, our goal for this data analysis is to answer whether CO2 emission be predicted based on technology use and energy consumption.

Exploratory Data Analysis

We conducted a preliminary analysis of the variables that we decided would be technology metrics in their individual relationships to CO2 emission. We generally found that higher values on the technology predictors would yield a high CO2 emission, and so, we would be able to hypothesize that a fairly strong relationship exists between CO2 emissions and the technology metrics that we chose. We've included the plots for the relationship of each technological predictor variable to the chosen CO2 emission response variable in the appendix.

Methodology

To answer our question, we produced classifications models based on supervised learning (with KNN and SVM as specific implementations):

Data Preparation

We collected a dataset that includes the attributes mentioned (such as alternative and nuclear energy, CO2 emissions, electricity production from different sources, broadband subscriptions, etc.) and then ensured that each data point has information on both the predictor variables (technology use and energy consumption) and the target variable (CO2 emissions).

Data Preprocessing

We cleaned the dataset by handling missing values, outliers, and normalizing the data (since KNN and SVM are distance-based algorithms, it's essential to ensure that the features are on similar scales).

Splitting the Data

We divided the dataset into training and testing sets (the training set will be used to train the models, while the testing set will be used to evaluate their performance) and then created a new variable in which the possible values would be 1 or 0 to indicate high or low emission (the prediction was compared with the mean of the original CO2 emission data so that what is greater than the mean would be high and what is less than the mean would be low).

K-Nearest Neighbors (KNN) for Classification

In KNN, the training process involves storing the training samples and their corresponding class labels (CO2 emissions levels). KNN does not learn explicit patterns but, rather, memorizes the training instances.

For feature selections, we found some attributes related to technology and electricity-production as our predictor variables (as mentioned before, we analyzed the relationship of these attributes with CO2 emissions by using EDA).

In determining the value of the K parameter (which represents the number of nearest neighbors to consider when making predictions), we could have used techniques like cross-validation to find the optimal value of K.

Model Fitting and Evaluation

We then trained the KNN model using the training dataset and the selected value of K.

Next, we used the testing dataset to evaluate the performance of the trained KNN model. Once the preferred model is selected, it can be used to predict CO2 emissions as high or low based on the technology use.

Support Vector Machines (SVM) for Classification

In SVM, the training process involves finding an optimal hyperplane that separates data points into different classes. In SVM, the training process involves finding the optimal hyperplane that maximizes the margin between classes while minimizing classification errors through the use of different kernels. SVM was applied to classify CO2 emissions levels as high or low based on the provided attributes.

We chose the relevant features related to technology use and energy consumption that we found have a significant impact on CO2 emissions.

Model Fitting and Evaluation

We trained the SVM model using the training dataset and the selected parameters.

We used the testing dataset to evaluate the performance of the trained SVM model and then calculated evaluation metrics such as accuracy.

We note that the SVM learning method has one notable advantage over the KNN method: the SVM method produces a parametric equation to model the two classifications of the observations, whereas the KNN method produces more of a black box that, while yielding a similarly accurate prediction, is not interpretable as a model. Still, we expect that both methods will complement each other in our intended prediction of CO2 emissions.

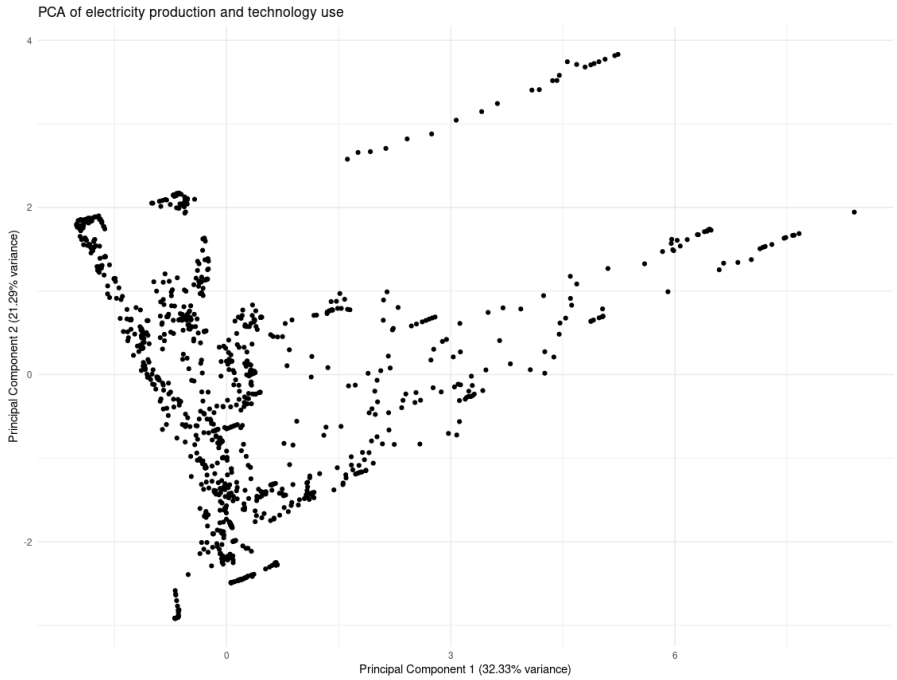
Data Analysis

In order to utilize the data for analysis, we first had to conduct extensive cleaning of the data: we had to remove unnecessary columns, rename columns, replace missing values with row means, deal with “NaN” values, and much else. We’ve included the original dataset as well as a file called Eurasia.csv (the dataset after cleaning) in the submitted folder.

Then, out of the 37 variables that we acquired, we narrowed our usage down to 10: we utilized a variable representative of CO2 emissions; as for predictors, we acquired the following variables:
a) Electricity production from coal sources (% of total)
b) Electricity production from natural gas sources (% of total)
c) Electricity production from oil sources (% of total)
d) Electricity production from oil, gas and coal sources (% of total)
e) Electricity production from hydroelectric sources (% of total)
f) Electricity production from nuclear sources (% of total)
g) Fixed broadband subscriptions
h) Fixed telephone subscriptions
i) Mobile cellular subscriptions

To help substantially reduce the number of dimensions that we would work with, we did a principle component analysis, and we used the principle components derived from this analysis in creating of our models. During the PCA, we found that scaling was necessary because we suspected that the spread and range of the variables that we possessed differed, and we did not want any variables to skew our analysis by any extremity of difference.

For the following picture, we can see the relationship between our CO2 emission and combination features of tech and electricity productions have an upward slope to the top of the graph, but mildly, which would may indicate a positive relationship between the two; means if our combinations features increases, so is CO2. But this is just an assumption based on “visual” aspect. In the next part we going to design and derive our predictions based on our SVM algorithm and draw the decision boundary SVM to determine where do these observations belong to.



For both models, we randomly divided the dataset after PCA into 80% training and 20% testing data.

SVM

For this SVM model, we applied our PCA into our SVM learning algorithm with 70:30 ratio testing and training set and used the tune function which performs a 10-fold cross validation to find the optimal parameters to run the SVM algorithm, our testing parameters have cost values 0.001,0.01,0.1,1,10,100 and gamma values from 1:8. After running the tune function, we found that our optimal cost =10 and gamma = 1 with SVM algorithm error of 1.12%, which indicates our SVM models performs the best with the parameter above. After running through SVM learning algorithm with the optimal parameter, this is our testing results for predictions, which have only 3 misclassifications, and most of the predictions are on the “0”, low emissions.

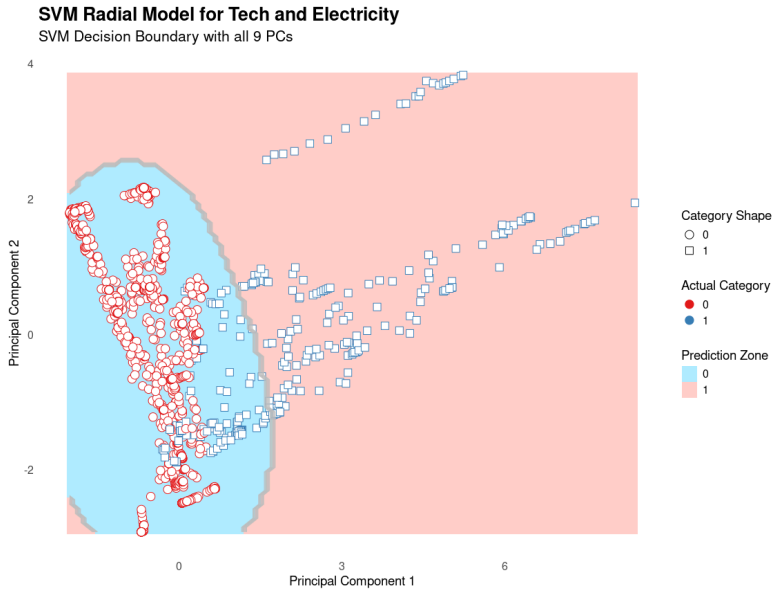
pred

true 0 1

0 1043 1

1 2 288

For the following picture looking at the SVM boundaries, we see that are seems like most of our PCs are in the low emission zone along our response variable just like our predictions in the "visual" aspect above and also the testing table predictions were correct



KNN

For our KNN model of life expectancy and CO2 we applied our PCA into our KNN learning algorithm with 80:20 ratio testing and training set. To find the best performance K value on our KNN algorithm we have performed cross-validation with k value from 1 through 25. After running through, we see that our optimal value for K = 10 with error = 6.05%. The reason we choose 10 is because if we choose our K = 1 that means the data is only comparing to itself and not with many neighbors. By choosing K = 10, we can eliminate the bias of the training error in our algorithm and the error on our KNN algorithm is more reliable and not under the risk of overfitting. After running through our KNN model, the prediction error is 7.08%.

pred

true 0 1

0 816 23

1 53 182

Based on this prediction table, we again also see that of the predictions are in the value of "0", low emissions.

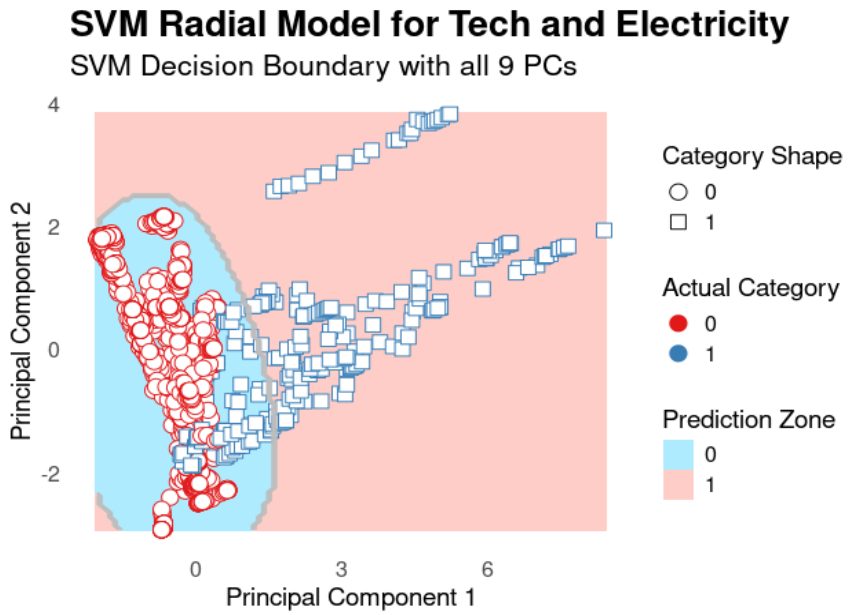
```
best_k = -1
err_k = 1000.0
for (i in 1:25){
  ele.knn = knn(train = data.frame(ele.train), test = data.frame(ele.test), cl= as.factor(co2.cl), k=i)
  m = mean(co2.cl2 != ele.knn)
  if(m < err_k){
    best_k = i
    err_k = m
  }
}
```

Comparison and Results

Based on the two results above, we concluded that the SVM clearly yields a lower test error rate than the KNN model, also it is much more reliable due to its mechanics on drawing high dimensional data which creates less bias than KNN, so we decided to fit the full data with SVM.

SVM on whole dataset

The following is both a plot of the newly fitted model (which is not noticeably different from the one produced from the model fitted on training data) as well a summary of it:




```
> summary(svmfit)

Call:
svm(formula = y ~ ., data = co2_techEnergy, cost = 10, gamma = 1, kernel = "radial", scale =


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  10

Number of Support Vectors:  292

 ( 124 168 )


Number of Classes:  2

Levels:
 0 1
```

For this whole dataset's SVM model, we applied our PCA into our SVM learning algorithm and used the tune function which performs a 10-fold cross validation to find the optimal parameters to run the SVM algorithm, our testing parameters have cost values 0.001,0.01,0.1,1,10,100 and gamma values from 1:8, same process with the SVM testing above. After running the tune function, we found that our optimal cost =10 and gamma = 1 with SVM algorithm error of 1.05%, which indicates our SVM models performs the best with the parameter above. After running through SVM learning algorithm with the optimal parameter, this is our testing results for predictions, which have only 4 misclassifications, which only increase 1 misclassification since we use the whole datset and most of the predictions are also on the "0", low emissions.

	pred	
true	0	1
0	1043	1
1	3	287

To say the least, we feel confident with the model we produced in predicting CO2 emissions correctly indicate that with the combination features of Electricity Productions of different fuel types and technology subscriptions, we can see that it will predict the observation on a low CO2 emissions

Conclusion

Based on our analysis, our developed model serves as a promising prototype for predicting future trends in CO2 emissions. This predictive capability enables the formulation of effective policies to mitigate and reduce CO2 emissions. By considering a combination of features such as Electricity Productions from different fuel types and technology subscriptions, our model accurately predicts CO2 emissions, especially in scenarios where emissions are expected to be low. This confidence in our model reinforces our belief in its ability to provide valuable insights into CO2 emission patterns. Leveraging this model empowers us to make informed decisions and implement proactive measures to optimize technology usage and minimize CO2 emissions, contributing to a more sustainable future. Ongoing refinement and validation are essential to continuously improve the model's accuracy and applicability, accounting for new data and technological advancements.

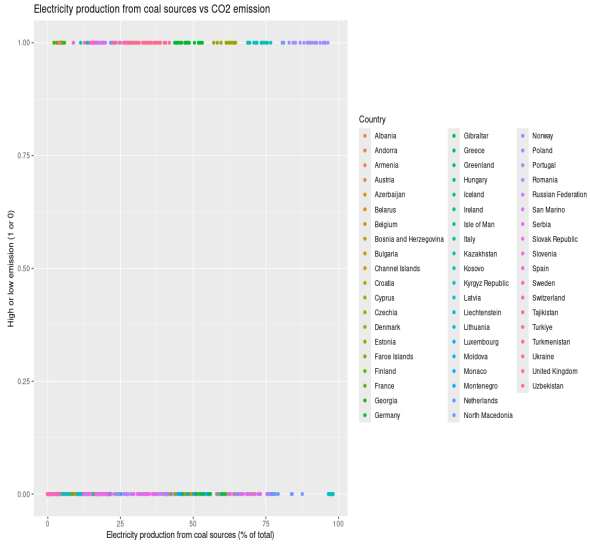
Student works: Andrew Ballance and Juan Manzano Vasquez work on KNN algorithm, Tung Kieu and Brandon Miramontes work on SVM, Kevin Kappadakkamadathil supervises on the project and brings everything together.

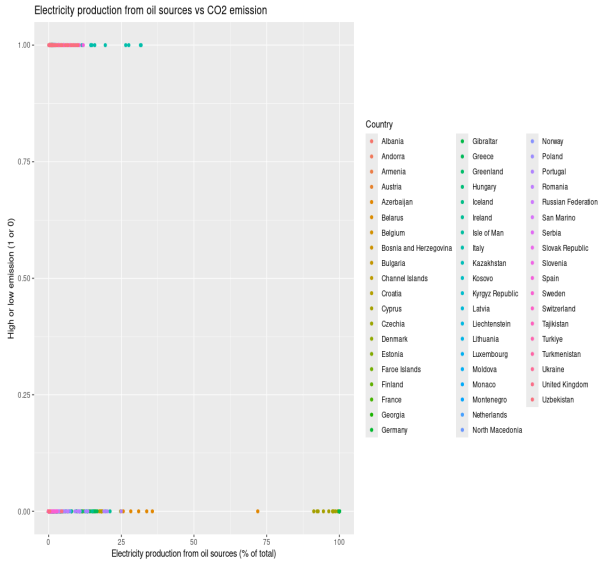
References

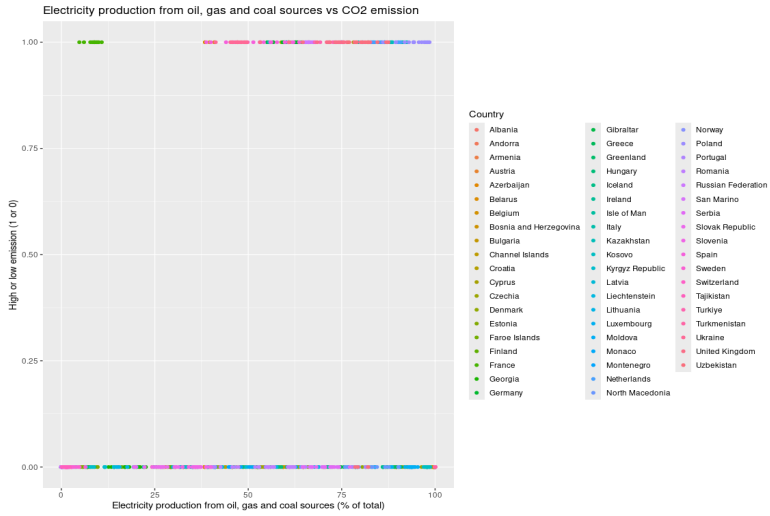
The direct link to the database filtering page is
https://databank.worldbank.org/source/world-development-indicators.

Appendix

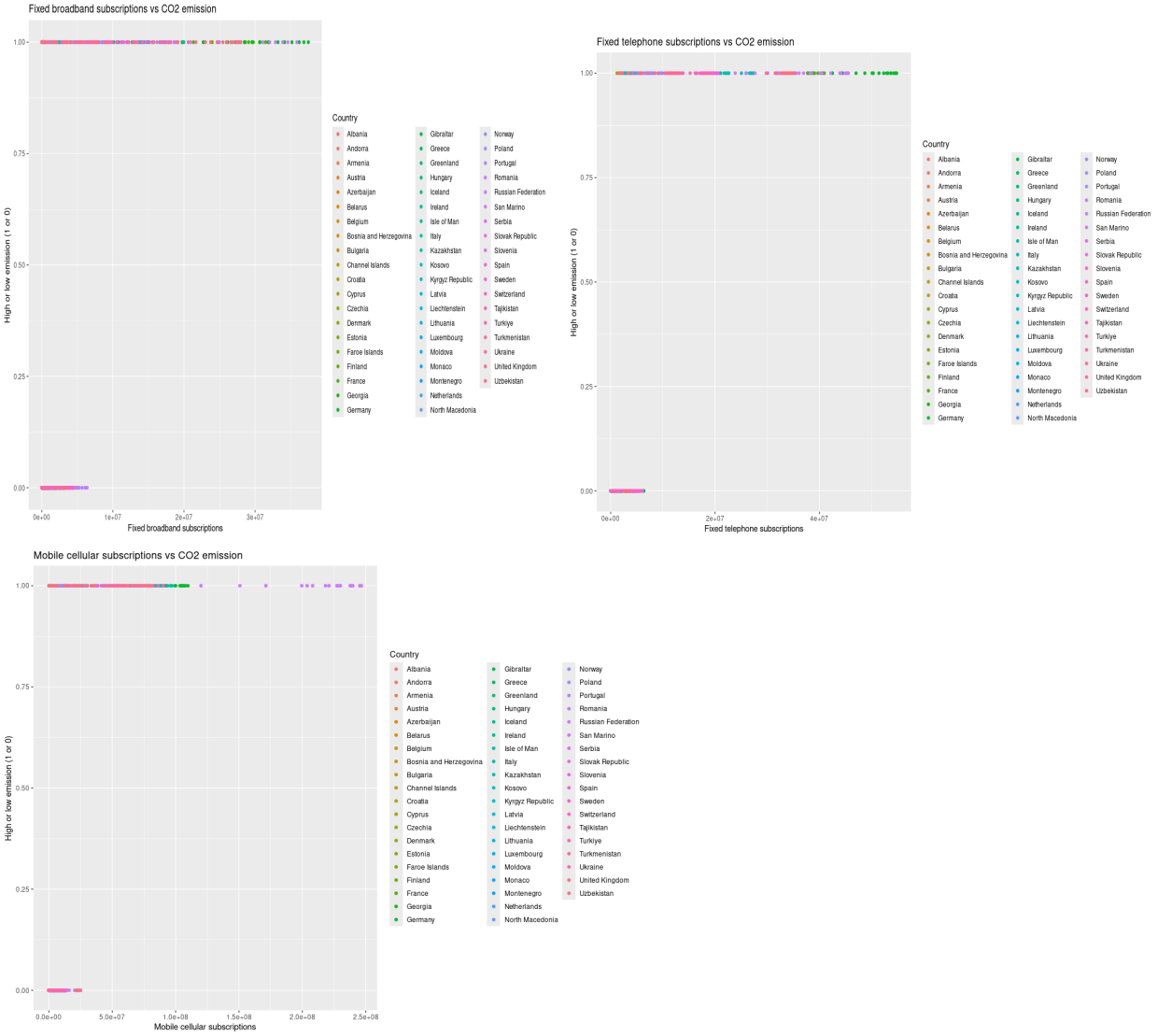
The following are the set of plots produced during our EDA:







Based on the 3 plots generated, the 3 plots based on the 3 features we use for electricity productions to predict on our CO2 emissions. Based in this 3 plots, it seems like our electricity productions based on these fuel types to have a 50/50 ratio where it does predict CO2 emissions is high or low. Like for coal sources it has a positive relationship where CO2 from coal sources is in the high of CO2 emissions, variable “1”. But in the oil sources, it seems to have an inverse relationship where CO2 emissions from oil is high but they are plotted in the low CO2 emissions, variable “0”. And for the last picture, where it combines coal, oil, gas has inverse relationship with CO2 emissions where most of the countries are labeled “0” indicate low emission based on CO2 emissions.



Based on the 3 plots generated, the 3 plots based on the 3 features we use for technology to predict on our CO2 emissions. Based in this 3 plots, it seems like our technology access based are heavily plotted where it does predict CO2 emissions is high, variable “1”. And most the countries to have low values of subscription features tend to be in the low CO2 emissions, which indicates these are trustworthy features to predict high CO2 emissions.