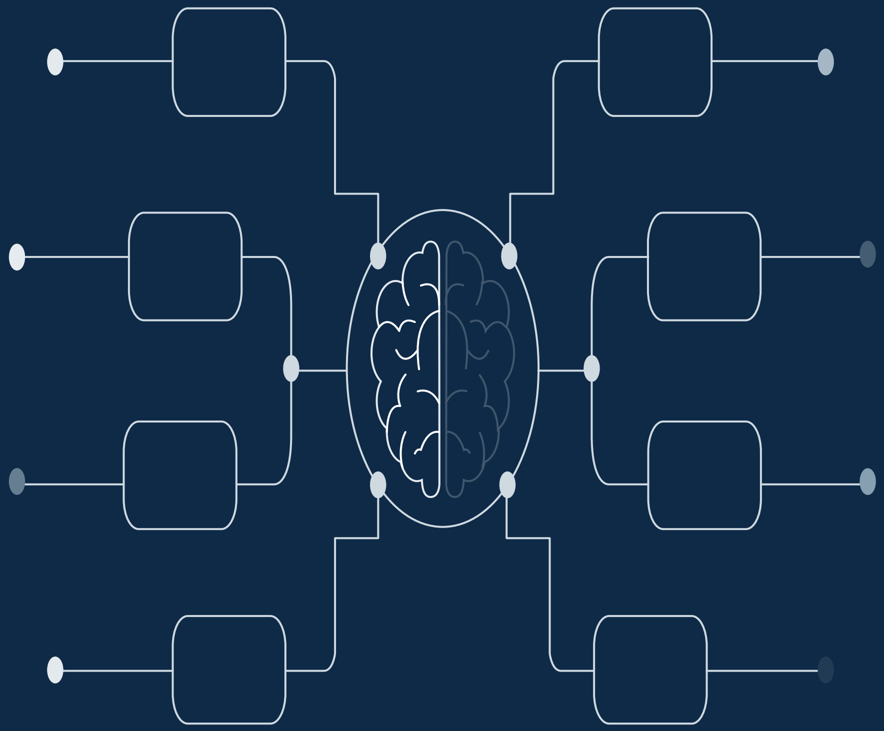
Can Data on Climate change Predict a Country's Socioeconomic class?

INTRODUCTION

- Data source: World Bank
- 2 tables of data
- 20,000 observations
- 217 countries
- 76 variables of climate change
- 1960 - 2021: 62 years
- 4 income classes:
  - High
  - Upper Middle
  - Lower Middle
  - Low

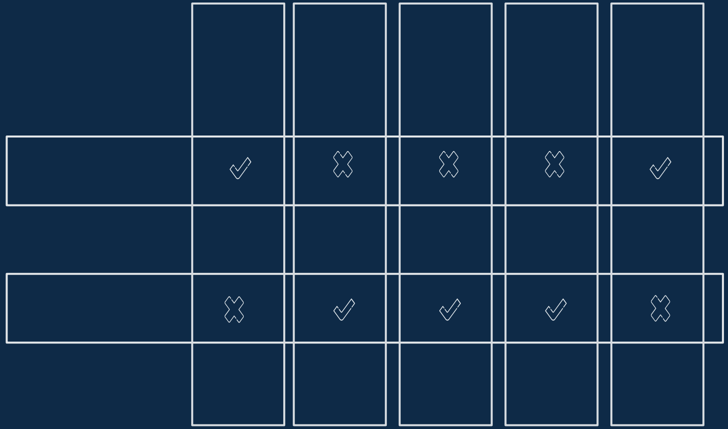


DATA PREPARATION

- **Dropping Columns**
  - Removed **1960 - 1990 (Politicization of climate change)**
  - Dropped columns that overlapped
- **Data Transpose and aggregation**
  - pd melt()
  - Switched variables to columns
  - Used pivot table and aggregated data for countries and years based on median for ML

Country Nar	Countr	Indicator Name	Indicator Code	1960	1961	1962
Aruba	ABW	Urban population (% of total population)	SP.URB.TOTL.IN.ZS	50.776	50.761	50.746
Aruba	ABW	Urban population	SP.URB.TOTL	27525	28139	28537
Aruba	ABW	Urban population growth (annual %)	SP.URB.GROW		2.206183184	1.404497644
Aruba	ABW	Population, total	SP.POP.TOTL	54208	55434	56234
Aruba	ABW	Population growth (annual %)	SP.POP.GROW		2.236462489	1.432843226
Aruba	ABW	Poverty headcount ratio at $2.15 a day (2017 PPP) (% of populatior	SI.POV.DDAY			
Aruba	ABW	Prevalence of underweight, weight for age (% of children under 5)	SH.STA.MALN.ZS			
Aruba	ABW	Community health workers (per 1,000 people)	SH.MED.CMHW.P3			
Aruba	ABW	Mortality rate, under-5 (per 1,000 live births)	SH.DYN.MORT			
Aruba	ABW	Primary completion rate, total (% of relevant age group)	SE.PRM.CMPT.ZS			
Aruba	ABW	School enrollment, primary and secondary (gross), gender parity inc	SE.ENR.PRSC.FM.ZS			
Aruba	ABW	Agriculture, forestry, and fishing, value added (% of GDP)	NV.AGR.TOTL.ZS			
Aruba	ABW	CPIA public sector management and institutions cluster average (1=	IQ.CPA.PUBS.XQ			
Aruba	ABW	Ease of doing business rank (1=most business-friendly regulations)	IC.BUS.EASE.XQ			
Aruba	ABW	Terrestrial and marine protected areas (% of total territorial area)	ER.PTD.TOTL.ZS			
Aruba	ABW	Marine protected areas (% of territorial waters)	ER.MRN.PTMR.ZS			
Aruba	ABW	Terrestrial protected areas (% of total land area)	ER.LND.PTLD.ZS			
Aruba	ABW	Annual freshwater withdrawals, total (% of internal resources)	ER.H2O.FWTL.ZS			
Aruba	ABW	Annual freshwater withdrawals, total (billion cubic meters)	ER.H2O.FWTL.K3			
Aruba	ABW	Population in urban agglomerations of more than 1 million (% of tota	EN.URB.MCTY.TL.ZS			
Aruba	ABW	Population living in areas where elevation is below 5 meters (% of tc	EN.POP.EL5M.ZS			
Aruba	ABW	Urban population living in areas where elevation is below 5 meters (	EN.POP.EL5M.UR.ZS			
Aruba	ABW	Rural population living in areas where elevation is below 5 meters (%	EN.POP.EL5M.RU.ZS			
Aruba	ABW	Droughts, floods, extreme temperatures (% of population, average 1	EN.CLC.MDAT.ZS			
Aruba	ABW	CO2 emissions from solid fuel consumption (% of total)	EN.ATM.CO2E.SF.ZS	0	0	0

DATA PREPARATION



- **Dealt with missing values**
    - Removed variables with more than 80% missing data
        - Removed 26 columns
    - Used MICE - Multivariate Imputation by Chained Equations to deal with the rest of the missing values
- **Merging tables**
    - Merged income class which had response feature to main table using outer join
- **Feature engineering and standardization of data**
    - Original response var: High income, upper middle, lower middle, and low income
    - Broke it down to 3 categories: high, middle and low
    - Used StandardScaler() for standardization

Leftover: 26 variables / features used for analysis and ML

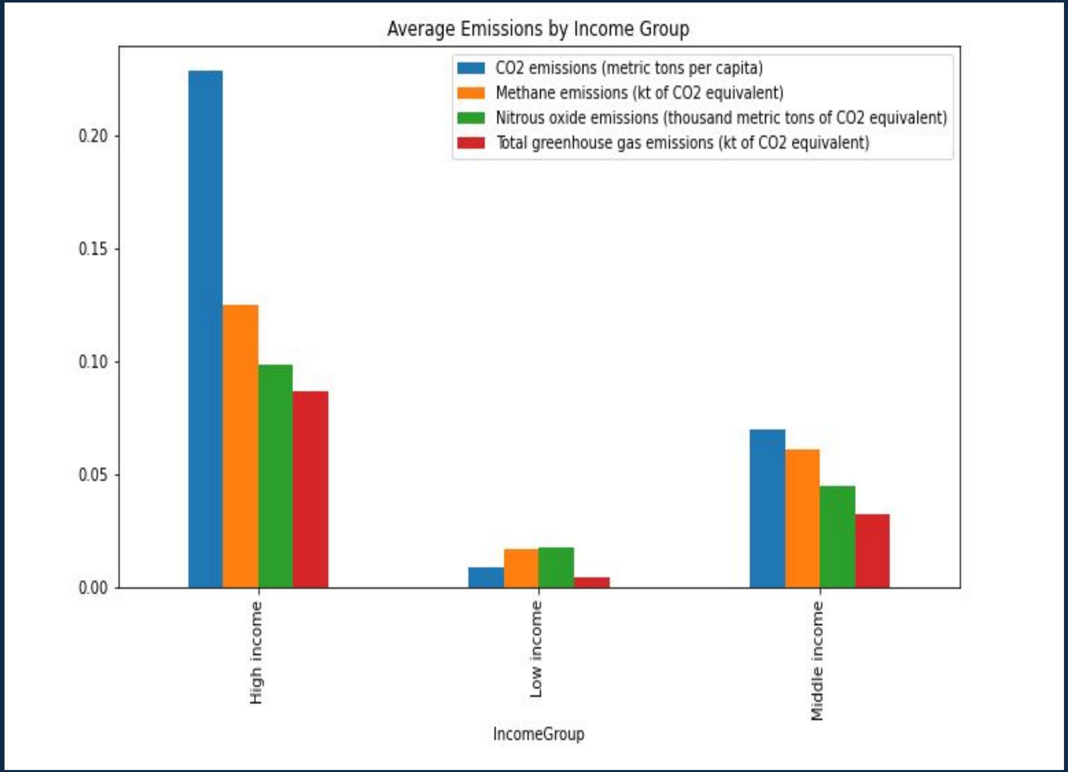
DATA ANALYSIS

- Trends for the top 6 most influential features for predicting Income Group

- Values are significantly different for these variables between the different income groups

- There is correlation between climate change variables and socio economic class

DATA ANALYSIS



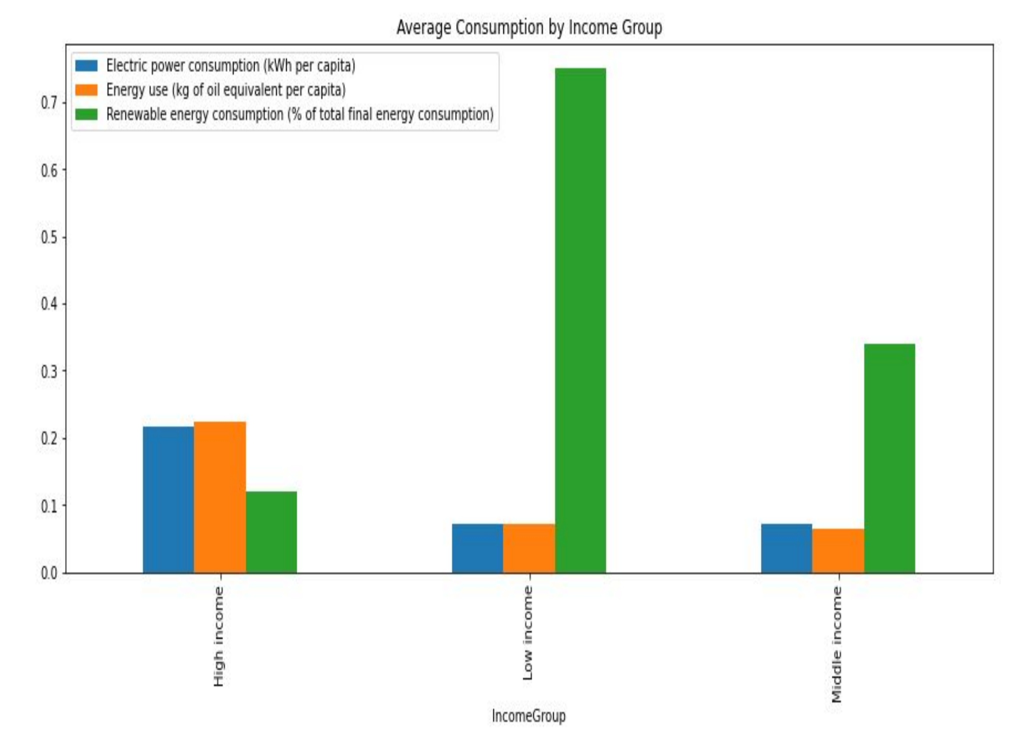
- Compare different gas emissions variables to see if there is a similarity in its trends
- High income countries have higher levels of gas emissions
- Low income countries have lower levels of gas emissions

DATA ANALYSIS

- Compare different energy consumption variables to see if there is a similarity in its trends

- High income countries consume more electric power and use more energy

- Low income countries consume significantly more renewable energy than high income countries



MODEL THEORY AND PREPARATION

```python
# importing sklearn train_test_split function(method) to split data into "training" and "test" set
from sklearn.model_selection import train_test_split
```

	Input 1	Input 2	Input 3	Input 4	Output
	5	1000	20	11	Low
	3	8	19	42	Medium
100%	17	47	83	1000	Low
	42	93.77	42	89	High
	47	83	149	98	High

X_train

y_train

	Input 1	Input 2	Input 3	Input 4	Output
	5	1000	20	11	Low
70%	3	8	19	42	Medium
	17	47	83	1000	Low
30%	42	93.77	42	89	High
	47	83	149	98	High

X_test

y_test

X train	y train	X test	y test

Continued...

MODEL THEORY AND PREPARATION

- Plain & Incomplex Model
- Data and Models were advanced as shown in the next slides...

```
# defined a function which takes in 4 inputs:
def gaussian_naive_bayes(X_train, X_test, y_train, y_test):

    # creating a variable/object for the GaussianNB() class that was imported from sklearn.naive_bayes module
    gnb = GaussianNB()

    # using the fit method to train the model with training data X_train, y_train
    gnb.fit(X_train, y_train)

    # testing the model to predict y-value
    # predicting using the predict method (on gnb object) to predict the data (X_test)
    y_pred = gnb.predict(X_test)

    # Compairing the actual y-values (y_test) with the model's predicted values (y_pred) using the accuracy_score() function
    # from sklearn.metrics
    acc = accuracy_score(y_test, y_pred)

# calling the function() to predict the data
gaussian_naive_bayes(X_train, X_test, y_train, y_test)

# the predicted model is average, the accuracy came to approx. 60%
```

```
accuracy: 0.60 ; custom test prediction: ['Low income']
```

Machine Learning

- Features were standardized so they were all on the same scale and the distance-based algorithms weren’t treating the variables with higher values as more important

- Used a variation of algorithms. For each algorithm we applied cross-validation to tune the hyperparameters and used the F1-score to pick the best model

- Models Used:
    - KNN
    - Support Vector Machine
    - Decision Tree
    - Random Forest
    - AdaBoost

Machine Learning - Raw Income Group

- There were 4 income groups - *High Income, Upper Middle Income, Lower Middle Income, Low Income*
- By using these income groups, here are the results from the models (F1 Scores)

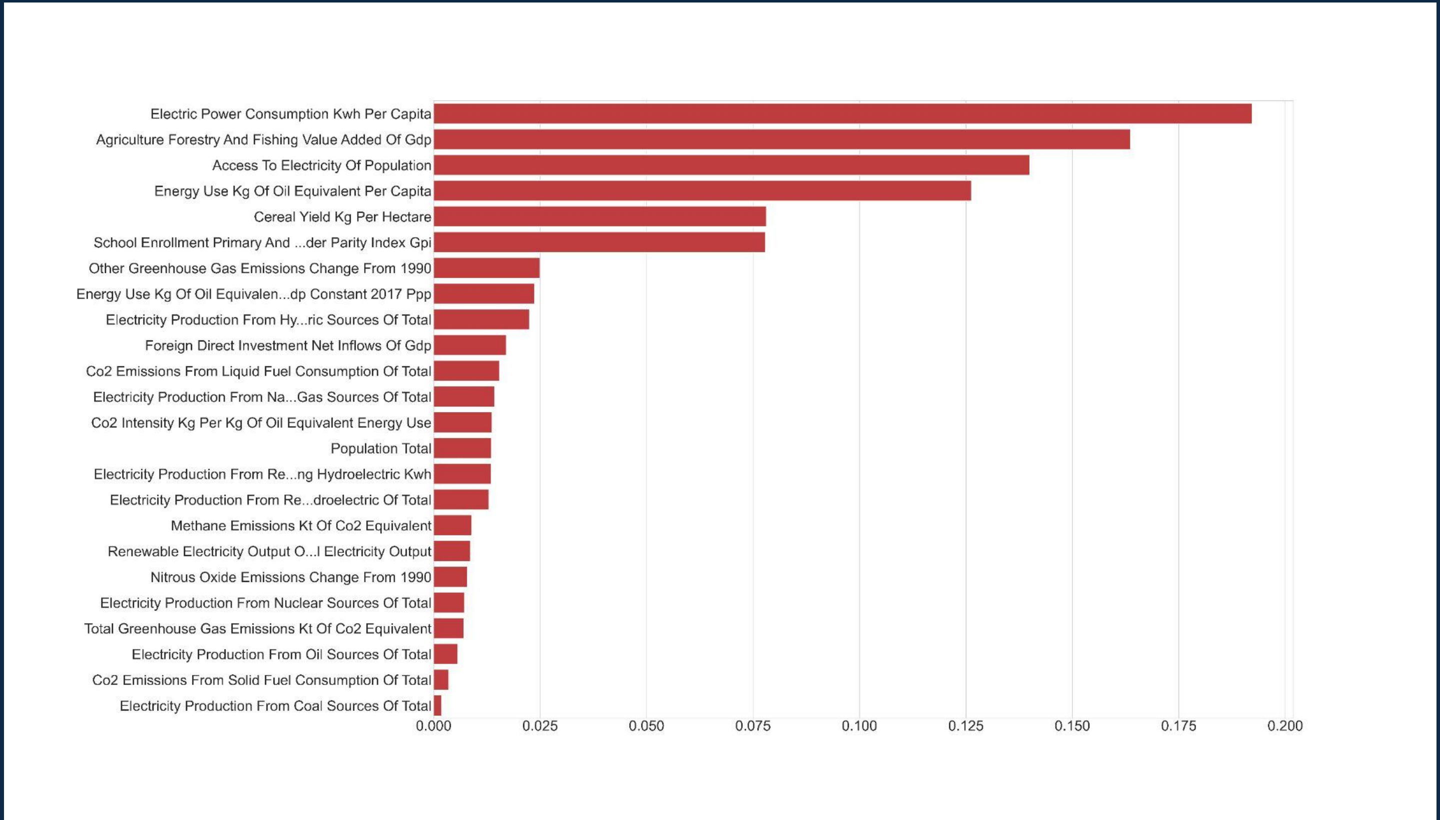
	SVM	Decision Tree	Random Forest	KNN
High Income	0.86	0.89	0.92	0.86
Low Income	0.93	0.88	0.93	0.93
Lower Middle	0.76	0.74	0.94	0.67
Upper Middle	0.67	0.77	0.93	0.74
Weighted Avg	0.79	0.82	0.93	0.80

Machine Learning - Grouped Income Group

- When talking about countries, we normally group them into three groups, so we decided to do that with our data too - we grouped the *Upper Middle Income* and the *Lower Middle Income* together to create *Middle Income*

	Random Forest	AdaBoost
High Income	0.96	0.96
Low Income	0.93	0.86
Middle Income	0.96	0.94
Weighted Avg	0.95	0.93

Machine Learning - Variables of Importance



Conclusion

Preparation

Analysis

Machine Learning Model

Learnings