Introduction

In this project, our research question was "Can we successfully predict average temperature from seven features (country population, renewable energy consumption percentage, number of extreme weather events, annual CO2 emissions, sea level rise, rainfall, and percentage of forest area). We were motivated to answer this question because climate change has led to an increase in the average global temperature in recent years, causing many harms, such as habitat loss for wildlife, droughts, food shortages, and wildfires. Through this project, we hoped that by successfully predicting average temperature, we could build a basis for the common goal of preventing the further increase in temperature in future years.

Our Kaggle dataset, named *Global Climate Change Indicators: A Comprehensive Dataset (2000-2024)*, contains 10 features, although we only used 7 predictive features and 1 target variable, and 1000 samples. We dropped the year and country because we wanted to focus on overall global average temperature predictions, not for a certain year or country. We renamed all of the columns to be one continuous character without spaces or symbols.

Methods

For the neural network model, we defined a Multi-Layer Perceptron class inherited from nn.Module where we initialize the model and progress to a forward pass. Then, we created a new class to fit the model, define the loss function (MSE) and optimizer (Adam), train it, and make predictions. Afterwards, we created a pipeline to first scale the training data and then instantiate the neural network. We used this pipeline in a grid search using 5-fold cross-validation to find the optimal values for the following hyperparameters: hidden layers, hidden units, dropout rate, learning rate, and epochs. Finally, we used the R2 score to assess model performance and define the optimal hyperparameters. We used MSE and R2 on the test set evaluation at the end of the project.

For the random forest model, we defined a pipeline where we first scaled the training data and then instantiated the random forest regressor. We performed a grid search using 5-fold cross-validation again, tuning the number of estimators, max depth, max features, and minimum samples split. We assessed the model performance using R2 again.

When we performed penalized linear regression, we applied StandardScalar to all features and our preprocessing was embedded inside a sklearn Pipeline to prevent data leakage. We used ridge regression (L2 penalty) that shrinks all coefficients towards zero but retains all features. Then, we used lasso regression (L1 penalty) that can shrink coefficients to zero, which performs feature selection. Lastly, we used ElasticNet, which balances ridge and lasso using the L1 ratio parameter. We used a five-fold cross-validation with shuffling, and GridSearchCV to tune the hyperparameters. We used R-squared as our scoring metric for hyperparameter selection. The

best model from these three types of penalized linear regression was the ElasticNet with an alpha of 1 and an L1 ratio of 0.3. Our best CV R-squared score was -0.017.

The Support Vector Regression (SVR) required scaling the data to fit the distance-based algorithm, and this was completed using the StandardScaler() function from Scikit Learn. The model was initiated using a pipeline to apply the scaling to the data and construct the initial model. The final SVR model was developed by performing cross-validation using GridSearchCV. This included the regularization parameter (C) , the kernel, and the kernel coefficient (gamma). The cross-validation process found the best parameters to be C = 0.1, gamma = 0.01, and the best kernel to be the radial basis function, equipped for looking at non-linear patterns in data. Using these three parameters yielded an r2 score of -0.015894269433369447 from cross-validation.

Findings and Results

Overall, the four models we tested were not successful in predicting temperature based on the input variables in the data. Each model showed an r2 score around -0.01, with the Support Vector Regression being the best-performing model of those tested.

Model	**R2 Score**	**Parameters Used**
Random Forest	-0.05016460483482486	max_depth = 10, max_features = sqrt, min_samples_split = 10, n_estimators = 300
Neural Network	-0.05130305691061303	dropout_rate = 0.3, hidden_layers = 3, hidden_units = 32, learning_rate = 0.1, n_epochs = 50
ElasticNet Penalized Regression	-0.0165	alpha = 1, l1_ratio = 0.3
Support Vector Regression	-0.015894269433369447	C = 0.1, gamma = 0.01, kernel = rbf

We used the support vector regression as our final model, built with the parameters that performed best in cross-validation. Unfortunately, this model did not perform as we had hoped when predicting on the test set. The model produced an r2 of -0.0003554235212928081 and an MSE of 69.30686828903669, showing weak to no relationship between the input variables and the target variable, temperature.

Discussion

Our results show that we were unable to accurately predict temperature using the resources and variables we had. Due to our use of several models of different types, it is likely that the data itself led to the observed results. The Support Vector Regression performed the best of the

models tested, albeit by a small margin, and this could be due to the model's design to fit non-linear data by using high dimensions to map the data.

One potential reason that the data may have not shown the correlation we expected was that the data begins in 2000, covering a span of just over two decades. Due to climate change being a phenomena occurring over hundreds to thousands of years, a twenty-year period reflecting different nations might not clearly show the change in global temperature occurring.