Part 1

Question 4

The dataset would be good for making predictions expected levels of emissions from countries and in coming years.

Problem Definition:

Global warming is a major concern in the world today. One of the main contributors to global warming is the human production of greenhouse gases from different industries. To determine what strategies should be taken in dealing with gas emissions, it is important to forecast what levels of gases can be expected from different countries of the world if operations continue without intervention. This is important in establishing urgency and setting up timelines for policy change. Machine learning algorithms such as regression and artificial neural networks and decision trees can be used to achieve this. It is therefore essential we evaluate different machine learning models in predicting greenhouse gas emission levels using independent variables such as country to determine which would be the best one to use. We aim to predict greenhouse gas emissions based on country of emission using emission data from different sectors, namely energy, industrial processes, solvent and other product use, agriculture, waste and forestry.

Question 5

Literature Review

Global warming and pollution are a pertinent topic in the world today. The production of greenhouse gases should, therefore, be monitored to determine if any intervention must be made and when.

In their paper, Saleh et al predicted carbon dioxide emissions using Support Vector Machines (SVM). The independent variables considered in predicting carbon dioxide emission were electricity energy consumption and coal energy consumption. This historical data was obtained from the alcohol industry. The data was normalized, and cross-validation used to split the data into training and testing sets. 90% of the data was assigned as training data used to make the model while 10% was testing data. The parameters of the SVM model, C and epsilon, were varied to determine which produces the optimal result for 10 epochs. The C and epsilon values are chosen by determining which value produces the smallest root mean square error (RMSE) values. The researchers were able, from their results, to relate high energy consumption with high carbon dioxide emission levels.

A paper written by Rehman et al evaluates classical statistical prediction models against Multi-Layered Perceptron Neural Networks (MLP) in forecasting carbon dioxide emissions from the energy and manufacturing sectors in Pakistan, projecting emissions until 2030. The statistical forecasting methods used include Auto-Regressive Integrated Moving Average (ARIMA) and Exponential Smoothing (ES). The methods were evaluated using mean absolute percentage error (MAPE), symmetric mean absolute square error (sMAPE) and mean absolute scaled error (MASE). From their results, it was shown that MLP performed better than both ARIMA and ES by having the smallest MAPE, MASE and sMAPE values. The MLP model was able to give useful information aimed to help environmental policymakers.

Greenhouse gas emission prediction is a valid learning topic in today’s world. Although many research papers exist using different models for predicting emission levels, there are still many areas for research to be conducted.

References

Saleh, Chairul, Nur Rachman Dzakiyullah and Jonathan Bayu Nugroho. "Carbon dioxide emission prediction using support." *IOP Conf. Series: Materials Science and Engineering 114* (2016).

Ur-Rehman, Hakeem, et al. "Forecasting CO2 emissions from energy, manufacturing and transport." (2018). <http://dx.doi.org/10.2139/ssrn.3292279>.

Question 6

- What are you trying to do?

  Evaluate the effectiveness of different models in predicting future carbon dioxide emissions from different countries.

- How is it done today? What are the limits of current practice?

  Forecasting emissions levels have been done using different machine learning techniques such as decision trees and neural networks. However, no literature was found doing a comparison of predicted emissions for different countries. Also, limited literature exists comparing the performance of different machine learning techniques in doing the prediction.

- What's new in your approach and why do you think it will be successful?

  Several papers have been done about forecasting on carbon emission levels using machine learning techniques such as neural networks, decision trees and different regression techniques. However, the focus has been on either some industry or country. It would, therefore, be insightful to compare emission projection across different countries and compare the similarities and differences between these countries.

- Who cares? If you're successful, what difference will it make?

  This is useful information for environmental policymakers and governments. This will allow us to see what levels of emissions can be expected from countries in the future if they would increase or decrease. In doing so, it can be determined if intervention is needed and by when.

- What are the risks and the payoffs?

  The risks include creating a model with inaccurate forecasts that propel policymakers to enact ineffective or even harmful policies. Enactment of these policies can be very expensive and can cause harm to either industry, the environment or both.

  Payoffs include being able to create a model with accurate forecasts that allow policymakers to produce policies that reduce global emissions and therefore slow global warming.

- How much will it cost?

Costs to consider in creating this model will be the cost of obtaining supporting information for the model, memory costs and processing power costs.

- How long will it take?

  Time will be needed in research factors that contribute to emission costs across countries. Depending on the results of that research, it may be necessary to gather supplementary data to form the model. This may take 4 to 6 weeks.

- What is the midterm and final “exams” to check for success?

  The results for the different models would be tested on a validation set as a midterm checking. Finally, the results would be tested against some final test set. The results would be compared using different methods such as Mean Absolute Error and Mean Squared Error.

Question 7

The emission.csv dataset provided will be used. Different machine learning algorithms will be evaluated and then we will determine which one is the best one to use. The dataset would be broken up into sections for training, validation and testing (divided possibly 70%, 10% and 20% of the data).

The dependent variable being predicted is the total greenhouse gas emission (excluding LUCF). Variables used to determine this include country, year and emissions from different sectors, including but not limited to energy, waste, agriculture and forestry.

Three main machine learning algorithms will be compared - feed-forward neural networks, decision trees and linear regression. The results of these algorithms can then be compared using root mean squared error (RMSE) and mean squared error (MSE). The better performing model would produce lower numbers of these metrics. Mean average percentage error (MAPE) was also considered but since MAPE is prone to zero division errors, it was dismissed. Bar graphs can be used to visually compare the results of each metric and easily evaluate which algorithm performed the best. Time graphs can also be created for significant countries within the data to help visualize the course of past emissions data to future emission predictions.

Part 2

Question 9

Topic modelling is an unsupervised machine learning technique that allows us to extract topics from documents. Through topic modelling, we can extract topics being discussed in textual data.

In their paper, Carron-Arthur et al. used topic modelling to determine what topics are most frequently discussed members of an online mental health support group. They compared topics discussed by two different types of users, dubbed “users” and “superusers. The topic model was built by using Latent Dirichlet Allocation (LDA), extracting twenty-five topics and using twenty-one based on coherency and specificity (distance from corpus). Two chi-square tests were then run for each topic. The first test determined, based on topic, if a user or superuser was the one who posted. The second test determined that, for a superuser, if the post was likely a response to a user or superuser. The study determined that superusers were most likely to write on thirteen of the twenty-one topics. However, they were more likely to write on five of their seven unpopular topics in response to regular users. Thus, it was discovered that superusers were likely to be providing support and help to regular users.

In a paper written by Moodley and Marivate, news coverage for two election cycles was analyzed using twenty topics for both LDA and Non-Negative Matrix Factorization (NMF). Topic similarity was determined using a pairwise cosine similarity comparison.   It was determined that coverage tends to be on similar topics across both election cycles. Topics about corruption were popular. It was determined that most of the topics discussed were negative and parties most talked about received lower votes.

Both these applications were able to find useful information that was relevant to the real world. Therefore, topic modelling is a useful machine learning tool, applicable in our society today.

References

Carron-Arthur, Bradley, et al. "What's all the talk about? Topic modelling in a mental health Internet support group." *BMC Psychiatry 16* 28 October 2016: 367.

Moodley, Avashlin Moodley and Vukosi Marivate. "Topic Modelling of News Articles for Two Consecutive Elections in South Africa." *6th International Conference on Soft Computing & Machine Intelligence (ISCMI)*. Johannesburg, South Africa, 2019. 131-134.