Julio Campos
Student ID: 011942624
jca2773@wgu.edu
04/09/2025

AI Optimization Task 1

Step B:

1. *Describe three AI algorithms you considered in your research.*

- Gradient Boosting Machine (GBM): Using the algorithms like XGBoost are effective for structured data like the environmental readings provided in task 1. Their predictive performance and ability to handle missing values are areas where GBMs excel. XGBoost does gradient boosting with pruning and regularization, and takes care to be efficient by organizing memory to avoid cache misses (Russell & Norvig, 2020, p.702). GBMs are strong when dealing with complex relationships between multiple environmental factors and health outcomes.

- Random Forest Regression: This model functions by building multiple decision trees and merges their predictions to provide more accurate and stable results. Random forests handle non-linear relationships well and can manage large datasets without a hitch. Russell and Norvig (2020) state, “The key idea is to randomly vary the attribute choices (rather than the training examples). At each split point in constructing the tree, we select a random sampling of attributes, and then compute which of those gives the highest information gain” (Russell & Norvig, 2020, p.697). For our environmental scenario, we can take advantage of its importance rankings which would help identify which pollutants or weather conditions have the most impact on health risks.

- Long Short-Term Memory Networks (LSTM): LSTMs are specialized recurrent neural networks that recognize patterns in sequential data and are effective for time series forecasting problems. The long-term memory component of an LSTM – the memory cell, is copied from time step to time step. New information enters the memory by adding updates, leading to gradient expressions to not accumulate multiplicatively over time (Russell & Norvig, 2020, p.775). With these properties, using an LSTM can capture dependencies in air quality and weather data and remember that over a long period of time, while filtering less relevant signals.

2. *Determine which algorithm you will use for the remainder of the task, and provide a justification.*

    a. Relationship:
    Random Forest Regression is a good choice for the optimization problem as it creates an ensemble of decision trees that can capture the relationship between the environmental factors like PM2.5, NO2, CO2, temperature, humidity and health risk scores. It can handle both the air quality forecasting and health risk prediction components by learning patterns from the data, which can then be used to make predictions that public health officials can use for preventative measures.

    b. Strengths:
    Random Forest Regression excels at handling the multiple input variables provided by the data without requiring complicated implementation. It can process diverse pollution metrics and weather conditions simultaneously. By the nature of random forests, the extra steps it takes to make the ensemble of trees more diverse reduces variance, unlike bagging decision trees that outputs trees that are too highly correlated (Russell & Norvig, 2020, p.697). The feature of importance rankings that can help identify which environmental factors most significantly impact health risks provides information for targeted intervention strategies.

    c. Limitations:
    Two limitations of Random Forest models in our scenario include their potential struggle with capturing temporal dependencies in air quality patterns over time, which can lead to missing

seasonal trends or long-term shifts in pollution behavior that would be relevant for forecasting. The other limitation is that it might not adapt quickly enough to sudden environmental changes or extreme events, like wildfires, that drastically alter air quality conditions beyond what is provided in the training data.

Step D:

1. *Identify two evaluation metrics that are appropriate for your chosen algorithm to assess its performance.*

   - RMSE (Root Mean Square Error): This measures the standard deviation of prediction errors, providing a value in the same units as our health risk score. The metric steers the model away from large errors and more into the smaller ones. It's a useful metric for our air quality scenario, making sure that there is an absence of large prediction errors that could have consequences on public health.

   - MAPE (Mean Absolute Percentage Error): MAPE displays the model's prediction accuracy as a percentage, making it easy to understand for people without background knowledge in statistics. This metric is valuable for our health risk prediction model for its simplicity in understanding the effects of prediction errors across different risk scores.

2. Implemented and pushed to GitLab Repo.

3. *Analyze the evaluation metrics results from part D2 and document your analysis in the narrative report.*

   - Our Random Forest model demonstrates an RMSE score of 0.1648, which is indicative of good accuracy in its predictions with low deviation from health risk scores. The MAPE score of 1.23% is another good indicator of good prediction accuracy, suggesting that the model's predictions are within 1% of the actual health risk values. The feature analysis shows that temperature is the most dominant factor with a weight of 40%, followed by wind speed at 19%, and CO2 levels at 14%.

References

Fundamentals of AI & ML: Metrics & Evaluation. (n.d.). *Emerging trends in AI/ML evaluation* [Video]. Percipio. https://wgu.percipio.com/videos/2bf9f05b-9eec-4b89-a5e7-25348834c02

Russell, S., & Norvig, P. (2020). *Artificial Intelligence* (4th ed.). Pearson Education (US). https://wgu.vitalsource.com/books/9780134671932