ENSF 611 Machine Learning for Software Engineers – Final Project Proposal

Group Members:
Cameron Dunn, Manuja Senanayake, Edmund Yu

Who – Proposed Client (Fictional)

A Calgary-based startup providing temperature forecasts for event planners, logistics companies, and construction firms. The client seeks to higher temperature prediction accuracy than the standard forecast.

Why – Question/Topic

Can a machine learning model find trends in the discrepancy between the forecasted weather 24hrs in advance and the observed temperature such that it can use the forecast to produce a more accurate temperature prediction.

How – Plan of Attack

We will collect two sets of data, we will collect the hourly 24hr in advance weather forecast data for (2022-2025) from the Open-Meteo Historical Forecast API and the hourly observed temperature data for (2022-2025) from Open-Meteo Historical Weather API. We will clean the data by splitting it into more granular features (hour, day, month) and adjusting it to match its predicted time, we will also apply standard cleaning and pre-processing stages. Three regression models Linear Regression (baseline), Support Vector Regressor and Gradient Boosting will be trained and tuned using k-fold cross-validation. Model performance will be compared using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

What – Dataset, Models, Framework, Components

Dataset:
Hourly Calgary 24hr in advance forecast data (latitude 51.05° N, longitude 114.08° W) from Open-Meteo, including temperature, relative humidity, dew point, cloud cover, precipitation, pressure, visibility, and wind variables.
https://open-meteo.com/en/docs/historical-forecast-api

Hourly Calgary observed temperature data (latitude 51.05° N, longitude 114.08° W) from Open-Meteo.
https://open-meteo.com/en/docs/historical-weather-api

Models:
Linear Regression, Support Vector Regressor, Gradient Boosting Regressor

Framework:
Preprocessing → Data Splitting → Feature Engineering → Model Training → Hyperparameter Tuning → Validation → Visualization

Components:

- Time encoding (sin–cos for hour/day)
- Dew-point spread = temperature − dew point
- Lag features (t − 1 hour) and wind direction encoding (cos, sin)
- Feature scaling and parameter search

Expected Outcome

We expect Support Vector and Gradient Boosting to outperform the linear baseline, achieving lower MAE/RMSE and producing reliable, bias-reduced local temperature predictions for Calgary.