



university of
 groningen

faculty of science
and engineering

Project Proposal

Intro to Machine Learning

Title of Deliverable:	Assignment 3
Group Number:	46
Author(s):	Alexandru Bumbu s5547725 Jonas Žemaitis s5611830
Delivery Date:	January 23, 2025

1 General Idea and Motivation

The goal of this project is to predict temperature of the following day based on historical weather data using machine learning techniques. Weather prediction is a practical and impactful problem. By using machine learning models, we want to explore how well we can predict future temperatures using features such as humidity, wind speed, and past temperatures.

2 Dataset

We will use the public dataset from Kaggle: Weather Prediction Dataset. This dataset includes various meteorological features such as temperature, humidity, wind speed, and pressure recorded over time. It is good for building and evaluating models for temperature prediction tasks. Link to the dataset: <https://www.kaggle.com/datasets/thedevastator/weather-prediction/data>

3 Task and Models

The task is to predict the temperature of the next day given the features in the dataset. We will focus on developing and training regression models to perform this task. The models we plan to explore:

Linear Regression: To serve as a baseline model.

Other regression-based models that can handle non-linear relationships, such as decision trees or ensemble methods (e.g., Random Forest).

Advanced models will not be focused on but may be considered for future exploration if time permits.

4 Goals

Our goal is to build a model that accurately predicts the temperature of the following day. We will evaluate model performance using metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Additionally, we want to detect and mitigate overfitting and ensure robust evaluation through the use of a validation and test set.

5 Methodology

5.1 Data Preprocessing

Clean the dataset by handling missing values and outliers.

Normalize features to improve model performance.

Split the dataset into training, validation, and test sets.

5.2 Model Training

Train baseline and regression-based models on the training set.

Optimize hyperparameters using the validation set.

Use the test set to evaluate final model performance.

5.3 Evaluation Metrics

Calculate MAE and RMSE to assess model accuracy.

Analyze residuals to understand prediction errors.

Examine model performance over time to identify patterns.

5.4 Report Preparation

Document the preprocessing steps, modeling process, and results.

Include visualizations such as feature importance and error distribution.

Discuss challenges, observations, and future improvements.