**Name: Tan Nguyen**

**CS 422**

**INTRODUCTION TO MACHINE LEAERNING**

**FALL 2023**

**ASSIGNMENT 6**

**Machine Learning Report: Heath Row Weather Temperature Regression with MLP**

# Dataset

This dataset, sourced from [Kaggle](https://www.kaggle.com/datasets/sujaykapadnis/whether-prediction-dataset?select=weather_prediction_dataset.csv), comprises daily weather measurements from the years 2000 to 2010, totaling 3,654 observations. The author collected multiple locations across Europe but I chose to focus on a specific location of Heathrow Airport, UK. The dataset includes a range of weather variables:

|  |  |  |  |
| --- | --- | --- | --- |
| Index | Field Name | Description | Dtype |
| 0 | Date | Date (YYYYMMDD) | int64 |
| 1 | MONTH | Month of the year (1-12) | int64 |
| 2 | HEATHROW\_cloud\_cover | Average cloud cover at Heathrow Airport | int64 |
| 3 | HEATHROW\_humidity | Average humidity at Heathrow Airport | float64 |
| 4 | HEATHROW\_pressure | Average atmospheric pressure at Heathrow Airport | float64 |
| 5 | HEATHROW\_global\_radiation | Average global radiation at Heathrow Airport | float64 |
| 6 | HEATHROW\_precipitation | Total precipitation at Heathrow Airport | float64 |
| 7 | HEATHROW\_sunshine | Total sunshine duration at Heathrow Airport | float64 |
| 8 | HEATHROW\_temp\_mean | Average temperature at Heathrow Airport | float64 |
| 9 | HEATHROW\_temp\_min | Minimum recorded temperature at Heathrow Airport for the month | float64 |
| 10 | HEATHROW\_temp\_max | Maximum recorded temperature at Heathrow Airport for the month | float64 |

# Hypotheses

1. **Question #1: There is a significant correlation between temperature and sunshine level.**

*Null Hypothesis (H0): There is no correlation.*

*Alternative Hypothesis (H1): There is a significant correlation.*

1. **Question #2: Mean temperature was increasing over the years from 2000 – 2010**

*Null Hypothesis (H0): The slope of the regression line is zero (no increase in mean temperature over the years).*

*Alternative Hypothesis (H1): The slope of the regression line is greater than zero (mean temperature has increased over the years).*

1. **Question #3: What is the set of features that best predict temperature?**
2. **Question #4: How well the MLPRegressor model can predict the temperature given the dataset?**

# Methodology

* Data preprocessing:
  + Format date: YYYYMMDD to YYYY-MM-DD using pd.to\_datetime()
  + Put into seasons (spring, summer, autumn, winter) using get\_season()
  + Cut and slice dataframe to suit different statistical tests
  + As there is seasonality in temperature, transform day and month into cyclical numerical variables
  + 80/20 data split for train and test
* Statistical tests:
  + For hypothesis #1, bivariate Pearson correlation test
  + For hypothesis #2, slope of regression line of Ordinary Least Squares test
  + For hypothesis #3, correlation table and feature\_importance from RandomForestRegressor model
* Temperature Regression:
  + MLPRegressor:



* + - 3 hidden layers with 100 neurons
    - As regression task, ReLU activation function for all hidden layers
    - Regulartion with alpha=0.01 to reduce overfitting
  + Other regressors as baselines:
    - Sklearn.linear\_model.Linear Regression()
    - Sklearn.neighbors.KNeighbotsRegressor()

# Results

## Finding 1: There is seasonality across variables

A graph with yellow bars

Description automatically generated A graph of different seasons

Description automatically generatedA graph of the seasons

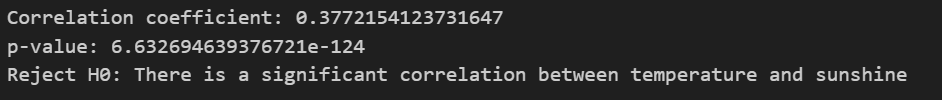
Description automatically generated

## Finding 2: Monthly temperature Distribution

A graph of a graph

Description automatically generated with medium confidence

## Finding 3: Ther is a significant correlation between temperature and sunshine



## Finding 4: There is no significant correlation between year and temperature – Mean temperature did not increase over the year

A screenshot of a computer

Description automatically generated

p-value = 0.126 > 0.05. Thus failed to reject null hypothesis

## Finding 5: Global Radiation, sunshine and humidity variables are the best factors to predict temperature.

A screenshot of a computer screen

Description automatically generated

Output of RandomForestRegressor.feature\_importance:

* + HEATHROW\_global\_radiation: 0.0040962768
  + HEATHROW\_sunshine: 0.0036967246
  + HEATHROW\_humidity: 0.0029268698
  + HEATHROW\_pressure: 0.0020540626
  + HEATHROW\_cloud\_cover: 0.0014867103
  + HEATHROW\_precipitation: 0.0010496996

## Finding 6: MLPRegressor underperforms compared to Linear Regression and KNN Regressor

A screenshot of a computer screen

Description automatically generated

Based on the graph, the MLPRegressor does not perform as well as the other two models on this dataset. The higher MSE and MAE and lower R2 score indicate that the MLPRegressor's predictions are less accurate than those of the Linear Regression and KNN models.

There is a possible explanation for the phenomenon of MLP being overfit is that MLP looking for global loss function while KNN’s nature fit better for seasonal data like weather