IBM Developer
SKILLS NETWORK

IBM COURSERA CAPSTONE PROJECT

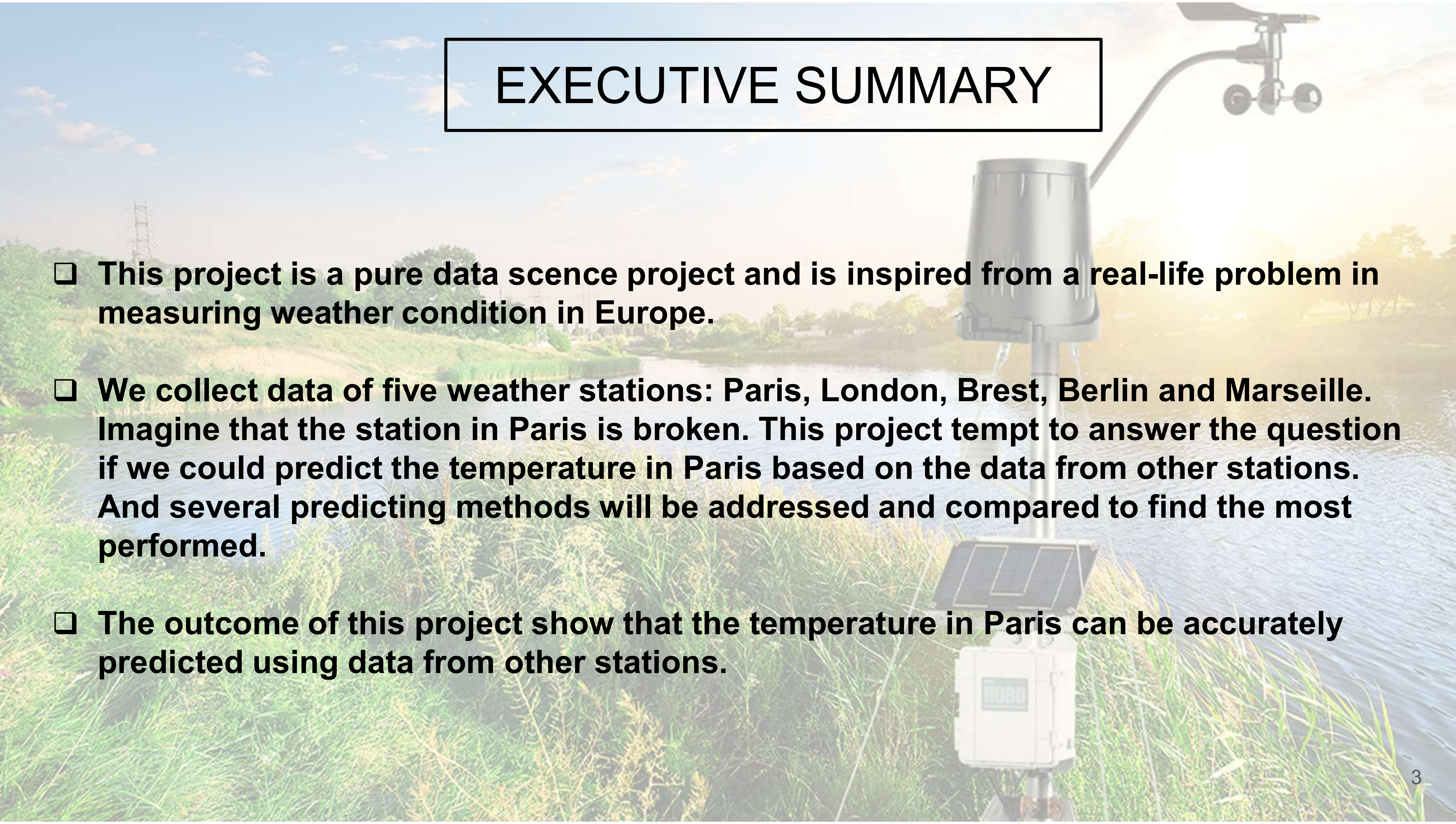
Weather Stations

Vo Thoi Nay

July 2nd, 2024

OUTLINE

- **Introduction**
- **Methodology**
- **Results**
  - 40 features
  - Dimension Reduction
  - 20 features
  - Comparison of all models
- **Discussion and Conclusion**



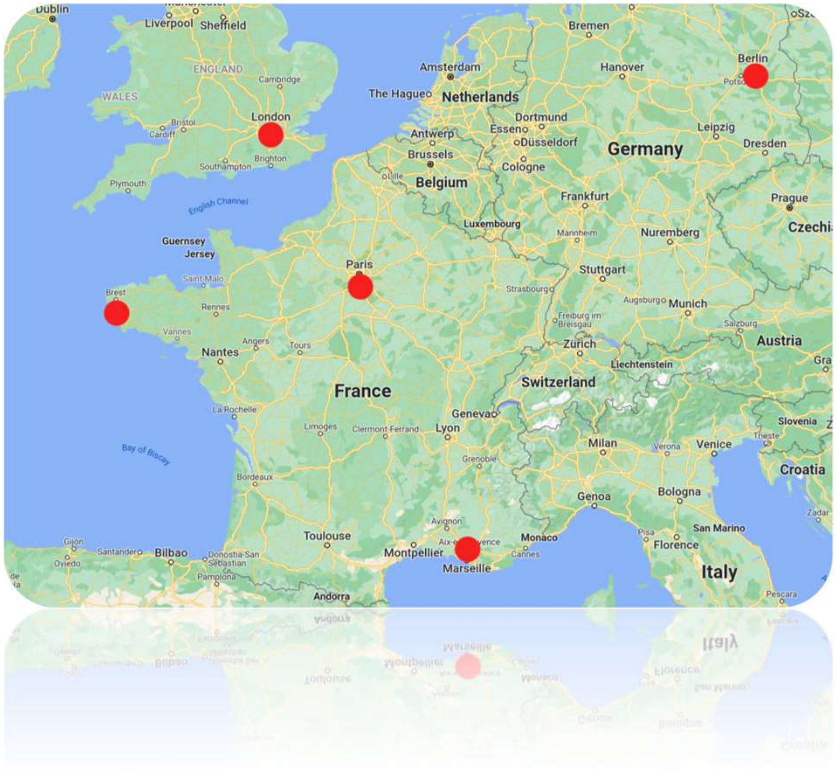
EXECUTIVE SUMMARY

- This project is a pure data scence project and is inspired from a real-life problem in measuring weather condition in Europe.

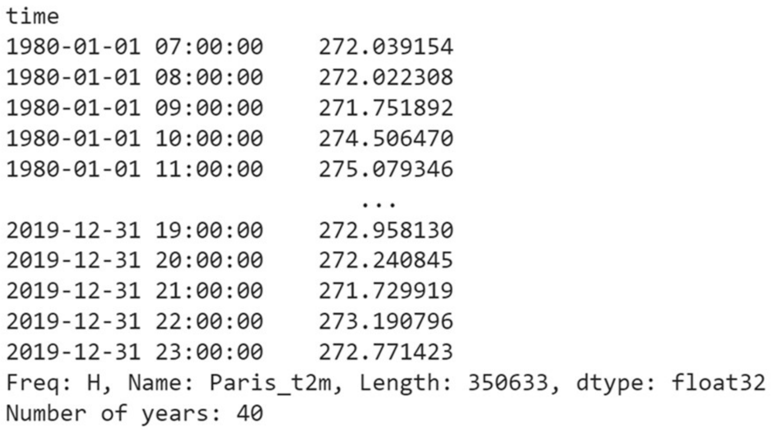
- We collect data of five weather stations: Paris, London, Brest, Berlin and Marseille. Imagine that the station in Paris is broken. This project tempt to answer the question if we could predict the temperature in Paris based on the data from other stations. And several predicting methods will be addressed and compared to find the most performed.

- The outcome of this project show that the temperature in Paris can be accurately predicted using data from other stations.

INTRODUCTION



- 5 stations, 10 different parameters each
- Station in Paris is broken => y: Temperature in Paris, X: 40 features
- Hourly data, 1980 - 2019



```
time
1980-01-01 07:00:00    272.039154
1980-01-01 08:00:00    272.022308
1980-01-01 09:00:00    271.751892
1980-01-01 10:00:00    274.506470
1980-01-01 11:00:00    275.079346
                          ...
2019-12-31 19:00:00    272.958130
2019-12-31 20:00:00    272.240845
2019-12-31 21:00:00    271.729919
2019-12-31 22:00:00    273.190796
2019-12-31 23:00:00    272.771423
Freq: H, Name: Paris_t2m, Length: 350633, dtype: float32
Number of years: 40
```

Temperature data in Paris

DATA COLLECTION

- Source:

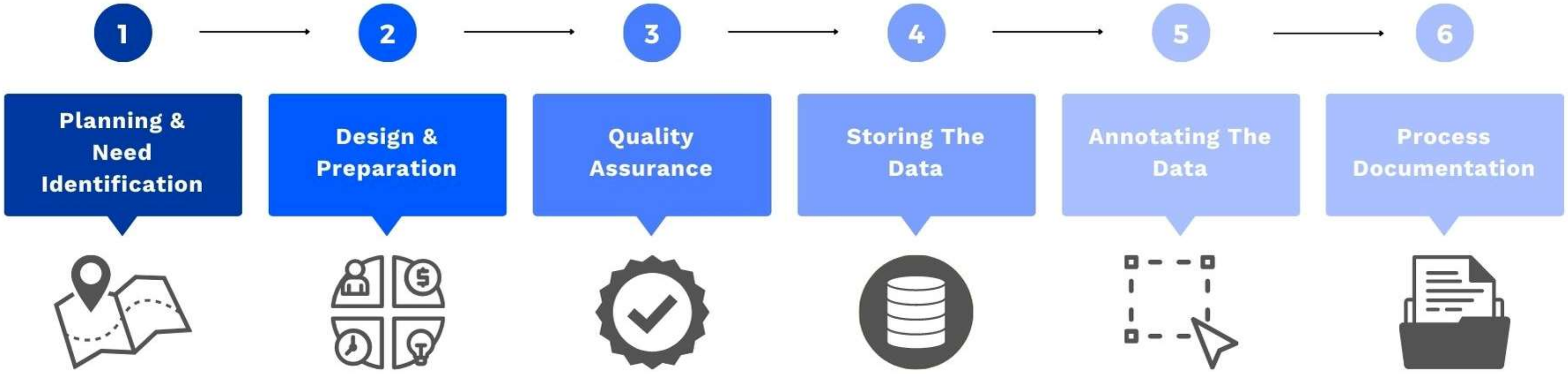
+ The Weather's Record Keeper: https://meteostat.net/en/
+ Personal Weather Station Network: https://www.wunderground.com/pws/overview

- Language used: SQL

- Data Structure:

+ Timeseries data, dated from 1980 to 2019
+ 5 weather stations
+ 10 features each station

DATA COLLECTION PROCESS



1 Planning & Need Identification

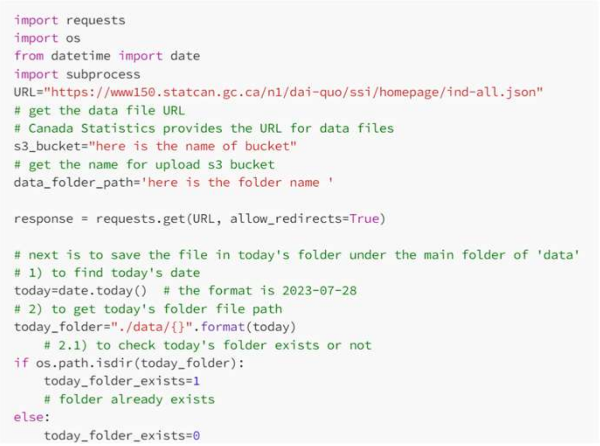
2 Design & Preparation

3 Quality Assurance

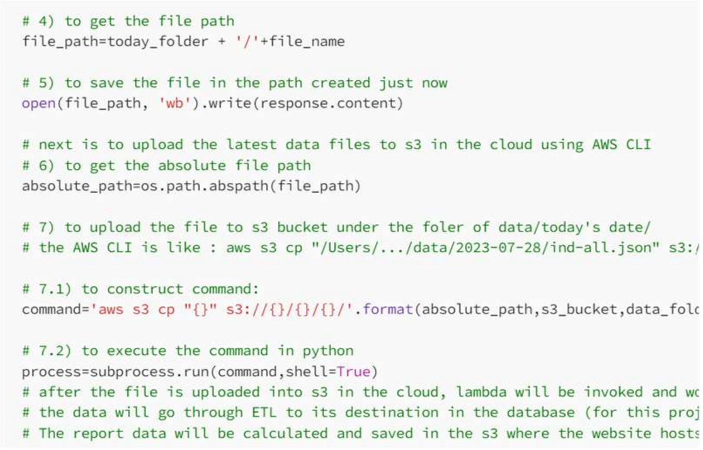
4 Storing The Data

5 Annotating The Data

6 Process Documentation

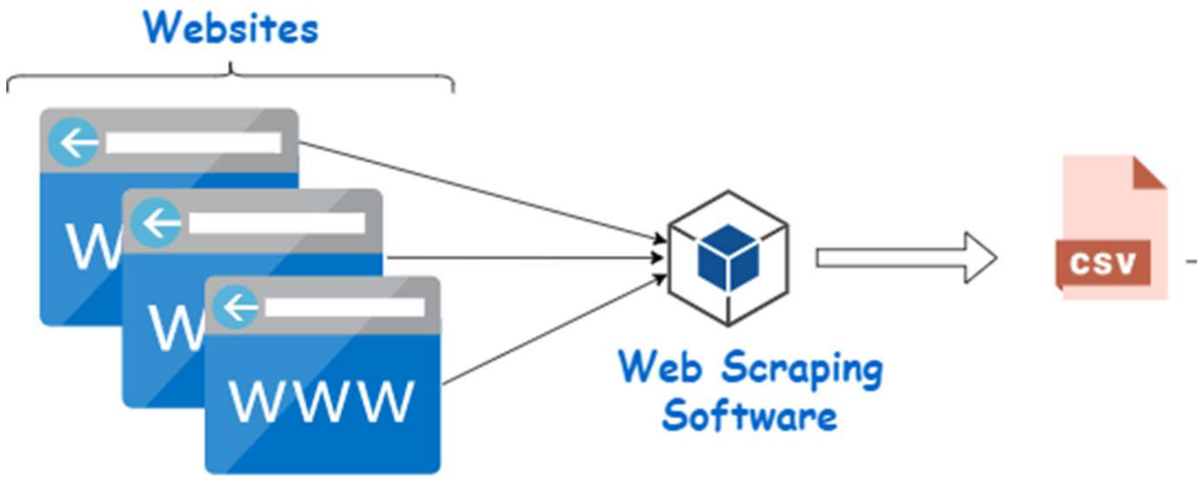


```
import requests
import os
from datetime import date
import subprocess
URL="https://www150.statcan.gc.ca/n1/dai-quo/ssi/homepage/ind-all.json"
# get the data file URL
# Canada Statistics provides the URL for data files
s3_bucket="here is the name of bucket"
# get the name for upload s3 bucket
data_folder_path='here is the folder name '

response = requests.get(URL, allow_redirects=True)

# next is to save the file in today's folder under the main folder of 'data'
# 1) to find today's date
today=date.today()  # the format is 2023-07-28
# 2) to get today's folder file path
today_folder="./data/{}".format(today)
    # 2.1) to check today's folder exists or not
if os.path.isdir(today_folder):
    today_folder_exists=1
    # folder already exists
else:
    today_folder_exists=0
```



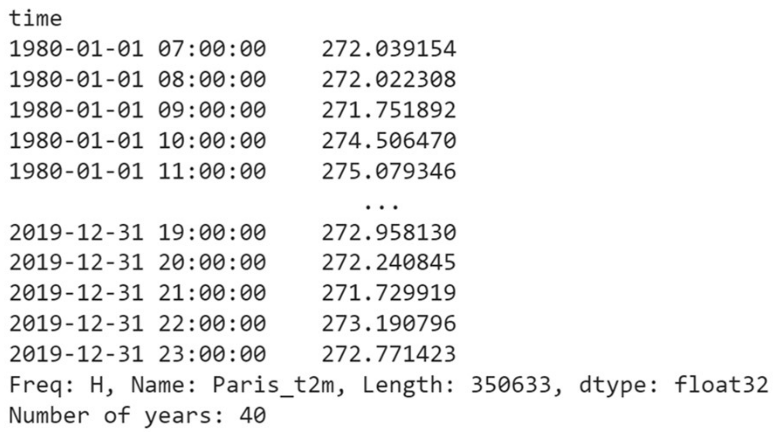
```
# 4) to get the file path
file_path=today_folder + '/'+file_name

# 5) to save the file in the path created just now
open(file_path, 'wb').write(response.content)

# next is to upload the latest data files to s3 in the cloud using AWS CLI
# 6) to get the absolute file path
absolute_path=os.path.abspath(file_path)

# 7) to upload the file to s3 bucket under the foler of data/today's date/
# the AWS CLI is like : aws s3 cp "/Users/.../data/2023-07-28/ind-all.json" s3:/

# 7.1) to construct command:
command='aws s3 cp "{}" s3://{}/{}/{}/'.format(absolute_path,s3_bucket,data_fold

# 7.2) to execute the command in python
process=subprocess.run(command,shell=True)
# after the file is uploaded into s3 in the cloud, lambda will be invoked and wo
# the data will go through ETL to its destination in the database (for this proj
# The report data will be calculated and saved in the s3 where the website hosts
```

DATA SCRAPING

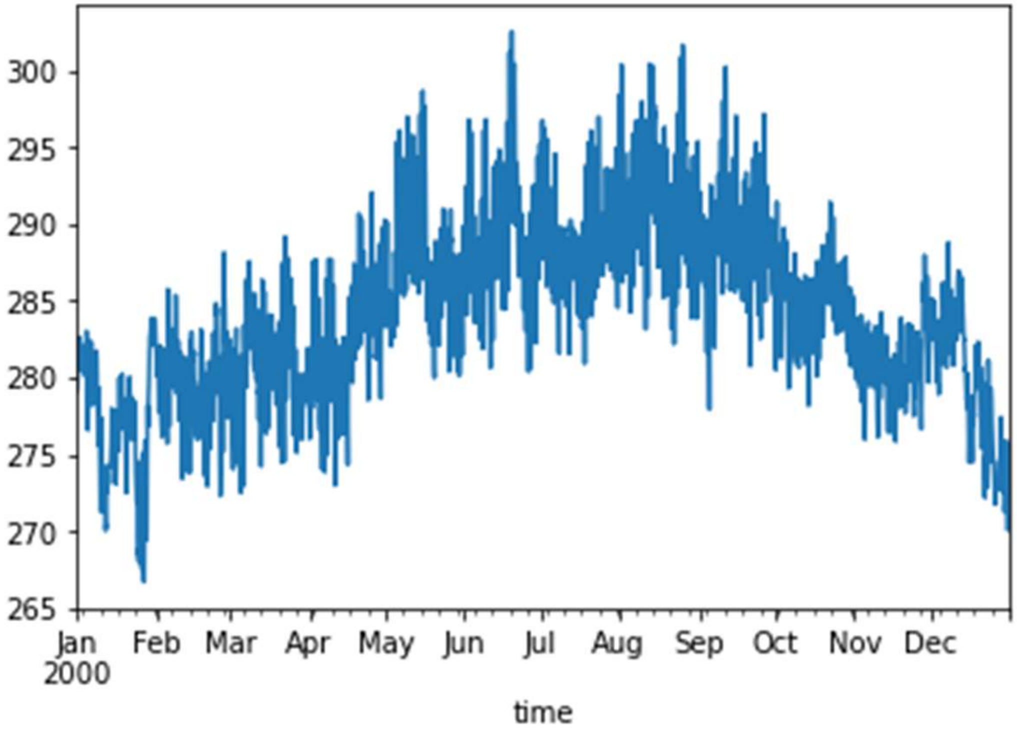


Data is stored in csv format and structured by SQL language

EDA AND DATA VISUALISATION

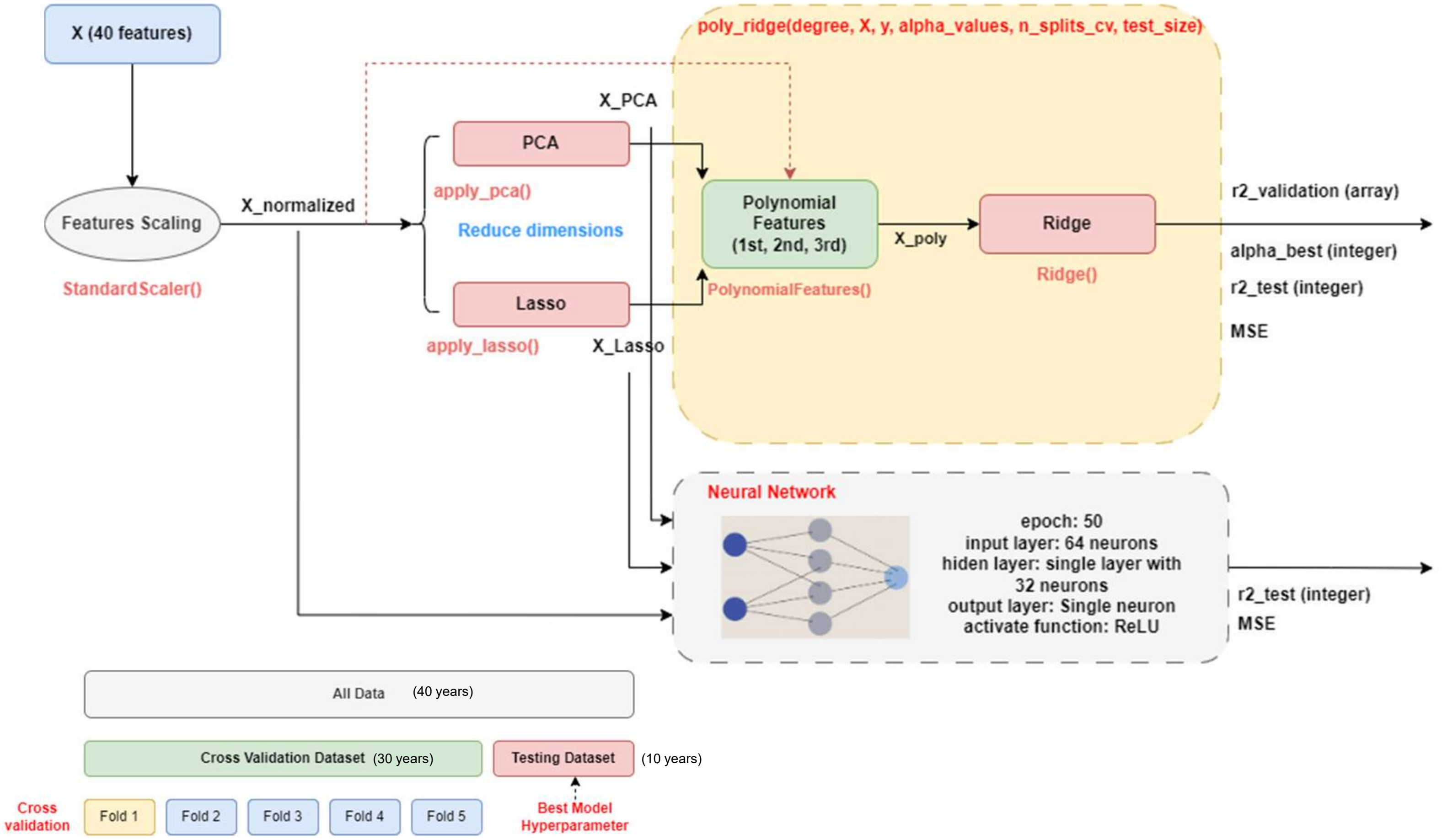


```
time
1980-01-01 07:00:00    272.039154
1980-01-01 08:00:00    272.022308
1980-01-01 09:00:00    271.751892
1980-01-01 10:00:00    274.506470
1980-01-01 11:00:00    275.079346
                          ...
2019-12-31 19:00:00    272.958130
2019-12-31 20:00:00    272.240845
2019-12-31 21:00:00    271.729919
2019-12-31 22:00:00    273.190796
2019-12-31 23:00:00    272.771423
Freq: H, Name: Paris_t2m, Length: 350633, dtype: float32
Number of years: 40
```



Time series data collected

METHODOLOGY



X (40 features)

Features Scaling

StandardScaler()

X_normalized

PCA

apply_pca()

Reduce dimensions

Lasso

apply_lasso()

X_PCA

X_Lasso

poly_ridge(degree, X, y, alpha_values, n_splits_cv, test_size)

Polynomial Features (1st, 2nd, 3rd)

PolynomialFeatures()

X_poly

Ridge

Ridge()

r2_validation (array)

alpha_best (integer)

r2_test (integer)

MSE

Neural Network

epoch: 50
input layer: 64 neurons
hiden layer: single layer with 32 neurons
output layer: Single neuron
activate function: ReLU

r2_test (integer)

MSE

All Data (40 years)

Cross Validation Dataset (30 years)

Testing Dataset (10 years)

Cross validation

Fold 1

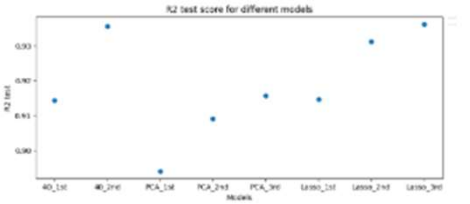
Fold 2

Fold 3

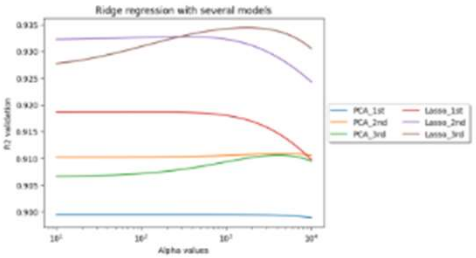
Fold 4

Fold 5

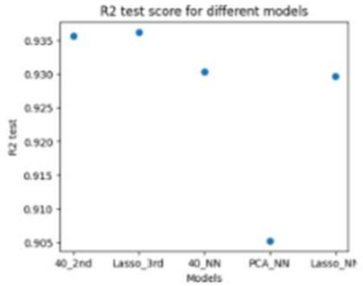
Best Model Hyperparameter



R2 test score for different models

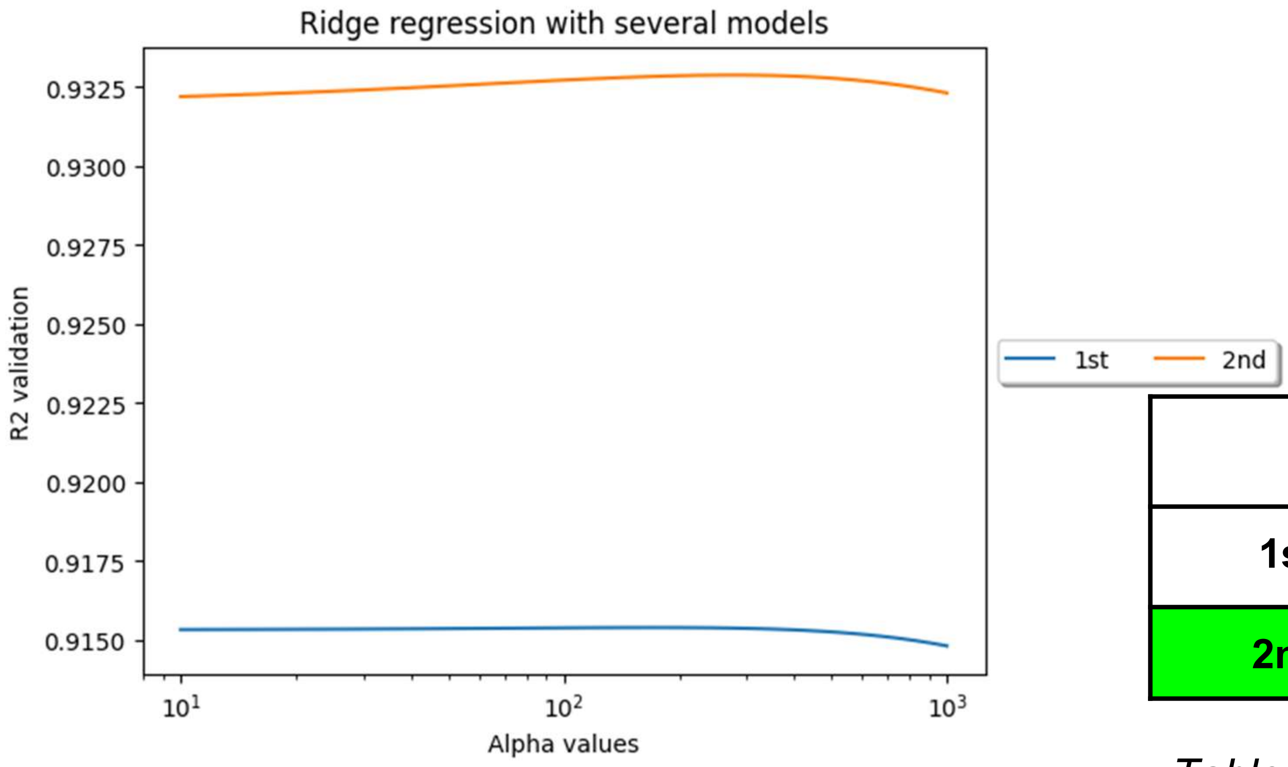


Ridge regression with several models



R2 test score for different models

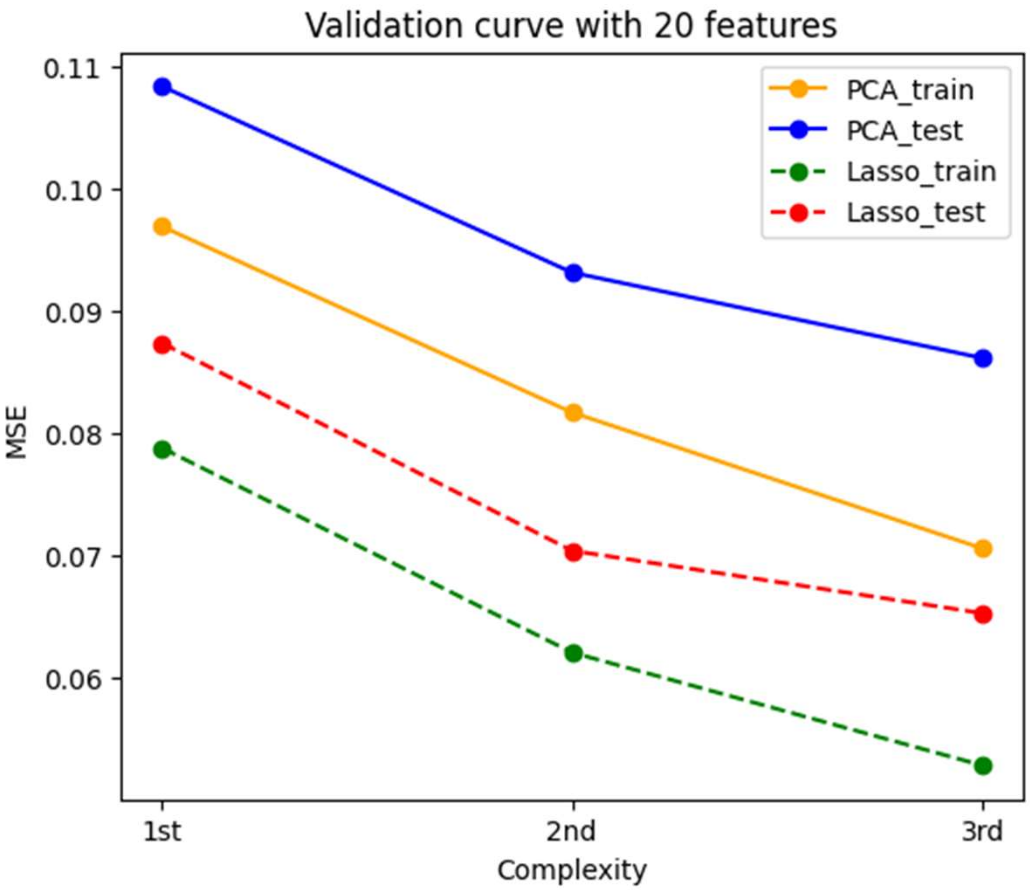
RESULTS - Comparison within 40 features



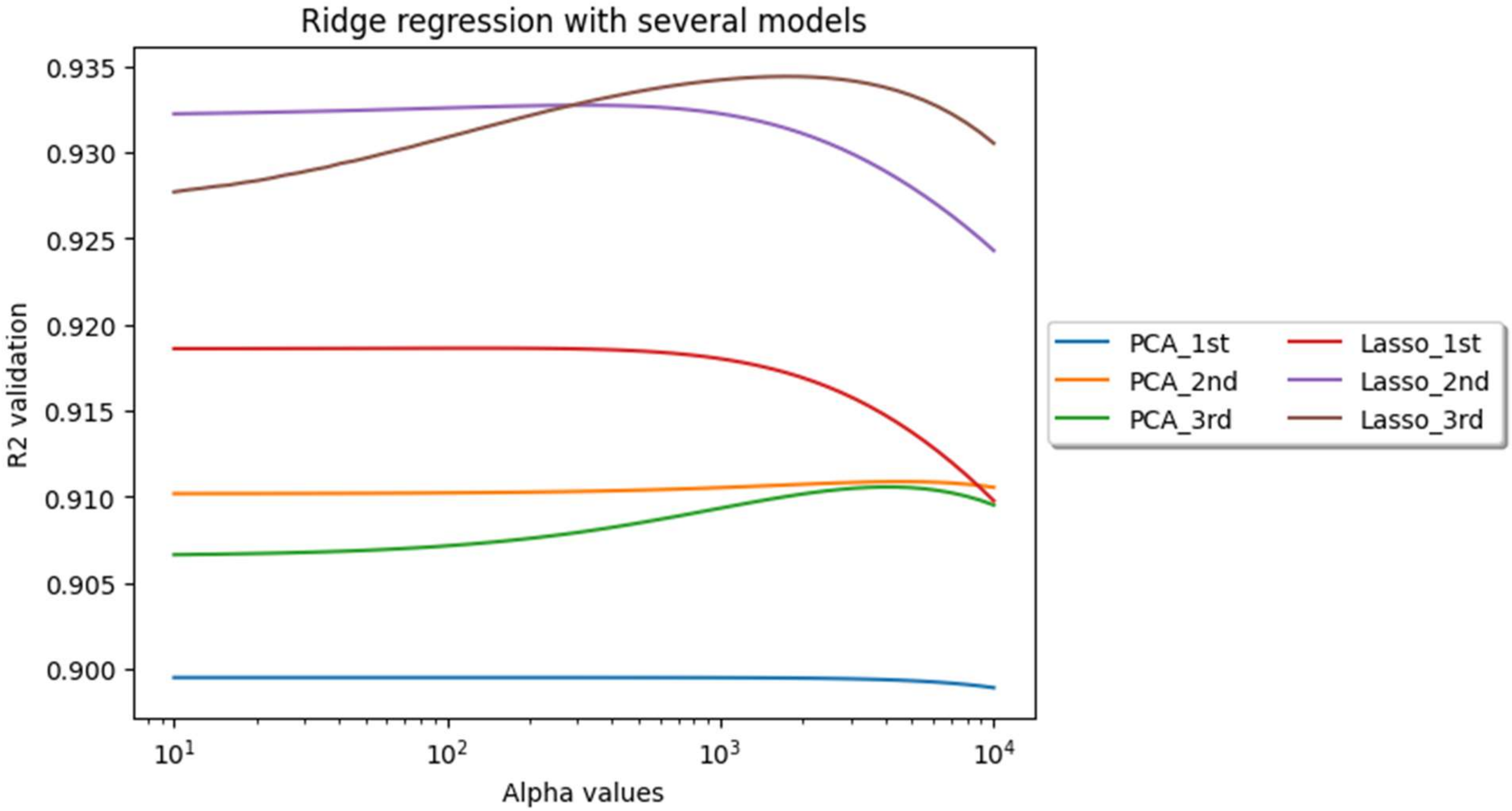
	MSE_train	MSE_test
1st	**0.078**	**0.088**
2nd	**0.056**	**0.067**

Table1. MSE of best alpha for two models

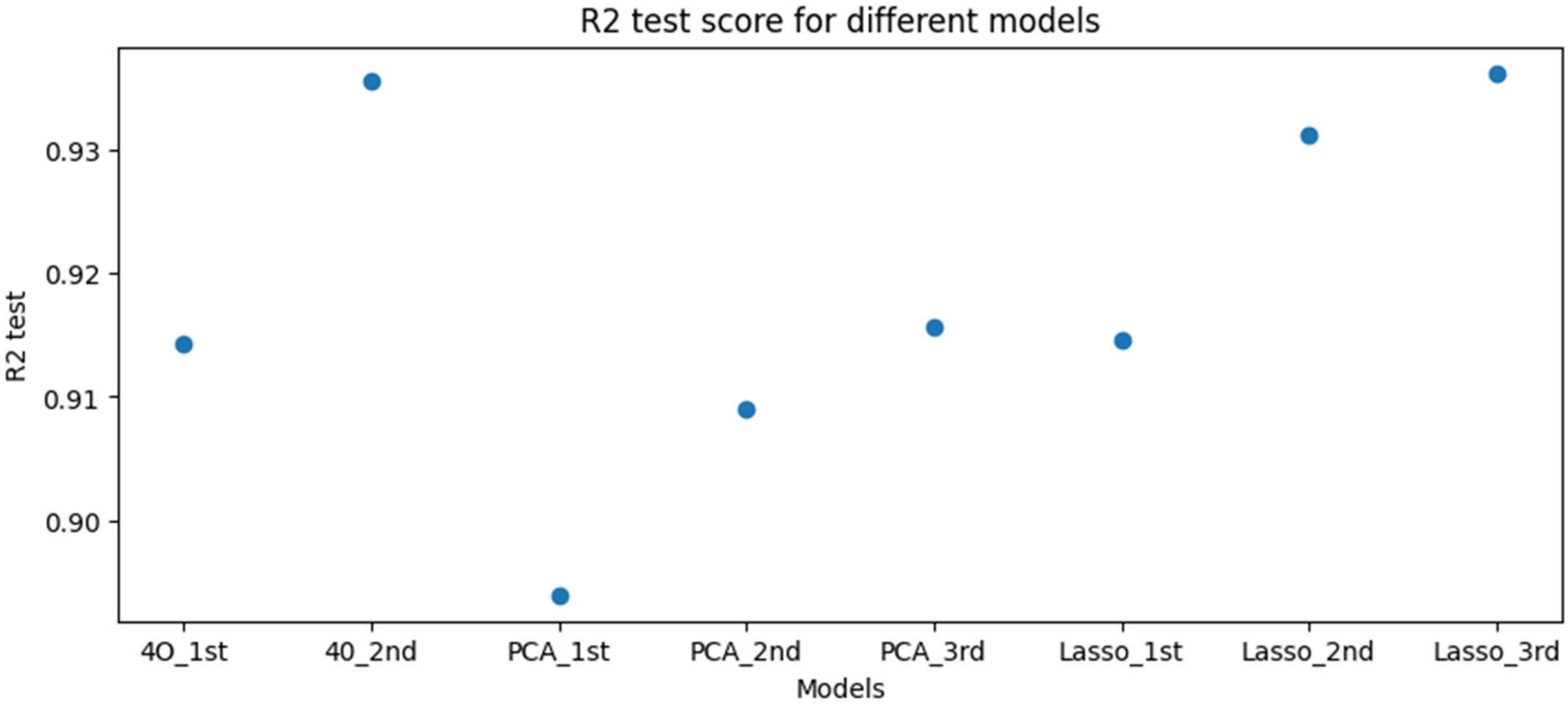
RESULTS - 20 Features Validation Curve (alpha = 0)



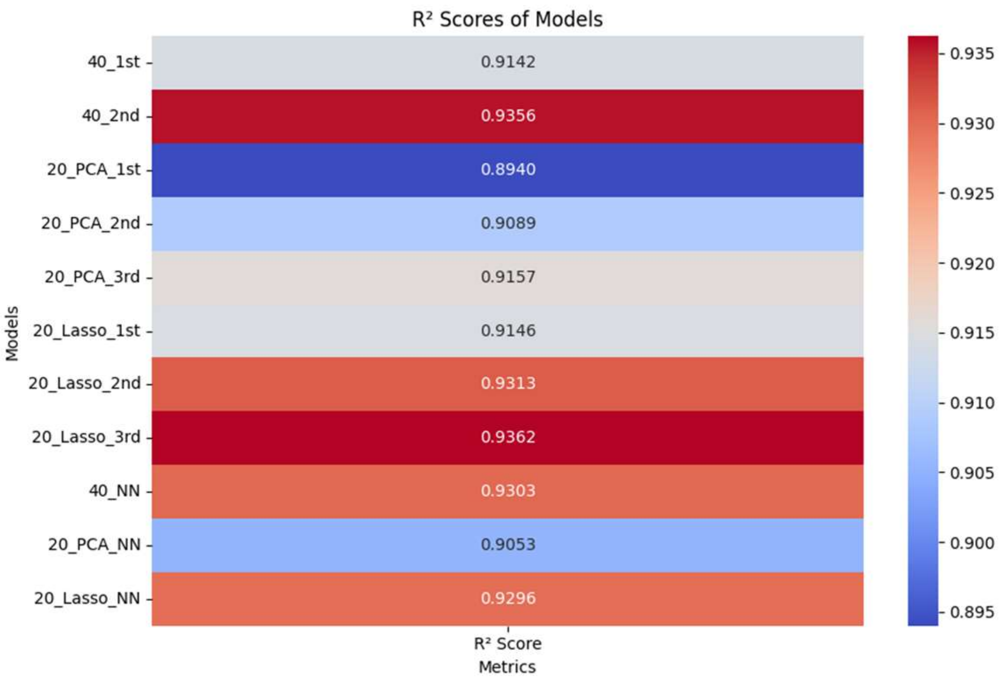
RESULTS - Comparison within 20 features

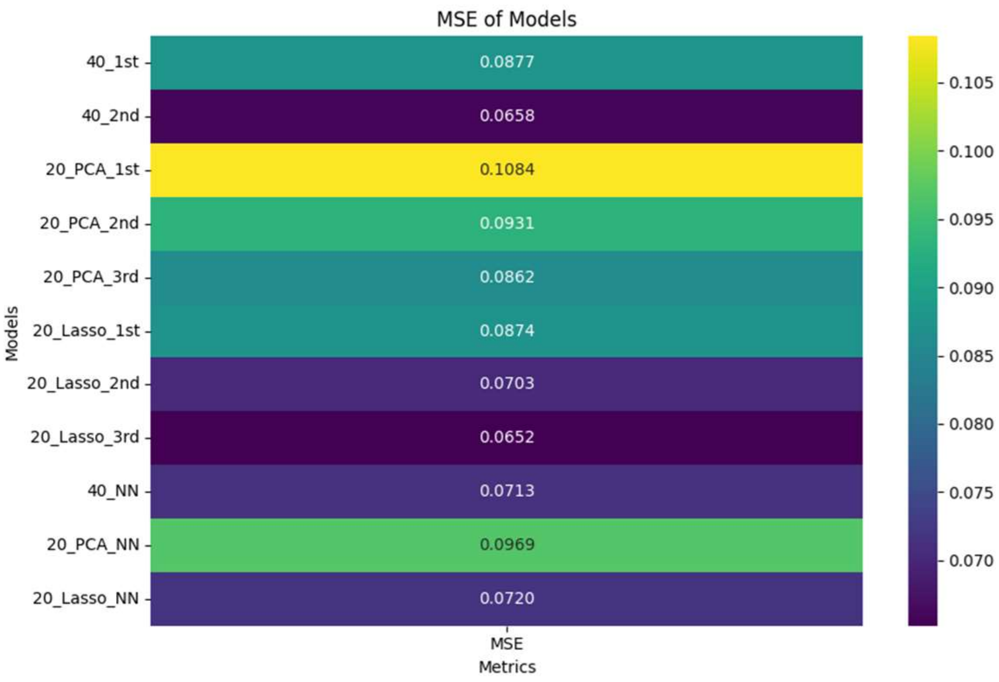


RESULTS - Comparison with different models

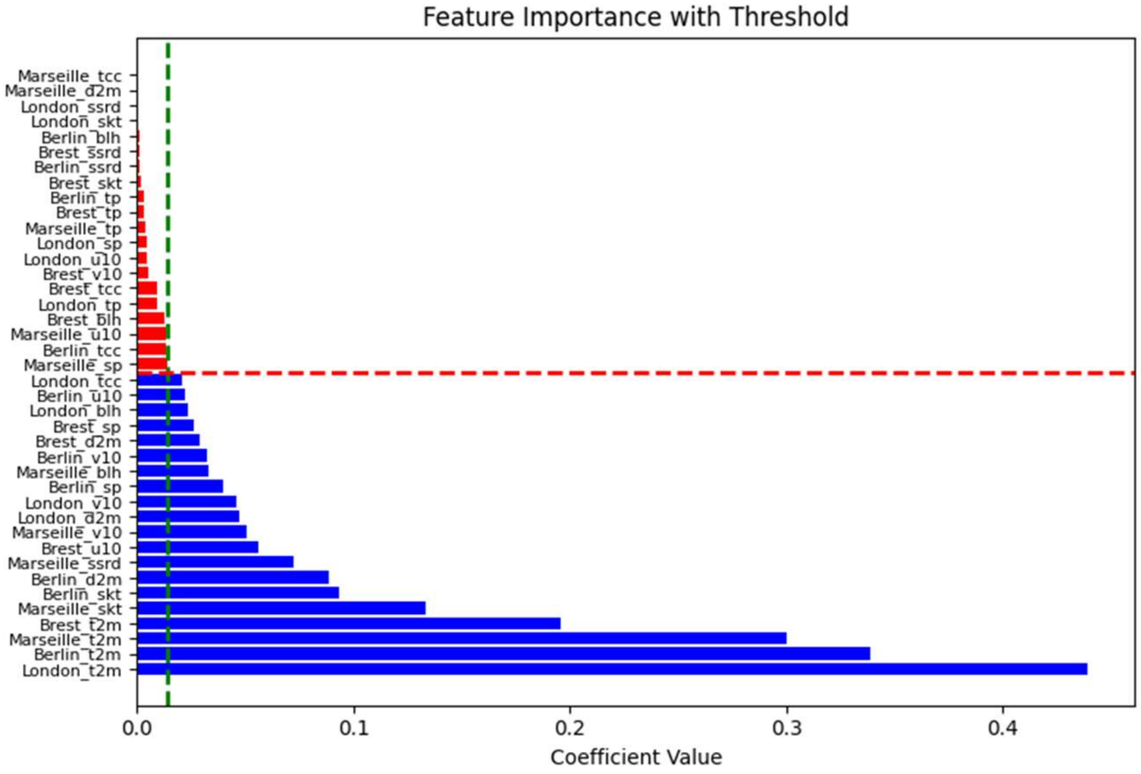


EDA WITH SQL RESULTS

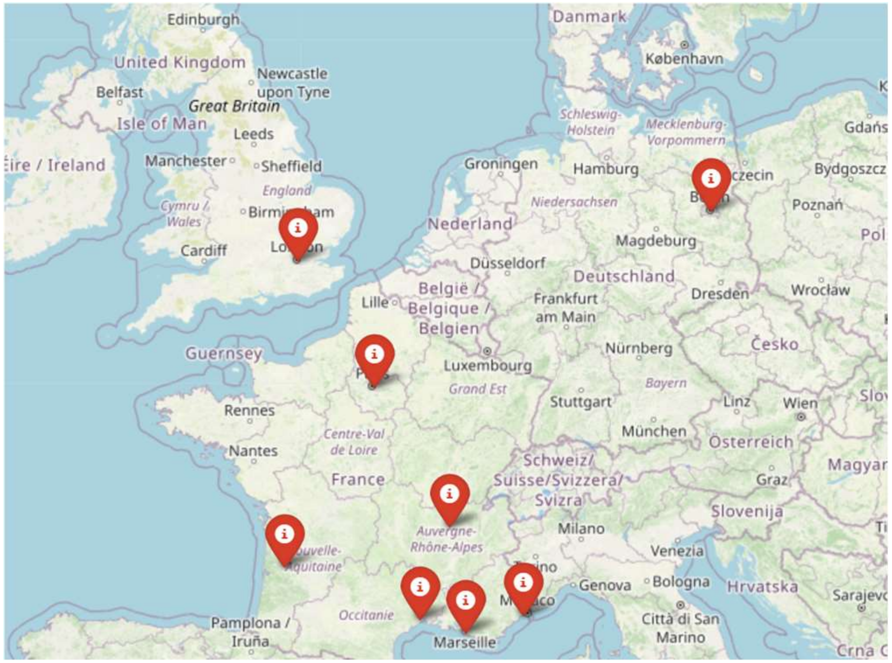




PLOTLY DASH DASHBOARD RESULTS



INTERACTIVE MAP WITH FOLIUM



- file:///C:/Users/Nay/Downloads/interactive_map_europe.html

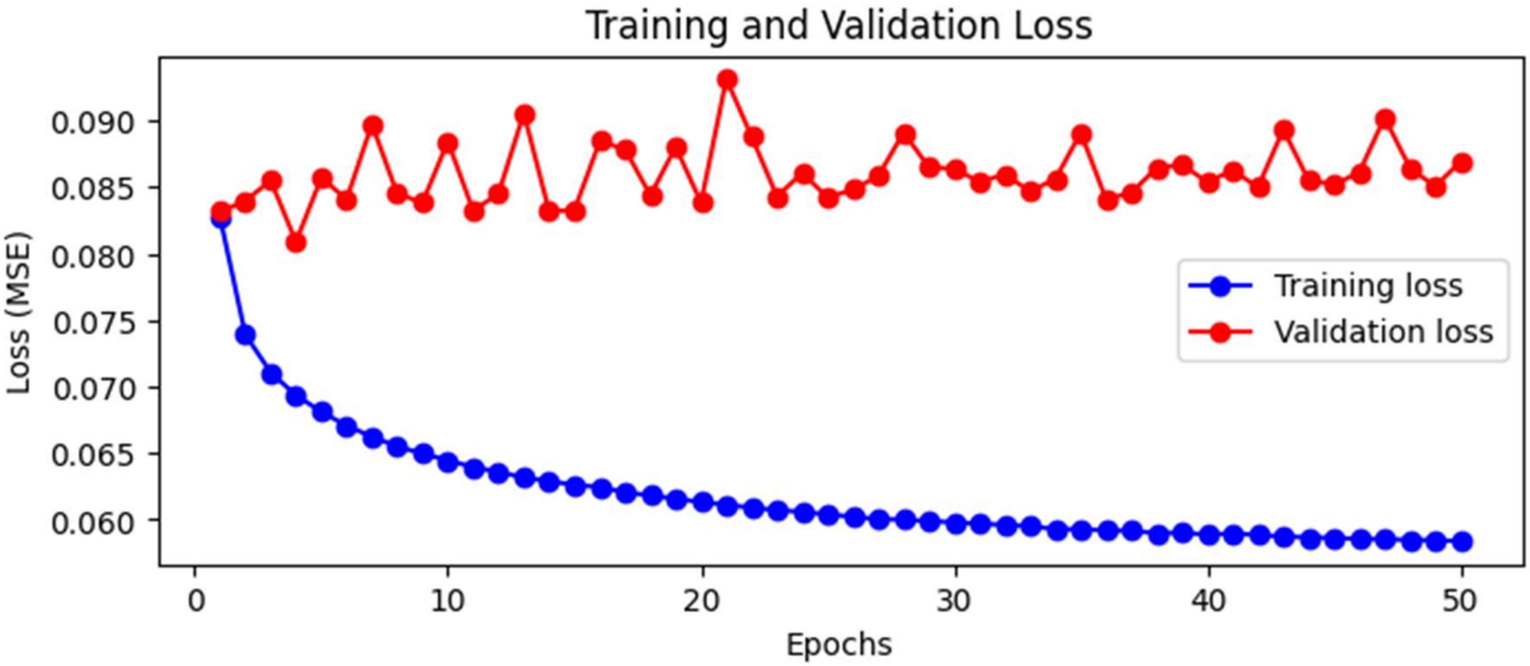
DISCUSSION & CONCLUSION

- **Data Splitting Methodology:** Validation set for parameter tuning, testing set for performance evaluation; nested cross-validation considered but not implemented.
- **Feature Reduction Insights:** Lasso's supervised feature reduction outperformed PCA's unsupervised approach, allowing for the precise selection of the top 20 features from 40, enhancing model effectiveness.
- **Best Evaluation with 40 features:** 2nd-degree polynomial model with alpha best = 268 and R2 test = 0.935.
- **Best Evaluation with 20 features:** 3rd-degree polynomial model reduced by Lasso with alpha best = 1727 and R2 test = 0.936.
- **Regularization Effects:** Lasso's regularization had a more substantial impact on performance alteration than PCA.
- **Neural Network Comparison:** Compared with the best models (40_2nd degree, Lasso_20_3rd degree) from polynomial families, neural networks perform worse.
- **The temperature in Paris can be accurately predicted using data from other stations.**

APPENDIX - Parameters

Abbreviation	Description
skt	Skin Temperature - The temperature of the land or sea surface.
u10	10-meter U Wind Component - The east-west (zonal) component of wind speed measured at 10 meters above the ground.
v10	10-meter V Wind Component - The north-south (meridional) component of wind speed measured at 10 meters above the ground.
t2m	2-meter Temperature - The air temperature measured at 2 meters above the ground.
d2m	2-meter Dewpoint Temperature - The dewpoint temperature measured at 2 meters above the ground, which indicates moisture.
tcc	Total Cloud Cover - The fraction of the sky covered by clouds.
sp	Surface Pressure - The atmospheric pressure at the earth's surface.
tp	Total Precipitation - The accumulated precipitation (rain, snow, etc.) over a specified period.
ssrd	Surface Solar Radiation Downwards - The amount of solar radiation reaching the ground.
blh	Boundary Layer Height - The height of the lowest part of the atmosphere where the earth's surface significantly influences temperature, moisture, and wind.

APPENDIX - Learning Curve for Neural Networks



Learning Curve for neural network prediction for 20 features (after PCA)