Short Report 3

**Data:id 287 Wine\_Quality**

#### **Dataset Overview**:

Contains **two datasets**: red and white wine samples.

Focuses on the Portuguese "Vinho Verde" wine.

Inputs: Results from **physicochemical tests** (e.g., pH values).

Output: **Sensory data**, a quality score (0–10) evaluated by wine experts (median of at least three evaluations).

Tasks: Suitable for **regression** or **classification** approaches.

Classes are **ordered but imbalanced** (normal wines are more common than excellent or poor wines).

Total instances:

Red wine: 1,599 instances.

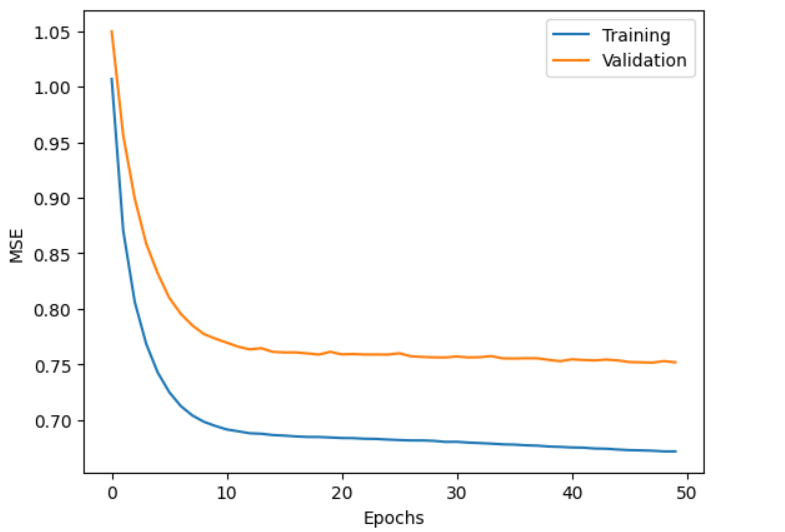
White wine: 4,898 instances (instances 1600–6497 in the combined dataset).

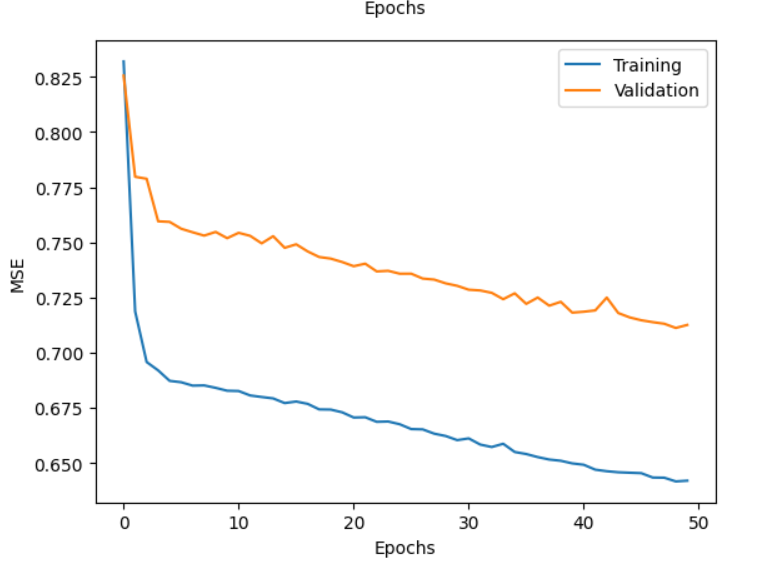
Attributes:

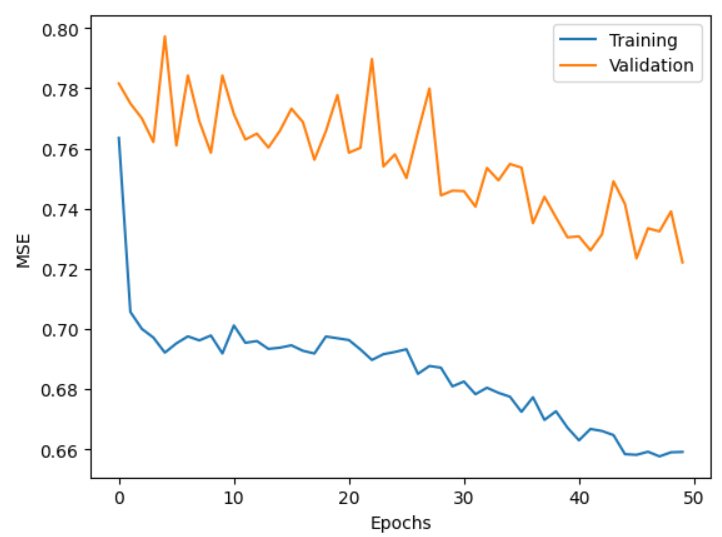
11 input attributes (physicochemical properties).

1 output attribute (quality score).

Missing Values: None.







Data:id 503 Wind

**Dataset Description**:

Contains **daily average wind speeds** from **1961 to 1978** at **12 synoptic meteorological stations** in the Republic of Ireland.

**Data Format**:

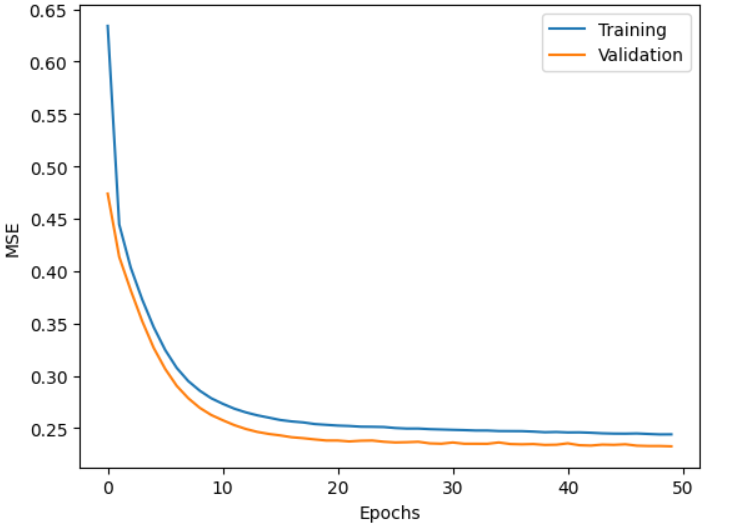
Columns: Year, Month, Day, and Average Wind Speed at 12 stations.

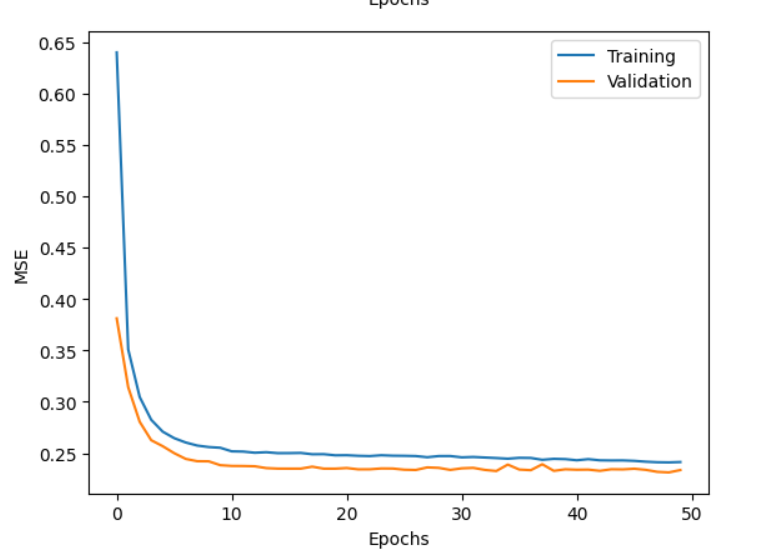
Stations (in order): RPT, VAL, ROS, KIL, SHA, BIR, DUB, CLA, MUL, CLO, BEL, MAL.

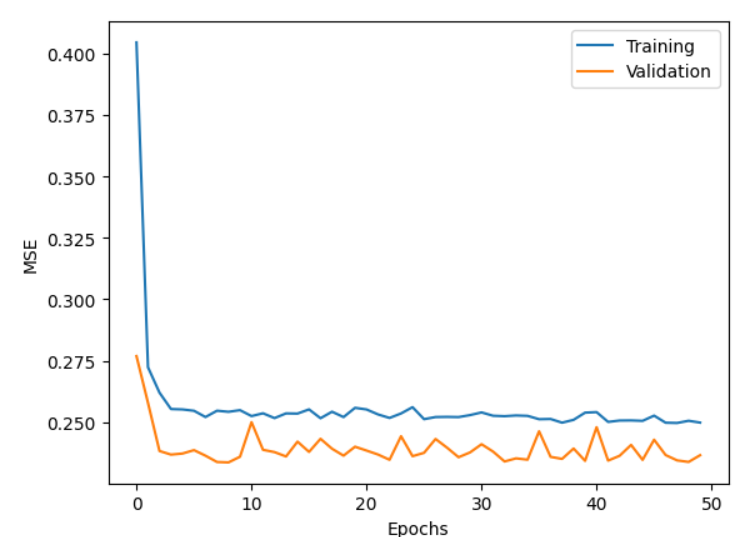
**Dataset Characteristics**:

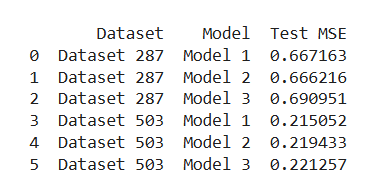
**CLASSTYPE**: Numeric.

**CLASSINDEX**: None specified.









Discussion

#### **Dataset: Wine Quality**

1. **Dataset Characteristics**:
   * Suitable for regression and classification tasks.
   * **Input Variables**: 11 physicochemical properties (e.g., pH, alcohol content).
   * **Output Variable**: Quality score (0–10, evaluated by experts).
   * Imbalanced classes: Normal quality wines significantly outnumber excellent or poor wines.
2. **Model Analysis**:
   * **Regression Models**: Analyze how physicochemical properties influence quality scores.
   * **Classification Models**: Classify wines into excellent, normal, or poor categories.
   * Support Vector Machines (SVM) performed best, but modern models (e.g., Random Forest, Gradient Boosting) may achieve better results.

**First Chart**: Both training and validation MSE decrease rapidly and stabilize, showing the best performance and strong generalization ability.

**Second Chart**: Training MSE continues to decline, but validation MSE decreases more slowly, with a noticeable gap, indicating slight overfitting.

**Third Chart**: Training MSE decreases significantly, but validation MSE fluctuates and remains high, suggesting severe overfitting.

#### **Dataset: Wind**

1. **Dataset Characteristics**:
   * Contains daily average wind speeds (1961–1978).
   * Data collected from 12 meteorological stations in Ireland.
   * Each row includes the year, month, day, and wind speeds from 12 stations.
2. **Model Analysis**:
   * **Time-Series Models**: Analyze seasonal and annual variations in wind speeds.
   * **Spatial Analysis Models**: Compare wind speed distributions across regions to assess wind energy potential.
   * Anomaly detection methods (e.g., removing extreme wind speeds) can improve model stability.

**First Chart**: The best-performing model with stable and close training and validation MSE, indicating strong generalization.

**Second Chart**: Stable performance with slower learning, suitable for simpler models.

**Third Chart**: Shows fluctuations in validation MSE, indicating potential overfitting; optimization is needed.