Laboratory Assignment

Component 1 + 2

Authors: Morar Cristina-Mirela & Matei Sonia
Group: HPC 242

November 16, 2024

Contents

The first requirement - Definition of the learning task(s)

1 Introduction

This document defines two machine learning tasks to analyze bike rental data: **Season Classification** and **Total Bike Rentals Prediction**. Each task is essential for understanding demand patterns and optimizing bike-sharing resources based on weather, time, and other contextual factors.

2 Task 1: Season Classification

2.1 Problem Definition

The goal of the *Season Classification* task is to predict the season (spring, summer, fall, or winter) in which a bike rental transaction occurred. The model relies on a range of weather and time-related features to determine the seasonal context.

2.2 Problem Specification

- **Input Data and Preconditions:**
  - Weather-related features: temperature, humidity, wind speed.
  - Time-based features: day of the week, month, and hour of the day.
  - Preconditions:
    - Comprehensive dataset covering all seasons.
    - Dataset includes labeled data indicating the actual season for each transaction.
- **Output and Postconditions:**
  - Output: Predicted season category (spring, summer, fall, or winter) for each bike rental instance.
  - Postconditions: Model should generalize well to unseen data, accurately predicting the season based on similar weather and time features.

2.3 Learning Task Specification

- **Task (T):** Season Classification - classify each instance of bike rentals into one of four seasonal categories.
- **Performance (P):** Evaluation metrics include accuracy, precision, recall, and F1 score, with consideration for model interpretability to identify feature influence on seasonal predictions.
- **Experience (E):** The model will train on historical bike rental data with labeled seasons, using supervised learning methods.

3 Task 2: Total Bike Rentals Prediction

3.1 Problem Definition

The aim of the *Total Bike Rentals Prediction* task is to predict the total number of bike rentals on a given day.

3.2 Problem Specification

- **Input Data and Preconditions:**
  - **Time-related features:** date, day of the week, season.
  - **Weather-related features:** temperature, humidity, general weather conditions (clear, rainy, cloudy).
  - **Other features:** whether the day is a holiday or a working day.
  - **Preconditions:**
    - Sufficient historical data with balanced representation across conditions.
- **Output and Postconditions:**
  - Output: Predicted number of bike rentals for a given date.
  - Postconditions: Model should generalize well to unseen data, producing reliable predictions across different conditions.

3.3 Learning Task Specification

- **Task (T):** Regression task - predict a continuous variable, representing the total number of bike rentals.
- **Performance (P):** Metrics include Mean Squared Error (MSE) and Mean Absolute Error (MAE) to evaluate prediction accuracy.
- **Experience (E):** The model will learn from historical data of past bike rentals, applying supervised learning techniques to predict rental count based on input features.

The second requirement - Data analysis

4 Bike Sharing Dataset Description

The **Bike Sharing Dataset** contains data collected from a bike-sharing system over the course of two years (2011 and 2012). Contains **731 instances**, representing daily data over those two years.

Attribute Information

1. **instant:** A unique identifier for each record.
2. **dteday:** The date of the record.
3. **season:** The season of the year (1 = Spring, 2 = Summer, 3 = Fall, 4 = Winter).
4. **yr:** The year (0 = 2011, 1 = 2012).
5. **mnth:** The month of the year (1 = January, ..., 12 = December).
6. **holiday:** Whether the day is a holiday (1 = Yes, 0 = No).
7. **weekday:** The day of the week (0 = Sunday, ..., 6 = Saturday).
8. **workingday:** Whether the day is a working day (1 = Yes, 0 = No).
9. **weathersit:** The weather condition:
   - 1 = Clear or partly cloudy.
   - 2 = Mist or cloudy.
   - 3 = Light rain or snow.
   - 4 = Heavy rain or snow.
10. **temp:** Normalized temperature (scaled from 0 to 1).
11. **atemp:** Normalized "feels like" temperature (scaled from 0 to 1).
12. **hum:** Normalized humidity (scaled from 0 to 1).
13. **windspeed:** Normalized windspeed (scaled from 0 to 1).
14. **casual:** Number of casual (non-registered) bike rentals.
15. **registered:** Number of registered user bike rentals.
16. **cnt:** Total number of bike rentals (sum of `casual` and `registered`).

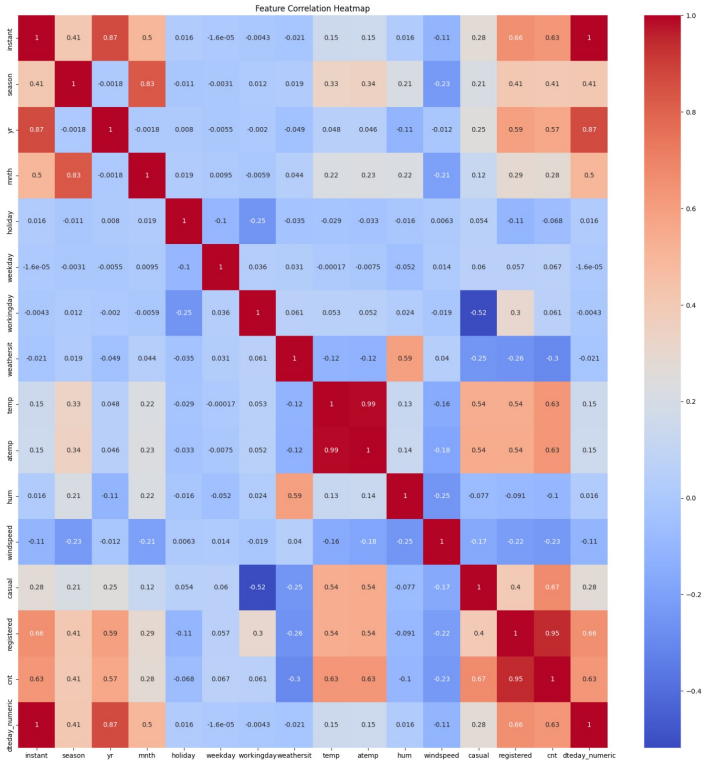
5 Analysis of the features used in learning

Description:

The **Count** for each field matches the total number of entries in the dataset, indicating that there are no missing values in any of the fields. The **Mean** values are close to the 50th percentile (Median), suggesting that the data is evenly distributed with minimal skewness and a limited number of outliers. The **Min** and **Max** values represent the minimum and maximum observations for each field, and they fall within valid ranges (e.g., for year, month, etc.). Additionally, the **Standard Deviation** is relatively small for all fields, indicating that the majority of the data points are close to the mean and do not exhibit significant variation.

Field	Count	Mean	Std Dev	Min	25%	50%	75%	Max
instant	731	366.00	211.17	1.00	183.50	366.00	548.50	731.00
season	731	2.50	1.11	1.00	2.00	3.00	3.00	4.00
yr	731	0.50	0.50	0.00	0.00	1.00	1.00	1.00
mnth	731	6.52	3.45	1.00	4.00	7.00	10.00	12.00
holiday	731	0.03	0.17	0.00	0.00	0.00	0.00	1.00
weekday	731	2.99	2.00	0.00	1.00	3.00	5.00	6.00
workingday	731	0.68	0.47	0.00	0.00	1.00	1.00	1.00
weathersit	731	1.40	0.54	1.00	1.00	1.00	2.00	3.00
temp	731	0.50	0.18	0.06	0.34	0.50	0.66	0.86
atemp	731	0.47	0.16	0.08	0.34	0.49	0.61	0.84
hum	731	0.63	0.14	0.00	0.52	0.63	0.73	0.97
windspeed	731	0.19	0.08	0.02	0.13	0.18	0.23	0.51
casual	731	848.18	686.62	2.00	315.50	713.00	1096.00	3410.00
registered	731	3656.17	1560.26	20.00	2497.00	3662.00	4776.50	6946.00
cnt	731	4504.35	1937.21	22.00	3152.00	4548.00	5956.00	8714.00
dteday_numeric	731	365.00	211.17	0.00	182.50	365.00	547.50	730.00

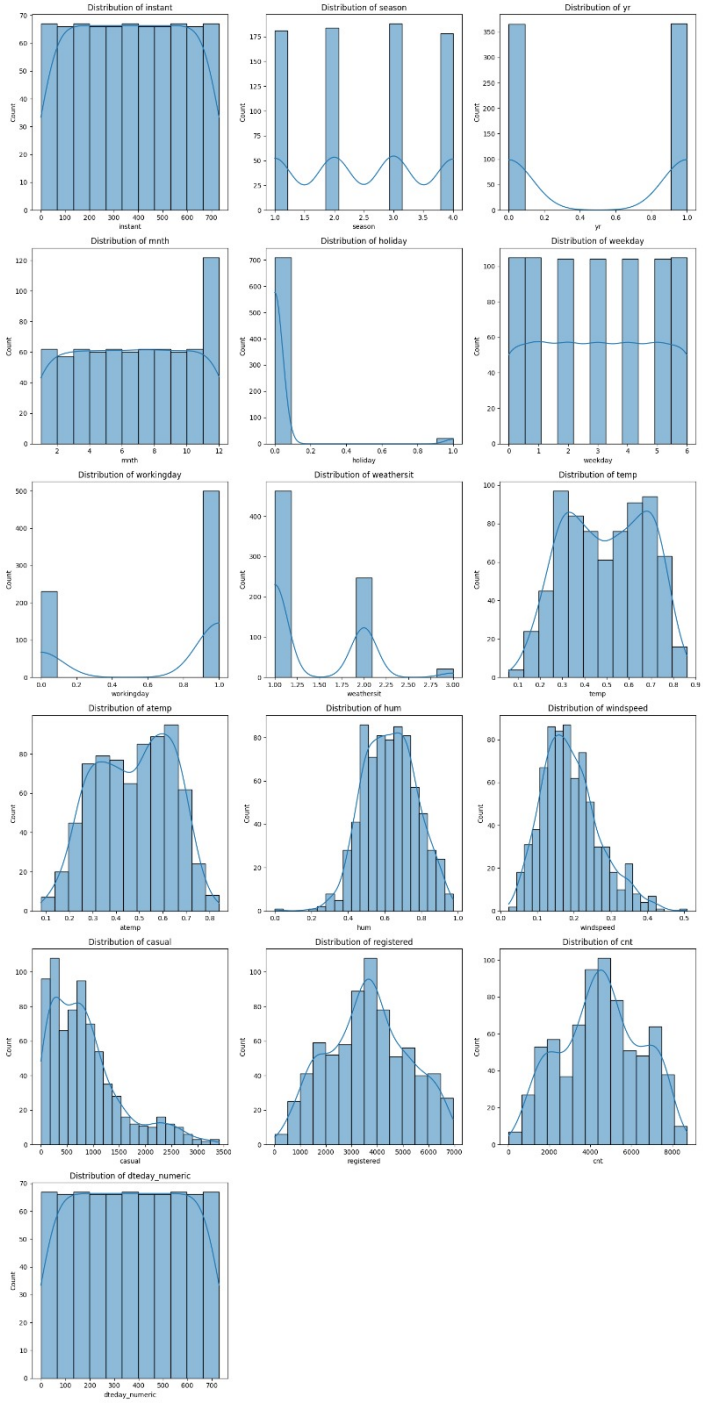
- **Correlation**



In the heatmap, we can observe the correlation of each field with all other fields. For example, the `cnt` field is highly correlated with the `registered` field and `temp` (temperature). The `temp` and `atemp` fields are also highly correlated, which is logical because `atemp` (how the temperature feels) depends on the `temp` field. Furthermore, we can say that the `atemp` field is redundant, as its information can be derived from the `temp` field.

Additionally, `windspeed` and `weathersit` negatively affect the `cnt` field (which is the target variable in one of the tasks). This indicates that fewer people rent bikes during high winds or poor weather conditions.

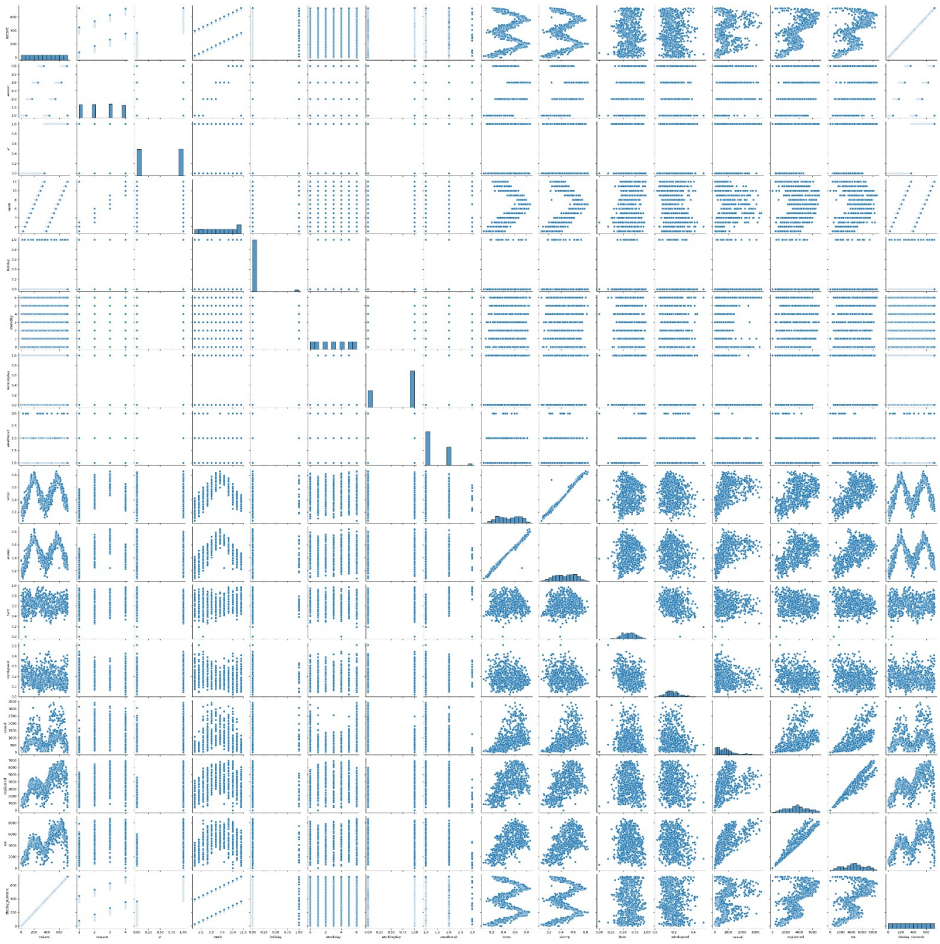
- **Data Distribution:**



The plot above illustrates the distribution of each feature in the dataset. Some key observations are as follows:

* `instant` and `dteday_numeric` are uniformly distributed.
* `season` exhibits four distinct peaks, indicating an equal number of registrations in each season. Similarly, `yr` shows a uniform distribution across the two years.
* The `holiday` feature reveals that most registrations occur on non-holiday days (value 0) compared to holidays (value 1).
* `hum` (humidity) and `temp` (temperature) follow normal distributions, suggesting a central mean for each variable.
* `windspeed` exhibits a right-skewed distribution, indicating that most rentals occur under low wind conditions.
* Categorical variables such as `holiday` and `weathersit` demonstrate imbalanced distributions (e.g., fewer holidays and adverse weather situations), which could influence the model's performance.
* Temporal features such as `season`, `mnth`, and `weekday` display reasonably uniform distributions, ensuring balanced data across time periods.

- **Data Visualization and Interpretation:**



The scatter plot matrix (pair plot) shown in the image visualizes the pairwise relationships between all numerical features in the dataset. Here are some key takeaways from this visualization:

The diagonal contains histograms of individual variables, providing insights into their distributions.

The scatter plots in the off-diagonal cells illustrate the relationships between pairs of variables:

- If the points form a clear linear trend, this indicates a strong correlation (positive or negative) between the two variables.
- If the points are scattered randomly, there is little to no correlation.

Features such as `temp` and `atemp` are likely to display a strong positive linear correlation, as they measure related concepts (temperature and perceived temperature).

Other features, such as `cnt` (total rentals), may exhibit relationships with environmental variables like `temp`, `atemp`, and `hum`, since rental activity is often influenced by weather conditions.

Additionally, some scatter plots reveal distinct clusters or cyclical patterns, suggesting periodic behavior (e.g., in time-based features such as `season` or `mnth`).

5.1 Task 1: Season Classification

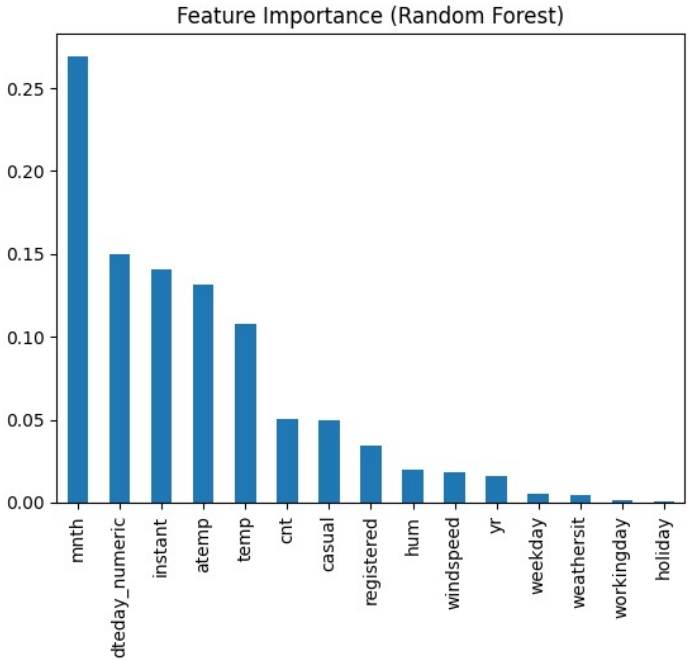
- **Independence:**

Feature	Mutual Information Score
instant	1.370269
dteday_numeric	1.370087
mnth	1.164685
atemp	0.591568
temp	0.561014
casual	0.253863
cnt	0.212872
registered	0.207865
windspeed	0.073409
hum	0.050348
yr	0.031917
workingday	0.024455
holiday	0.021844
weathersit	0.004399
weekday	0.000000

Table 1: Mutual Information Scores of Features

The mutual dependency values indicate the contribution of each feature to the target variable, which in this case is the `season` field. The features `instant`, `dteday_numeric`, and `mnth` have the highest mutual information scores, suggesting that time-based information is highly relevant for predicting the season. Moderately impactful features include temperature-related variables, such as `temp` and `atemp`. Features with low impact are those with the smallest mutual information scores, listed at the bottom of the table. In the context of a classification task, this means that the most impactful features should be prioritized for predicting the `season`.

- **Feature Importance:**



This bar chart illustrates the feature importance of various fields for predicting the target field `season` using a Random Forest classifier. The most significant feature is `mnth`, followed by `dteday_numeric` and `instant`, as well as other statistics mentioned above.

5.2 Task 1: Season Classification

The same data analysis applies to the characteristics of the data for task two, as described previously.

- **Independence:**

Feature	Mutual Information Score
registered	1.665601
instant	0.905276
dteday_numeric	0.903527
casual	0.669468
atemp	0.464677
temp	0.389529
mnth	0.375189
yr	0.278808
season	0.215862
weathersit	0.065937
windspeed	0.055770
hum	0.045847
weekday	0.044802
workingday	0.023912
holiday	0.011006

Table 2: Mutual Information Scores of Features

- **Feature Importance:**