Academic Year	Module	Assignment Number	Assignment Type
2	5CS037/HJ1: Concepts and Technologies of AI	Final Project	Individual Coursework

Classification Analysis Report

Student Id : 2408008

Student Name : Nayanika Dubey

Section : L5CG14

Module Leader : Mr. Siman Giri

Tutor : Ms. Sunita Parajuli

Submitted on : 11-02-2025 (February 11, 2025)

Table of Contents

Abstract

This project explored the use of machine learning classification models, including Logistic Regression, Ridge, SVC, and Random Forest, to predict air quality. The main goal was to create a reliable model that could correctly categorize air quality levels into 4 categories: good, moderate, poor, and hazardous, according to industrial and environmental characteristics. To enhance model performance, hyperparameter tuning techniques such as RandomizedSearchCV and GridSearchCV, along with feature selection methods like Recursive Feature Elimination (RFE) and SelectFromModel, were employed.

Among the four models built for this task, Random Forest model outperformed the others. The initial Random Forest model demonstrated exceptional performance, achieving an F1 score of 0.959. However, after selecting key features — including temperature, NO_2, SO_2, CO, and proximity to industrial areas — and fine-tuning hyperparameters, the optimized model's F1 score saw a slight reduction to 0.942.

This outcome suggests that the initial Random Forest configuration was already well-suited to the dataset. The study highlights the value of thoroughly assessing initial models, carefully selecting features, and remaining flexible to achieve reliable and meaningful results.

Introduction

Automated analytical model building for data analysis, focusing on identifying patterns from data, learning from it, and making decisions with minimal human intervention is Machine Learning.

In general, machine learning problems fall under one of the two:

- Supervised Learning
- Unsupervised Learning

Supervised Learning trains a computer algorithm based on input data that has been labeled for a particular task/ output. Model training continues until it can detect an underlying pattern or connection between the provided input data and the output labels, in a manner that allows it to provide accurate labelling results for unseen data. There are several ways to approach supervised learning, however in this section of the final assessment we will focus on Classification.

The process of identifying a function that separates a dataset into groups according to several parameters is known as classification. A dataset is used to train an algorithm, which then uses that training to classify the data into several groups.

Problem Context

The Problem Statement

The project's focus is to implement logistic regression with Python and Scikit-Learn. The dataset "Air Quality and Pollution Assessment" is used for this project, to build various models that can predict the target "Air Quality".

Dataset Overview and Connection of Dataset to SDG goal:

"Air Quality and Pollution Assessment" dataset holds over 5000 rows of unique data related to environmental factors and pollution metrics, and was obtained from Kaggle.

This dataset directly correlates to SDG 3 and SDG11: Good Health and Well Being and Sustainable Cities and Communities. The dataset directly helps understand the correlation between various metrics and how they affect the air quality, hence directly relating to these goals, leading us to better understand the impacts of the indicators and helping us reflect on better practices for the environment.

Analytical strategy

The primary analytical strategy is to build, train, and evaluate classification models to predict the 'air_quality' using a variety of explanatory variables. Data preparation, model construction and implementation (fitting), model evaluation, hyperparameter finetuning, feature selection, and developing the final model with the best features and optimum hyperparameters are all steps in this process.

Data Preprocessing

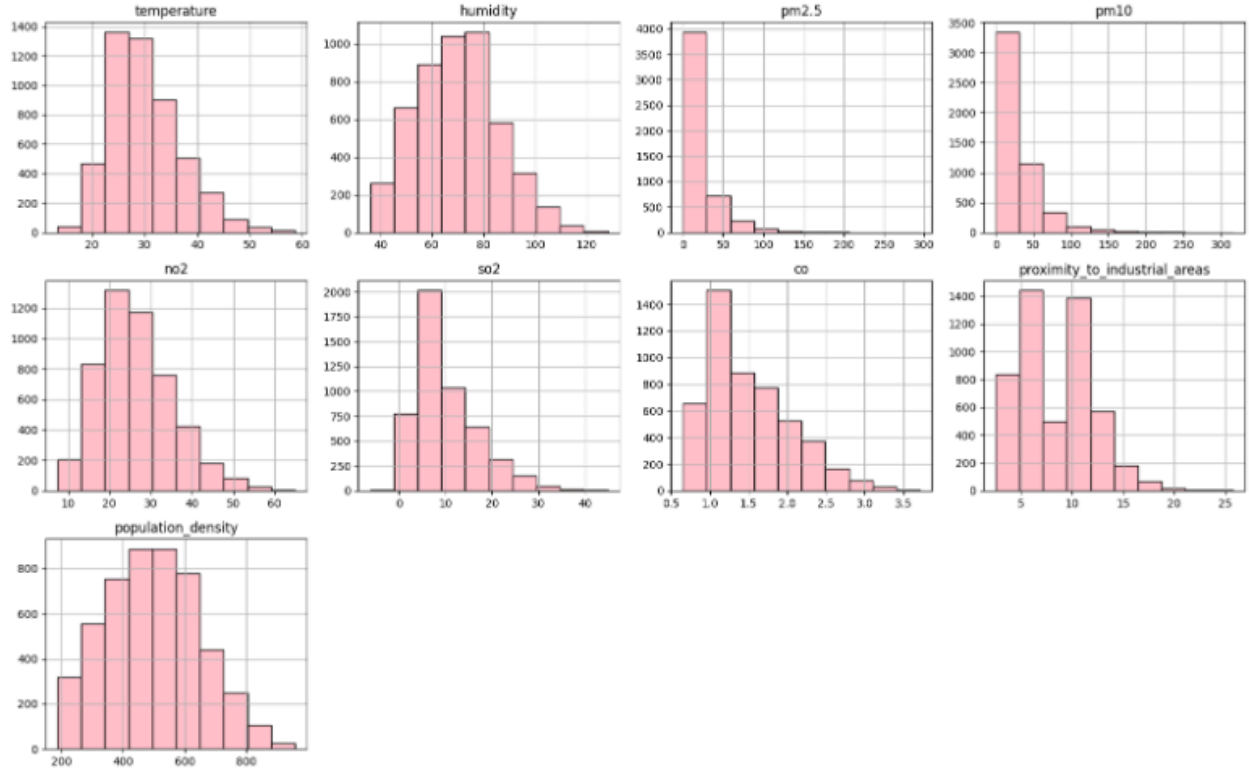
Several functions, including df.head(), df.describe(), and df.info(), are used to examine the dataset after it has been loaded. For improved implementation flow, the column names are standardized to lowercase, underscore separated values. Null values are handled by substituting the mean for numerical values and the mode for categorical ones once a fundamental comprehension of the dataset has been attained. In order to normalize feature data, the mean is subtracted and then divided by the standard deviation.

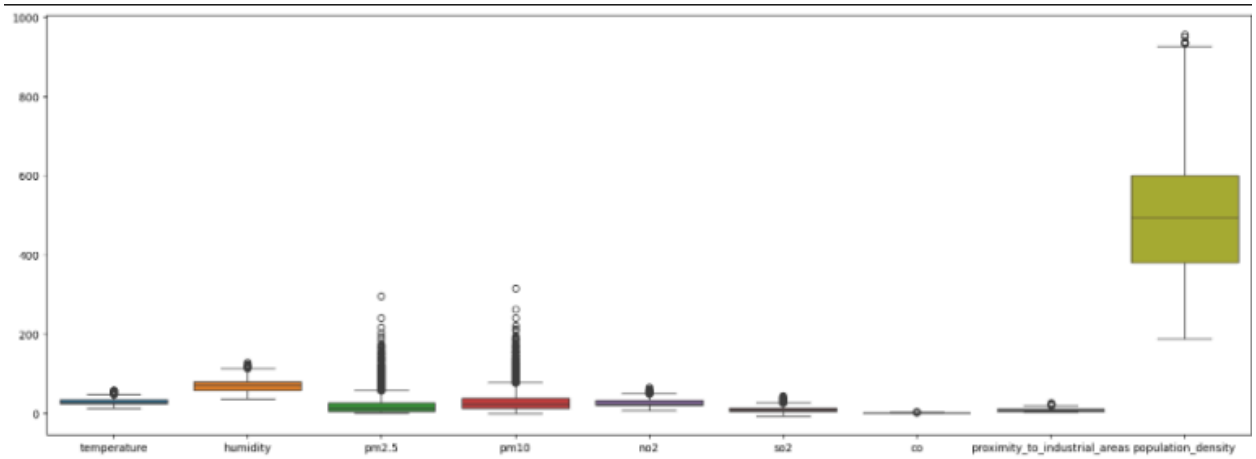
```
# Fill or drop missing values (example: using mean for numerical columns)
for column in df.select_dtypes(include=np.number).columns:
    if df[column].isnull().sum() > 0:
        df[column].fillna(df[column].mean(), inplace=True)

# Fill or drop missing categorical values (example: using mode)
for column in df.select_dtypes(include=['object']).columns:
    if df[column].isnull().sum() > 0:
        df[column].fillna(df[column].mode()[0], inplace=True)
```

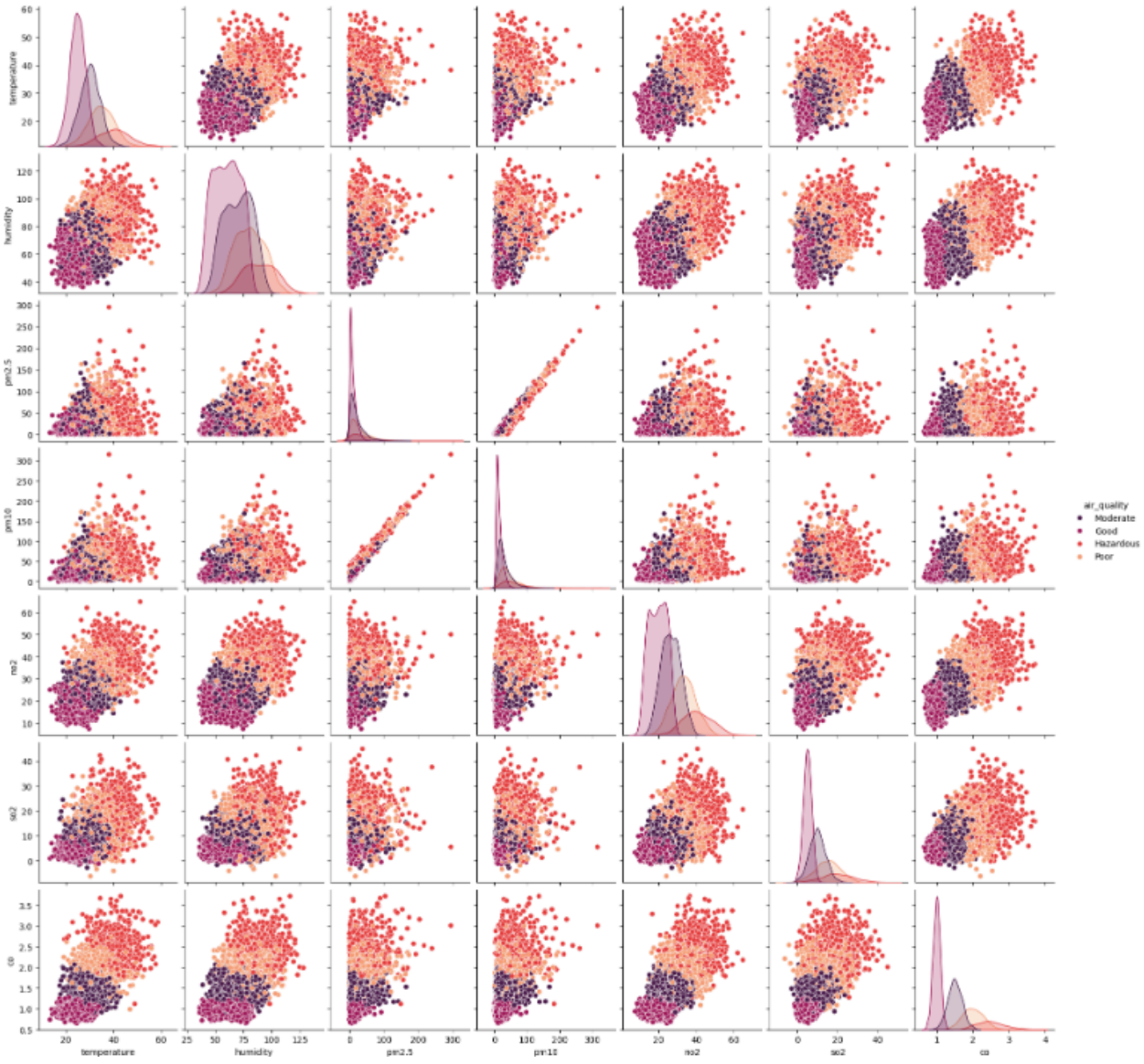
```
# Remove duplicates if any
df.drop_duplicates(inplace=True)

# Standardize column names (optional but helpful for consistency)
df.columns = df.columns.str.strip().str.lower().str.replace(' ', '_')

print("\nData after basic cleaning:")
print(df.head())
```

Visualization

Various methods such as pairplots, histograms, boxplots, and heatmaps are used to visualize data and explore relationships between variables. Here are some of the visualizations:






temperature
humidity
pm2.5
pm10
no2
so2
co
air_quality
Moderate
Good
Hazardous
Poor

Model Development

A total of 4 classification models are built in this project, namely a logistic regression model from scratch, a Ridge Classification Model, a Support Vector Classifier Model, and Random Forest Classification Model.

Model from scratch:

A logistic regression model was built using the softmax function to deal with the multivariate labels present in the target feature. The model was evaluated using accuracy, precision, recall and F1-score. Confusion matrices were also built to get better visualization of the results.

Pre-Built models:

Ridge Classification, Support Vector Classification (SVC) and Random Forest Classification were built and fitted using sklearn. These models were used to do detailed comparative analyses.

The goal of classification was to find a model that did an adequate job at predicting the air quality label, either good, moderate, poor or hazardous, for unseen data.

Evaluation Metric and Results

The models were evaluated based on their various scores such as

Accuracy:

TP + TN / TP + FN + TN + FP

Precision:

TP / (TP +FP)

Recall:

TP / (TP +FN)

F1 - Score:

2 * precision * recall / precision + recall

Where,

TP = True Positive
FP = False Positive
TN = True Negative
FN = False Negative

Model Metrics:

Below are the evaluation metrics of the different models:

	Accuracy	Precision	Recall	F1 Score
Logistic Regression (Scratch)	0.92	0.91	0.92	0.91
Ridge Classification	0.738	0.722	0.738	0.7126
Support Vector Classification	0.942	0.942	0.942	0.9418
Random Forest Classification	0.959	0.959	0.959	0.959

Based on the above observations, we can draw the conclusion that Random Forest Classification did the best job at the logistic regression task when compared to the other models, as it has a F1 score of 0.959.

Model Enhancement

Hyperparameter Tuning

Hyperparameters refer to the parameter used during machine learning that are chosen before the training process starts and isn't learned by the model, but affects its learning process and performance.

RandomizedSearchCV was used to tune hyperparameters for Ridge Classifier and Support Vector Classifier, to tune 'Alpha' for ridge and C, gamma and the kernel for SVC. For Random Forest, GridSearchCV was used to find the optimal hyperparameter value.

Feature Selection

The features considered for selection across all the models were: 'temperature', 'humidity', 'pm2.5', 'pm10', 'no2', 'so2', 'co', 'proximity_to_industrial_areas', and 'population_density'. Recursive Feature Elimination (RFE) was used to select the top 3 features for Ridge Classification and Support Vector Classification (SVC). SelectFromModel was used to pick features based on importance scores for Random Forest Classification, with median being the threshold.

Final Model Building

The final model was built with the initial performance comparison, hyperparameter tuning and feature selection in mind. A Random Forest Model was built with the following hyperparameters:

- **n_estimators:** 50 was chosen as the number of trees in the forest.
- **max_depth:** 15 was the best value for maximum depth of the trees.
- **min_samples_split:** 3 was the minimum number of samples required to split an internal node.
- **min_samples_leaf:** 2 was the minimum number of samples required to be at a leaf node.

The features selected for the model were **temperature, no2, so2, co, and proximity_to_industrial_areas.**

The updated model returned the evaluation metrics as follows:

- Accuracy: 0.942

- Recall: 0.942

- Precision: 0.942

- F1 Score: 0.942

Despite the hyperparameter tuning and feature selection, the F1 score for the final random forest was slightly lower than the initial model.

Conclusive remarks

In conclusion, this project aimed to predict air quality using various machine learning classification models and optimize their performance through hyperparameter tuning and feature selection. While these techniques are typically essential for enhancing model performance, the final optimized Random Forest Classifier did not outperform the original model. The initial Random Forest model demonstrated the strongest performance with the highest F1 score, whereas the feature selection and tuning unexpectedly reduced the final model's effectiveness. This study emphasizes how machine learning is iterative and that optimization alone is unlikely to result in improved outcomes. It emphasizes how crucial it is to assess several models and combinations in order to determine the best one.