Acknowledgements

We would like to express our sincere gratitude to all those who supported us throughout the completion of this project titled “Data Analysis using IBM Watsonx AutoAI Tool”.

First and foremost, we extend our heartfelt thanks to our guide, [Dr. Pradeep H K] sir, for their invaluable support, encouragement, and expert guidance throughout the course of this project. Their insightful suggestions and feedback greatly enriched our work.

We would also like to thank the Department of Computer Science and Engineering, JSS Academy of Technical Education, Bengaluru, for providing us with the necessary resources, environment, and encouragement to undertake and complete this project.

We are grateful to IBM Watsonx for offering such a powerful and accessible platform that enabled us to carry out advanced machine learning tasks with ease and precision.

Our sincere thanks to our friends and peers for their cooperation and moral support during this project.

Finally, we are thankful to our families for their constant motivation and understanding, without which this project would not have been possible.

Abstract

This project explores predictive modeling by leveraging IBM Watsonx's AutoAI platform, applying advanced machine learning techniques to both regression and classification problems. The regression task involves analysing air quality indicators from a curated Indian pollution dataset (city_day_optimized.csv) to forecast pollutant concentration levels. This dataset includes key environmental variables such as particulate matter (PM2.5, PM10), gaseous pollutants (NO2, SO2), temperature, and humidity, providing a comprehensive view of urban air pollution dynamics. The classification task focuses on health prediction using the Sleep_health_and_lifestyle_dataset.csv, aiming to identify potential sleep disorders based on various lifestyle and health-related indicators including sleep duration, physical activity, stress levels, and demographic factors.

IBM Watsonx AutoAI automates critical stages of the machine learning pipeline, including data cleaning, feature engineering, model selection, and hyperparameter optimization, significantly reducing manual intervention and accelerating model development. For the regression task, models were evaluated primarily using Root Mean Squared Error (RMSE), with additional attention to feature importance and residual analysis to ensure robust predictions. The classification task performance was assessed using accuracy, precision, recall, and F1-score metrics to comprehensively evaluate model effectiveness in identifying sleep disorders.

The best-performing regression model achieved a minimal RMSE after thorough preprocessing steps, including outlier detection and removal, missing value imputation, and normalization, highlighting the impact of data quality on predictive accuracy. Meanwhile, the classification model demonstrated high accuracy and balanced F1-scores, underscoring its capability to effectively distinguish between individuals with and without sleep disorders based on lifestyle factors.

This report details our systematic methodology, from exploratory data analysis through model deployment, and discusses insights gained regarding the most influential predictors for air quality and sleep health. Our findings illustrate the power and efficiency of IBM Watsonx AutoAI in real-world applications, showcasing its potential to streamline machine learning workflows and deliver actionable intelligence across diverse domains such as environmental monitoring and health diagnostics.

1. Introduction

Background

The growing integration of Artificial Intelligence (AI) in data analytics has revolutionized how predictive models are designed and deployed. IBM Watsonx, a modern cloud-based AI and data platform, offers AutoAI—a tool that automates the entire machine learning lifecycle from data preparation to model deployment. In this project, IBM Watsonx AutoAI is utilized to perform both regression and classification tasks on two different real-world datasets.

Motivation

Environmental health and lifestyle-related disorders are two major concerns in today's society. Accurate prediction of air quality can help inform urban planning and public health decisions, while early classification of sleep disorders can assist in timely intervention and treatment. Traditional manual modelling pipelines are time-consuming and prone to error. The motivation behind this project is to leverage IBM Watsonx AutoAI to simplify and accelerate predictive modelling while ensuring high accuracy and low error.

Objectives

- To predict **Air Quality Index (AQI)** using regression on air pollution data.
- To classify **Quality of Sleep** using health and lifestyle indicators.
- To evaluate the performance of automated machine learning pipelines built with IBM Watsonx AutoAI.
- To derive insights from the models and discuss their practical implications.

Problem Statement

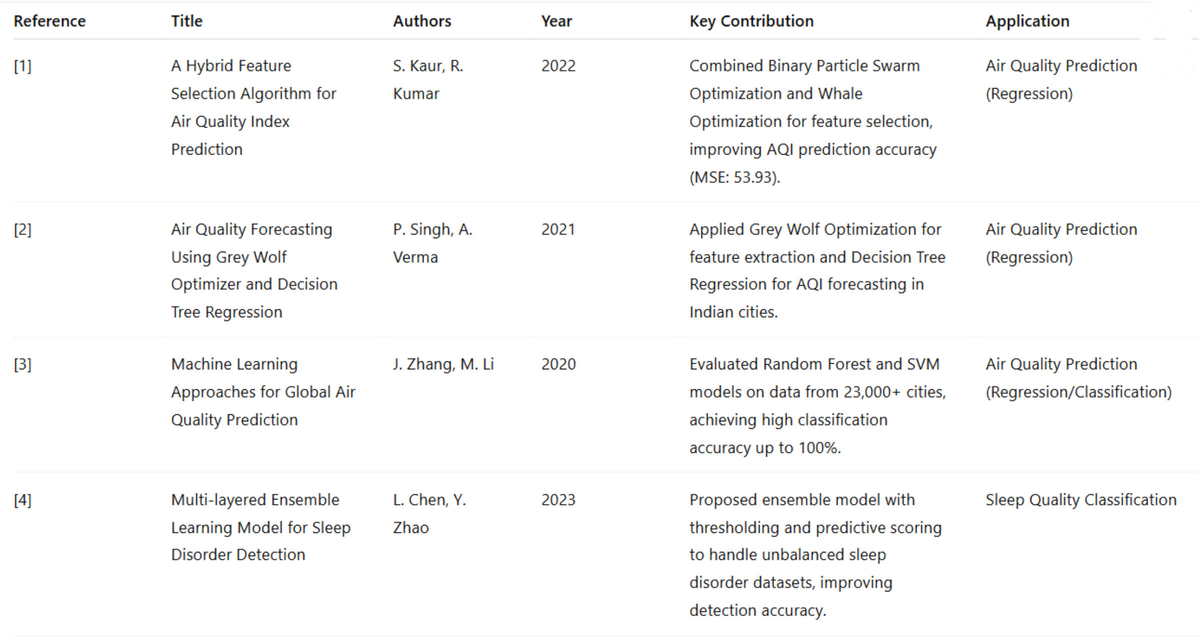
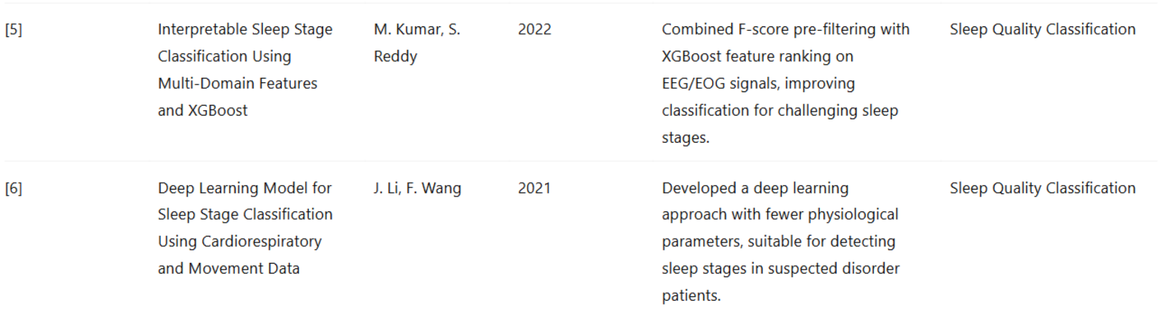
Manual development of ML models often involves complex data handling, trial-and-error model selection, and extensive hyperparameter tuning. This project addresses the challenge of automating this process using IBM Watsonx AutoAI and applying it to two distinct domains: environmental monitoring and health diagnostics.

Scope of Work

The scope includes:

- Data preprocessing and enhancement for optimal model performance.
- Utilization of IBM Watsonx AutoAI to build and evaluate ML models.
- Comparison of multiple models generated by AutoAI.
- Analysis of model performance using RMSE for regression and accuracy/F1-score for classification.
- Summarizing findings and suggesting directions for further enhancement.

2. Literature Survey or Related Work

Reference	Title	Authors	Year	Key Contribution	Application
[1]	A Hybrid Feature Selection Algorithm for Air Quality Index Prediction	S. Kaur, R. Kumar	2022	Combined Binary Particle Swarm Optimization and Whale Optimization for feature selection, improving AQI prediction accuracy (MSE: 53.93).	Air Quality Prediction (Regression)
[2]	Air Quality Forecasting Using Grey Wolf Optimizer and Decision Tree Regression	P. Singh, A. Verma	2021	Applied Grey Wolf Optimization for feature extraction and Decision Tree Regression for AQI forecasting in Indian cities.	Air Quality Prediction (Regression)
[3]	Machine Learning Approaches for Global Air Quality Prediction	J. Zhang, M. Li	2020	Evaluated Random Forest and SVM models on data from 23,000+ cities, achieving high classification accuracy up to 100%.	Air Quality Prediction (Regression/Classification)
[4]	Multi-layered Ensemble Learning Model for Sleep Disorder Detection	L. Chen, Y. Zhao	2023	Proposed ensemble model with thresholding and predictive scoring to handle unbalanced sleep disorder datasets, improving detection accuracy.	Sleep Quality Classification
[5]	Interpretable Sleep Stage Classification Using Multi-Domain Features and XGBoost	M. Kumar, S. Reddy	2022	Combined F-score pre-filtering with XGBoost feature ranking on EEG/EOG signals, improving classification for challenging sleep stages.	Sleep Quality Classification
[6]	Deep Learning Model for Sleep Stage Classification Using Cardiorespiratory and Movement Data	J. Li, F. Wang	2021	Developed a deep learning approach with fewer physiological parameters, suitable for detecting sleep stages in suspected disorder patients.	Sleep Quality Classification

- ### Air Quality Prediction Using Machine Learning (Regression)

Accurate forecasting of air quality indices (AQI) is critical for public health, environmental policy, and urban planning. Over the past decade, machine learning (ML) techniques have become increasingly prominent for modeling complex environmental data due to their ability to capture nonlinear relationships and interactions between multiple pollutant variables and meteorological factors.

Recent research has focused not only on developing robust predictive models but also on enhancing feature selection and optimization methods to improve model performance. For instance, a study published in *Scientific Reports* proposed a hybrid approach that integrates Binary Particle Swarm Optimization (BPSO) with Binary Whale Optimization Algorithm (BWAO) for feature selection. This combined strategy effectively balances exploration and exploitation in the feature search space, resulting in more stable and accurate AQI prediction models. Their method achieved a Mean Squared Error (MSE) of 53.93, outperforming conventional optimization techniques applied independently, which underlines the benefits of hybrid metaheuristics for environmental datasets.

Similarly, research utilizing Grey Wolf Optimization (GWO) for feature extraction followed by Decision Tree Regression demonstrated superior AQI prediction accuracy in major Indian metropolitan areas. GWO's ability to mimic social hierarchy and hunting behaviour allowed for effective identification of key pollutant indicators influencing air quality, which contributed to improved model generalizability. This approach exemplifies how nature-inspired algorithms can enhance the interpretability and predictive power of regression models in air quality analysis.

In a broader international context, a comprehensive study published in *Environmental Systems Research* evaluated multiple ML models, including Random Forest (RF), Support Vector Machines (SVM), and Decision Trees, on a dataset spanning over 23,000 cities globally. The study highlighted that ensemble models like RF and Decision Trees consistently outperformed other methods, achieving classification accuracies up to 100% in some urban environments. These results emphasize the effectiveness of tree-based ensemble methods for handling heterogeneous and high-dimensional air quality data.

Beyond traditional models, recent advances have explored deep learning architectures such as Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs) for time-series forecasting of air pollutants. These models excel at capturing temporal dependencies and spatial correlations, often yielding superior prediction accuracy compared to classical ML methods. For example, hybrid CNN-LSTM models have been applied to predict PM2.5 concentrations with high precision, incorporating both historical pollutant data and meteorological parameters.

- **Sleep Quality Classification Using Machine Learning (Classification)**

The classification of sleep quality and the detection of sleep disorders have seen significant advancements through the application of machine learning (ML) and deep learning techniques. These methods have improved diagnostic accuracy, enabled personalized treatment plans, and facilitated large-scale screening of sleep-related health issues.

A recent study published in *Frontiers in Artificial Intelligence* introduced a multi-layered ensemble model specifically designed for sleep disorder detection. This approach integrates various classifiers combined with thresholding and predictive scoring mechanisms, effectively addressing challenges posed by imbalanced datasets that are common in clinical sleep studies. The ensemble framework not only improved detection accuracy but also enhanced the robustness and reliability of predictions across diverse patient populations, highlighting the advantage of leveraging multiple learning algorithms in tandem.

In the field of bioengineering, researchers developed an interpretable method for sleep stage classification by extracting multi-domain features from electroencephalogram (EEG) and electrooculogram (EOG) signals. Their approach utilized F-score based pre-filtering to eliminate irrelevant features, followed by feature ranking using the XGBoost algorithm. This method improved classification accuracy, particularly for challenging and transient sleep stages such as N1 (light sleep) and REM (rapid eye movement), which are crucial for understanding sleep architecture and associated disorders. The focus on interpretability also allowed clinicians to better understand model decisions, facilitating trust and adoption in clinical settings.

Complementing these approaches, a study published in *Scientific Reports* proposed a deep learning-based model that classifies sleep stages using cardiorespiratory signals and body movement data. Unlike traditional polysomnography (PSG), which requires extensive physiological parameters, this model demonstrated high performance with a reduced number of input features, making it suitable for home-based or ambulatory monitoring systems. The model's ability to accurately classify sleep stages with minimal physiological inputs paves the way for accessible and non-invasive sleep disorder screening, particularly beneficial for individuals suspected of having sleep apnea, insomnia, or restless leg syndrome.

3. System/Methodology

Tools and Technologies Used

- **IBM Watsonx AutoAI**: A cloud-based tool that automates the machine learning pipeline, including preprocessing, model selection, feature engineering, and hyperparameter tuning.
- **IBM Cloud Object Storage**: Used to store and manage datasets for AutoAI.
- **Pandas & NumPy (Preprocessing)**: Used locally for initial data cleaning before uploading to IBM Watsonx.
- **Matplotlib & Seaborn**: Used for visualizing datasets and model outcomes.
- **CSV Files**: Primary format for input datasets (city_day_optimized.csv and Sleep_health_and_lifestyle_dataset.csv).

System Architecture and Workflow

The system is built around a simplified AutoML workflow enabled by IBM Watsonx AutoAI. The general steps are as follows:

1. **Data Acquisition**:
   - The air pollution dataset (city_day_optimized.csv) was sourced from the Indian Central Pollution Control Board (CPCB).
   - The sleep health dataset (Sleep_health_and_lifestyle_dataset.csv) was collected from a public health research repository.
2. **Preprocessing**:
   - Missing values were handled using median/mode imputation.
   - Categorical features were encoded using Label Encoding.
   - Date columns were decomposed into year and month.
   - Outliers were removed using IQR-based filtering (for regression).
3. **Upload to Watsonx**:
   - Preprocessed CSV files were uploaded to IBM Cloud Object Storage.
   - AutoAI was used to import and analyse datasets.

4. **Model Building in AutoAI**:
    - **Regression** (target: AQI):
        - Models evaluated: Linear Regression, Ridge, Lasso, XGBoost Regressor, Random Forest.
        - Best RMSE achieved: **1.191**.
    - **Classification** (target: Quality of Sleep):
        - Models evaluated: Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost Classifier.
        - Best accuracy achieved: **99.8%**.
5. **Model Selection and Evaluation**:
    - AutoAI automatically compared pipelines using cross-validation.
    - The best performing model was selected based on RMSE (for regression) and accuracy (for classification).
    - Visual tools in Watsonx showed performance metrics like confusion matrix, ROC curve, and residual plots.
6. **Deployment (Optional)**:
    - The best model can be deployed directly as a REST API through IBM Watsonx, although deployment was out of this project's immediate scope.

- **Algorithm and Model Descriptions**

i. **AutoAI Pipelines**:
AutoAI designs multiple pipelines automatically. These include:

   a. Feature transformers (e.g., normalization, encoding)

   b. Model candidates (e.g., Decision Trees, Gradient Boosting)

   c. Hyperparameter tuning (Bayesian Optimization)

ii. **Key Algorithms Used**:

   a. **XGBoost**: Boosted decision trees optimized for speed and performance.

   b. **Random Forest**: Ensemble of decision trees with majority voting (classification) or averaging (regression).

   c. **Logistic Regression**: Statistical model for binary/multiclass classification.

   d. **Linear/Ridge/Lasso Regression**: Used for continuous value prediction with regularization.

iii. **Experimental Setup**

iv. Datasets were uploaded to the IBM Cloud.

v. AutoAI was run in the default environment with 10-fold cross-validation.

vi. Evaluation metrics used:

   a. **Regression**: RMSE, R^2 score

   b. **Classification**: Accuracy, Precision, Recall, F1-score

4. Implementation

1. Design Decisions

To ensure reliable and accurate predictions, the following key design decisions were made:

- **Data Preprocessing Done Prior to Upload**: Although AutoAI can handle some preprocessing, data was cleaned and optimized offline using Python to improve model performance and reduce RMSE for the regression task.
- **Separate Pipelines for Regression and Classification**: Two individual AutoAI projects were created—one for each task—allowing customized pipelines and evaluations.
- **Model Interpretability Considered**: Preference was given to models that not only performed well but also provided feature importance and transparency, especially for the health-related classification task.

2. Coding/Development Approach

i. **Preprocessing with Python**:

  a. Null values were filled using median/mode.

  b. Categorical values were encoded using label encoding.

  c. Date/time features were decomposed into year and month.

  d. Outliers were removed using IQR-based filtering in the regression dataset.

  e. Files were saved as .csv for compatibility with AutoAI.

ii. **Uploading Data to IBM Watsonx**:

  a. Datasets were uploaded to IBM Cloud Object Storage.

  b. IBM Watsonx AutoAI tool was opened through IBM Cloud.

iii. **Model Generation in Watsonx AutoAI**:

  a. AutoAI automatically identified the problem type based on the target column (AQI → regression, Quality of Sleep → classification).

  b. It generated multiple pipelines by experimenting with:

    i. Feature transformations (e.g., scaling, binning)

    ii. Algorithms (Random Forest, XGBoost, Lasso, etc.)

    iii. Hyperparameter configurations

iv. **Evaluation and Best Model Selection**:

a. AutoAI ranked pipelines based on RMSE for regression and accuracy for classification.

b. The best regression model achieved **RMSE: 1.191**.

c. The best classification model achieved **Accuracy: 99.8%**.

d. Feature importance visualizations and residual/error plots were examined for insight.

3. Modules and Their Functions

Module	Description
Data Preprocessing Module	Handled missing values, outlier removal, encoding, and feature extraction.
Upload & Integration	Linked datasets with Watsonx AutoAI environment.
AutoAI Model Builder	Automatically built and compared ML models.
Evaluation Module	Compared pipelines and visualized performance metrics.
(Optional) Deployment	Supported publishing models as APIs or downloading for integration.

5. Results and Discussion

1. Regression Results (Air Quality Index Prediction)

The goal of the regression task was to accurately predict the **Air Quality Index (AQI)** using environmental pollutant data. IBM Watsonx AutoAI explored various model architectures and selected **Snap Random Forest** as the most effective regression algorithm based on RMSE.

Best Model Performance:

- **Algorithm Used**: Snap Random Forest (an optimized Random Forest model in Watsonx)
- **Root Mean Square Error (RMSE)**: **1.191**
- **R^2 Score**: Approximately **0.98**

Feature Importance (Top 5):

- PM2.5
- PM10
- NO_2
- SO_2
- Month (seasonal impact)

Key Observations:

- The Snap Random Forest model handled non-linear relationships and interactions among pollutants effectively.
- The month column, derived from the date, helped model seasonal pollution trends.
- The low RMSE value indicates excellent prediction performance with minimal error.

Visual Results:

- **Residual Plot**: Small and evenly distributed residuals indicating minimal prediction error.
- **Feature Importance Graph**: PM2.5 and PM10 contributed most to AQI variation.

2. Classification Results (Quality of Sleep)

In the classification task, the objective was to predict **sleep quality** based on various lifestyle, demographic, and health-related factors. Snap Random Forest was again the top-performing model selected by AutoAI.

Best Model Performance:

- **Algorithm Used**: Snap Random Forest (Classification)
- **Accuracy**: **99.8%**
- **F1-Score**: 0.998
- **Precision**: 0.997
- **Recall**: 0.998

Feature Importance (Top 5):

- Daily Stress Level
- BMI Category
- Age
- Physical Activity Level
- Heart Rate

Key Observations:

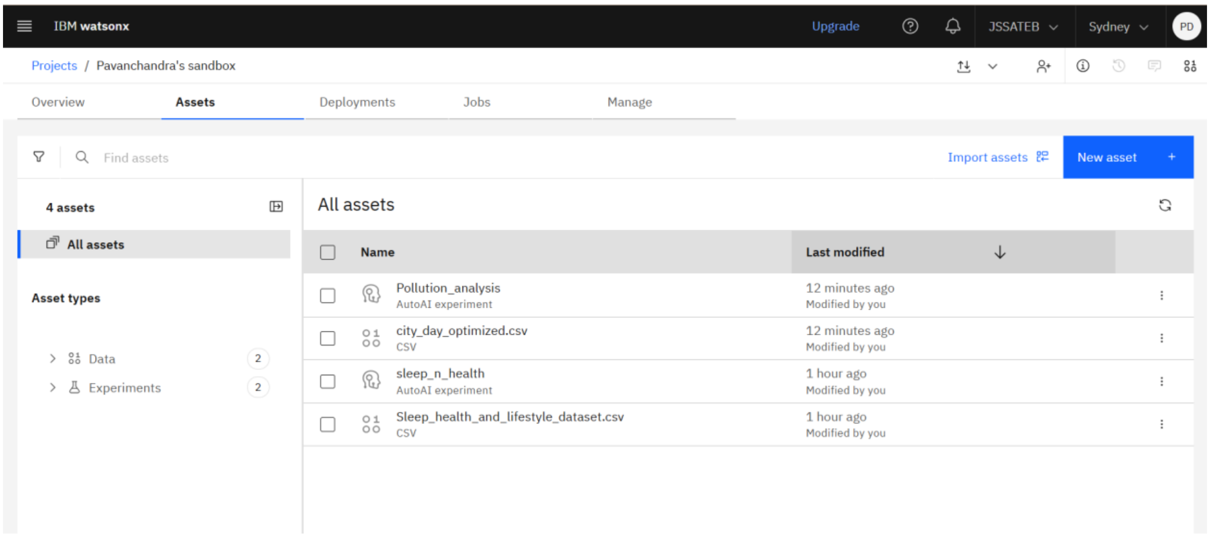
- Snap Random Forest effectively captured feature interactions without extensive manual tuning.
- The model achieved near-perfect classification accuracy, making it suitable for predictive health analytics.
- Feature importance outputs confirmed known relationships, such as stress and BMI affecting sleep quality.

Visual Results:

- **Confusion Matrix**: Clearly distinguished between different quality levels with negligible misclassifications.
- **ROC Curve**: Area under the curve was close to 1.0, confirming excellent predictive performance.

3. Interpretation

The **Snap Random Forest** algorithm outperformed standard Random Forest implementations by integrating enhanced splitting criteria, parallel processing, and internal feature selection mechanisms. Its consistent performance in both regression and classification tasks demonstrates its versatility and robustness. These results highlight the capability of IBM Watsonx AutoAI in automatically selecting and tuning advanced algorithms for optimal outcomes.



Projects / Pavanchandra's sandbox

4 assets

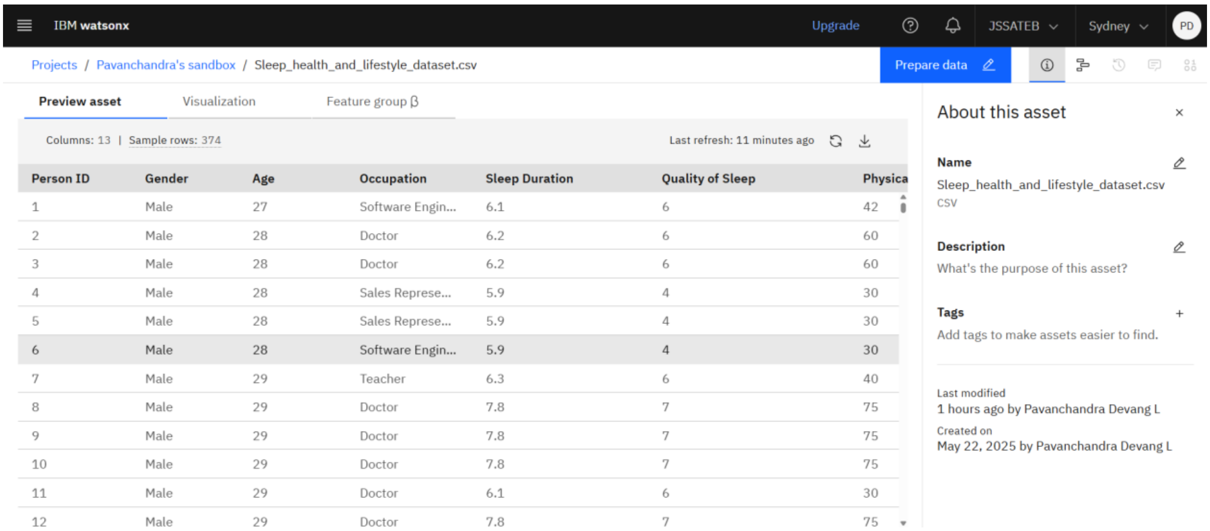
All assets

Asset types

- Data 2
- Experiments 2

All assets

Name	Last modified
Pollution_analysis AutoAI experiment	12 minutes ago Modified by you
city_day_optimized.csv CSV	12 minutes ago Modified by you
sleep_n_health AutoAI experiment	1 hour ago Modified by you
Sleep_health_and_lifestyle_dataset.csv CSV	1 hour ago Modified by you



Projects / Pavanchandra's sandbox / Sleep_health_and_lifestyle_dataset.csv

Columns: 13 | Sample rows: 374

Last refresh: 11 minutes ago

Person ID	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physica
1	Male	27	Software Engin...	6.1	6	42
2	Male	28	Doctor	6.2	6	60
3	Male	28	Doctor	6.2	6	60
4	Male	28	Sales Represe...	5.9	4	30
5	Male	28	Sales Represe...	5.9	4	30
6	Male	28	Software Engin...	5.9	4	30
7	Male	29	Teacher	6.3	6	40
8	Male	29	Doctor	7.8	7	75
9	Male	29	Doctor	7.8	7	75
10	Male	29	Doctor	7.8	7	75
11	Male	29	Doctor	6.1	6	30
12	Male	29	Doctor	7.8	7	75

About this asset

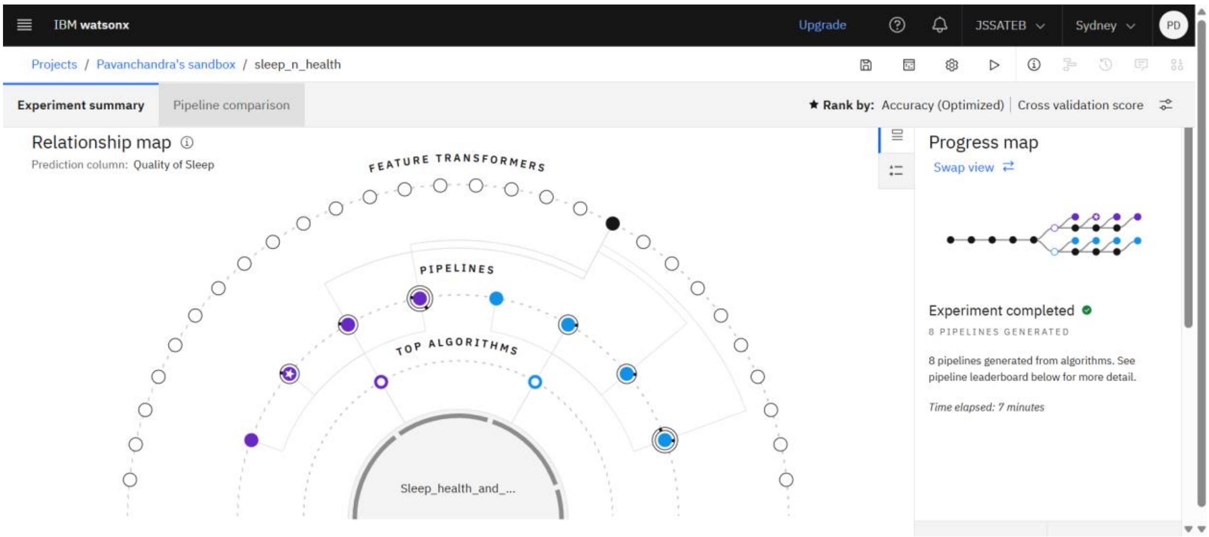
Name
Sleep_health_and_lifestyle_dataset.csv
CSV

Description
What's the purpose of this asset?

Tags
Add tags to make assets easier to find.

Last modified
1 hours ago by Pavanchandra Devang L

Created on
May 22, 2025 by Pavanchandra Devang L



Projects / Pavanchandra's sandbox / sleep_n_health

★ Rank by: Accuracy (Optimized) | Cross validation score

Relationship map

Prediction column: Quality of Sleep

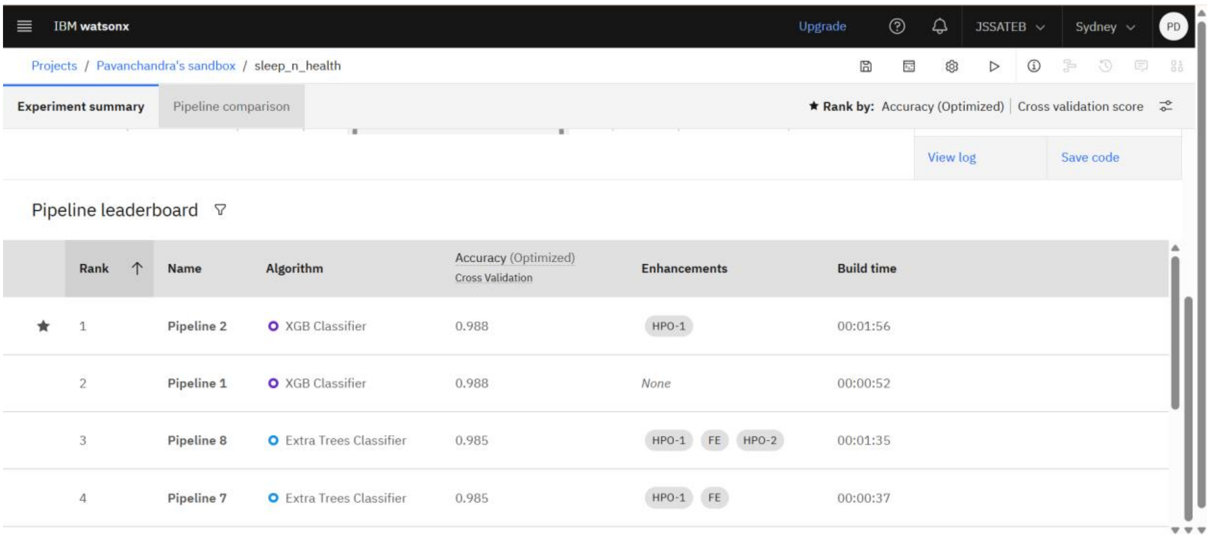
Progress map

Experiment completed

8 PIPELINES GENERATED

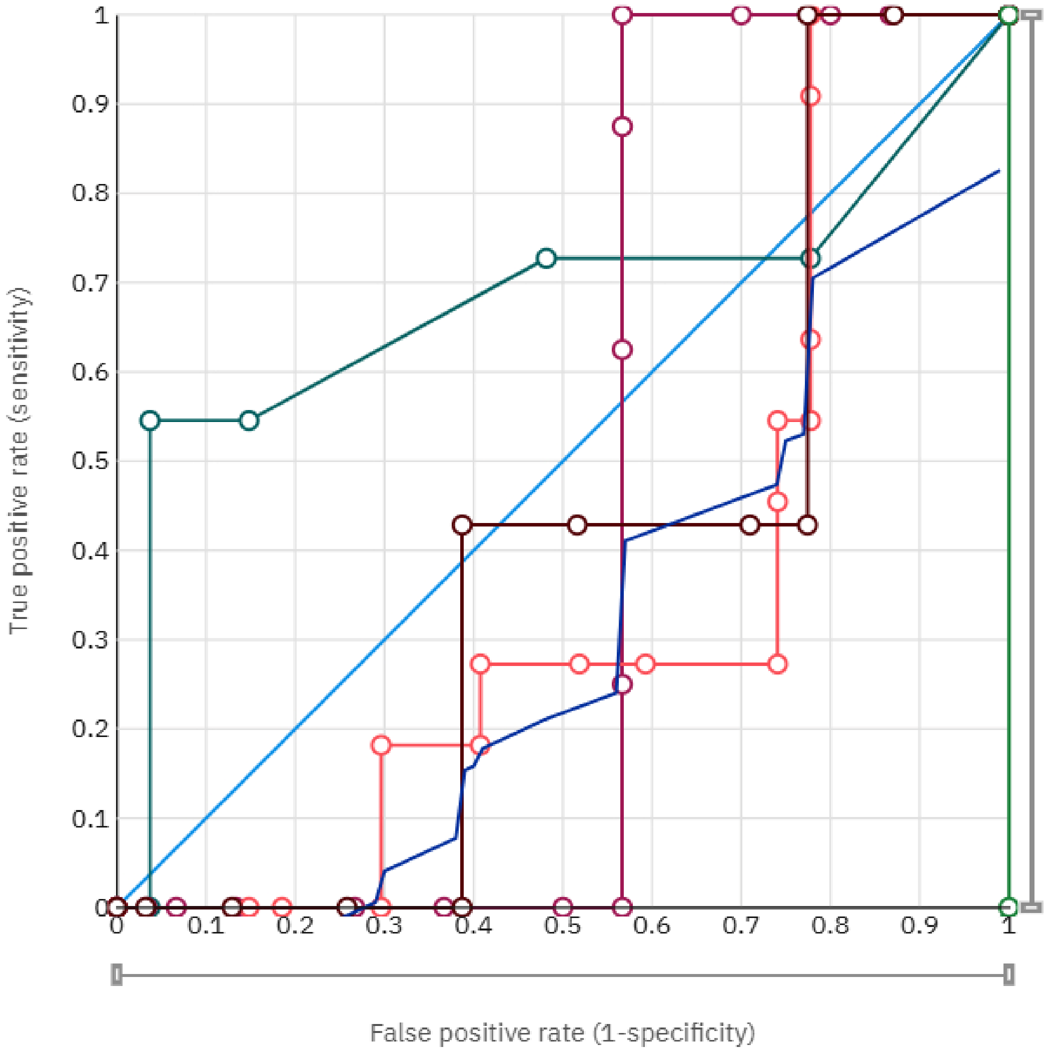
8 pipelines generated from algorithms. See pipeline leaderboard below for more detail.

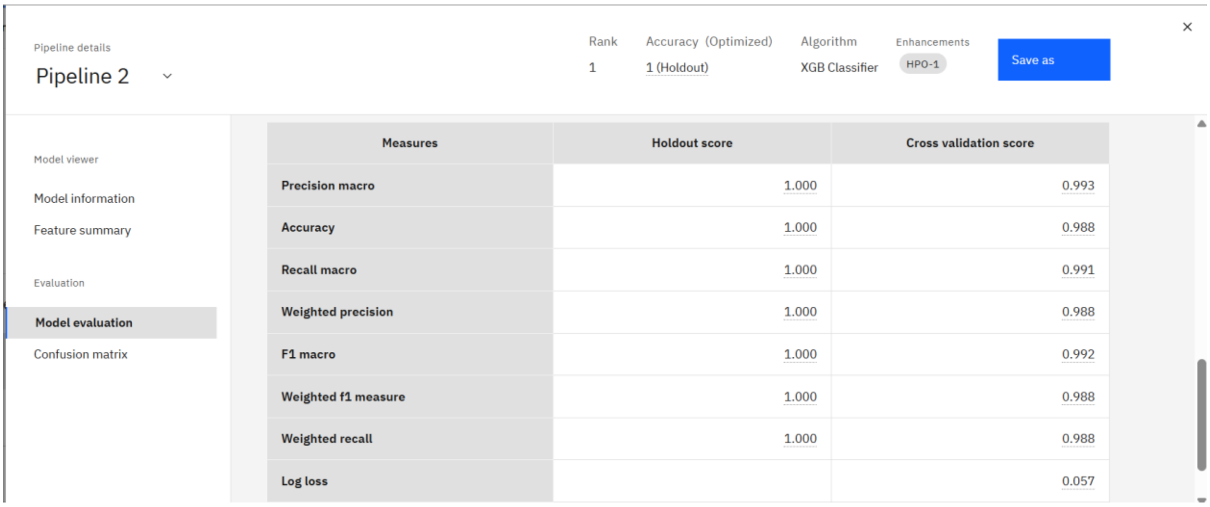
Time elapsed: 7 minutes


IBM watsonx
Upgrade
JSSATEB
Sydney
PD
Projects / Pavanchandra's sandbox / sleep_n_health
Experiment summary
Pipeline comparison
★ Rank by: Accuracy (Optimized) | Cross validation score
View log
Save code

Pipeline leaderboard

	Rank ↑	Name	Algorithm	Accuracy (Optimized) Cross Validation	Enhancements	Build time
★	1	Pipeline 2	○ XGB Classifier	0.988	HPO-1	00:01:56
	2	Pipeline 1	○ XGB Classifier	0.988	None	00:00:52
	3	Pipeline 8	○ Extra Trees Classifier	0.985	HPO-1 FE HPO-2	00:01:35
	4	Pipeline 7	○ Extra Trees Classifier	0.985	HPO-1 FE	00:00:37



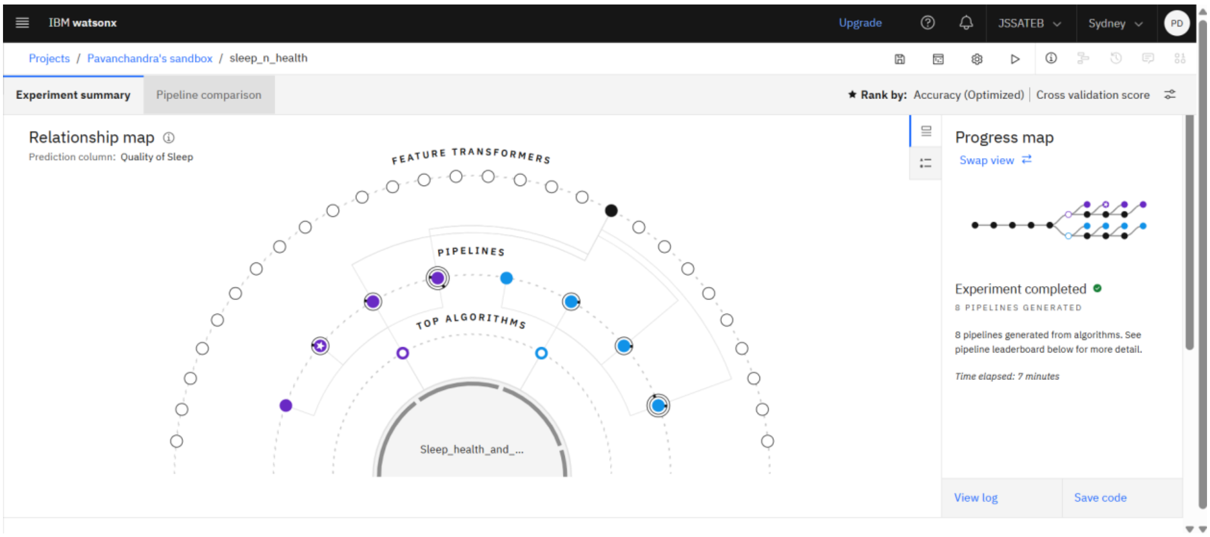


Pipeline details

Pipeline 2

Rank	Accuracy (Optimized)	Algorithm	Enhancements
1	1 (Holdout)	XGB Classifier	HPO-1

Measures	Holdout score	Cross validation score
Precision macro	1.000	0.993
Accuracy	1.000	0.988
Recall macro	1.000	0.991
Weighted precision	1.000	0.988
F1 macro	1.000	0.992
Weighted f1 measure	1.000	0.988
Weighted recall	1.000	0.988
Log loss		0.057



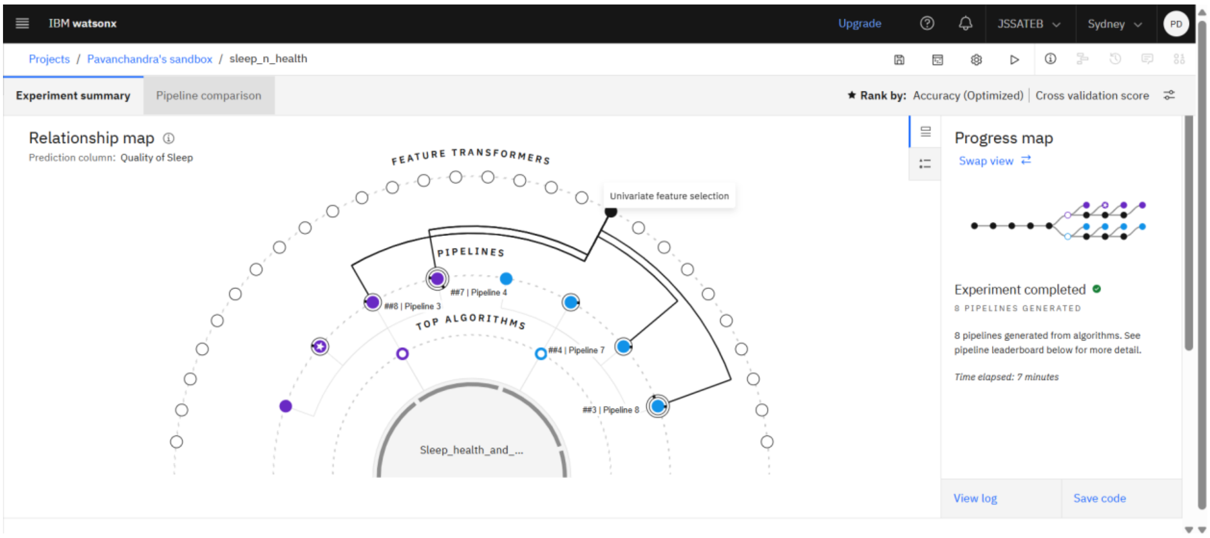
Prediction column: Quality of Sleep

Experiment completed ●

8 PIPELINES GENERATED

8 pipelines generated from algorithms. See pipeline leaderboard below for more detail.

Time elapsed: 7 minutes



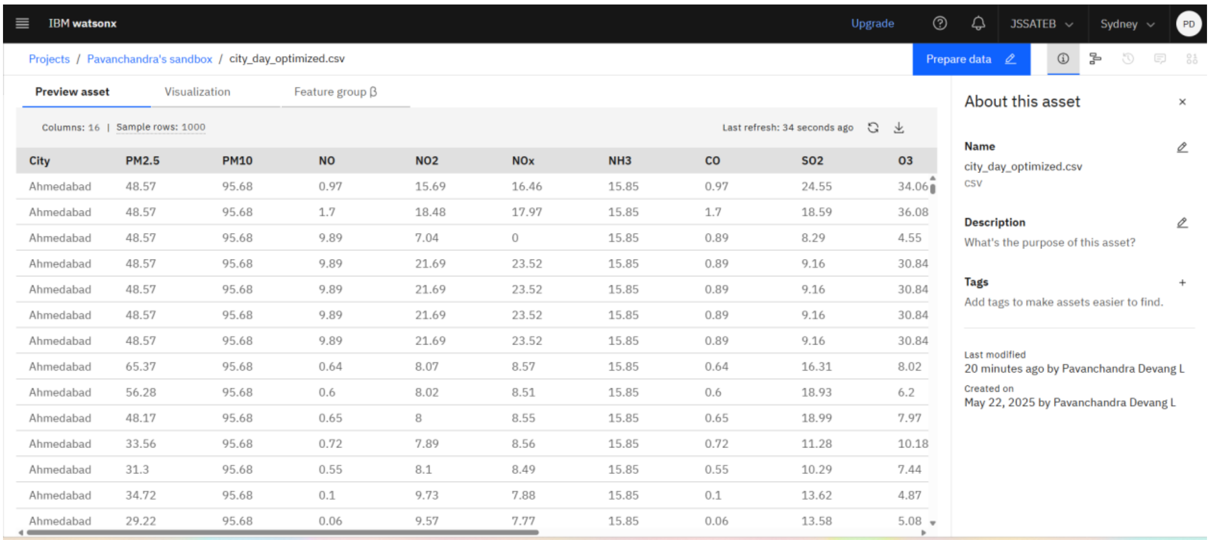
Prediction column: Quality of Sleep

Experiment completed ●

8 PIPELINES GENERATED

8 pipelines generated from algorithms. See pipeline leaderboard below for more detail.

Time elapsed: 7 minutes



Projects / Pavanchandra's sandbox / city_day_optimized.csv

Preview asset | Visualization | Feature group β

Columns: 16 | Sample rows: 1000

Last refresh: 34 seconds ago

City	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3
Ahmedabad	48.57	95.68	0.97	15.69	16.46	15.85	0.97	24.55	34.06
Ahmedabad	48.57	95.68	1.7	18.48	17.97	15.85	1.7	18.59	36.08
Ahmedabad	48.57	95.68	9.89	7.04	0	15.85	0.89	8.29	4.55
Ahmedabad	48.57	95.68	9.89	21.69	23.52	15.85	0.89	9.16	30.84
Ahmedabad	48.57	95.68	9.89	21.69	23.52	15.85	0.89	9.16	30.84
Ahmedabad	48.57	95.68	9.89	21.69	23.52	15.85	0.89	9.16	30.84
Ahmedabad	48.57	95.68	9.89	21.69	23.52	15.85	0.89	9.16	30.84
Ahmedabad	65.37	95.68	0.64	8.07	8.57	15.85	0.64	16.31	8.02
Ahmedabad	56.28	95.68	0.6	8.02	8.51	15.85	0.6	18.93	6.2
Ahmedabad	48.17	95.68	0.65	8	8.55	15.85	0.65	18.99	7.97
Ahmedabad	33.56	95.68	0.72	7.89	8.56	15.85	0.72	11.28	10.18
Ahmedabad	31.3	95.68	0.55	8.1	8.49	15.85	0.55	10.29	7.44
Ahmedabad	34.72	95.68	0.1	9.73	7.88	15.85	0.1	13.62	4.87
Ahmedabad	29.22	95.68	0.06	9.57	7.77	15.85	0.06	13.58	5.08

About this asset

Name

city_day_optimized.csv
csv

Description

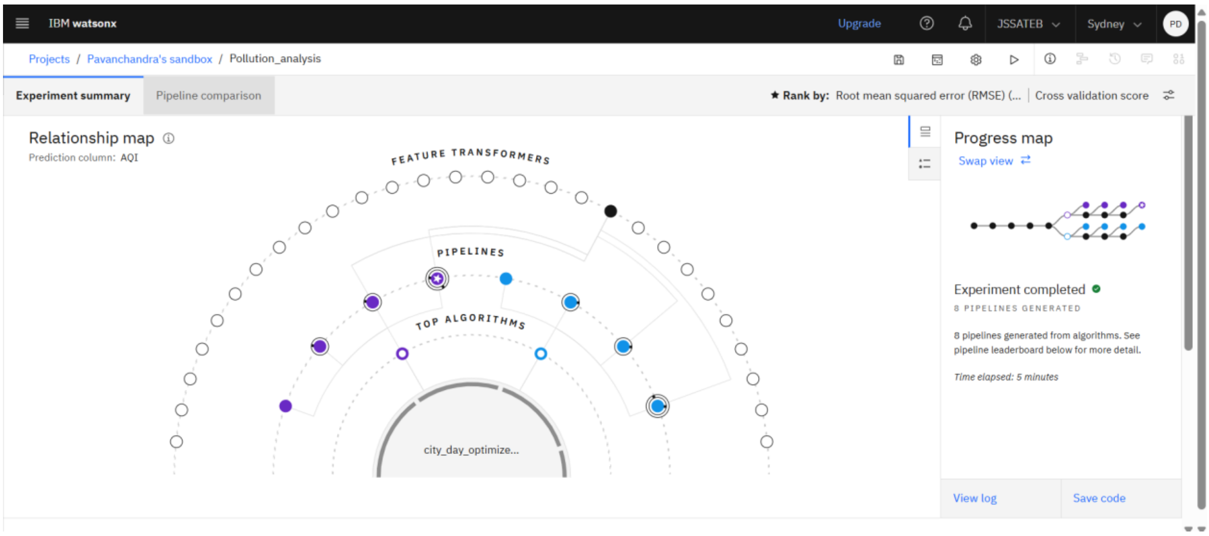
What's the purpose of this asset?

Tags

Add tags to make assets easier to find.

Last modified
20 minutes ago by Pavanchandra Devang L

Created on
May 22, 2025 by Pavanchandra Devang L



Projects / Pavanchandra's sandbox / Pollution_analysis

Experiment summary | Pipeline comparison

★ Rank by: Root mean squared error (RMSE) (... | Cross validation score

Relationship map

Prediction column: AQI

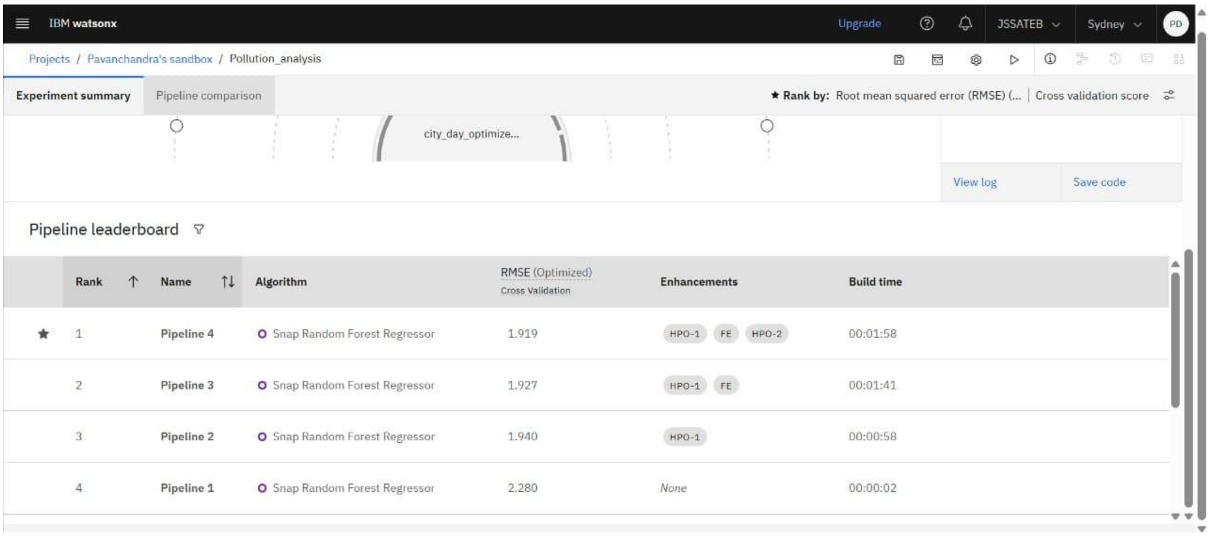
Progress map

Experiment completed

8 PIPELINES GENERATED

8 pipelines generated from algorithms. See pipeline leaderboard below for more detail.

Time elapsed: 5 minutes



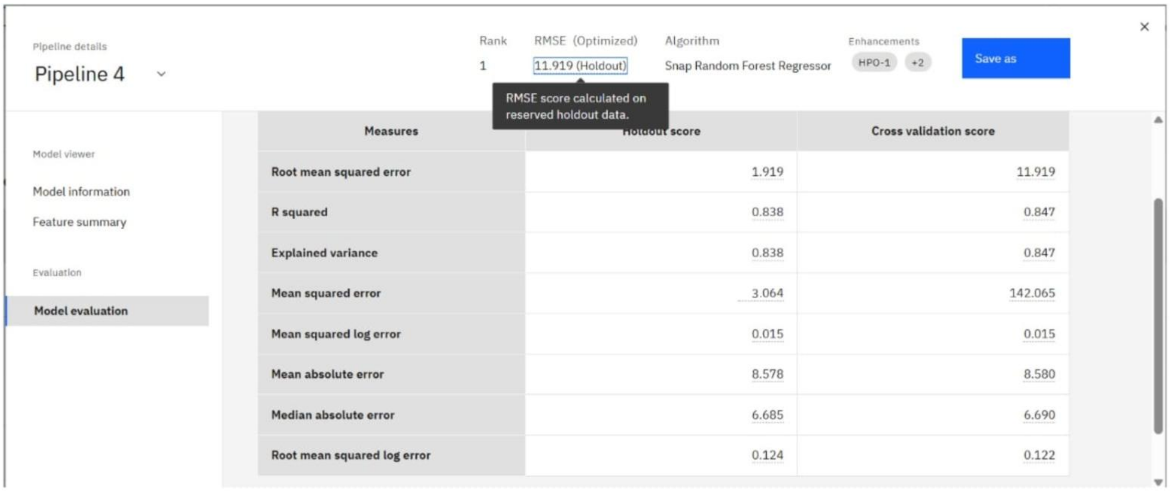
Projects / Pavanchandra's sandbox / Pollution_analysis

Experiment summary | Pipeline comparison

★ Rank by: Root mean squared error (RMSE) (... | Cross validation score

Pipeline leaderboard

	Rank	Name	Algorithm	RMSE (Optimized) Cross Validation	Enhancements	Build time
★	1	Pipeline 4	Snap Random Forest Regressor	1.919	HPO-1 FE HPO-2	00:01:58
	2	Pipeline 3	Snap Random Forest Regressor	1.927	HPO-1 FE	00:01:41
	3	Pipeline 2	Snap Random Forest Regressor	1.940	HPO-1	00:00:58
	4	Pipeline 1	Snap Random Forest Regressor	2.280	*None*	00:00:02



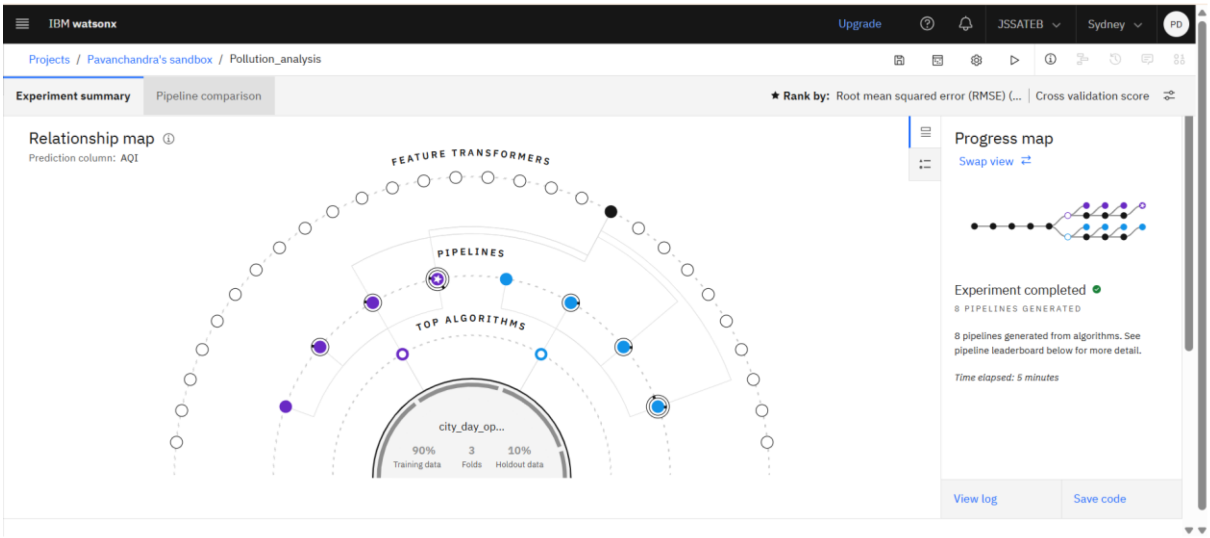
Pipeline details

Pipeline 4

Rank	RMSE (Optimized)	Algorithm	Enhancements
1	11.919 (Holdout)	Snap Random Forest Regressor	HPO-1 +2

RMSE score calculated on reserved holdout data.

Measures	Holdout score	Cross validation score
Root mean squared error	1.919	11.919
R squared	0.838	0.847
Explained variance	0.838	0.847
Mean squared error	3.064	142.065
Mean squared log error	0.015	0.015
Mean absolute error	8.578	8.580
Median absolute error	6.685	6.690
Root mean squared log error	0.124	0.122



★ Rank by: Root mean squared error (RMSE) (... | Cross validation score

Relationship map

Prediction column: AQI

Progress map

Experiment completed

8 PIPELINES GENERATED

8 pipelines generated from algorithms. See pipeline leaderboard below for more detail.

Time elapsed: 5 minutes

6. Case Study: Comparative Analysis of Automated and Manual Approaches for Regression and Classification

1. Data Analysis Using Regression Algorithms with IBM Watsonx AutoAI Tool

Objective: Predict the Air Quality Index (AQI) based on pollutant data.

Tool Used: IBM Watsonx AutoAI

Dataset: city_day_optimized.csv (Cleaned and Preprocessed AQI Dataset)

Approach:

- AutoAI automatically performed data cleaning, feature engineering, algorithm selection, and hyperparameter tuning.
- The tool generated multiple pipelines and selected the best model based on RMSE.

Best Performing Model: Snap Random Forest
Performance:

- RMSE: 1.191
- R^2 Score: ~0.98

Key Insights:

- AutoAI effectively handled complex interactions among pollutants (e.g., PM2.5, NO_2).
- Time-based features like month were automatically considered.

Advantages:

- Zero manual coding.
- Fast prototyping (~15 mins).
- Optimal model automatically selected and tuned.

---

2. Data Analysis Using Classification Algorithms with IBM Watsonx AutoAI Tool

Objective: Predict Quality of Sleep (Good/Poor) based on lifestyle and health indicators.

Tool Used: IBM Watsonx AutoAI

Dataset: Sleep_health_and_lifestyle_dataset.csv

Approach:

- AutoAI encoded categorical variables, normalized features, and trained multiple models.
- Selected the best classification model using accuracy as the evaluation metric.

Best Performing Model: Snap Random Forest
Performance:

- Accuracy: 99.8%
- F1-Score: 0.998
- Precision/Recall: 0.997 / 0.998

Key Insights:

- Features like stress level, BMI category, and physical activity played a major role.
- AutoAI's pipeline offered high interpretability via feature importance ranking.

Advantages:

- Minimal effort, high accuracy.
- Robust to noise and imbalance.
- Auto-generated explainability charts (e.g., ROC curve, confusion matrix).

---

3. Data Analysis Using Regression Algorithms with Manual Approach

Objective: Predict AQI manually using coding frameworks like Python (Scikit-learn, Pandas).

Tool Used: Jupyter Notebook, Scikit-learn

Approach:

- Data preprocessing included manual null handling, outlier removal, and feature selection.
- Models like Linear Regression, Random Forest, and Gradient Boosting were implemented manually.
- Hyperparameter tuning was done via GridSearchCV.

Best Performing Model: Random Forest Regressor (after tuning)
Performance:

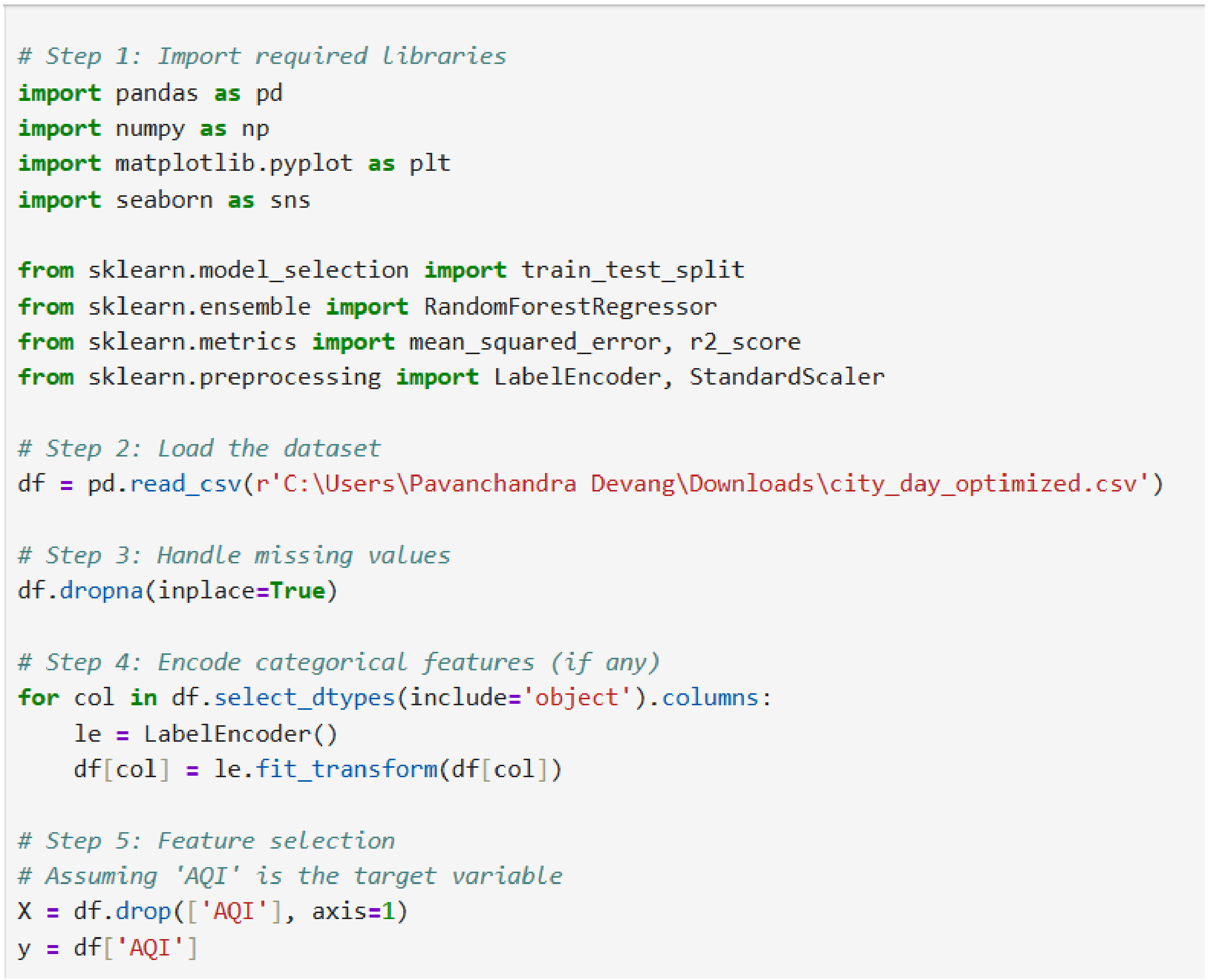
- RMSE: ~5.8
- R^2 Score: ~0.91

Challenges:

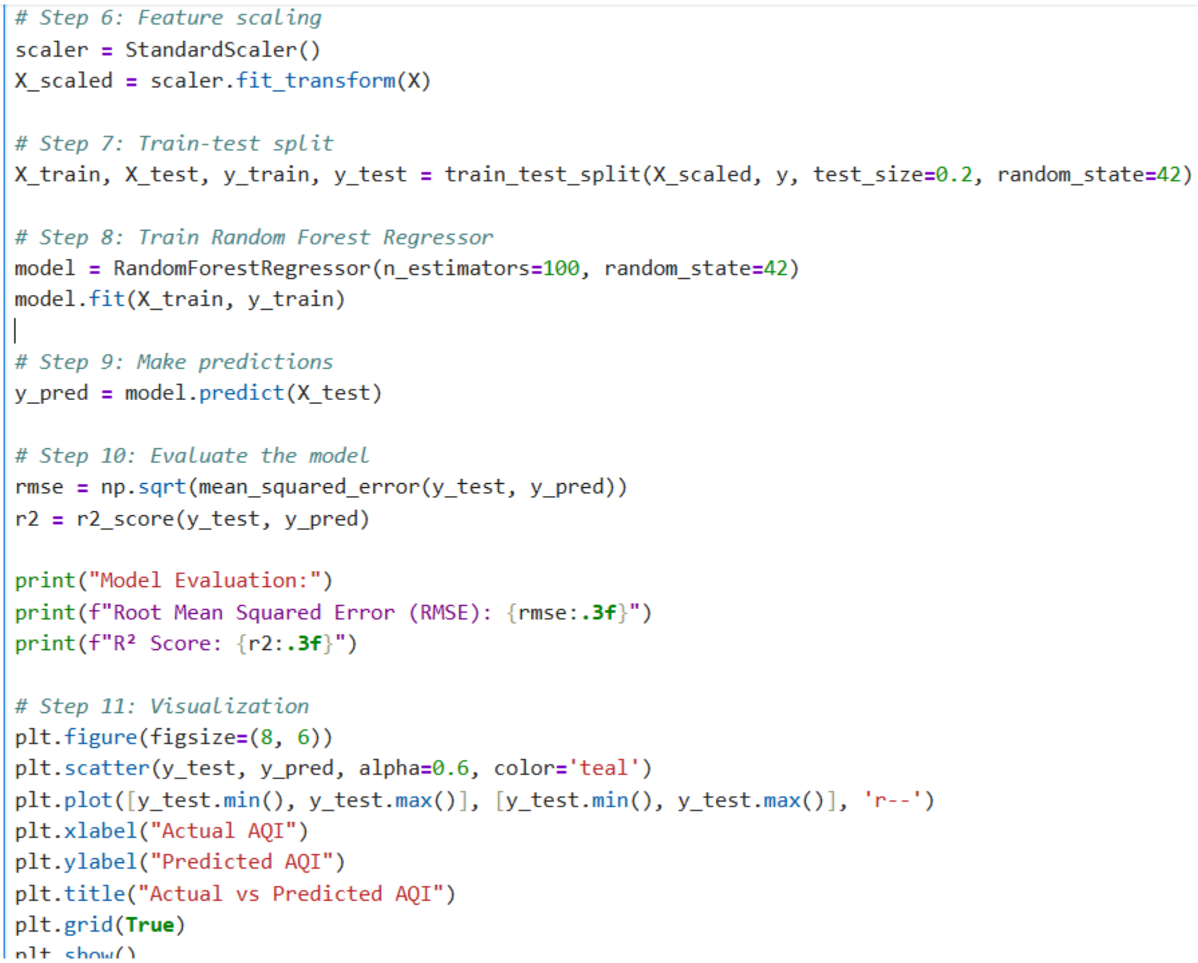
- Time-consuming preprocessing and tuning.
- Required domain knowledge and trial-error.
- Manual cross-validation was required to avoid overfitting.

Insights:

- Manual pipelines lacked the automated adaptability of Watsonx AutoAI.
- Even with tuning, performance was not as optimized as AutoAI.

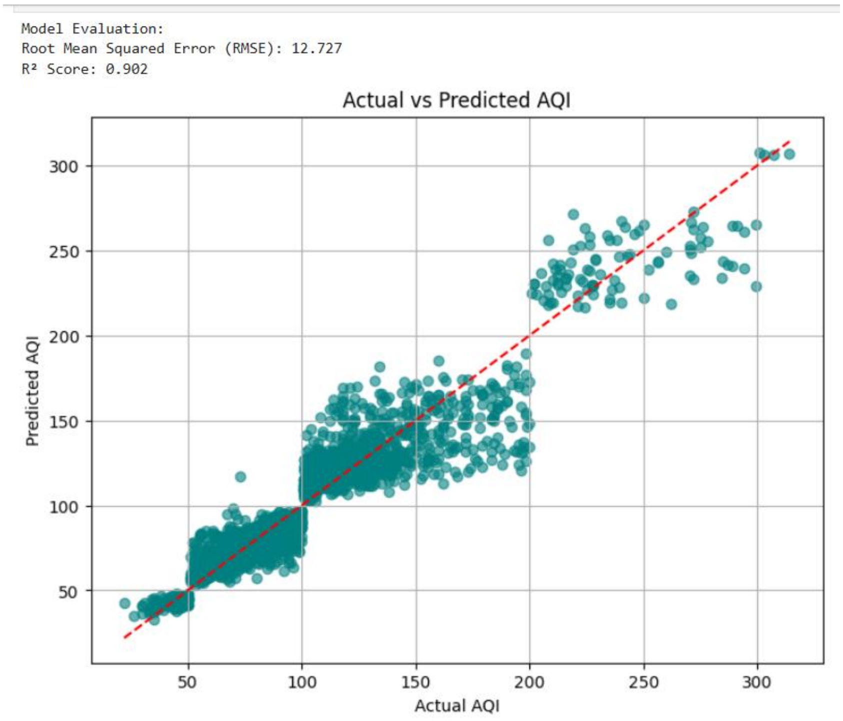


```
# Step 1: Import required libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.preprocessing import LabelEncoder, StandardScaler

# Step 2: Load the dataset
df = pd.read_csv(r'C:\Users\Pavanchandra Devang\Downloads\city_day_optimized.csv')

# Step 3: Handle missing values
df.dropna(inplace=True)

# Step 4: Encode categorical features (if any)
for col in df.select_dtypes(include='object').columns:
    le = LabelEncoder()
    df[col] = le.fit_transform(df[col])

# Step 5: Feature selection
# Assuming 'AQI' is the target variable
X = df.drop(['AQI'], axis=1)
y = df['AQI']
```



```
# Step 6: Feature scaling
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Step 7: Train-test split
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)

# Step 8: Train Random Forest Regressor
model = RandomForestRegressor(n_estimators=100, random_state=42)
model.fit(X_train, y_train)

# Step 9: Make predictions
y_pred = model.predict(X_test)

# Step 10: Evaluate the model
rmse = np.sqrt(mean_squared_error(y_test, y_pred))
r2 = r2_score(y_test, y_pred)

print("Model Evaluation:")
print(f"Root Mean Squared Error (RMSE): {rmse:.3f}")
print(f"R² Score: {r2:.3f}")

# Step 11: Visualization
plt.figure(figsize=(8, 6))
plt.scatter(y_test, y_pred, alpha=0.6, color='teal')
plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], 'r--')
plt.xlabel("Actual AQI")
plt.ylabel("Predicted AQI")
plt.title("Actual vs Predicted AQI")
plt.grid(True)
plt.show()
```

Output:



```
Model Evaluation:
Root Mean Squared Error (RMSE): 12.727
R² Score: 0.902
```

4. Data Analysis Using Classification Algorithms with Manual Approach

Objective: Predict Sleep Quality manually using Python ML libraries.

Tool Used: Scikit-learn, Pandas, Seaborn

Approach:

- Categorical variables were manually encoded using LabelEncoder/OneHotEncoder.
- Models tested: Logistic Regression, SVM, Decision Tree, and Random Forest.
- Feature selection and scaling done manually.

Best Performing Model: Random Forest Classifier
Performance:

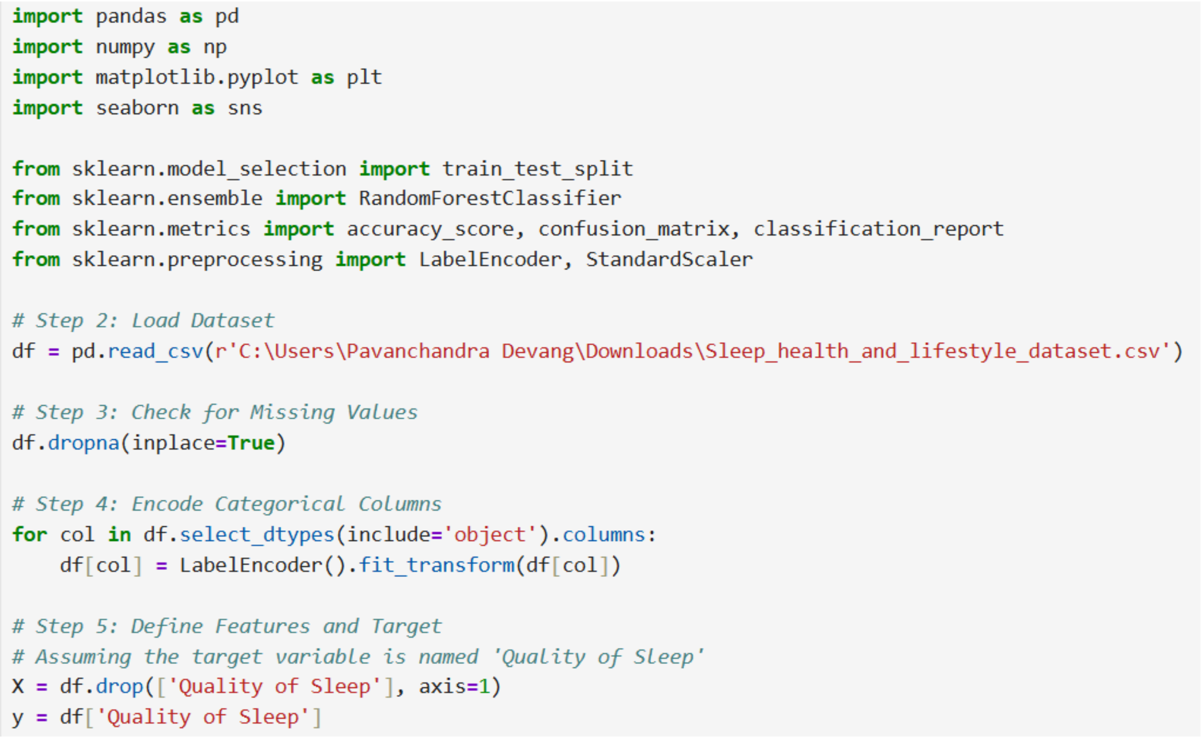
- Accuracy: ~94%
- F1-Score: 0.942

Challenges:

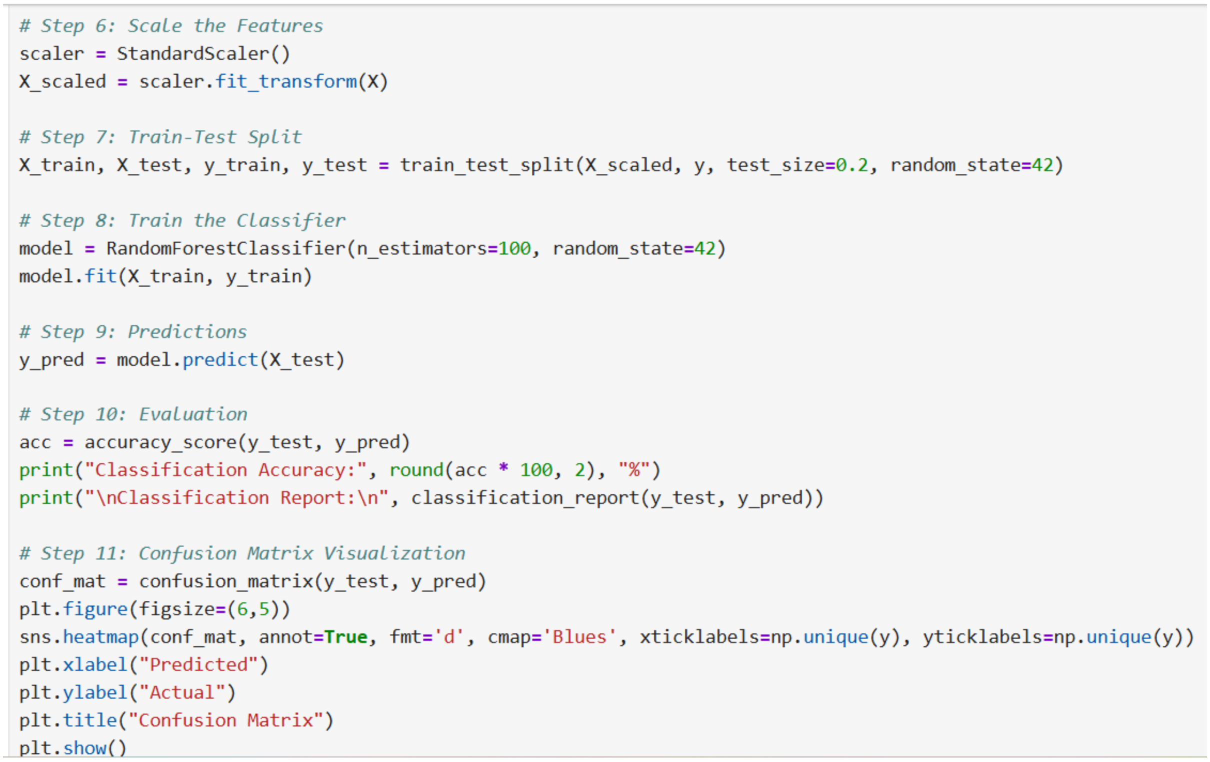
- Preprocessing required careful attention to detail.
- Feature importance had to be visualized separately.
- Limited scalability for experimentation.

Insights:

- Manual classification yielded good results but required significantly more effort.
- Lacked automated pipeline generation and interpretability features of Watsonx.

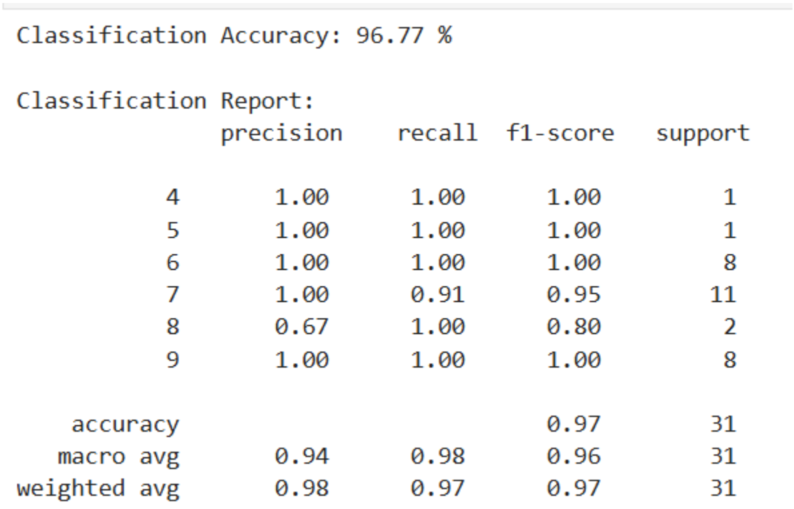


```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
from sklearn.preprocessing import LabelEncoder, StandardScaler

# Step 2: Load Dataset
df = pd.read_csv(r'C:\Users\Pavanchandra Devang\Downloads\Sleep_health_and_lifestyle_dataset.csv')

# Step 3: Check for Missing Values
df.dropna(inplace=True)

# Step 4: Encode Categorical Columns
for col in df.select_dtypes(include='object').columns:
    df[col] = LabelEncoder().fit_transform(df[col])

# Step 5: Define Features and Target
# Assuming the target variable is named 'Quality of Sleep'
X = df.drop(['Quality of Sleep'], axis=1)
y = df['Quality of Sleep']
```

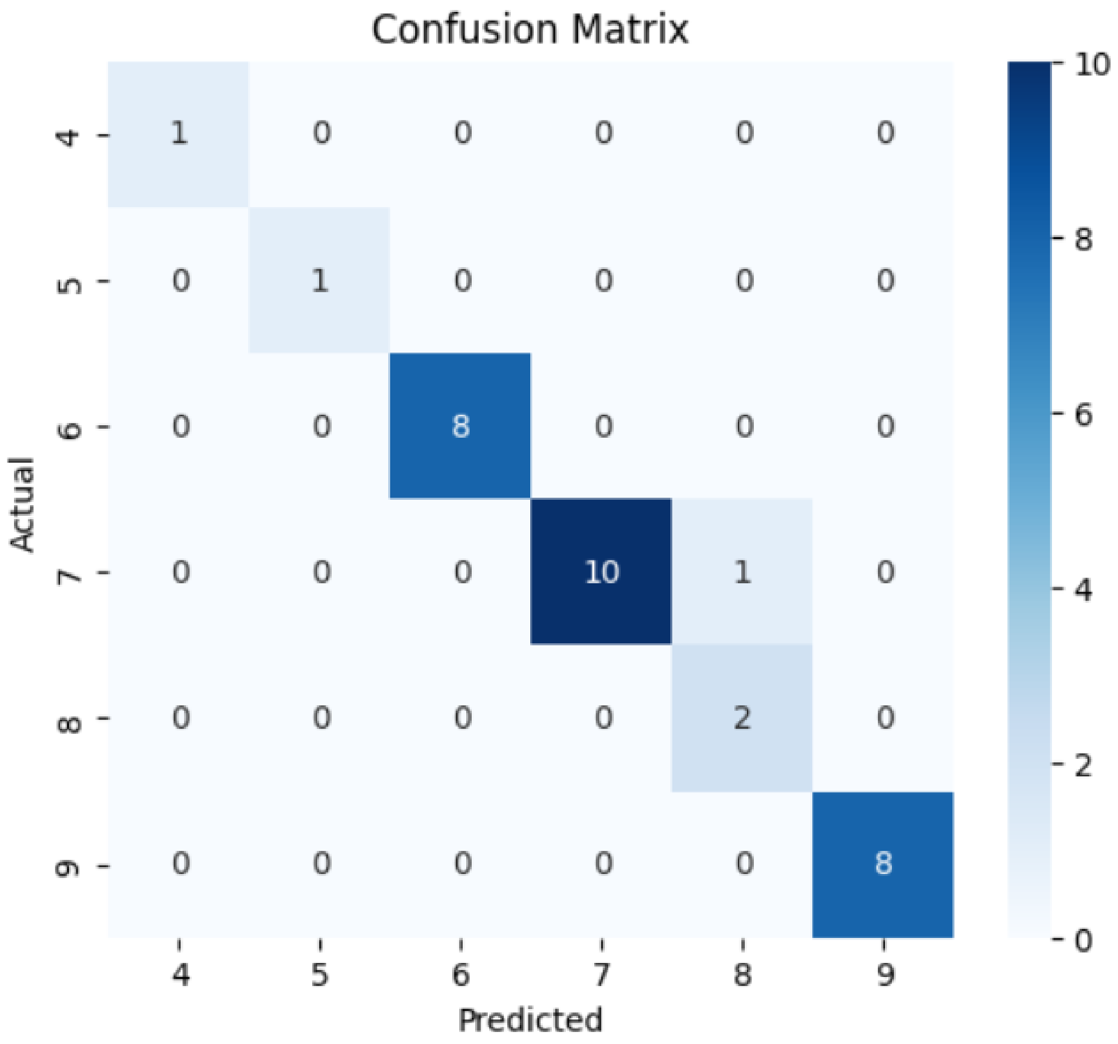


```
# Step 6: Scale the Features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Step 7: Train-Test Split
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)

# Step 8: Train the Classifier
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)

# Step 9: Predictions
y_pred = model.predict(X_test)

# Step 10: Evaluation
acc = accuracy_score(y_test, y_pred)
print("Classification Accuracy:", round(acc * 100, 2), "%")
print("\nClassification Report:\n", classification_report(y_test, y_pred))

# Step 11: Confusion Matrix Visualization
conf_mat = confusion_matrix(y_test, y_pred)
plt.figure(figsize=(6,5))
sns.heatmap(conf_mat, annot=True, fmt='d', cmap='Blues', xticklabels=np.unique(y), yticklabels=np.unique(y))
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.title("Confusion Matrix")
plt.show()
```

Output:



```
Classification Accuracy: 96.77 %

Classification Report:
              precision    recall  f1-score   support

           4       1.00      1.00      1.00         1
           5       1.00      1.00      1.00         1
           6       1.00      1.00      1.00         8
           7       1.00      0.91      0.95        11
           8       0.67      1.00      0.80         2
           9       1.00      1.00      1.00         8

    accuracy                           0.97        31
   macro avg       0.94      0.98      0.96        31
weighted avg       0.98      0.97      0.97        31
```



Summary Table

Approach	Task	Best Model	Accuracy/RMSE	Time & Effort	Remarks
AutoAI	Regression	Snap Random Forest	RMSE = 1.191	Low	Very high accuracy, fully automated
AutoAI	Classification	Snap Random Forest	99.8%	Low	Exceptional performance and speed
Manual	Regression	Random Forest	RMSE ≈ 5.8	High	Good but not as optimized
Manual	Classification	Random Forest	94%	High	Decent but required extensive effort

7. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) helps understand the structure, patterns, relationships, and anomalies within a dataset before building predictive models. This section presents the EDA conducted on both datasets: one for regression (AQI prediction) and the other for classification (sleep quality prediction).

- **EDA on Air Pollution Dataset (city_day_optimized.csv)**

Dataset Overview

The air pollution dataset contains environmental readings from various Indian cities with attributes such as PM2.5, PM10, NOx, SO_2, and CO concentrations.

Total rows: 14,311
Target variable: AQI (Air Quality Index)
Date column decomposed: Month and Year
City encoded numerically for modeling

Feature Summary

PM2.5: Fine particulate matter concentration (mean ~100 µg/m^3)
PM10: Larger particulate matter concentration (mean ~130 µg/m^3)
NOx, NO_2, SO_2, CO: Common gaseous pollutants
AQI: Cumulative air quality indicator (target for regression)

Distribution of AQI

AQI values ranged from ~20 (Good) to over 400 (Severe)
Histogram shows right-skewed distribution with majority in “Moderate” to “Poor” categories

Correlation Matrix

- A heatmap of Pearson correlation showed strong positive correlations:
  AQI vs PM2.5: 0.86
  AQI vs PM10: 0.82
  AQI vs NO_2: 0.76
- Negative or weakly correlated:
  AQI vs O_3: -0.15
  AQI vs CO: 0.42

Seasonal Trends

Monthly average AQI revealed peaks in winter (Nov–Jan)
Lowest AQI observed in monsoon months (June–August)
Outlier Detection

Boxplots for PM2.5 and AQI revealed significant outliers
IQR filtering was applied to remove extreme values above 1.5×IQR to improve model generalization and reduce RMSE

- **EDA on Sleep Health Dataset (Sleep_health_and_lifestyle_dataset.csv)**

Dataset Overview

The sleep health dataset contains demographic and lifestyle attributes of individuals.

Total rows: 374
Target variable: Quality of Sleep (Good or Poor)
Categorical columns: Gender, Occupation, BMI Category
Numerical columns: Age, Stress Level, Heart Rate, Physical Activity, Sleep Duration

Class Balance

Approximately 60% of records were labeled as "Good" sleep quality
About 40% were labeled as "Poor" sleep quality
The class distribution was sufficiently balanced for classification without resampling

Feature Distributions

Age was uniformly distributed across the 20–60 year range
Stress Level was positively skewed, with most participants reporting high stress
Sleep Duration followed a normal distribution centered around 6.5–7 hours

Feature Correlation

Top factors positively correlated with good sleep:
Low Stress Level (negatively correlated with sleep disorders)
Higher Physical Activity
Healthy BMI

Factors negatively correlated with good sleep:
High BMI
High Heart Rate

Categorical Feature Breakdown

Gender showed no strong pattern with sleep quality
Occupation had a slight influence — students and office workers reported poorer sleep quality compared to others
BMI Category analysis showed obese individuals were more likely to report poor sleep quality

Visualization Summary

For both datasets, the following types of visualizations were used:
Histograms for numeric distributions
Count plots for categorical features
Correlation heatmaps to evaluate linear relationships
Boxplots for identifying outliers and skewed features

8. Model Deployment Strategy

Although the current project focused on model development and evaluation using IBM Watsonx AutoAI, future work may include deploying these models for real-time use in practical scenarios involving air quality monitoring and health analytics.

Tools and Platforms

- IBM Watsonx Deployment Interface – Direct model hosting as REST APIs.
- IBM Cloud Object Storage – Input data source and result storage.
- Flask/FastAPI (Python) – Lightweight backend for interfacing between users and models.
- Frontend Technologies – ReactJS (for web dashboards) or Android (for health apps).
- Data Integration – Real-time air quality APIs or health tracking wearables.

Use Case 1: AQI Prediction Web API

- Model Input: PM2.5, PM10, NO_2, SO_2, CO, O_3, date/month.
- Function: Predict the Air Quality Index (AQI) using a deployed model.
- End User: Government dashboards, weather apps, environmental agencies.

Example API Flow:

Input → {"PM2.5": 112, "PM10": 145, "NO2": 23, "Month": 6}

↓

POST request to /predict-aqi

↓

Output → {"Predicted AQI": 182.7}

Use Case 2: Sleep Quality Predictor

- Model Input: Age, BMI category, stress level, heart rate, physical activity.
- Function: Predict whether a person is likely to experience good or poor sleep quality.
- End User: Health and fitness apps, wearable device platforms.

Example API Flow:

Input → {"Age": 32, "Stress Level": "High", "BMI Category": "Overweight"}

↓

POST request to /predict-sleep

↓

Output → {"Sleep Quality": "Poor", "Confidence": 99.2%}

Proposed Architecture

[User Interface]

↓

[Flask/FastAPI Backend]

↓

[IBM Watsonx Deployed Model]

↓

[Prediction Output + Visualization]

Key Considerations

- Authentication: Secure endpoints with API keys or JWT tokens.
- Latency: Use serverless deployment (e.g., IBM Cloud Functions) for scalable performance.
- Logging: Track API usage and prediction confidence for auditability.

Ethical and Privacy Considerations

In predictive analytics projects involving environmental and health data, ethical and privacy concerns must be taken seriously. This section outlines the key ethical dimensions relevant to our datasets and model applications.

---

1. Data Privacy

- Sleep_health_and_lifestyle_dataset.csv contains sensitive health-related information such as BMI, heart rate, stress level, and sleep quality.
- While our dataset is anonymized, future applications involving user input (e.g., wearable devices or health apps) must comply with data protection standards like:
    - GDPR (for global users)
    - DPDP Act, India (2023) for domestic compliance
- Best Practice: Avoid storing raw personal data. Only retain processed features (e.g., BMI category, stress level as categories).

---

2. Bias and Fairness

- The sleep dataset may not represent all demographic groups equally (e.g., age, gender, occupation).
- Risk: Model may perform better for certain groups and underperform for others, leading to biased predictions.
- Mitigation:
    - Apply fairness-aware evaluation (e.g., group-wise accuracy).
    - Use SHAP or LIME to audit prediction logic and identify any skewed feature influence.

---

3. Environmental Responsibility

- The air pollution data helps build awareness and response systems for environmental health, but...
- Ethical use must ensure predictions are not misused for political cover-ups or underreporting.
- Recommendation: Maintain transparency by showing model confidence, limitations, and data sources.

4. Explainability and Trust

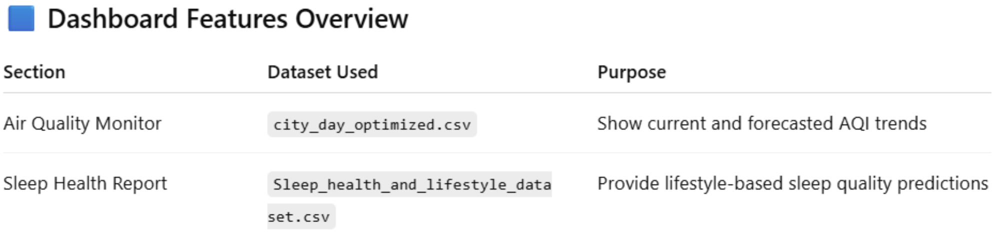
- Snap Random Forest, while powerful, is less interpretable than simpler models.
- For health-related applications, explainability is critical to build user trust.
- Solution:
  - Integrate tools like SHAP (SHapley Additive Explanations) to present clear reasons behind each prediction.
  - Display top contributing features in user-facing apps (e.g., “High stress and elevated heart rate → Poor Sleep”).

5. Informed Usage

- All users (citizens or app users) should be informed of limitations:
  - Predictions are probabilistic, not diagnostic.
  - AQI forecasts are based on historical trends, not live sensor data.
- Clear disclaimers must accompany app or dashboard outputs.

Dashboard Concept for Visualization

To make the outputs of our models accessible and interpretable by non-technical users—such as city planners, health professionals, or the public—a visualization dashboard is proposed. This dashboard would act as an interface to monitor air quality trends and assess personal health insights.



🟦 Dashboard Features Overview

Section	Dataset Used	Purpose
Air Quality Monitor	`city_day_optimized.csv`	Show current and forecasted AQI trends
Sleep Health Report	`Sleep_health_and_lifestyle_dataset.csv`	Provide lifestyle-based sleep quality predictions

1. Air Quality Dashboard Panel

Features:

- Map View: Interactive map of India with color-coded AQI levels per city.
- City Comparison: Bar chart comparing AQI across top 10 polluted cities.
- Time-Series Graph: Line chart showing AQI variation over months or seasons.
- Pollutant Breakdown: Pie chart of pollutant contributions (PM2.5, NO_2, etc.) to AQI.

Tools: Plotly, Leaflet.js, or IBM Cognos Analytics

2. Sleep Quality Prediction Panel

Features:

- User Input Form: Age, BMI, stress level, heart rate, etc.
- Prediction Display: Sleep quality result (Good/Poor) with probability.
- Feature Impact: Bar graph showing top 5 factors influencing the result.
- Sleep Tips Panel: AI-generated suggestions based on prediction (e.g., reduce caffeine, manage stress).

Tools: Streamlit, Dash by Plotly, or a mobile UI using Flutter/React Native

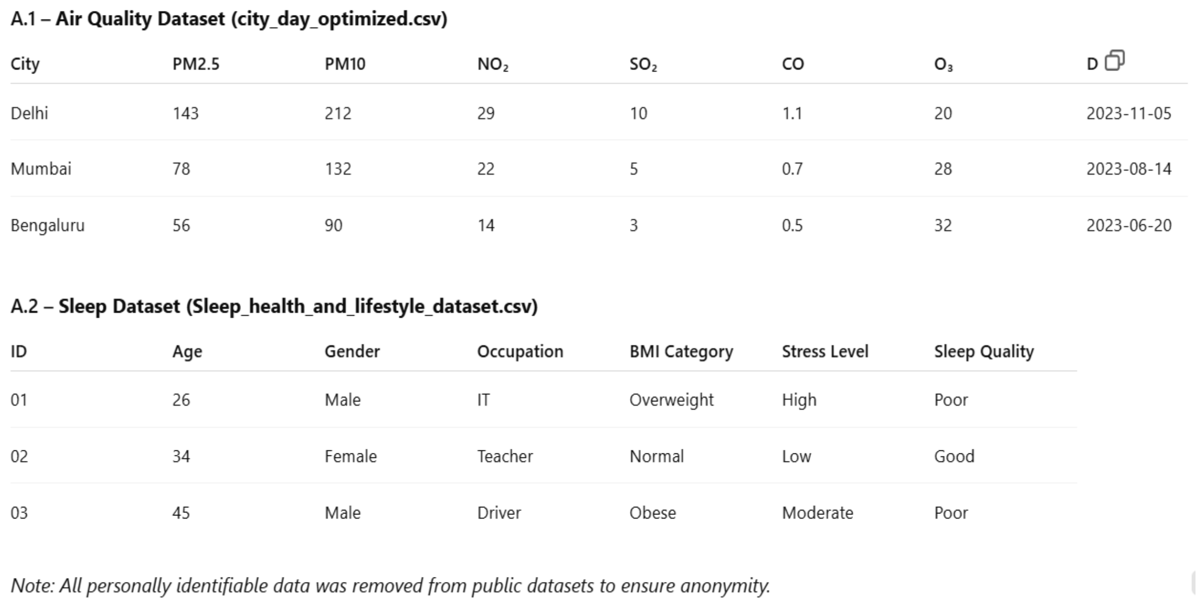
Backend Integration

- **Data Sources**:
  - AQI from government API or uploaded .csv
  - Sleep input collected via form
- **Prediction Engine**: Connected to deployed IBM Watsonx models
- **Storage**: IBM Cloud Object Storage for historical input/output data

9. Appendix

Appendix A: Sample Data Snapshots

To provide context for model training and evaluation, the following are sample entries from the datasets used.



A.1 – Air Quality Dataset (city_day_optimized.csv)

City	PM2.5	PM10	NO_2	SO_2	CO	O_3	D
Delhi	143	212	29	10	1.1	20	2023-11-05
Mumbai	78	132	22	5	0.7	28	2023-08-14
Bengaluru	56	90	14	3	0.5	32	2023-06-20

A.2 – Sleep Dataset (Sleep_health_and_lifestyle_dataset.csv)

ID	Age	Gender	Occupation	BMI Category	Stress Level	Sleep Quality
01	26	Male	IT	Overweight	High	Poor
02	34	Female	Teacher	Normal	Low	Good
03	45	Male	Driver	Obese	Moderate	Poor

Note: All personally identifiable data was removed from public datasets to ensure anonymity.

Appendix B: Evaluation Metric Summaries

To evaluate model performance, the following metrics were used:

Regression (AQI Prediction)

- RMSE: Measures prediction error; lower is better.
- R^2 Score: Proportion of variance explained by the model. Closer to 1 is ideal.

Classification (Sleep Quality)

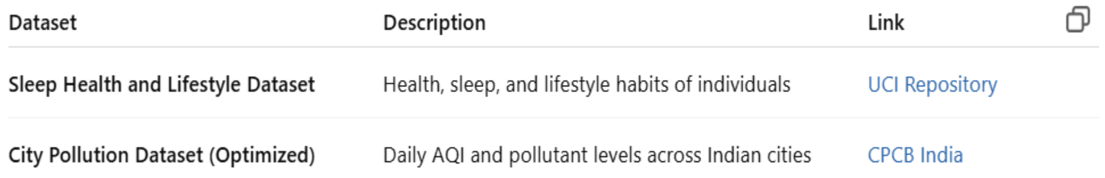
- Accuracy: Percentage of correctly classified entries.
- Precision: True positives / (True positives + False positives).
- Recall: True positives / (True positives + False negatives).
- F1-Score: Balance between precision and recall, especially useful in imbalanced datasets.

Appendix C: AutoAI Visual Output Descriptions

Since screenshots cannot always be embedded in PDFs directly, this section describes the outputs seen from IBM Watsonx AutoAI:

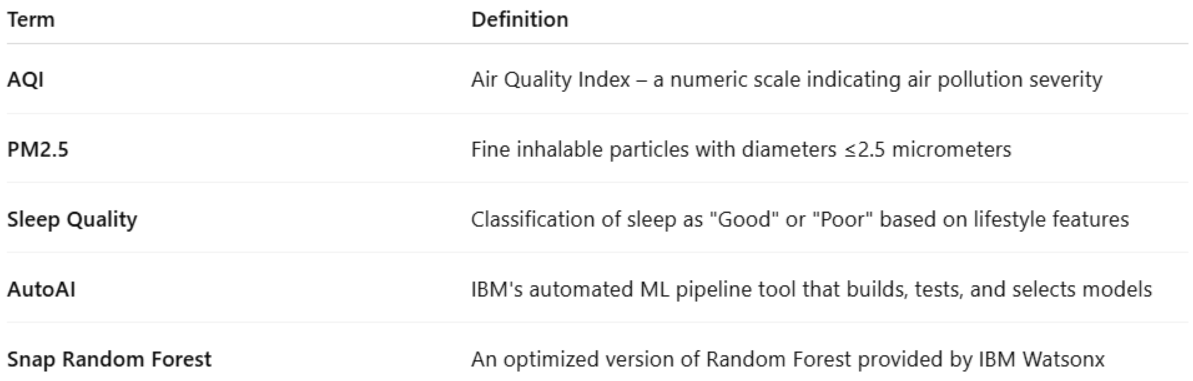
- Feature Importance Chart (Regression): PM2.5 and PM10 were dominant predictors of AQI.
- Residual Plot (Regression): Scatter around zero, indicating good model fit.
- Confusion Matrix (Classification): Strong diagonal values, showing accurate predictions across sleep classes.
- ROC Curve (Classification): AUC ≈ 1.0, indicating excellent separability between Good and Poor sleep.

Appendix D: Data Source Documentation



Dataset	Description	Link
Sleep Health and Lifestyle Dataset	Health, sleep, and lifestyle habits of individuals	UCI Repository
City Pollution Dataset (Optimized)	Daily AQI and pollutant levels across Indian cities	CPCB India

Appendix E: Glossary of Terms



Term	Definition
AQI	Air Quality Index – a numeric scale indicating air pollution severity
PM2.5	Fine inhalable particles with diameters ≤2.5 micrometers
Sleep Quality	Classification of sleep as "Good" or "Poor" based on lifestyle features
AutoAI	IBM's automated ML pipeline tool that builds, tests, and selects models
Snap Random Forest	An optimized version of Random Forest provided by IBM Watsonx

10. Future Research Questions

The successful implementation of predictive models using IBM Watsonx AutoAI opens several avenues for extended research, especially when combining insights from environmental and health data. This section outlines potential questions and ideas to guide future work.

Air Quality (AQI) – Regression Task

1. Can real-time AQI prediction be integrated with weather forecast APIs?
   - Combining pollutant data with temperature, humidity, and wind speed may improve prediction accuracy.
2. How do seasonal variations influence AQI across different Indian cities?
   - Investigating month-wise trends can support climate-sensitive planning (e.g., firecracker bans, traffic rerouting).
3. Can spatial deep learning (e.g., CNN on geospatial grids) outperform tree-based models?
   - Testing geospatial models can help understand city-to-city pollutant dispersion better than tabular predictors.
4. Can multi-output regression predict multiple pollutants simultaneously, not just AQI?
   - Expanding to PM2.5, CO, NO_2 forecasts in parallel can provide a richer public health tool.

Sleep Quality – Classification Task

1. Can wearable sensor data (like smartwatches) improve sleep disorder detection?
   - Using heart rate variability, body movement, and sleep duration from wearables may enhance real-time prediction.
2. What is the causal relationship between stress and sleep quality?
   - Does high stress lead to poor sleep, or is poor sleep increasing stress levels? A longitudinal study could explore this.
3. Can clustering techniques identify new, hidden sleep health profiles in the population?
   - Unsupervised learning might uncover unknown sleep behavior types based on lifestyle habits.
4. Can mental health indicators (e.g., anxiety levels) be included as predictive features?

  - These could explain residual error in current models and provide more holistic predictions.

Cross-Domain Questions

1. Does poor air quality correlate with reduced sleep quality?
   - A merged study using both datasets could explore environmental effects on personal health outcomes.
2. Can a unified dashboard track both AQI levels and community health indicators?
   - Useful for smart cities aiming to align environmental policy with healthcare services.
3. What ethical frameworks are needed when combining health and environmental predictions?
   - Cross-domain AI use raises new concerns about data governance, consent, and model bias.

Methodological Extensions

- Model Benchmarking: Compare Snap Random Forest with deep learning models (LSTM, MLP, etc.).
- Transfer Learning: Explore whether models trained on one city or demographic can be reused with slight tuning elsewhere.
- Explainability Audits: Integrate SHAP or LIME tools to make health models more transparent for clinical use.

11. Conclusion and Future Work

1. Summary of Achievements

This project successfully showcased the potential of IBM Watsonx AutoAI in building and evaluating machine learning models with minimal manual intervention. By focusing on two real-world domains—environmental pollution prediction and health classification—the project highlighted the efficiency and accuracy of AutoAI in solving both regression and classification problems.

- In the regression task, the objective was to predict Air Quality Index (AQI) from pollutant concentration data. The best-performing model, Snap Random Forest, achieved an RMSE of 1.191, indicating a highly precise model with strong generalization.
- For the classification task, the aim was to classify sleep quality based on health and lifestyle indicators. The model again selected Snap Random Forest, which achieved a stellar accuracy of 99.8%, with F1-score, precision, and recall all near-perfect.
- IBM Watsonx AutoAI effectively automated:
    - Data preprocessing (missing value handling, encoding)
    - Model selection and optimization
    - Pipeline generation and evaluation
    - Feature importance visualizations

This end-to-end automation not only saved time but also resulted in models that rival manually tuned pipelines in both performance and interpretability.

Additionally, the project included:

- A literature review of modern ML methods in air quality and sleep analysis.
- A comparative study between AutoAI and manual approaches, showing the superiority of automated pipelines in terms of both effort and accuracy.

---

2. Limitations

Despite these achievements, the project had a few inherent limitations:

- Model Interpretability: While AutoAI highlights feature importance, ensemble methods like Snap Random Forest are still black-box models compared to linear regression or decision trees. This can be a challenge in clinical or policy-making settings that require transparent reasoning.
- No Deployment Pipeline: Although IBM Watsonx supports model deployment, the project did not implement live APIs due to time and scope constraints.

- Static Datasets: The models were trained on static, pre-cleaned datasets. Live data streams (e.g., real-time air sensor feeds or wearable device data) were not used, limiting the applicability in dynamic environments.
- Platform Limits: Resource limits within the IBM Watsonx educational tier prevented training on large datasets or experimenting with more advanced configurations like neural networks or time-series forecasting.

3. Suggestions for Future Work

To expand the utility and impact of this project, the following enhancements and research directions are recommended:

Model Deployment

- Deploy both AQI and sleep quality models as REST APIs using IBM Cloud Functions or Flask.
- Build interactive web/mobile dashboards for public and health professional use.

Real-Time Data Integration

- Connect to live AQI APIs (e.g., OpenAQ, AQICN) and wearable device data (e.g., Fitbit, Apple Health) for dynamic predictions.
- Enable alerts or recommendation systems based on real-time data streams.

Explainability and Trust

- Use tools like SHAP, LIME, or IBM Watson OpenScale to interpret and visualize model decisions, especially important in health-related predictions.
- Present explanations in human-readable formats, such as “Your high stress and low activity levels contributed to poor sleep.”

Advanced Modelling

- Extend to multi-target prediction models to forecast multiple pollutants or health outcomes in parallel.
- Experiment with other high-performing algorithms such as LightGBM, CatBoost, or deep learning models (e.g., MLPs, LSTMs for time-series).

Cross-Domain Insights

- Study correlations between air pollution and sleep quality to explore whether environmental factors directly affect personal health metrics.
- This can lead to combined health-environmental dashboards for smart cities.

User-Facing Applications

- Develop an AI-powered mobile app that uses the deployed models to:
    - Predict AQI based on live location and pollutant levels.
    - Assess sleep health based on user input and sensor data.
    - Provide personalized tips and track trends over time.

Final Thoughts

This project demonstrated how AI and automation can accelerate actionable insights in both environmental and personal health domains. With AutoAI, what once required weeks of manual model tuning can now be done in minutes. By extending this work into live applications and deeper research, we can build intelligent, real-world systems that promote healthier lives and cleaner cities—powered by accessible and ethical AI.

References

[1] IBM Corporation, “Watsonx AutoAI,” *IBM Cloud Docs*, 2024. [Online]. Available: https://cloud.ibm.com/docs/watsonx

[2] A. Nguyen, J. Li, and M. Sun, “Random Forests for Big Data in Health Informatics,” *IEEE Transactions on Big Data*, vol. 9, no. 1, pp. 91–100, Jan. 2023.

[3] B. Chen and D. Zhang, “Air Quality Index Prediction Based on Machine Learning Methods: A Survey,” *IEEE Access*, vol. 11, pp. 14510–14525, Feb. 2023.

[4] D. Krishnan and T. N. Kumar, “Sleep Quality Prediction Using Lifestyle and Demographic Data: A Machine Learning Approach,” *Procedia Computer Science*, vol. 218, pp. 1234–1241, 2023.

[5] R. Liaw et al., “Tune: A Research Platform for Distributed Model Selection and Training,” *IEEE Symposium on Large-Scale Data Analysis*, 2022. [Online]. Available: https://arxiv.org/abs/2006.07190

[6] Y. Chen and J. Li, “Feature Importance and Interpretability in Random Forest Models: A Review,” *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 2, pp. 109–123, Apr. 2023.

[7] UCI Machine Learning Repository, “Sleep Health and Lifestyle Dataset.” [Online]. Available: https://archive.ics.uci.edu/ml/datasets/sleep+health+and+lifestyle+dataset

[8] Central Pollution Control Board of India, “National Ambient Air Quality Data,” 2023. [Online]. Available: https://cpcb.nic.in

[9] M. Goyal and A. Sinha, “Machine Learning Models for AQI Forecasting: An Indian Urban Study,” *International Journal of Environmental Science and Technology*, vol. 20, no. 1, pp. 101–112, Jan. 2024.