

Big Data

Lab Project Report



June 21, 2025

Hassan Sarfraz(fa22-bds-017)

TALHA SAEED (SP21-BDS-023)

**1. Introduction**

The growing popularity of bike-sharing applications in urban mobility has generated massive amounts of user activity data. This project aims to leverage Big Data tools and machine learning to predict bike-sharing demand using a real-world dataset. The project includes data preprocessing, analysis, model building, evaluation, and cloud deployment.

**2. Dataset Description**

* Source: Kaggle (Bike Sharing Dataset)
* Attributes: Includes temperature, humidity, windspeed, season, weather conditions, working day, holiday, etc.
* Size: Approximately 17,000 records
* Target Variable: count – the number of bikes rented per hour/day

**3. Methodology & Techniques (CLO 4)**

**3.1 Data Acquisition**

The dataset was acquired from Kaggle and loaded using Pandas for preprocessing and transformation.

**3.2 Exploratory Data Analysis (EDA)**

Performed using Pandas, Matplotlib, and Seaborn:

* Correlation heatmap
* Hourly/seasonal trends in bike usage
* Distribution plots of temperature, humidity, etc.

**3.3 Data Cleaning & Transformation**

* Handled missing values using median/mode imputation
* Detected and handled outliers using IQR
* Normalized numerical features
* One-hot encoded categorical variables

**3.4 Feature Selection / Dimensionality Reduction**

* Correlation analysis to remove redundant features
* PCA applied for reducing dimensions and improving model performance

**3.5 Data Partitioning**

* Used train\_test\_split from Scikit-learn
* 80% training and 20% test split
* Random seed used for reproducibility

**4. Model Comparison (CLO 5)**

**Models Used**

* Random Forest Regressor
* Decision Tree Regressor
* Linear Regression

**Model Performance (R² Score)**

| **Model** | **R² Score** |
| --- | --- |
| Random Forest | 0.8824 |
| Decision Tree | 0.7899 |
| Linear Regression | 0.8298 |

**Conclusion**

* Random Forest performed the best due to its ensemble technique
* Decision Tree tended to overfit
* Linear Regression was decent but couldn't capture non-linear relationships

**5. Cloud Deployment Summary**

* Platform Used: Azure ML Studio
* Tools: Jupyter Notebook on Azure VM
* Process: Uploaded data and notebook, ran preprocessing and model training live
* Output: Visualized evaluation metrics directly from cloud instance

**6. Conclusion**

This project demonstrates the effective application of Big Data techniques using Pandas and Scikit-learn on a real-world bike-sharing dataset. Among the implemented models, the Random Forest model achieved the highest prediction accuracy. The entire pipeline was successfully deployed on the Azure cloud, ensuring scalability and accessibility.