Detailed Project Evaluation and Technical Assessment Report

A Structured Study of Machine Learning Workflows for Classification and Regression

Author: Sarang K A

Role: Data Science Intern

Date: 08 January 2026

1) ABSTRACT AND PROJECT FOUNDATIONS

1.1 Abstract

This report presents a detailed technical evaluation of a machine learning project conducted over a series of sequential experimental stages. The project systematically addresses two fundamental predictive modeling tasks: binary classification for health diagnostics and multi-algorithm regression for student performance prediction. The study emphasizes methodological rigor, data integrity, and progressive skill development through hands-on experimentation. This document critically analyzes each stage of the workflow, from data ingestion to model optimization.

1.2 Introduction

Machine learning solutions require structured workflows, consistent evaluation, and robust validation mechanisms to ensure reliability. During this Data Science Internship, a multi-stage project was undertaken to develop and refine applied machine learning skills. The work is sequenced to reflect a learning progression, beginning with foundational exploratory data analysis (EDA) and gradually incorporating preprocessing strategies, classical modeling techniques, and evaluation metrics.

1.3 Technical Environment

The project was executed using a standardized Python-based data science stack:

- **Pandas (v2.3.3):** For data manipulation and DataFrame management.
- **NumPy (v2.2.6):** Utilized for high-performance numerical operations.

- **Matplotlib (v3.10.8) & Seaborn (v0.13.2):** Core libraries for statistical visualization.
  +1
- **Scikit-Learn (v1.7.2):** Employed for model building, training, and performance evaluation

2) DATA EXPLORATION AND PREPROCESSING

2.1 Classification Task: Diabetes Dataset

The classification component utilized a dataset comprising 768 observations with 9 attributes

Feature Inventory: Key predictors include Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and Age.
Data Quality: Initial analysis using df.isnull().sum() confirmed that the dataset contained zero missing values.
Statistical Insights:

- **Glucose:** Ranged from 0 to 199, with a mean of approximately 120.89.
- **Insulin:** Showed extreme variance with a maximum value of 846, indicating significant outliers.
- **Target Distribution:** The Outcome column revealed that approximately 34.9% of the population tested positive for diabetes.

2.2 Regression Task: Student Performance Dataset

The second dataset focused on predicting a continuous Performance Index based on student activities.

- **Primary Features:** Analysis identified Hours Studied and Previous Scores as the most influential features for predicting academic outcomes.
- **Data Preparation:** The data was partitioned into feature matrices (**X**) and target vectors (**y**) to facilitate supervised learning

3) MODELING METHODOLOGY AND EVALUATION

3.1 Classification Strategies

Two primary algorithms were implemented and compared for the diagnostic classification task:

- **Logistic Regression:** Implemented as a baseline for binary classification performance.
- **Decision Tree Classifier:** Explored for its ability to capture non-linear feature interactions.

3.2 Regression Strategies

A multi-algorithm approach was taken to identify the optimal regressor for student performance:

- **Linear Regression:** Used for its simplicity and strong baseline performance in modeling linear dependencies.
- **K-Nearest Neighbors (KNN) Regressor:** Evaluated for its instance-based learning capabilities.
- **Decision Tree Regressor:** Tested for its ability to map complex relationships, though noted for potential overfitting risks.

3.3 Evaluation Metrics

The models were assessed using a standardized suite of metrics to ensure a holistic view of accuracy and error:

- **Regression Metrics:** Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-Squared (***R^2***)
- **Classification Metrics:** Accuracy, Precision, and Recall scores.

4) INSIGHTS, OPTIMIZATION, AND CONCLUSION

4.1 Key Observations

- **Algorithm Sensitivity:** The performance of the KNN model was found to be highly dependent on the selection of the optimal K value.
- **Risk Management:** Decision Trees demonstrated a tendency to overfit when applied to smaller datasets without proper pruning.
- **Predictive Strength:** In the student performance model, study duration and prior academic history were confirmed as the strongest predictors of the final index.

4.2 Model Optimization and Validation

To ensure reliability, several advanced validation strategies were discussed:

- **Hyperparameter Tuning:** Suggested use of grid-based search strategies to find optimal model settings.

- **Stratified Validation:** Recommended Stratified K-Fold for classification to preserve class distribution across folds.

4.3 Conclusions

This project demonstrates a structured and methodical approach to applied machine learning from the perspective of a Data Science Intern. By transitioning from exploratory analysis to model comparison, the study identifies that the model with the **highest R^2 and lowest RMSE** is the most reliable for practical intervention. The gradual refinement of preprocessing and modeling practices reflects strong technical growth and provides a solid framework for data-driven decision-making.

End of Report