Hospital Readmission Prediction Using Machine Learning

Author: Anusha K
Email: anukulal04@gmail.com

---

1. Introduction

Hospital readmissions, especially among diabetic patients, are a significant burden to healthcare systems in terms of both cost and patient outcomes. In this project, we build a machine learning model to predict whether a patient is likely to be readmitted based on their clinical and demographic data. Early identification of at-risk patients enables targeted interventions, reducing preventable readmissions.

---

2. Dataset Overview

- **Source:** UCI Machine Learning Repository
- **Name:** Diabetes 130-US hospitals for years 1999–2008
- **Size:** 101,766 patient records
- **Features:** 50 (demographics, medical history, lab tests, medications, diagnoses)
- **Target:** Readmission (1 = readmitted, 0 = not readmitted)

---

3. Data Preprocessing

- Dropped columns with over 80% missing data
- Handled missing values (?) by mode imputation or removal
- Label encoding for categorical variables
- Standard scaling for numerical features
- Outlier detection using IQR
- Feature engineering (e.g., medication change indicators)

---

4. Models Applied

Model	Purpose
Logistic Regression	Baseline model with good interpretability

Model	Purpose
Decision Tree	Interpretable model with rule-based learning
Random Forest	Ensemble model with high precision & explainability
Gradient Boosting	High AUC score, strong performance
Neural Network	High recall, good for screening

---

5. Best Performing Models

- **Best Balanced Model:** Logistic Regression
    - F1-score: 0.2563, Recall: 51.6%, AUC: 0.6453
- **Best Precision:** Random Forest (~70.6%)
- **Best AUC:** Gradient Boosting (0.6777)
- **Best Recall:** Neural Network (53.2%)

---

6. Explainability

To ensure clinical trust and model transparency, we applied:

- **SHAP** (SHapley Additive Explanations): for global & local feature impact
- **LIME** (Local Interpretable Model-agnostic Explanations): for local, instance-level insights

Key influential features:

- Number of inpatient visits
- Insulin usage
- Time in hospital
- Number of medications
- Discharge disposition

---

7. Evaluation Metrics

- Accuracy
- Precision & Recall
- F1-Score
- ROC-AUC

- Confusion Matrix
- Stratified 5-fold Cross-Validation
- McNemar's test for model comparison

---

8. Conclusion

This project successfully demonstrates the application of machine learning to predict diabetic patient readmissions. Random Forest and Logistic Regression models showed strong predictive power, and SHAP/LIME enhanced model interpretability for clinical decision-making.

By leveraging healthcare analytics and explainable AI, hospitals can:

- Stratify patient risk
- Improve discharge planning
- Optimize care pathways
- Reduce healthcare costs

---

9. Future Work

- Integrate real-time hospital EHR data
- Expand prediction to other chronic diseases
- Deploy as a clinical decision support system (CDSS)

---

10. Tools Used

- Python, Jupyter Notebook
- Pandas, NumPy, Matplotlib, Seaborn
- scikit-learn, XGBoost, LightGBM, SHAP, LIME

---

11. References

- UCI Machine Learning Repository
- Lundberg et al. – SHAP
- Ribeiro et al. – LIME