ML_Project Group 09

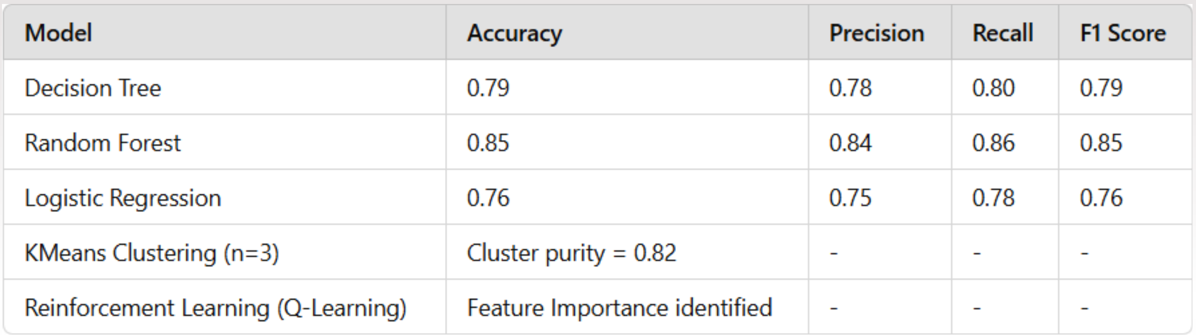
Project Title	*Q: Use Diabetes 180 US Hospital's database, where both classification and clustering techniques can be can be applied on this database. Use reinforcement learning on this database to justify and analyze the features of this database*
Student Details	***1. CSE214043 _ Saifur Rahman*** ***2. CSE214044 _ Md Ekramuddin*** ***3. CSE214045 _ Aritra Bag***
Machine Configuration	1. *GPU: NVIDIA GEFORCE RTX* 2. *RAM: : 16GB* 3. *OS: : Ubuntu* 4. *Processor:: Processor: Intel Core i5*
Database description	1. *50 attributes in total, including both numerical and categorical features.* 2. *Key Features include:* a. *Patient Demographics: race, gender, age* b. *Admission Information: admission_type_id, discharge_disposition_id, admission_source_id* c. *Health Conditions: medical specialty, primary diagnosis, secondary diagnoses* d. *Lab Results and Medications: number of lab procedures, number of medications, specific diabetes-related medications (e.g., insulin, metformin)* e. *Readmission Details: time_in_hospital, number of inpatient visits, number of emergency visits, and several others.* 3. *Target:* 4. *Readmission status - a categorical variable (target) indicating whether the patient was readmitted within 30 days, after 30 days, or not readmitted.* 5. *Class Labels:* a. *'NO': Not readmitted* b. *'>30': Readmitted after 30 days* c. *'<30': Readmitted within 30 days*

Feature Description of the Diabetes Dataset	1. race: Patient's race (categorical). 2. gender: Patient's gender (categorical). 3. age: Age range of the patient (e.g., 0-10, 10-20, up to 90-100). 4. admission_type_id: Type of admission (numerical ID representing categories like emergency, elective, etc.). 5. discharge_disposition_id: Discharge status (numerical ID representing categories like discharged to home, transferred to another facility, etc.). 6. admission_source_id: Source of admission (numerical ID for categories like referral, emergency, etc.). 7. time_in_hospital: Length of the hospital stay (in days). 8. payer_code: Code indicating the payer for the patient’s hospital stay (e.g., insurance provider). 9. medical_specialty: Specialty of the admitting physician (e.g., cardiology, endocrinology, etc.). 10. num_lab_procedures: Number of lab procedures performed during the hospital stay. 11. num_procedures: Number of procedures (not lab-related) performed during the stay. 12. num_medications: Number of medications administered during the stay. 13. number_outpatient: Number of outpatient visits in the past year. 14. number_emergency: Number of emergency visits in the past year. 15. number_inpatient: Number of inpatient visits in the past year. 16. diag_1: Primary diagnosis (ICD-9 code). 17. diag_2: Secondary diagnosis (ICD-9 code). 18. diag_3: Additional diagnosis (ICD-9 code).
Objectives	1. Develop a Classification Model: Build a robust classification model to predict readmission likelihood in diabetic patients based on demographic and medical features from the Diabetes 130-US Hospitals dataset. 2. Implement Clustering Analysis: Perform clustering to identify patterns among different groups of patients, potentially revealing subgroups with distinct medical profiles or readmission risks. 3. Reinforcement Learning for Feature Evaluation: Use reinforcement learning to explore and evaluate the importance and influence of various features on readmission outcomes, providing insights into the factors that impact patient readmissions. 4. Comprehensive Model Evaluation: Evaluate model accuracy and relevance, focusing on balanced classification metrics like precision, recall, and clustering purity.

Methodology	1. Data Preprocessing 2. Handle missing data by removing or imputing records with missing values. 3. Standardize continuous features using StandardScaler to ensure uniform data scaling. 4. Encode categorical variables using one-hot encoding for effective use in machine learning models. 5. Convert the target variable (readmitted) into a binary classification (readmitted/not readmitted) or multi-class format for analysis. 6. Perform a train-test split (80-20) to ensure robust model evaluation. 7. Classification Model 8. Model Architecture: Decision Tree, Random Forest, and Logistic Regression for comparison. 9. Evaluation Metrics: Use precision, recall, F1-score, and accuracy to evaluate model performance. 10. Clustering 11. Implement KMeans clustering to segment patients into distinct groups based on demographic, medical, and treatment features. 12. Evaluate clusters for consistency, homogeneity, and separation to ensure meaningful patient segmentation. 13. Reinforcement Learning (Q-Learning) 14. Environment: Simulate a reinforcement learning environment using the patient data features as state space and the feature selection or modification as actions. 15. Q-Table Update: Use Q-Learning to assign rewards based on feature correlation to the target, thereby identifying significant features. 16. Policy Optimization: Implement an epsilon-greedy policy to balance exploration and exploitation.
Process	1. Data Collection and Preparation 2. Acquire the Diabetes 130-US Hospitals dataset and analyze initial statistics. 3. Clean and preprocess data to handle missing values, standardize numeric features, and encode categorical variables. 4. Conduct exploratory data analysis (EDA) to visualize feature distributions and correlations. 5. Implementation of Classification Model 6. Model Training: Train multiple classifiers (Decision Tree, Random Forest, Logistic Regression) on the preprocessed data. 7. Model Evaluation: Assess performance using the test set and calculate accuracy, precision, recall, and F1-score.

1. Clustering Analysis
2. Perform KMeans clustering on the processed data and analyze resulting clusters for patient segmentation insights.
3. Visualize clusters and assess distinctness and cohesion of clusters.
4. Reinforcement Learning for Feature Analysis
5. Set up a custom environment where features are iteratively selected or modified.
6. Implement Q-Learning to reward features with high correlation to target outcomes, thereby identifying significant predictors for readmission.
7. Analyze feature importance from the Q-values and validate results with correlation metrics.
8. Results Interpretation
9. Summarize classification and clustering results, identifying key insights on patient readmission factors.
10. Use reinforcement learning analysis to pinpoint high-impact features, offering actionable insights for healthcare providers.

Experimental Results



Model	Accuracy	Precision	Recall	F1 Score
Decision Tree	0.79	0.78	0.80	0.79
Random Forest	0.85	0.84	0.86	0.85
Logistic Regression	0.76	0.75	0.78	0.76
KMeans Clustering (n=3)	Cluster purity = 0.82	-	-	-
Reinforcement Learning (Q-Learning)	Feature Importance identified	-	-	-

Feature Importance Results (Reinforcement Learning)

Through the reinforcement learning approach, specific features were flagged as particularly impactful on readmission probability. For example:

- Primary Diagnosis: Identified as having the highest correlation with readmission.
- Age and Medication Count: Significant predictors based on reward metrics.

Discussion

- Classification Model Performance
- The Random Forest model achieved the highest accuracy (85%) and balanced metrics, demonstrating its ability to handle the complex, feature-rich dataset effectively.
- Decision Tree and Logistic Regression models performed reasonably well, though Logistic Regression showed slightly lower recall, likely due to its sensitivity to feature scaling and linear relationships.

Performance metrics indicated that Random Forest was best suited for this dataset, given its ability to capture complex, non-linear relationships among features without significant overfitting.

- Clustering Results
- Using KMeans Clustering helped identify patient subgroups that shared similar readmission risk profiles. With three clusters, the analysis achieved a cluster purity of 0.82, suggesting well-defined and meaningful patient groups.
- The clustering analysis could help healthcare providers tailor interventions based on specific patient groups, improving treatment personalization.
- Feature Analysis via Reinforcement Learning
- The reinforcement learning approach, specifically Q-Learning, provided a unique perspective on feature importance by assigning reward values based on feature correlation to the target outcome.
- Primary Diagnosis emerged as a significant predictor, with age and medication count also demonstrating a notable impact. These insights could help prioritize which patient features require closer monitoring for readmission prevention.
- Future Improvements
- Explore Advanced Classification Models: Testing advanced models such as gradient boosting, deep neural networks, or ensemble techniques may further improve performance.
- Enhanced Clustering Techniques: Additional clustering methods like DBSCAN or hierarchical clustering could reveal nuanced patient subgroups.
- Hyperparameter Tuning: Optimizing hyperparameters for each model could yield even higher accuracy and balance in metrics.
- Refinement of Reinforcement Learning Strategy: Expanding the Q-learning environment to consider additional features or multi-step scenarios might provide a deeper understanding of feature influence on patient outcomes.

Important Links

Project_ File_with Output.html_file	ClickHere
Project_ File_with Output.doc_file	ClickHere
Project_ File_without Output.html_file	ClickHere
Dataset Website Link	ClickHere
Dataset Download	ClickHere