MACHINE LEARNING PROJECT REPORT

Topic: A Streamlit-Based Diabetes Anomaly Detection System Using Machine Learning

➢ Introduction:

Diabetes is a chronic health condition affecting millions worldwide. Identifying unusual patient records in diabetes datasets can help healthcare professionals detect potential errors in data collection, identify patients with exceptional conditions, or flag cases requiring special attention.

This project implements a web-based anomaly detection system for diabetes patient records using machine learning algorithms. The application provides an intuitive interface for healthcare analysts to upload patient data, configure detection parameters, visualize results, and document findings.

➢ Project Overview:

- Objectives

  - ✓ Develop an interactive tool for detecting anomalies in diabetes patient records
  - ✓ Implement multiple machine learning algorithms for comprehensive detection
  - ✓ Create visualization tools to explore and understand anomalies
  - ✓ Build a case review system for documenting findings

- Technologies Used

  - ✓ Frontend: Streamlit (Python web framework)
  - ✓ Machine Learning: Scikit-learn (OneClassSVM, LocalOutlierFactor)
  - ✓ Data Processing: Pandas, NumPy
  - ✓ Visualization: Plotly, Matplotlib, Seaborn

➢ Technical Implementation:

- Data Pipeline

  - ✓ **Data Input**: Users can upload CSV files or use a provided sample dataset
  - ✓ **Preprocessing**: Automatic missing value detection and standardization
  - ✓ **Feature Scaling**: StandardScaler normalizes features for ML algorithms

- Anomaly Detection Algorithms

1. One-Class SVM:
   - ✓ Kernel-based approach that learns a decision boundary
   - ✓ Parameters: nu (expected anomaly fraction), gamma (kernel coefficient)
2. Local Outlier Factor (LOF):
   - ✓ Density-based approach comparing local density of points
   - ✓ Parameters: number of neighbors, expected contamination

- User Interface Components

1. Data Exploration Section:
   - ✓ Data preview and statistical summaries
   - ✓ Missing value analysis
2. Detection Configuration:
   - ✓ Algorithm selection (SVM, LOF, or both)
   - ✓ Parameter tuning via sidebar controls
3. Results Visualization:
   - ✓ Interactive and static scatter plots
   - ✓ Feature selection for axis variables
   - ✓ Anomaly highlighting
4. Case Review System:
   - ✓ Filterable anomaly browser
   - ✓ Note-taking functionality for individual cases
   - ✓ Export capability for documentation

➢ **Results and Analysis:**

- Sample Findings

1. Algorithm Comparison:
   - ✓ One-Class SVM tends to find more global anomalies
   - ✓ LOF detects local density variations better

- ✓ Consensus anomalies (flagged by both) are most likely true anomalies

2. Visual Patterns:
   - ✓ Anomalies often appear at extremes of feature distributions
   - ✓ Some anomalies show unusual combinations of features (e.g., high glucose with normal BMI)
3. Clinical Relevance:
   - ✓ Detected anomalies could represent:
     - Data entry errors
     - Patients with unusual symptom combinations
     - Potential cases requiring special attention

➢ Conclusion:

The Diabetes Patient Anomaly Detection System provides a valuable tool for healthcare data analysis. By combining multiple detection algorithms with an intuitive interface, it enables healthcare professionals to:

- Identify potential data quality issues
- Discover patients with unusual characteristics
- Document findings for further investigation

The system's flexibility allows adaptation to different datasets while maintaining clinical relevance through its case review features.