IT5006 Fundamentals of Data Analytics: Healthcare Readmission Analysis Project

Academic Year 2025/26, Semester 1

📋 Project Overview

This capstone project requires student teams to conduct a comprehensive analytics workflow using real-world healthcare data. Teams will analyze the **Diabetes 130-US Hospitals (1999–2008)** dataset to predict hospital readmissions and understand key factors influencing patient outcomes.

Learning Objectives

By completing this project, students will:

- Apply end-to-end data analytics methodology
- Implement classification and regression algorithms learned in IT5006
- Develop interactive dashboards for stakeholder communication
- Generate actionable business insights from complex healthcare data
- Practice professional reporting and presentation skills

Dataset Information

- **Source**: UCI Machine Learning Repository - Diabetes 130-US Hospitals dataset
- **Access**: https://github.com/uci-ml-repo/ucimlrepo
- **Domain**: Healthcare/Hospital Management
- **Size**: 101,766 patient encounters across 130 hospitals
- **Time Period**: 1999-2008
- **Features**: 50+ attributes covering demographics, clinical data, medications, and outcomes

Dataset Features Overview

Category	Feature	Data Type	Description
Identifiers	encounter_id	Integer	Unique identifier for each hospital encounter
	patient_nbr	Integer	Unique identifier for each patient
Demographics	race	Categorical	Patient race (Caucasian, AfricanAmerican, Asian, Hispanic, Other)
	gender	Categorical	Patient gender (Male, Female, Unknown/Invalid)
	age	Categorical	Patient age in 10-year intervals ([0-10), [10-20), ..., [90-100))
	weight	Categorical	Patient weight in ranges ([0-25), [25-50), ..., [200-300), >300)
Admission Details	admission_type_id	Integer	Type of admission (1=Emergency, 2=Urgent, 3=Elective, etc.)
	discharge_disposition_id	Integer	Discharge destination (1=Home, 3=SNF, 6=Home with service, etc.)
	admission_source_id	Integer	Source of admission (1=Physician referral, 7=Emergency room, etc.)
	time_in_hospital	Integer	Number of days in hospital (1-14 days)
Healthcare Provider	payer_code	Categorical	Insurance/payment information (BC, MC, HM, SP, etc.)
	medical_specialty	Categorical	Specialty of attending physician (73 categories)
Clinical Metrics	num_lab_procedures	Integer	Number of laboratory tests performed
	num_procedures	Integer	Number of procedures (other than lab tests) performed
	num_medications	Integer	Number of distinct generic names administered
	number_outpatient	Integer	Number of outpatient visits in year preceding encounter
	number_emergency	Integer	Number of emergency visits in year preceding encounter
	number_inpatient	Integer	Number of inpatient visits in year preceding encounter
	number_diagnoses	Integer	Number of diagnoses entered to system
Diagnoses	diag_1	Categorical	Primary diagnosis (ICD-9 codes, 700+ categories)
	diag_2	Categorical	Secondary diagnosis (ICD-9 codes)
	diag_3	Categorical	Additional secondary diagnosis (ICD-9 codes)
Laboratory Results	max_glu_serum	Categorical	Glucose serum test result (None, Normal, >200, >300)
	A1Cresult	Categorical	HbA1c test result (None, Normal, >7, >8)

Category	**Feature**	**Data Type**	**Description**
Medications	metformin	Categorical	Diabetes medication dosage (No, Steady, Up, Down)
	repaglinide	Categorical	Diabetes medication dosage
	nateglinide	Categorical	Diabetes medication dosage
	chlorpropamide	Categorical	Diabetes medication dosage
	glimepiride	Categorical	Diabetes medication dosage
	glipizide	Categorical	Diabetes medication dosage
	glyburide	Categorical	Diabetes medication dosage
	pioglitazone	Categorical	Diabetes medication dosage
	rosiglitazone	Categorical	Diabetes medication dosage
	insulin	Categorical	Insulin dosage (No, Steady, Up, Down)
	[Additional medications]	Categorical	23 total diabetes medications tracked
Treatment Changes	change	Categorical	Whether diabetes medication was changed (Ch, No)
	diabetesMed	Categorical	Whether diabetes medication was prescribed (Yes, No)
Target Variables	readmitted	Categorical	Hospital readmission status (NO, <30, >30)

Example Analyses

Regression Tasks: Predict length of stay using relevant patient and clinical features

- Outcome variable: time_in_hospital

Classification Tasks: Predict 30-day readmission risk

- Outcome variable: readmitted (converted to binary: readmitted within 30 days vs. not)

👥 Team Formation & Registration

Requirement	**Details**
Team Size	3 (min) - 4 (max); No individual projects allowed
Registration Deadline	**Sunday, August 24, 2025 at 23:59** (end of Week 2)
Submission	Team member list + project topic confirmation via Canvas
Late Registration	Not permitted - affects final grade

📅 Project Timeline & Deliverables

Phase/Milestone 1: Foundation - Literature Review & Exploratory Data Analysis

Deadline: Sunday, September 14, 2025 at 23:59 (end of Week 5) - **Weight: 20%**

Deliverables:

1. **Literature Review Report** (2 pages)
2. **Exploratory Data Analysis Report** (2-3 pages)
3. **Interactive Dashboard**
   - Built using Streamlit, Tableau Public, or Power BI
   - Submit as live link (include link in submitted report)

Submission Format:

- Combined PDF report (Literature Review + EDA) with dashboard link included
- GitHub repository with all raw code/notebooks

Phase/Milestone 2: Analytics Implementation - Model Building & Evaluation

Deadline: Sunday, October 12, 2025 at 23:59 (end of Week 8) - **Weight: 40%**

Deliverables:

1. **Problem Definition & Data Preparation** (1-2 pages)
2. **Model Implementation & Training** (3-4 pages)
3. **Model Evaluation & Comparison** (2-3 pages)
4. **Model Interpretation & Business Insights** (1-2 pages)

Submission Format:

- Technical report (6-8 pages) as PDF
- GitHub repository with Jupyter notebooks and Python scripts
- Model performance summary tables

Phase/Milestone 3: Integration & Communication - Final Report & Presentation

Weight: 40%

- **Final Report & Presentation Slides Deadline: Sunday, November 2, 2025 at 23:59** (Submit as ZIP file containing: final report PDF, presentation slides, and GitHub repository link)
- **Live Presentations: November 6-13, 2025** (Weeks 12-13, presentation order will be decided by the instructors)
- **Peer Evaluation Deadline: Sunday, November 16, 2025 at 23:59**

Deliverables:

1. **Comprehensive Final Report** (10-12 pages; submit on Canvas)
   - **Due: Sunday, November 2, 2025 at 23:59**
   - **Submit as ZIP file containing:**
     - Final report PDF with GitHub repository link
     - Presentation slides (PowerPoint or PDF format)
2. **Live Presentation** (10 minutes + 5 minutes Q&A)
   - Problem statement and business context
   - Dataset overview and EDA highlights
   - Model development approach and methodology
   - Results and key findings with statistical significance
   - Business recommendations and expected impact
   - **Presentation order will be decided by instructors during Weeks 12-13 (November 6-13, 2025)**
3. **Code Repository** (Required)
   - Well-documented GitHub repository with all project code
   - Clear README with setup instructions and project overview
   - Organized folder structure (data/, notebooks/, src/, docs/)
   - All Jupyter notebooks or Python scripts
   - **GitHub repository link must be included in final report**
4. **Presentation Slides** (Required)
   - Professional slides for live presentation
   - PowerPoint or PDF format
   - 10-minute presentation structure
   - **Submit with final report in ZIP file**
5. **Peer Evaluation** (Confidential; Mandatory to submit)
   - Individual contribution assessment
   - May be used for individual grade adjustment
   - Submitted separately via Canvas survey
   - **Due: Sunday, November 16, 2025 at 23:59**

   - Individual contribution assessment
   - May be used for individual grade adjustment
   - Submitted separately via Canvas survey
   - **Due: Sunday, November 16, 2025 at 23:59**

Submission Format:

- ZIP file containing final report PDF (with GitHub repository link) and presentation slides (due November 2)
- Live presentation during assigned session (November 6-13, order decided by instructors)

🎯 Assessment Criteria

Phase 1 (20%)

- Literature review quality: relevance, credibility, and synthesis of sources
- EDA depth and thoroughness: comprehensive analysis with meaningful insights
- Dashboard functionality: interactivity, design, and usability
- Report clarity: professional writing, organization, and presentation

Phase 2 (40%)

- Problem framing: clear definition of regression and classification tasks
- Data preparation: preprocessing quality and thoughtful feature engineering
- Model implementation: technical correctness and appropriate algorithm selection
- Evaluation rigor: proper metrics, cross-validation, and statistical analysis
- Business interpretation: actionable insights and limitations discussion

Phase 3 (40%)

- Report integration: coherent narrative combining all phases professionally
- Presentation quality: clear communication, engagement, and time management
- Business impact: actionable recommendations with practical value
- Technical excellence: code quality, GitHub repository organization, and thorough documentation

---

🛠️ Technical Requirements

Required Tools

- **Python**: pandas, scikit-learn, matplotlib/seaborn, numpy
- **Dashboard**: Streamlit (recommended), Tableau Public, or Power BI
- **Version Control**: GitHub repository
- **Documentation**: Jupyter Notebooks with markdown explanations

Coding Guidelines

- Clean, well-commented code with meaningful variable names

- Reproducible analysis with random seeds set
- Error handling and data validation checks
- Clear separation of data preparation, modeling, and evaluation phases
- All code organized in GitHub repository with descriptive commit messages
- Comprehensive README file with project setup and execution instructions
- Repository must be accessible when final report is submitted

Data Usage Guidelines

- Use provided UCI dataset as primary data source
- Allowed to use external data sources with proper justification in conjunction with main dataset
- Properly handle missing values and outliers
- Document all data transformations and assumptions
- Ensure reproducibility of results

📚 Support Resources

Academic Support

- **Consultation Hours**: Available on request basis (recommended once every 2 weeks)
- **Canvas Discussion Forum**: For technical questions and peer assistance
- **Email Support**: Course instructors available for guidance

---

⚠️ Important Policies

Academic Integrity

- Cite all sources and reference materials appropriately
- Original analysis and interpretation required
- Collaboration within teams encouraged; between teams discouraged
- Use of AI tools must be declared and appropriately credited

Late Submission Policy (Applicable to ALL phases)

- **0-24 hours late**: 10% penalty
- **24-48 hours late**: 20% penalty
- **>48 hours late**: 50% penalty
- **>7 days late**: Zero marks

Formatting Requirements

- All reports in PDF format with 12pt font, single-spaced
- Figure captions and table labels required
- Professional formatting with consistent styling
- File naming convention: TeamX_MilestoneY_IT5006_AY2526.pdf
- GitHub repository naming: TeamX_IT5006_Healthcare_Analytics_AY2526
- ZIP file naming: TeamX_Milestone3_IT5006_AY2526.zip

Useful Pointers

1. **Start Early**: Begin data exploration immediately after team formation
2. **Regular Meetings**: Schedule weekly team check-ins with documented minutes
3. **Version Control**: Commit code frequently with descriptive messages to your GitHub repository
4. **Seek Guidance**: Schedule consultation sessions once every 2 weeks to stay on track
5. **Focus on Business Value**: Always connect technical findings to practical implications
6. **Practice Presentations**: Rehearse live presentations multiple times and prepare clear, professional slides
7. **Document Everything**: Maintain clear documentation in both notebooks and GitHub README

Good luck with your analytics journey! We look forward to seeing your innovative insights and professional growth.

For questions or clarifications, please contact the course instructors via email or schedule consultation sessions as needed. Students are encouraged to use Canvas discussion forums for technical questions that may benefit all teams.