Institute of Software Engineering

ITS 2122: Python for Data Science & AI - Group Project Specification (Semester 3, 2025)

Project Title: Strategic Patient Risk Stratification & Readmission Predictive Modeling for Vitality Health Network

1. Executive Context: The Pedagogical & Industry Landscape

1.1 The Convergence of Data Science and Value-Based Care

The contemporary healthcare landscape is undergoing a fundamental structural transformation, shifting from a fee-for-service model—where providers are compensated based on the volume of care delivered—to a value-based care paradigm. In this new economic reality, financial incentives are inextricably linked to patient outcomes, operational efficiency, and the long-term management of chronic conditions. For the aspiring data scientist, this domain offers a uniquely rigorous testing ground. Unlike the relatively structured datasets found in retail or finance, healthcare data is characterized by its high dimensionality, sparsity, and deep semantic complexity. It requires the practitioner to navigate cryptic coding systems (such as ICD-9 and CPT), handle sensitive demographic variables with ethical precision, and integrate disparate information sources to form a cohesive patient narrative.

This project specification is designed to simulate this precise environment. By positioning your team as consultants for the fictitious **"Vitality Health Network" (VHN)**, the module moves beyond abstract coding exercises into the realm of decision support systems. The central challenge—reducing hospital readmissions—is not merely an academic exercise but a reflection of the **Hospital Readmissions Reduction Program (HRRP)** established by the Centers for Medicare & Medicaid Services (CMS) **[1]**. Under the HRRP, hospitals face significant financial penalties if they exhibit excessive readmission rates for conditions such as diabetes, heart failure, and pneumonia. This regulatory pressure has transformed readmission prediction from a clinical curiosity into a mission-critical business objective.

The dataset selected for this module, the "Diabetes 130-US Hospitals" dataset from the UCI Machine Learning Repository, serves as an ideal pedagogical instrument **[2]**. Spanning ten years of clinical care (1999–2008) and encompassing over 100,000 patient encounters, it forces your team to confront the "messiness" of real-world data: non-standard missing values, imbalanced target variables, and complex categorical features that require mapping. Furthermore, the requirement to integrate external data via web scraping ensures that your team develops the capability to enrich local datasets with global knowledge—a key

competency in modern data engineering.

2. Project Specification: Student Mandate

2.1 Introduction: The Role of the Health Informatics Consultant

For this capstone project, your group will operate as a high-performance team of **Health Informatics Consultants**. You have been engaged by the executive leadership of **"Vitality Health Network" (VHN)**, a large, integrated healthcare delivery network operating across the United States. VHN is currently facing a "perfect storm" of clinical and financial challenges. The implementation of the Hospital Readmissions Reduction Program (HRRP) by CMS has placed a spotlight on the network's performance metrics.

Specifically, VHN's 30-day readmission rate for diabetic patients has climbed to **18%**, a figure that significantly exceeds the national benchmark. This underperformance is resulting in millions of dollars in annual penalties and, more critically, signals a failure in the continuity of care for the network's most vulnerable chronic disease patients. The Chief Medical Information Officer (CMIO) has provided your team with a de-identified extract of their clinical database and has issued a clear mandate: **"Stop reporting on what happened. Tell us why it happened, and identify who is at risk next."**

Your mission is to leverage the Python data science stack to ingest this raw clinical data, enrich it with external medical context, and build a risk stratification framework that can guide nursing staff and discharge planners. You are expected to deliver a "Strategic Patient Risk Stratification System" that moves beyond descriptive statistics into diagnostic and prescriptive analytics.

2.2 The Core Business Challenge: From Data to Care Management

The VHN leadership has identified several "blind spots" in their current operational view. They have tasked your team with answering the following strategic questions through rigorous data analysis:

- **The Readmission Drivers:** What are the primary factors driving diabetic readmissions? Is the risk driven by clinical severity (e.g., high number of lab tests), operational factors (e.g., discharged to home vs. skilled nursing), or patient history (e.g., number of prior

emergency visits)?

- **Medication Efficacy & Protocol:** How do medication regimens correlate with outcomes? Are patients on Insulin therapy at higher risk compared to those on oral antidiabetics like Metformin? Does a change in medication dosage during the stay indicate stability or instability?
- **Demographic & Social Determinants:** Are there disparities in readmission rates across different age groups, races, or genders? Identifying these trends is crucial for ensuring equitable care delivery.
- **The Complexity Puzzle:** Can we quantify the "complexity" of a patient? The nursing staff needs a simple score—a "Vitality Complexity Index"—to identify high-needs patients at a glance. Can you build an algorithm to calculate this based on the available data?
- **Diagnostic Clarity:** The dataset contains thousands of numerical diagnosis codes (ICD-9). These are meaningless to the administrators. Can you interpret these codes to identify the top co-occurring conditions (comorbidities) that complicate diabetes management?

2.3 Project Objectives

This project is a comprehensive evaluation of the skills acquired in ITS 2122:

- **Data Sanitation (Modules 5 & 6):** You must stabilize a "dirty" dataset. This involves handling non-standard null values (recorded as ?), converting data types (e.g., object to category), and filtering out invalid records (e.g., deceased patients).
- **Web Scraping & Enrichment (Module 8):** You will use the requests and BeautifulSoup libraries to scrape a public ICD-9 coding repository. This external data must be merged with your internal dataset to provide human-readable disease descriptions for the top diagnoses.
- **Feature Engineering (Modules 2 & 6):** You will implement the **Vitality Complexity Index (VCI)**, a custom risk scoring algorithm based on the medical literature (specifically the LACE index concept) **[3]**. This requires writing complex Python functions with conditional logic.
- **Exploratory Data Analysis (Module 7):** You will use Seaborn and Matplotlib to visualize

distributions, correlations, and trends, providing the evidence base for your recommendations.

- **Strategic Reporting:** You will synthesize your technical work into a professional business report that offers actionable recommendations to the VHN board.

3. The Primary Asset: Clinical Data Architecture

3.1 The Dataset: Diabetes 130-US Hospitals

Your analysis will be based on the "Diabetes 130-US Hospitals" dataset [2], a rich repository of clinical data spanning ten years (1999-2008). The dataset characterizes **diabetic care at 130 hospitals across the United States**.

Your first task, which directly tests your file handling capabilities, will be to load the provided data file (in .csv format), **diabetic_data.csv**, into a Pandas DataFrame to begin your analysis. This dataset is unique because it captures the "encounter" level of detail. It includes patient demographics, admission details, laboratory utilization, medication administration, and discharge outcomes.

3.2 Data Dictionary & Feature Semantics

A deep understanding of the features is prerequisite to any analysis. The following table provides a definitive dictionary for the dataset's columns. Your team must familiarize itself with these fields and their specific nuances before proceeding.

Variable Name	Data Type	Description	Key Considerations & Analysis Notes
encounter_id	Int64	Unique identifier for the admission.	Primary key for the dataset.
patient_nbr	Int64	Unique identifier	A single patient may have multiple

		for the patient.	encounters (rows). Crucial for identifying frequent flyers.
race	Object (str)	Patient's race (e.g., Caucasian, AfricanAmerican).	Contains missing values recorded as ?.
gender	Object (str)	Gender (Male, Female, Unknown).	'Unknown/Invalid' entries effectively represent missing data.
age	Object (str)	Age grouped in 10-year intervals.	
weight	Object (str)	Patient's weight in 25-pound ranges.	**Audit Task:** Expected missingness >90%. If found, students must drop this column and document it as a "Data Quality Limitation"
admission_type_id	Int64	Numeric code indicating the type of admission.	**Crucial:** Used to calculate the **Acuity Score (A)** for the VCI. ID 1 (Emergency) = 3

			points
discharge_disposition_id	Int64	Numeric code indicating discharge destination.	**Requires Mapping:** ID 1=Home, ID 11=Expired. **Critical:** Patients who expired (died) cannot be readmitted and must be filtered out.
admission_source_id	Int64	Numeric code indicating who referred the patient.	**Requires Mapping:** ID 7=Emergency Room, ID 1=Physician Referral.
time_in_hospital	Int64	Number of days between admission and discharge.	A key metric for resource utilization and a component of the complexity score.
payer_code	Object (str)	The insurance/payment source (e.g., Medicare, Self-Pay).	**Audit Task:** Often missing because it isn't always mandatory for emergency workflows
medical_specialty	Object (str)	The specialty of the	**Audit Task:** Check

		admitting physician (e.g., Cardiology).	for high missingness (?). This reflects EHR data being a byproduct of billing systems.
num_lab_procedures	Int64	Number of lab tests performed.	A proxy for clinical investigation intensity.
num_procedures	Int64	Number of non-lab procedures (surgeries, scans).	
num_medications	Int64	Number of distinct medications administered.	
number_outpatient	Int64	Outpatient visits in the preceding year.	
number_emergency	Int64	Emergency visits in the preceding year.	**Critical Risk Factor:** High numbers indicate patient instability and chronic "super-utilizer" status.
number_inpatient	Int64	Inpatient visits in the preceding year.	

diag_1	Object (str)	The primary diagnosis (ICD-9 code).	E.g., "250.8" or "428". **Target for Web Scraping Enrichment.**
diag_2	Object (str)	Secondary diagnosis (ICD-9 code).	
diag_3	Object (str)	Tertiary diagnosis (ICD-9 code).	
number_diagnoses	Int64	Total number of diagnoses entered.	A proxy for comorbidity burden (sickness level).
max_glu_serum	Object (str)	Glucose serum test result.	Values: >200, >300, Norm, None.
A1Cresult	Object (str)	HbA1c test result.	Values: >8, >7, Norm, None. Indicates long-term diabetes control.
insulin	Object (str)	Indicates if insulin was prescribed during the encounter and whether the dosage was adjusted.	Values: Up (dose increased), Down (dose reduced), Steady, No.
[22 Medications]	Object (str)	Individual columns for oral drugs (e.g.,	**Efficacy Task:** Use these to compare

		metformin, glimepiride, pioglitazone).	against Insulin users. Values indicate dose changes: "Up", "Down", "Steady", or "No".
change	Object (str)	Did medication change?	Ch (Yes) or No.
diabetesMed	Object (str)	Was any diabetes medication prescribed?	Yes or No.
readmitted	Object (str)	**The Target Variable.**	NO (not readmitted), >30 (readmitted after 30 days), <30 (readmitted within 30 days). **Focus on <30.**

3.3 The ID Mapping Files

Alongside the dataset, you are provided with **IDs_mapping.csv**. This file acts as a relational key, containing the descriptions for admission_type_id, discharge_disposition_id, and admission_source_id. You must programmatically merge or map these descriptions into your main DataFrame. Analyzing "Admission Source 7" is useless; analyzing "Admission Source: Emergency Room" provides insight.

4. The Analytical Blueprint: A Phased Execution

To ensure a methodical and scientifically sound analysis, your team will adhere to the following phased lifecycle. This structure is non-negotiable and ensures that all technical requirements are met systematically.

Phase 1: Data Ingestion & Clinical Sanitation

Objective: Transform raw, "dirty" clinical data into a reliable analytical dataset.
Context: Clinical data is notoriously messy. Systems use different conventions for "missing," and human error is common.

1. **Ingestion & Assessment:** Ingest the given **diabetic_data.csv** file into a Pandas DataFrame. Use .info(), .describe(), and .head() to perform an initial audit. Identify columns that have incorrect data types (e.g., IDs reading as integers instead of categories). Use .columns to audit the full schema. Note that the clinical database contains supplementary medication columns not listed individually (they are listed as '**[22 Medications]**'" in the above given data dictionary (section 3.2); treat these as the 'Oral Medication' cohort for your efficacy analysis.
2. **Standardizing Nulls:** The dataset utilizes the ? character to represent missing values in columns such as race, weight, payer_code, and medical_specialty.
    - ***Task:*** Reload or process the dataframe to interpret ? as a standard NumPy NaN value.
    - ***Analysis:*** Investigate the weight column. If the missingness exceeds 90% (as is typical in this legacy data), you must drop the column entirely. Document this decision in your report as a "Data Quality Limitation".
3. **Handling Deceased Patients:** A patient who dies during their hospital stay cannot be readmitted. Including them in a readmission prediction model introduces noise.
    - ***Task:*** Consult the **IDs_mapping.csv** to identify the discharge_disposition_id codes corresponding to "Expired" (e.g., 11, 19, 20). Filter these rows out of your dataset before proceeding to readmission analysis.
4. **Deduplication Strategy:** While multiple encounters per patient are valid, exact duplicate rows indicate a data entry error. Check for and remove any exact duplicates.

Phase 2: Data Enrichment via Web Scraping

Objective: Contextualize the cryptic ICD-9 diagnosis codes using external knowledge.

Context: The column diag_1 contains codes like "428" or "250.02". To a business executive, these are meaningless. To a doctor, "428" means "Heart Failure." You must build a bridge between these two worlds.

1. **Target Identification:** It is inefficient to scrape thousands of codes. Calculate the frequency of all codes in diag_1 and identify the **top 20 most frequent diagnoses**. These will be your scraping targets.
2. **Scraper Architecture:**
   - Identify a reliable public ICD-9 lookup resource. Options include http://icd9.chrisendres.com/ or the Wikipedia List of ICD-9 codes.
   - Construct a Python script using requests to fetch the HTML and BeautifulSoup to parse it.
   - ***Requirement:*** Your script must dynamically retrieve the "Long Description" or "Disease Name" for the target codes.
   - ***Constraint:*** Implement a time delay (time.sleep(1)) between requests to adhere to ethical web scraping practices and avoid being blocked by the server.
3. **Integration:** Create a new column Primary_Diagnosis_Desc. Map the scraped descriptions to the codes in your dataframe. For codes outside the top 20, label them as "Other" or "Not in Top 20".

Phase 3: Exploratory Data Analysis (EDA)

Objective: Uncover the hidden trends and correlations driving readmissions.

Context: Before building algorithms, you must understand the "shape" of the data.

1. **The Readmission Landscape:** Generate a count plot for the readmitted variable (NO, >30, <30).
   - *Insight:* Identify the class imbalance. Is the <30 group (the HRRP penalty group) a minority? How does this affect your analysis?

2. **Demographic Profiling:**
    - Visualize the distribution of age. Is diabetes predominantly affecting the elderly in this cohort?
    - Analyze readmission rates stratified by race and gender. Create a **grouped bar chart** showing "Readmission by Race and Gender". Analyze if intersectional disparities exist that suggest unequal care outcomes.
3. **Medication Efficacy Analysis:**
    - Compare the readmission rates of patients on **Insulin** versus those on oral medications (i.e one or more of [22 Medications]) or no medication. Does the necessity of Insulin imply a higher severity and thus higher risk?
    - Analyze the change column. Does a change in medication dosage during the encounter correlate with higher readmission, perhaps indicating an unstable patient?
4. **Operational Metrics:**
    - Visualize the relationship between time_in_hospital and num_lab_procedures. Is there a linear correlation?
    - Generate a **correlation heatmap** of numerical features, including num_medications, num_lab_procedures, and time_in_hospital, to identify multicollinearity.
    - Create **box plots** of time_in_hospital grouped by readmitted status (NO vs. <30). Investigate if readmitted patients had longer initial stays and more outliers.
    - Examine discharge_disposition. Do patients discharged to "Skilled Nursing Facilities" have higher return rates than those discharged to "Home"?

Phase 4: Feature Engineering - The "Vitality Complexity Index"

Objective: Translate clinical intuition into a programmable algorithm.

Context: VHN nursing leadership has requested a simplified "Complexity Score" to flag high-risk patients on the floor. You will implement a variant of the LACE Index (Length of Stay, Acuity, Comorbidities, Emergency Visits), a validated tool for predicting readmission risk.

Task: Programmatically create a new column VCI_Score for every patient using the following scoring logic:

1. **L - Length of Stay Score:**
   - time_in_hospital < 1 day: **0 points**
   - 1–4 days: **1 point**
   - 5–13 days: **4 points**
   - = 14 days: **7 points**
2. **A - Acuity of Admission Score:**
   - Check admission_type_id. If the patient was admitted via "Emergency" (ID 1) or "Trauma Center" (ID 7): **3 points**.
   - All other admission types: **0 points**.
3. **C - Comorbidity Burden Score (Proxy):**
   - Using the Charlson Index is too complex for this module. Instead, use number_diagnoses as a proxy for sickness burden.
   - number_diagnoses < 4: **0 points**
   - 4–7 diagnoses: **3 points**
   - >= 8 diagnoses: **5 points**
4. **E - Emergency Visit Intensity Score:**
   - Check number_emergency (visits in the prior year).
   - 0 visits: **0 points**
   - 1–4 visits: **3 points**
   - > 4 visits: **5 points**

Calculation: VCI_Score = L + A + C + E

Stratification & Analysis:

- Categorize patients into three strata:
  - **Low Risk:** VCI < 7
  - **Medium Risk:** VCI 7–10
  - **High Risk:** VCI > 10
- Create a visualization showing the **Readmission Rate (<30)** for each of these three risk categories. Does your algorithm successfully segregate the high-risk patients? This

effectively "validates" your model.

5. Final Deliverables

Your submission will consist of two distinct components: a professional business report and a technical appendix.

5.1 The Strategic Insight Report (PDF)

Audience: VHN Executive Board (CMIO, CEO, Nursing Directors).

Length: 2,500 – 3,000 words (Strict Requirement).

Tone: Professional, persuasive, evidence-based. No code snippets in the main text.

Required Structure:

1. **Executive Summary:** A one-page standalone synopsis. Define the problem (HRRP penalties), your approach (Data Science), the key finding (e.g., "Insulin users are 2x more likely to be readmitted"), and the top recommendation.
2. **Introduction:** Outline the business context. Why is reducing readmissions critical for VHN?
3. **Data Methodology:** Briefly describe the data cleaning and the logic behind the Web Scraping enrichment. Explain *why* you removed deceased patients (methodological rigor) (Phase 1 & Phase 2).
4. **Clinical Insights & Visualizations:** Present the findings from Phase 3. Discuss the demographic and operational drivers of risk. Use your best visualizations here.
5. **The Vitality Complexity Index (VCI):** Explain the logic of your algorithm in plain English. Present the results: Did the High-Risk group actually have higher readmissions? Discuss the utility of this score for nurse staffing (Phase 4).
6. **Strategic Recommendations:** Provide three concrete, data-driven recommendations.
    - *Example:* "Given the high risk associated with Emergency admissions (Acuity Score), implement a mandatory 48-hour follow-up call for all patients admitted via the Emergency Room."

5.2 The Technical Appendix (Jupyter Notebook)

Audience: Lead Data Scientist.

Format: .ipynb file (answers to all phases should be in a single .ipynb file)

Requirements:

- **Reproducibility:** The notebook must run from top to bottom without errors.
- **Documentation:** Extensive Markdown cells must explain the *rationale* behind every step. (e.g., "I am dropping column X because...").
- **Code Quality:** Clean, modular code. Use functions for the VCI calculation and the scraper.

6. Assessment Rubric

The project is graded out of **100 points**.

Criteria	Weight	Evaluation Details
Data Sanitation & Wrangling	**20%**	Correct handling of ? values; appropriate type conversion; filtering of deceased patients (discharge_disposition_id).
Web Scraping Implementation	**15%**	Functional requests/BeautifulSoup script; correct extraction of ICD-9 descriptions; error handling/delays implemented.
Algorithm Logic (VCI)	**15%**	Accurate translation of LACE logic into Python; correct

		use of conditionals/loops; effective risk stratification.
Exploratory Analysis (EDA)	**15%**	Depth of insight; discovery of non-obvious trends; correlation of clinical factors with readmission.
Visualizations	**10%**	Clarity, aesthetics, labeling, and appropriate chart selection (Seaborn/Matplotlib).
Strategic Report Quality	**15%**	Professionalism; business impact; clarity of recommendations; coherence of narrative.
Code Quality & Documentation	**10%**	PEP-8 compliance; use of functions; extensive Markdown commentary explaining the "Why".

7. Student's Reference & Domain Knowledge Base

7.1 The Medical Context: Diabetes and Readmissions

To generate a "rich insight" report, you must understand the underlying medical reality.

- **ICD-9 Coding System:** The International Classification of Diseases (9th Revision) was the standard during the dataset's timeframe. Codes starting with 250 represent Diabetes Mellitus. 250.00 is Type 2 Diabetes; 250.01 is Type 1. Codes starting with 428 (Heart

Failure) or 414 (Ischemic Heart Disease) are common comorbidities. Your scraper will reveal these connections.

- **The LACE Index:** The LACE index is a validated risk assessment tool. Studies show that scores >10 are associated with a significantly higher probability of unplanned readmission. By replicating this logic, you are simulating the deployment of a clinical decision support system.

7.2 Guidance on Data Imbalance

The readmitted column usually shows a class imbalance:

Example:

- NO: ~50-60%
- >30: ~30-35%
- <30: ~10-15% (The Target Class)

  When visualizing, simple count plots will make the <30 bars look insignificant. Students should be encouraged to calculate rates. For example, "12% of female patients were readmitted <30 days, compared to 11% of males." This requires using groupby and value_counts (normalize=True).

7.3 Feature Engineering Nuances

The VCI score logic requires mapping **discrete numerical variables** (time_in_hospital) to specific point values (0, 1, 4, 7).

- **Inefficient:** Iterating through rows with for index, row in df.iterrows().
- **Efficient:** Defining a function **def calculate_L_points(days):** and applying it via **df['time_in_hospital'].apply(calculate_L_points)**.

References:

[1]. Centers for Medicare & Medicaid Services. (2025). *Hospital Readmissions Reduction Program (HRRP)*. U.S. Department of Health & Human Services. Retrieved from https://www.cms.gov/medicare/payment/prospective-payment-systems/acute-inpatient-pps/hospital-readmissions-reduction-program-hrrp

[2]. "Diabetes 130-US Hospitals" dataset from the UCI Machine Learning Repository https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008

[3]. LACE index for readmission https://www.mdcalc.com/calc/3805/lace-index-readmission