**Patient Consultation Data Analysis: Initial Exploration**

**1. Project Overview**

**1.1 Problem Definition**

This analysis aims to develop a machine learning model to predict the probability of a variable called "nbe" based on patient consultation data. Each patient (identified by a unique accident\_number) receives multiple consultations where consultants assess:

* **Pain score (p\_score)**: Values range from 0 (no pain) to 4 (maximum pain)
* **Pain status (p\_status)**: Values are 0 (worse), 1 (unchanged), or 2 (better)
* **Function limitation score (fl\_score)**: Values range from 0 (no limitation) to 4 (maximum limitation)
* **Function limitation status (fl\_status)**: Values are 0 (worse), 1 (unchanged), or 2 (better)

Based on these assessments and established medical guidelines, consultants determine whether a patient is within the expected normal range of recovery (nbe=1), outside this range (nbe=0), or when there's insufficient information (nbe=2).

The goal is to build a machine learning model that can accurately predict the probability of nbe=1 or nbe=0 for new patient consultations.

**1.2 Data Source**

The dataset was extracted from a SQL database and contains 7,491 records representing patient consultations. Each record includes patient identifiers, consultation dates, assessment scores, and the target variable (nbe).

**1.3 Steps Covered**

This first phase covered the "Data Understanding and Exploration" stage of our analysis plan, including:

1. Dataset overview and basic statistics
2. Descriptive statistics and distributions of key variables
3. Data quality assessment
4. Feature relationships analysis
5. Temporal pattern analysis
6. Initial feature importance evaluation

**2. Data Exploration Results**

**2.1 Dataset Overview**

* **Dataset size**: 7,491 consultations for 2,379 unique patients
* **Consultations per patient**: Average of 3.15 consultations per patient (range: 1-16)
* **Data completeness**: No missing values detected in any fields
* **Date range**: Consultations span from 2021 to 2025

**2.2 Key Variable Distributions**

**2.2.1 Target Variable (nbe)**

The target variable shows:

* 55.88% of consultations are within normal expected range (nbe=1)
* 27.95% have insufficient information (nbe=2)
* 16.17% are outside the expected range (nbe=0)

This distribution shows a moderate class imbalance that may need to be addressed during model training.

**2.2.2 Assessment Scores**

* **Pain scores (p\_score)**: Mean of 1.44 (on scale 0-4)
* **Function limitation scores (fl\_score)**: Mean of 1.97 (on scale 0-4)
* **Pain status (p\_status)**: Most consultations show improvement (58%)
* **Function limitation status (fl\_status)**: Most consultations show improvement (57%)

**2.2.3 Consultation Types**

* 51.54% are follow-up consultations
* 27.61% are first contacts
* 16.30% represent case closures
* 3.54% represent cases where patients weren't reached
* 1.01% are classified as complex cases

**2.3 Temporal Patterns**

* **Days since accident**: Average of 119.21 days between accident and consultation
* **Days between consultations**: Average of 34.73 days between consecutive consultations

**2.4 Key Relationships**

**2.4.1 Feature Correlations with Target**

* **Positive correlations with nbe=1**:
  + Higher p\_status (pain improving) correlates with higher likelihood of nbe=1
  + Higher fl\_status (function limitation improving) correlates with higher likelihood of nbe=1
* **Negative correlations with nbe=1**:
  + Higher p\_score (more pain) correlates with lower likelihood of nbe=1
  + Higher fl\_score (more function limitation) correlates with lower likelihood of nbe=1

**2.4.2 Temporal Effects**

Days since accident appears to have a substantial relationship with nbe status, suggesting recovery timelines play an important role in determining whether a patient is within normal expected ranges.

**2.5 Feature Importance**

Based on initial decision tree analysis, the top features for predicting nbe are:

1. Days since accident (37.21%)
2. Pain status (32.19%)
3. Function limitation score (12.95%)
4. Function limitation status (12.15%)
5. Pain score (3.98%)

This suggests that recovery time and improvement trends (particularly for pain) are more influential than absolute scores in determining whether a patient is within normal expected ranges.

**2.6 Telephone Category Impact**

Different consultation types show varying distributions of nbe:

* First contacts (0): 60.35% within normal range
* Follow-up consultations (1): 51.52% within normal range
* Not reached (2): 67.92% within normal range
* Case closed (3): 60.93% within normal range
* Complex cases (4): Only 32.89% within normal range

Notably, complex cases have a much higher rate of being outside normal expectations (38.16%).

**3. Insights and Next Steps**

**3.1 Key Insights**

1. **Temporal factors are critical**: Days since accident appears to be the strongest predictor of whether a patient is within normal expected ranges.
2. **Status more important than absolute scores**: The trend in a patient's condition (improving vs. worsening) appears more predictive than their absolute pain or limitation levels.
3. **Complex cases present challenges**: The relatively small number of complex cases (1% of dataset) have substantially different nbe distributions and may need special consideration.
4. **Class imbalance exists**: With only 16% of cases being outside normal range (nbe=0), strategies to address class imbalance will be important.

**3.2 Next Steps: Feature Engineering**

Based on our data exploration, we'll proceed to feature engineering with a focus on:

1. **Enhanced temporal features**:
   * Create features representing recovery timeline stages
   * Generate consultation sequence-related features
   * Calculate rate-of-change metrics between consultations
2. **Patient-level aggregations**:
   * Create features summarizing a patient's consultation history
   * Develop trajectory features showing improvement/deterioration patterns
   * Calculate deviation from expected recovery timelines
3. **Interaction features**:
   * Generate interaction terms between scores and statuses
   * Create composite recovery indicators combining pain and function limitation
   * Develop features capturing relationship between injury severity and recovery time
4. **Categorical encodings**:
   * Apply appropriate encoding for telephone category
   * Convert date-based features into meaningful time periods

These engineered features will be used to develop models for predicting the probability of a patient being within or outside normal expected ranges in the next phase of our analysis.