**Feature Engineering for NBE Prediction: Documentation**

**Project Overview**

This documentation covers the feature engineering phase of a machine learning project aimed at predicting the Normal-Based Evaluation (NBE) status of patients based on consultation data. The NBE status indicates whether a patient's recovery is within expected norms (1), outside expected norms (0), or has insufficient information (2).

**Objective**

The objective of this phase was to transform the raw patient consultation data into a rich set of features that can better capture the underlying patterns influencing NBE outcomes, thereby improving the predictive performance of subsequent machine learning models.

**Methodology**

**Raw Data Overview**

* 7,491 consultation records for 2,379 unique patients
* 12 original features including identifiers, dates, pain scores, function limitation scores, and their status metrics

**Feature Engineering Categories**

1. **Time-based Features (17 created)**
   * Recovery timeline stages (very\_early, early, mid, late, very\_late)
   * Consultation timing metrics (frequency, density)
   * Date components and time differences
2. **Score-based Features (12 created)**
   * Combined scores (total\_score, score\_ratio)
   * Status indicators (is\_improving, is\_worsening)
   * Normalized scores and comparisons to expected values
3. **Sequential Features (18 created)**
   * Score differences between consultations
   * Cumulative scores and trends
   * Rate of change metrics
4. **Categorical Features (13 created)**
   * Encoded telephone categories
   * Recovery stage indicators
   * Combined categorical variables
5. **Aggregate Features (16 created)**
   * Patient-level statistics
   * Historical NBE proportions
   * Consultation pattern metrics

**Missing Value Handling**

* 31.76% of sequential features had missing values (first consultations)
* 0.32% of recovery stage features had missing values
* Applied appropriate imputation strategies:
  + Median imputation for numerical features
  + Mode imputation for categorical features

**Feature Selection**

Evaluated feature importance using Random Forest and selected the top 50 features based on importance scores.

**Key Findings**

**Most Important Features**

1. **Historical NBE metrics**:
   * patient\_prop\_nbe\_1 (36.65%): Proportion of a patient's consultations with NBE=1
   * prev\_nbe (18.39%): Previous consultation's NBE value
   * patient\_prop\_nbe\_0 (16.12%): Proportion of a patient's consultations with NBE=0
2. **Temporal metrics**:
   * days\_since\_accident (2.02%)
   * days\_per\_consult (1.42%)
   * consult\_density (1.30%)
3. **Recovery indicators**:
   * patient\_prop\_improving (1.09%)
   * p\_score\_cumsum (0.90%)

**Insights**

1. Patient history is the strongest predictor of current NBE status
2. Temporal patterns of recovery significantly influence outcomes
3. Improvement trends appear more predictive than absolute scores
4. Different consultation types show varying NBE distributions

**Technical Implementation**

**Platform and Environment**

* Python 3.13 in PyCharm IDE
* Key libraries: pandas, numpy, scikit-learn, matplotlib, seaborn

**Data Storage**

* Intermediate and final datasets stored as pickle (.pkl) files
* Feature importance metrics saved as CSV
* Visualizations saved as PNG files

**Code Structure**

1. Project structure creation
2. Data loading and initial preprocessing
3. Feature creation by category
4. Missing value handling
5. Feature importance evaluation
6. Final dataset preparation

**Performance Metrics**

* Processed 7,491 records
* Created 81 new features (93 total)
* Selected 55 features for the final model
* Processing completed in seconds

**Results and Deliverables**

1. **Engineered Feature Dataset**
   * 7,491 records with 55 selected features + target variable
   * Stored as engineered\_features.pkl
2. **Feature Importance Analysis**
   * Comprehensive ranking of all features
   * Saved as feature\_importance.csv
3. **Visualization Outputs**
   * Feature importance plots
   * Feature distribution by NBE class
   * Correlation matrices

**Next Step: Model Selection and Development**

The next phase will focus on developing and evaluating machine learning models using the engineered features:

1. **Data Preparation**
   * Split data into training, validation, and test sets
   * Ensure patient-level separation (no patient appears in multiple sets)
   * Handle class imbalance (only 16% of cases are NBE=0)
2. **Model Selection**
   * Test a variety of algorithms:
     + Tree-based models (Random Forest, Gradient Boosting)
     + Linear models (Logistic Regression)
     + Sequential models (if appropriate)
   * Evaluate models on appropriate metrics (AUC-ROC, precision-recall, log loss)
3. **Hyperparameter Tuning**
   * Perform grid or random search for optimal parameters
   * Use cross-validation to ensure robustness
   * Optimize for probability calibration
4. **Model Evaluation**
   * Assess performance on held-out test data
   * Analyze confusion matrices and classification reports
   * Evaluate probability calibration
5. **Model Interpretation**
   * Analyze feature contributions using SHAP values
   * Identify key decision boundaries
   * Create patient-level case studies
6. **Final Model Selection**
   * Compare performance across models
   * Select the best model based on performance and interpretability
   * Document model strengths and limitations

The model development phase will focus on creating a reliable predictor of NBE probabilities that can be used to support clinical decision-making during patient consultations.