Predictive Model Technical Details: Female Farmers Health Study

This document focuses specifically on the predictive elements of the Female Farmers Health Risk Assessment model, including the machine learning approach, training methodology, NLP components, and the dynamic vs. static aspects of implementation.

Machine Learning Approach

Core Prediction Model

The health risk assessment system uses a **Random Forest Regressor** as its core prediction model, configured with the following parameters:

```
RandomForestRegressor(
    n_estimators=100,
    max_depth=10,
    random_state=42
)
```

This model was selected after evaluating several alternatives for the following reasons:

1. **Feature Importance Capabilities**: Random Forests provide built-in feature importance metrics, essential for explaining risk factors
2. **Non-linear Relationship Modeling**: Agricultural health risks involve complex non-linear relationships that Random Forests excel at capturing
3. **Robustness to Outliers**: The dataset contains some outliers in health metrics and work patterns

4. **Ensemble Advantage**: Random Forests' ensemble approach reduces overfitting risk with our limited sample size

The model predicts a continuous risk score (0-100) rather than discrete categories, enabling more nuanced risk assessment and intervention planning.

Hybrid Knowledge-Enhanced Approach

The system uses a **hybrid approach** that combines:

1. **Data-driven ML component**: Uses the Random Forest to learn patterns from the dataset
2. **Domain knowledge rules**: Applies expert-defined weightings for certain variables
3. **Causal factor integration**: Incorporates established agricultural health risk factors

This hybrid approach addresses the limitations of pure machine learning with our small dataset (n=81) while maintaining the benefits of data-driven insights.

Model Training Methodology

Training Data Preparation

The training process uses the following approach:

1. **Feature Engineering**:
    - Conversion of categorical variables (e.g., education level) to ordinal or one-hot encodings
    - Computation of derived features (protection score, chemical risk score)
    - Normalization of numerical features using StandardScaler
2. **Target Variable Construction**:
    - Since direct "health risk" labels don't exist, we constructed a synthetic target variable
    - This target combines known risk factors with weights derived from literature

   - Base formula:

     ```
     risk = 20 + (age > 50) * 15 + (work_hours > 8) * 10 +       (work_days >
     6) * 10 + (10 - protection_score) * 3 +       has_respiratory_conditions *
     15 + has_skin_conditions * 10 +       has_neurological_conditions * 12
     ```

3. **Train-Test Split**:
   - 80% training, 20% testing data
   - Stratified by age groups to maintain demographic distribution

Training Process

The model training follows these steps:

1. **Feature Selection**: Initial features selected based on multivariate analysis results
2. **Model Fitting**: Random Forest trained on the engineered features and synthetic target
3. **Feature Importance Extraction**: Extraction and storage of importance values
4. **Hyperparameter Tuning**: Grid search for optimal n_estimators and max_depth
5. **Model Persistence**: Serialization using joblib to disk for API usage

The training code is implemented in the `train_model()` function in `risk_prediction.py`, which is called by the `/train_model` endpoint when new data is available.

Static vs. Dynamic Elements

The risk assessment system contains both static and dynamic elements:

Static Elements

1. **Domain Knowledge Tables**:
   - Chemical severity ratings (fixed scale 0-10)
   - Protection equipment effectiveness ratings
   - Health condition risk coefficients

2. **Categorical Mapping Dictionaries**:
    - Marital status encodings
    - Socioeconomic status values
    - Employment status categories
3. **Core Risk Formula Structure**:
    - Base risk calculation formula
    - Domain-specific risk adjustment formulas

These static elements encode expert knowledge that doesn't change with new data and ensure consistency in risk assessment.

Dynamic Elements

1. **Learned Feature Importance**:
    - Feature importance weights from the Random Forest
    - Updated when model is retrained with new data
2. **Scaling Parameters**:
    - Feature standardization parameters
    - Recalculated with each model retaining
3. **Risk Thresholds**:
    - Category boundaries for risk levels
    - Can be adjusted based on population distribution

The dynamic elements allow the model to adapt to new data patterns while maintaining the fundamental risk assessment framework.

NLP Components for Text Analysis

Text Processing Pipeline

The system includes sophisticated NLP capabilities for extracting structured information from free-text descriptions, implemented in `nlp_processor.py`:

1. **Tokenization and Preprocessing**:
    - Text normalization (lowercase, whitespace normalization)
    - Tokenization using NLTK's word_tokenize
    - Stopword removal using French stopwords from NLTK
2. **Domain-Specific Keyword Extraction**:
    - Patterns matching for 4 specialized domains:
        - Chemical products (14 terms with severity ratings)
        - Agricultural tasks (13 terms with risk ratings)
        - Health conditions (22 terms with severity ratings)
        - Protection equipment (13 terms with effectiveness ratings)
3. **Entity Recognition and Extraction**:
    - Regular expression patterns for extracting:
        - Age (`\b(\d{1,2})\s*ans\b`)
        - Work experience (`\b(\d{1,2})\s*ans?\s*d\'(expérience|ancienneté)`)
        - Work hours (`\b(\d{1,2})\s*heures?\s*(par jour|\/jour)`)
        - Number of children (`\b(\d{1,2})\s*enfants?\b`)
4. **Context Analysis**:
    - Proximity analysis for related terms (e.g., health issues near chemical mentions)
    - Temporal pattern detection for chronic exposure

Feature Extraction from Text

The text analysis component converts unstructured text into structured features through these steps:

1. **Domain Segmentation**: Text is analyzed in context-specific segments:
    - General description
    - Chemical usage description

   - Task description
   - Health condition description
   - Protection behavior description

2. **Keyword Mapping**: Domain-specific dictionaries map terms to standardized values

   ```
   CHEMICAL_KEYWORDS = {
       "pesticides": 9.0,
       "herbicides": 7.5,
       "fongicides": 7.0,
       # ...other chemicals with severity ratings
   }
   ```

3. **Feature Computation**: Extracted keywords are transformed into model features:
   - Binary features (e.g., has_respiratory_conditions)
   - Numeric features (e.g., work_hours_per_day)
   - Composite scores (e.g., chemical_risk_score)

4. **Confidence Assessment**: Extraction confidence is calculated based on:
   - Pattern match quality
   - Keyword frequency
   - Context clarity

This NLP pipeline enables the model to accept natural language descriptions of farmers' situations and convert them to the same structured format used by the form-based input.

Risk Visualization and Recommendation Components

Risk Calculation and Visualization

The risk calculation produces these key outputs for visualization:

1. **Overall Risk Score (0-100)**:
   - Comprehensive measure of occupational health risk
   - Visualized with a gauge chart showing severity zones
2. **Domain-specific Risk Scores**:
   - Respiratory risk (0-100)
   - Skin risk (0-100)
   - Neurological risk (0-100)
   - Displayed as comparative bar charts
3. **Risk Factors**:
   - Contributing elements to the risk score
   - Presented as ranked list with magnitude indicators
4. **Confidence Interval**:
   - Statistical bounds on the risk estimation (±5%)
   - Displayed as error bars or confidence bands

Dynamic Risk Assessment Elements

The risk assessment includes these dynamic, interactive elements:

1. **Real-time Input Validation**:
   - Immediate feedback on input quality
   - Warning indicators for high-risk combinations
2. **Interactive Prediction Updates**:
   - Form inputs trigger immediate risk recalculation
   - Visual indicators show which factors increase/decrease risk
3. **What-If Scenario Explorer**:
   - Interactive modification of input parameters
   - Side-by-side comparison of risk scores under different scenarios

- Quantified risk reduction for each intervention

Recommendation Generation Engine

The recommendation engine dynamically generates personalized interventions:

1. **Risk Factor Analysis**:
    - Identification of top risk contributors
    - Classification by domain and severity
2. **Intervention Matching**:
    - Mapping of risk factors to potential interventions
    - Prioritization by impact potential
3. **Contextual Customization**:
    - Adaptation based on socioeconomic status
    - Consideration of regional factors
    - Adjustment for practical constraints
4. **Recommendation Formatting**:
    - Presentation of top 3-5 actionable recommendations
    - Classification by implementation difficulty
    - Estimation of risk reduction potential

This recommendation system transforms the risk assessment from informational to actionable, providing concrete steps to reduce occupational health risks.

Integration Points and Data Flow

Form-Based Input Flow

The structured data input follows this flow:

1. User completes the Risk Assessment Form
2. Form data is validated and formatted
3. JSON payload is sent to `/predict_risk` endpoint

4. Backend processes input through feature engineering pipeline
5. Random Forest model generates prediction
6. Risk calculation formulas apply domain adjustments
7. Recommendation engine generates interventions
8. Complete response returned to frontend
9. Dashboard components visualize the results

Text-Based Input Flow

The unstructured text input follows this alternative flow:

1. User provides text descriptions in the Text Analysis Interface
2. Text is segmented by domain (general, chemicals, tasks, health, protection)
3. Frontend sends text to `/predict_risk_from_text` endpoint
4. NLP processor extracts structured features
5. Extracted features follow the same prediction pipeline as form data
6. Confidence indicators reflect extraction certainty
7. Results are returned with the same structure as form-based prediction
8. Dashboard visualizes results with confidence indicators

What-If Scenario Data Flow

The exploration of intervention impacts follows this flow:

1. Initial risk assessment results are displayed
2. User modifies parameters in the What-If Scenario Explorer
3. Frontend calculates preliminary impact estimates
4. Significant changes trigger new API requests
5. Backend calculates precise scenario outcomes
6. Response contains comparison between scenarios
7. Dashboard updates to show intervention impacts

8. Cost-benefit metrics highlight most effective interventions

Machine Learning Model Evaluation

The machine learning component of the system was evaluated using these metrics:

1. **Input-Output Consistency**:
   - Verification that risk increases with known risk factors
   - Confirmation that protective measures reduce risk
   - Testing of edge cases for reasonable outputs
2. **Feature Importance Validation**:
   - Comparison with literature-established risk factors
   - Examination for counterintuitive importance rankings
   - Verification of stability across different model runs
3. **Domain Expert Review**:
   - Expert assessment of risk scores for sample cases
   - Evaluation of recommendation relevance
   - Verification of what-if scenario plausibility
4. **Cross-Validation Performance**:
   - 5-fold cross-validation on the training data
   - Evaluation of mean squared error and R^2 metrics
   - Assessment of prediction variance across folds

These evaluation approaches ensured the model provides reliable, actionable risk assessments despite the constraints of limited labeled outcome data.

Conclusion

The predictive elements of the Female Farmers Health Study dashboard combine sophisticated machine learning with domain expertise to create a comprehensive risk assessment system. The hybrid approach balances data-driven insights with established agricultural health knowledge, while the NLP capabilities extend accessibility to users without technical expertise.

The system's combination of static domain knowledge with dynamically learned patterns creates a robust foundation that can evolve as more data becomes available. The interactive visualization and recommendation components transform complex risk calculations into actionable insights, supporting the project's goal of improving occupational health for female agricultural workers.