學生輟學預測機器學習項目 - 功能規格書（理論基礎版）

1. 項目概述與理論基礎

1.1 研究背景

學生輟學預測是教育數據挖掘的重要應用，透過機器學習技術可以早期識別高風險學生，提供及時的介入措施。

1.2 理論假設

1. **多因素影響假設**：學生的學業成功受個人背景、家庭環境、學業表現、經濟狀況等多重因素影響
2. **早期預警假設**：第一、二學期的學業表現能有效預測最終的畢業狀態
3. **環境因素假設**：社會經濟環境（失業率、通膨、GDP）對學生完成學業有顯著影響

2. 數據集規格

2.1 數據集概要

- **總資料筆數**: 4,424筆
- **訓練集**: 3,524筆（約80%）
- **測試集**: 900筆（約20%）
- **特徵數量**: 36個輸入特徵
- **目標變數**: 1個（三分類）

2.2 目標值定義

- **Dropout（輟學）**: 32.1%
- **Enrolled（仍在學）**: 17.9%
- **Graduate（畢業）**: 49.9%

2.3 特徵類型分析

根據數據特性，將36個特徵分為：

- **類別特徵**: 18個（需要編碼處理）
- **數值特徵**: 18個（部分需要標準化）

- **序數特徵**: 部分類別特徵具有序數性質

3. 特徵工程理論架構

3.1 理論基礎

採用「由粗到細」的三階段特徵選擇策略，結合統計檢定、機器學習和性能驗證，確保選出的特徵具有：

1. 統計顯著性
2. 預測能力
3. 實際效用

3.2 階段1：Filter（過濾法）- 統計篩選

理論依據: 基於統計檢定理論，評估每個特徵與目標變數的相關性

方法選擇理由:

1. **卡方檢定**（類別特徵 vs 類別目標）
   - 適用於類別變數間的獨立性檢定
   - 對大樣本（n=4,424）特別有效
   - 可處理多類別目標變數
2. **ANOVA F-test**（數值特徵 vs 類別目標）
   - 檢定不同類別間數值特徵的均值差異
   - 適合三分類問題的特徵篩選
   - 符合方差分析的理論假設
3. **皮爾森相關係數**（數值特徵之間）
   - 識別高度相關的特徵，避免共線性
   - 用於特徵冗餘檢測
4. **互信息（MI）**（通用）
   - 捕捉非線性關係
   - 不受變數類型限制
   - 理論上更全面的相關性度量

執行策略:

```
1. 卡方檢定：評估18個類別特徵與目標的關係
2. ANOVA：評估18個數值特徵與目標的關係
```

3. 相關性分析：檢測數值特徵間的共線性
4. 互信息：補充非線性關係的檢測

篩選標準:

- p-value < 0.05（統計顯著性）
- 相關係數絕對值 > 0.7（共線性檢測）
- MI分數排名前75%

需要程式碼:

- `feature_eng_01.py`: Filter階段的統計檢定實施

3.3 階段2：Embedded（嵌入法）- 模型驅動篩選

理論依據: 利用具有內建特徵選擇能力的模型，考慮特徵間的交互作用

方法選擇理由:

1. **Lasso回歸**
   - L1正則化的稀疏性原理
   - 自動特徵選擇（係數收縮至0）
   - 對共線性特徵的處理能力
   - 適合作為多類別分類的基礎
2. **隨機森林特徵重要性**
   - 基於不純度減少的重要性計算
   - 對非線性關係敏感
   - 不受特徵尺度影響
   - 提供穩定的重要性評估
3. **Elastic Net**（建議新增）
   - 結合L1和L2正則化優點
   - 處理高度相關特徵群組
   - 比純Lasso更穩定

執行策略:

1. `Lasso`：識別線性重要特徵
2. 隨機森林：捕捉非線性重要性
3. `Elastic Net`：處理相關特徵群
4. 綜合三種方法的結果，取交集或加權平均

篩選標準:

- Lasso/Elastic Net：非零係數
- 隨機森林：重要性排名前N個或累積重要性80%
- 綜合評分：多方法共識

需要程式碼:

- `feature_eng_02.py`: Embedded階段的多模型特徵選擇

3.4 階段3：Wrapper（包裝法）- 性能優化

理論依據: 基於預測性能的特徵子集搜索，找到最優組合

方法選擇理由:

1. **遞迴特徵消除（RFE）with CV**
   - 系統化的後向消除策略
   - 結合交叉驗證避免過擬合
   - 提供特徵數量vs性能的權衡分析
2. **前向逐步選擇**（可選）
   - 從空集開始逐步添加特徵
   - 對比RFE的結果
   - 驗證特徵選擇的穩定性

執行策略:

```
1. 使用隨機森林作為RFE的基礎模型
2. 5-fold交叉驗證評估每個特徵子集
3. 繪製性能曲線找出最佳特徵數
4. 考慮計算成本與性能的平衡點
```

需要程式碼:

- `feature_eng_03.py`: RFE實施與最優特徵集確定

4. 機器學習模型選擇理論

4.1 模型選擇原則

基於「無免費午餐定理」，選擇多種不同類型的模型進行比較：

1. 線性模型（基準）
2. 樹基模型（非線性）
3. 距離基模型（區域性）
4. 神經網絡（深度學習）

4.2 具體模型與理論依據

4.2.1 邏輯回歸（基準模型）

理論基礎: 最大似然估計、線性決策邊界 **適用性分析**:

- 提供可解釋的基準性能
- 對特徵影響的直接量化
- 多類別擴展（softmax）
- 計算效率高

需要程式碼: `prediction_01.py`

4.2.2 XGBoost/LightGBM

理論基礎: 梯度提升、二階泰勒展開、正則化 **適用性分析**:

- 處理混合類型特徵
- 自動處理缺失值
- 對不平衡數據的內建支持
- 高效的特徵重要性計算

需要程式碼:

- `prediction_02.py`: XGBoost實施
- `prediction_03.py`: LightGBM實施

4.2.3 隨機森林

理論基礎: Bootstrap聚合、隨機子空間 **適用性分析**:

- 降低過擬合風險
- 並行計算效率
- 不需要特徵縮放
- OOB誤差估計

需要程式碼: `prediction_04.py`

4.2.4 支持向量機（SVM）

理論基礎: 結構風險最小化、核技巧 **適用性分析**:

- 高維特徵空間優勢
- RBF核處理非線性
- 對異常值的魯棒性
- 理論保證的泛化界

需要程式碼: `prediction_05.py`

4.2.5 多層感知器（MLP）

理論基礎: 通用逼近定理、反向傳播 **適用性分析**:

- 自動特徵學習
- 複雜模式識別
- 端到端優化
- 可擴展架構

需要程式碼: `prediction_06.py`

5. 模型評估與比較框架

5.1 評估指標體系

基於多類別分類的評估理論，採用以下指標：

5.1.1 基礎指標

- **準確率（Accuracy）**: 整體正確率
- **混淆矩陣**: 詳細分類結果
- **分類報告**: 各類別的精確率、召回率、F1

5.1.2 進階指標

- **Macro-F1**: 類別平衡的F1分數
- **Weighted-F1**: 考慮類別比例的F1
- **Cohen's Kappa**: 考慮隨機一致性
- **Matthews相關係數**: 平衡的二元分類指標

5.1.3 概率指標

- **多類別AUC**: One-vs-Rest策略
- **對數損失**: 概率預測的質量
- **Brier分數**: 概率校準度量

5.2 交叉驗證策略

- **分層5折交叉驗證**: 保持類別比例
- **時間序列分割**: 如果數據有時間特性
- **嵌套交叉驗證**: 超參數調優

需要程式碼:

- `DA_term_01.py`：模型評估框架
- `DA_term_02.py`：結果視覺化與報告

6. 實驗設計與執行計劃

6.1 第一階段實驗

1. 執行完整的三階段特徵選擇
2. 訓練所有五種模型
3. 進行初步性能評估
4. 識別最佳模型和特徵集

6.2 第二階段優化（基於第一階段結果）

1. 針對最佳模型進行超參數優化
2. 探索集成學習方法
3. 處理類別不平衡問題
4. 特徵工程的進一步優化

6.3 第三階段部署

1. 最終模型的訓練
2. 模型解釋性分析
3. 部署準備與文檔
4. 監控指標設計

7. 程式碼架構設計

7.1 特徵工程模組

```
feature_eng_01.py  # Filter階段實施
feature_eng_02.py  # Embedded階段實施
feature_eng_03.py  # Wrapper階段實施
feature_eng_04.py  # 特徵工程管道整合
```

7.2 預測模型模組

```
prediction_01.py   # 邏輯回歸
prediction_02.py   # XGBoost
prediction_03.py   # LightGBM
prediction_04.py   # 隨機森林
prediction_05.py   # SVM
prediction_06.py   # 神經網絡
```

7.3 輔助功能模組

```
DA_term_01.py      # 數據預處理與探索
DA_term_02.py      # 模型評估框架
DA_term_03.py      # 結果視覺化
DA_term_04.py      # 報告生成
DA_term_05.py      # 超參數優化
```

8. 預期成果與可交付物

1. **特徵選擇報告**：三階段的詳細分析結果
2. **模型比較報告**：五種模型的性能對比
3. **最佳模型文檔**：包含參數設定與性能指標
4. **可視化儀表板**：展示預測結果與特徵重要性
5. **部署指南**：模型應用的技術文檔

9. 程式碼能力確認

回答: Yes

我可以產生所有規格書中提到的程式碼，每個模組都將包含：

- 完整的功能實現
- 詳細的註解說明
- 錯誤處理機制

- 單元測試範例
- 使用說明文檔

所有代碼將遵循Python最佳實踐，確保可讀性、可維護性和可擴展性。