CRISP-DM 報告

CRISP-DM 報告草稿：葡萄酒品質預測專案

==========

1. 商業理解 (Business Understanding)

1.1 專案目標:
本專案的主要商業目標是利用葡萄酒的物理化學數據，建立一個能夠預測其品質分數的機器學習模型。次要目標是透過模型分析，理解哪些化學特性是影響葡萄酒品質的關鍵因素。

1.2 應用場景:

- **品質輔助評估**: 幫助釀酒師在釀造過程中，透過分析化學成分數據，初步評估或預測最終成品的品質等級。
- **品質改善建議**: 透過理解關鍵特徵的重要性，指導生產過程中的參數調整，以期提高葡萄酒的品質。
- **市場決策**: 根據預測的品質等級，輔助制定價格策略或市場定位。

1.3 成功標準:

- **模型效能**: 建立的迴歸模型在測試集上的 R^2 分數應顯著優於基準模型（例如，僅預測平均值），並提供一個可接受的預測誤差 (RMSE)。
- **業務洞察**: 模型分析結果需能清晰地指出至少 3-5 個對葡萄酒品質有顯著影響的化學指標，並提供可解釋的業務建議。

==========

2. 資料理解 (Data Understanding)

2.1 資料來源:
本專案使用的資料來自 UCI 機器學習資料庫的「紅酒品質資料集」(Red Wine Quality dataset)。

2.2 資料描述:

- 資料集包含 1599 筆紅酒樣本。
- 每筆樣本包含 11 個物理化學特徵（如 fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol）和 1 個感官品質分數（`quality`，範圍從 0 到 10）。

2.3 初步探索 (Initial Exploration):

- **資料結構**: 載入資料後，確認資料維度為 (1599, 12)。
- **缺失值檢查**: 經檢查，整個資料集非常乾淨，沒有任何缺失值。
- **離群值分析**: 透過箱形圖 (Box Plot) 視覺化，發現多數特徵（如 `total sulfur dioxide`）存在顯著的離群值。這些極端值可能會對線性模型的訓練產生不成比例的影響。
- **目標變數分佈**: `quality` 分數主要集中在 5 和 6，呈現出類似常態分佈的趨勢，但其本質為離散的整數，這暗示了此問題未來也可被視為一個分類問題。

==========

3. 資料準備 (Data Preparation)

此階段的目標是將原始資料轉換為適合模型訓練的格式。所有步驟均已在 `app.py` 中實現。

3.1 離群值處理:

- **方法**: 採用 IQR (四分位距) 方法。
- **執行**: 對於每個特徵，計算其 Q1、Q3 和 IQR。將任何低於 `Q1 - 1.5 * IQR` 或高於 `Q3 + 1.5 * IQR` 的值，裁剪 (clip) 到對應的邊界上。此舉有效降低了極端值對模型的影響。

3.2 特徵標準化:

- **方法**: 採用 Z-score 標準化 (StandardScaler)。
- **執行**: 對所有 11 個自變數特徵進行標準化，使其平均值為 0，標準差為 1。這確保了不同單位和量級的特徵在模型訓練中具有平等的貢獻機會，對於線性模型尤其重要。

3.3 資料集切分:

- **方法**: 將處理後的資料集切分為訓練集和測試集。
- **執行**: 採用 80/20 的比例進行切分，並設定 `random_state=42` 以保證實驗結果的可重現性。訓練集用於模型學習，測試集用於評估模型在未見過資料上的泛化能力。

==========

4. 模型建立 (Modeling)

4.1 特徵選擇:

- **方法**: 採用 `SelectKBest` 搭配 `f_regression` 統計檢定。

- **執行**: `f_regression` 計算每個特徵與目標變數 `quality` 之間的線性關係強度 (F-score) 和統計顯著性 (p-value)。根據 F-score 從高到低排序，選出前 `k` 個最重要的特徵。此步驟有助於簡化模型、降低噪音並可能提升預測效能。
- **互動性**: 在 Streamlit 應用中，使用者可以透過滑桿動態調整 `k` 值，觀察不同特徵子集對模型的影響。

4.2 模型選擇與訓練:

- **模型**: 選擇「多元線性回歸 (Multiple Linear Regression)」作為初始的基準模型。此模型簡單、易於解釋，非常適合用來探索特徵間的線性關係。
- **訓練**: 使用經過特徵選擇後的訓練集 (`X_train_selected`, `y_train`) 來訓練線性回歸模型。

==========

5. 模型評估 (Evaluation)

5.1 定量評估:

模型在測試集上得到以下評估指標：

- **R^2 (R-squared)**: 分數偏低（約 0.3-0.4，取決於 k 值），表示目前的線性模型僅能解釋葡萄酒品質約 30-40% 的變異性，預測能力有限。
- **RMSE (Root Mean Squared Error)**: 提供了平均預測誤差的量級，其單位與 `quality` 相同。
- **MAE (Mean Absolute Error)**: 提供了誤差絕對值的平均值，更易於直觀理解平均偏離程度。

5.2 視覺化評估:

- **預測值 vs. 實際值散佈圖**: 圖中的數據點分佈較為發散，並未緊密貼合 45 度完美預測線，再次證實了 R^2 的結論，即模型預測的精確度不高。
- **殘差圖**: 殘差（實際值 - 預測值）大致在 0 線上下隨機分佈，但可以看出在預測值較高或較低時，誤差有變大的趨勢。大部分殘差落在 95% 預測區間內，但仍有部分離群點，表示模型對某些樣本的預測偏差較大。

5.3 綜合結論:

目前的線性模型作為一個起點，成功地識別出如 `alcohol`, `sulphates`, `volatile acidity` 等關鍵特徵。然而，其整體預測能力有限。這很可能是因為葡萄酒品質與化學成分之間的關係並非純粹的線性關係。

==========

6. 部署 (Deployment)

6.1 部署形式:

本專案已透過 **Streamlit** 部署為一個互動式的 Web 應用程式。此應用程式不僅僅是一個模型預測工具，而是將整個 CRISP-DM 流程（從資料載入到模型評估）進行了完整的視覺化呈現。

6.2 應用價值:

- **透明化**: 讓非技術背景的利害關係人（如釀酒師、管理層）也能輕易理解資料探勘的每一步驟。
- **互動性**: 使用者可以親自操作滑桿來選擇特徵數量，即時看到模型結果的變化，增強了對模型的信任感和理解深度。
- **快速原型**: Streamlit 的方式非常適合快速搭建數據科學專案的原型，用於展示和收集回饋。

6.3 未來部署建議:

- **模型迭代**: 根據「評估」階段的建議（如改用隨機森林等非線性模型），在應用中加入更多模型選項供使用者比較。
- **API 化**: 對於生產環境，可以將訓練好的最佳模型封裝成 REST API，以便與其他生產系統（如釀造監控系統）進行整合。
- **定期再訓練**: 建立一個自動化的流程，定期使用新的葡萄酒數據對模型進行再訓練和更新，以保持其時效性。