**專案標題 (Project Title)**

**天、地、人：天氣變化如何影響葡萄酒的風味與品質**

**摘要 (Abstract)**

葡萄酒品質的影響因素通常歸納為「天、地、人」三大要素。「天」指的是氣候條件，包括陽光、溫度和降雨量等環境因素；「地」則代表葡萄生長的土壤類型、地形和地理位置，這些自然條件對葡萄的成熟和風味有深遠的影響；「人」則是指釀酒過程中，釀酒師的技術、選擇和經驗，從葡萄的採收到發酵和陳年方式都至關重要。這三者共同作用，決定了葡萄酒的品質和風味特徵。

本研究旨在通過觀察葡萄產地的天氣變化，深入分析其對葡萄酒品質的影響，並進一步預測葡萄酒的品質。研究將重點放在氣候條件如溫度、降水量和日照時間等因素，如何在不同生長季節中影響葡萄的成熟度與風味，進而影響釀製出的葡萄酒品質。同時，本研究將應用數據分析和機器學習技術，建立預測模型，準確預測葡萄酒的品質。透過此分析，研究期望提供有助於葡萄農和釀酒師優化栽培及釀造策略的實用洞見，以提升葡萄酒的整體品質。

**引言 (Introduction)**

葡萄酒的品質和風味深受多種因素影響，而在傳統釀酒理論中，這些因素常歸納為「天、地、人」三者。「天」代表氣候條件，包括陽光、溫度、降水等自然環境；「地」則指葡萄生長的土壤、地形與地理位置，影響葡萄的風味和成熟度；「人」則是釀酒師的技藝與經驗，從栽培到釀造的過程都決定了葡萄酒的最終品質。

其中，氣候變化對葡萄酒的影響尤為顯著。隨著全球氣候變遷，氣溫上升、降水模式改變，這些天氣因素直接影響了葡萄的生長週期、糖分含量以及酸度平衡，從而塑造了葡萄酒的風味特徵。現今，越來越多的研究試圖通過深入分析天氣數據，來預測葡萄酒的品質，並幫助釀酒師更好地適應氣候變化，提升釀酒技術。

本研究的目的是透過觀察和分析葡萄產區的氣候變化，探討天氣對葡萄酒品質的具體影響，並進一步建立預測模型，為未來的葡萄栽種與釀造策略提供參考依據。透過結合傳統「天、地、人」的釀酒理念，研究旨在找出氣候變化與葡萄酒風味之間的具體關聯性，以期在現代釀酒工藝中應用這些新發現，提升葡萄酒的整體品質與市場競爭力。

**相關研究 (Related Work)**

**在葡萄酒品質與風味的研究中，現有文獻大致可分為兩個主要方向：一是通過分析葡萄酒的評論和評分，二是通過分析葡萄酒的成分來預測品質。**

**1. 透過評論和評分進行的研究：**  
第一類研究著重於收集消費者對葡萄酒的評價，並使用這些數據建立機器學習模型來預測葡萄酒的評分。例如，某研究通過從 Vivino.com 抓取資料，創建了一個包含不同類型葡萄酒（紅酒、白酒、玫瑰酒和氣泡酒）的數據集，數據集包含八個特徵欄位，如評分數量、評分值和葡萄酒種類。研究者發現，由於一些葡萄酒的評分數量過少，評分可能缺乏代表性，從而難以反映其真實品質。基於這一點，研究者構建了機器學習模型，嘗試預測那些評分數量較少但具有潛力的葡萄酒的評分了消費者評價在葡萄酒市場中的重要性，並指出新產品往往因缺乏評價而無法獲得市場關注，導致潛在銷量損失。因此，預測葡萄酒評分模型不僅有助於消費者更好地選擇產品，還能幫助生產商提高新產品的曝光率和市場需求。

**2. 透過成分分析進行的研究：**  
另一類研究則側重於葡萄酒的物理和化學成分分析，以評估其品質和風味。這些研究通常通過分析葡萄酒的成分，如酒精含量（ABV）、葡萄品種、風味特徵、年份、產區等，來預測其最終品質。某研究中的數據集涵蓋來自世界各地的葡萄酒，詳細記錄了葡萄酒的名稱、描述、價格、葡萄品種、瓶塞類型、酒精含量、地區、風格及其風味特徵等。這類研究的目的是利用這些特徵來構建預測模型，評估葡萄酒的風味、香氣及市場價值 。

這類研究的析葡萄酒的成分，可以對其品質和市場價值進行更加精確的預測，這對葡萄酒生產商在市場定位和定價策略上具有重要意義。

綜上所述，無論是基於評論與評分，還是基於成分的分析，這兩種方向的研究均試圖通過數據驅動的方式來預測和提升葡萄酒的品質與市場表現。然而，這些研究大多集中於葡萄酒的結果評估，而對於影響葡萄生長的外部自然條件，如天氣變化，探討仍然不足。本研究旨在彌補這一空白，通過觀察葡萄生產地的氣候變化，來分析天氣如何影響葡萄酒的風味與品質，並建立相關的預測模型。

**擬定計劃 (Proposed Work)**

本研究旨在通過結合葡萄酒評分數據與氣象數據，探討天氣變化如何影響葡萄酒的風味與品質。具體來說，本研究將利用兩個包含不同產地葡萄酒的數據集（紅酒與白酒），並根據每個產地的地理位置，查詢當地的歷史天氣資料，進行相關分析。以下為本研究的具體步驟：

**1. 數據集整合：**  
研究將首先對兩個數據集進行清理與整合，確保數據的一致性與可用性。每個數據集包括葡萄酒名稱、國家、地區、酒莊名稱、評分（Rating）、評分數量（NumberOfRatings）、價格（Price）及年份（Year）等關鍵變數。這些變數將用於後續的統計分析和模型訓練。

**2. 產地座標查詢：**  
為了將葡萄酒產地與氣象數據進行匹配，本研究將基於數據集中提供的地區名稱，利用地理編碼技術查詢每個地區的經緯度座標。

**3. 氣象站匹配與天氣數據獲取：**  
接下來，本研究將使用產地的經緯度，查詢產地的歷史天氣數據。具體的天氣變數將包括：

* 日平均溫度
* 降雨量
* 日照時數

這些天氣變數將對應到葡萄的生長季節，進一步分析它們對葡萄酒品質的影響。

**4. 數據分析與模型建構：**  
在數據整合完成後，研究將運用統計分析方法和機器學習模型來探索天氣變量與葡萄酒評分之間的關聯性。具體分析步驟包括：

* **相關性分析：** 探討不同天氣變數（如溫度、降水量等）與葡萄酒評分之間的相關性。
* **回歸模型：** 建立多元回歸模型，分析天氣變數對葡萄酒評分的預測能力。
* **機器學習模型：** 使用隨機森林（Random Forest）機器學習方法，進行葡萄酒品質的預測。

**5. 結果驗證與討論：**  
最後，研究將對模型的預測結果進行驗證與討論，分析天氣條件對葡萄酒品質的具體影響，並根據結果提出相應的釀酒與栽培建議。特別是，研究將強調如何根據氣候變化調整葡萄的種植和釀造技術，以提高葡萄酒的品質。

本研究的成果將有助於釀酒師和葡萄農理解天氣變化對葡萄酒品質的影響，並為未來應對氣候變化提供科學依據。

### **評估與驗證 **(Evaluation and Validation)****

**Lasso 回歸結果**：在進行 Lasso 回歸時，我們選擇了葡萄酒各個品類的季節性天氣數據進行模型訓練，結果顯示出選取的主要特徵集中在天氣數據中的**日照時間 (tsun)**，包括**Spring\_tsun**、**Summer\_tsun** 和 **Fall\_tsun**。Lasso 模型對不同品類葡萄酒的預測表現如下：

* **Red 葡萄酒**：最終 MSE 為 0.0903。
* **White 葡萄酒**：最終 MSE 為 0.0615。
* **Rose 葡萄酒**：最終 MSE 為 0.0805。
* **Sparkling 葡萄酒**：最終 MSE 為 0.0711。

根據 Lasso 回歸的結果，我們發現**日照時間 (tsun)** 對葡萄酒評分具有較高的影響力，這可能是因為日照時間直接影響到葡萄的成熟度和風味，進而影響葡萄酒的品質評分。

**多重共線性 (VIF)**：根據**方差膨脹因子 (VIF)** 的結果，不同季節的天氣變量，如**Spring\_tavg**、**Summer\_tavg** 等，具有非常高的 VIF 值，這意味著它們之間可能存在強烈的線性相關性。這是由於季節性天氣數據高度相似，特別是同一季節內的溫度和降水量等特徵。

為了處理這個問題，我們使用了**PCA** 降維技術來減少這些特徵之間的相似性，並且使用了 Lasso 回歸進行特徵選擇。透過這些方法，我們成功地降低了多重共線性的影響，並保留了對評分預測影響最大的天氣變數。

在葡萄酒評分預測專案中，運用了三種不同的模型來評估天氣對葡萄酒評分的影響 ：**主成分分析 (PCA)**、**隨機森林回歸**、以及**深度學習模型**。每種方法都使用了**交叉驗證**來進行驗證，以確保模型的穩健性。

* **效能指標**：主要使用**均方誤差 (Mean Squared Error, MSE)** 來評估模型的預測效能。PCA 模型的 MSE 為 0.067，隨機森林的 MSE 為 0.068，而深度學習模型的 MSE 則為 0.266。相比之下，隨機森林和 PCA 的效能表現更為優秀。
* **效率評估**：三個模型在處理大量數據時的效率不同。隨機森林和 PCA 的計算速度較快，而深度學習模型則需要更多的運算資源和時間。儘管深度學習模型能夠捕捉更為複雜的非線性關係，但其在本專案中的效果並不顯著，可能是因為資料特徵間的線性關聯較強。
* **交叉驗證**：使用**5 折交叉驗證**來驗證模型的表現，以避免過度擬合。隨機森林的交叉驗證 MSE 為 0.072，而 PCA 的交叉驗證 MSE 為 0.071，均顯示了較穩健的模型表現。

### **時間規劃與挑戰 **(Timeline and Challenges)****

* **時間表設置**：
  1. 資料整理與前處理：2 週
  2. 模型選擇與調整：2 週
  3. 交叉驗證與結果分析：1 週
  4. 報告撰寫與提交：1 週
* **挑戰預估**：
  1. **資料特徵工程**：不同季節的天氣數據之間存在高度的線性相關性，這會導致多重共線性問題，影響模型的準確性和解釋力。為了解決這個問題，我們使用了**PCA** 降維技術，並結合 Lasso 進行特徵選擇，篩選出與葡萄酒評分相關性最高的天氣特徵。
  2. **運算資源**：深度學習模型在訓練時需要更多的計算資源，尤其是處理較大的數據集時，計算時間顯著增加。應準備額外的運算資源以加快處理。

### **結論與未來工作 **(Conclusion and Future Work)****

* **總結**：本專案通過對天氣數據進行分析，成功地構建了幾個模型來預測葡萄酒的評分。經過比較，**隨機森林回歸**和**主成分分析 (PCA)** 模型在效能與計算資源之間取得了良好平衡，顯示出較低的 MSE 值。
* **未來工作**：
  1. **擴展資料集**：未來可以納入更多的天氣變量（如風速、濕度）來進一步提高模型預測的準確度。
  2. **其他模型探索**：可以嘗試**支持向量回歸 (SVR)** 或**增強學習**來捕捉更複雜的數據模式，並優化深度學習模型的結構以提高效能。