# ワイン品質予測における機械学習手法の比較研究

学籍番号:2515111031 氏名: 小西隆一

2025 年 9 月 21 日

## 1　序論

近年、機械学習技術の発展により、食品の品質評価における客観的な予測手法が注目されている。特に、ワインの品質評価は伝統的に専門家による官能評価に依存してきたが、理化学的な成分分析データから品質を予測することができれば、品質管理の効率化や客観性の向上が期待できる。

本研究の目的は、ワインの理化学的特性から品質を予測する回帰モデルの構築と評価である。具体的には、酸度、糖分、アルコール度数などの 11 の理化学的特徴量を用いて、3 つの異なる機械学習手法（線形回帰、多層パーセプトロン、サポートベクター回帰）による品質予測性能を比較検討する。これにより、ワイン品質予測における各手法の特徴と有効性を明らかにし、実用的な品質予測システムの基盤となる知見を得ることを目指す。

## 2　データと前処理

### 2.1　データセットの概要

本研究では、UCI Machine Learning Repository で公開されている Portuguese ”Vinho Verde” wine dataset の赤ワインデータを使用した。このデータセットは、ポルトガルの「ヴィーニョ・ヴェルデ」地域で生産された赤ワイン 1,599 本について、理化学的特性と品質評価を記録したものである。

データセットには以下の 11 個の特徴量が含まれている：

- **固定酸度（fixed acidity）**：主にタルタル酸などの不揮発性酸の濃度（g/dm$^3$）
- **揮発性酸度（volatile acidity）**：主に酢酸の濃度で、高すぎると不快な酢酸臭を生じる（g/dm$^3$）
- **クエン酸（citric acid）**：少量で酸味と清涼感を与える（g/dm$^3$）
- **残留糖分（residual sugar）**：発酵後に残った糖分の量（g/dm$^3$）
- **塩化物（chlorides）**：塩分濃度（g/dm$^3$）
- **遊離亜硫酸（free sulfur dioxide）**：微生物増殖と酸化を防ぐ（mg/dm$^3$）
- **総亜硫酸（total sulfur dioxide）**：遊離形と結合形の亜硫酸の総量（mg/dm$^3$）
- **密度（density）**：ワインの密度（g/cm$^3$）
- **pH**: 酸性度を示す指標（0-14 スケール）
- **硫酸塩（sulphates）**：抗酸化剤として使用される硫酸カリウム等（g/dm$^3$）
- **アルコール度数（alcohol）**：エタノール含有量（% vol）

目的変数である品質（quality）は、ワイン専門家による官能評価により 0-10 のスケールで評価されているが、実際のデータでは 3-8 の範囲に分布している。品質の分布を詳細に見ると、5 と 6 の評価が最も多く、極

端に良い（8）や悪い（3）評価は少数である。

## 2.2 前処理手順

データの前処理は以下の手順で実施した。

**データ分割**: 全データを学習用（70%）、検証用（15%）、テスト用（15%）に分割した。具体的には、1,599本のワインを学習用 1,119 本、検証用 240 本、テスト用 240 本に分割した。この分割比率は、十分な学習データを確保しつつ、検証とテストで適切な評価を行うために設定した。

**学習・検証・テストデータの役割**: 学習データはモデルのパラメータを最適化するために使用し、検証データは学習中の早期終了判定とハイパーパラメータ調整に使用した。テストデータは最終的なモデル性能評価にのみ使用し、学習過程では一切使用しないことで、未知データに対する汎化性能を客観的に評価した。

**標準化処理**: 各特徴量は平均 0、標準偏差 1 になるよう標準化を実施した。これは、特徴量間でスケールが大きく異なるため（例：pH は 3-4 の範囲、総亜硫酸は 6-289 の範囲）、学習の収束性を向上させ、各特徴量の寄与を公平に評価するために必要である。標準化は学習データの統計量を用いて実施し、検証・テストデータには同じ変換を適用した。

# 3 手法

## 3.1 評価対象モデル

本研究では、回帰問題における代表的な 3 つの手法を比較対象とした。

### 3.1.1 線形回帰（Linear Regression）

線形回帰は最も基本的な回帰手法であり、入力特徴量の線形結合により出力を予測する。本実装では、PyTorch を用いてニューラルネットワーク形式で実装し、入力層から出力層への直接的な線形変換を行う。活性化関数は使用せず、純粋な線形回帰モデルとして構築した。重みの初期化には Xavier uniform 初期化を採用し、安定した学習を実現した。

### 3.1.2 多層パーセプトロン（MLP: Multi-Layer Perceptron）

多層パーセプトロンは、複数の隠れ層を持つフィードフォワードニューラルネットワークである。本研究では、入力層（11 次元）$\rightarrow$ 隠れ層 1（128 次元）$\rightarrow$ 隠れ層 2（64 次元）$\rightarrow$ 出力層（1 次元）の 4 層構造を採用した。各隠れ層には ReLU 活性化関数を適用し、非線形性を導入することで複雑なパターンの学習を可能にした。過学習を抑制するため、各隠れ層後にドロップアウト層を配置した。

### 3.1.3 サポートベクター回帰（SVR: Support Vector Regression）

サポートベクター回帰は、サポートベクターマシンの回帰版であり、$\varepsilon$-insensitive 損失関数を用いてロバストな予測を行う。本研究では、scikit-learn の実装を使用し、線形カーネルを採用した。SVR は外れ値に対してロバストな特性を持ち、高次元データにおいても効果的に動作することが知られている。

## 3.2 ハイパーパラメータ調整

各モデルについて、グリッドサーチによる系統的なハイパーパラメータ調整を実施した。

**線形回帰のハイパーパラメータ**:

- 学習率（learning rate）：{0.001, 0.005, 0.01, 0.02, 0.05}
- ドロップアウト率（dropout rate）：{0.0, 0.05, 0.1, 0.2}
- 重み減衰（weight decay）：{0.0, 0.001, 0.01}

**MLP のハイパーパラメータ**:

- 学習率: {0.001, 0.002, 0.005, 0.01}
- ドロップアウト率: {0.0, 0.1, 0.2}
- 重み減衰: {0.001, 0.01, 0.1}
- 隠れ層構造: (128, 64) で固定

**SVR のハイパーパラメータ**:

- 正則化パラメータ（C）：{0.1, 1.0, 10.0, 100.0}
- $\varepsilon$-tube parameter: {0.01, 0.1, 0.2, 0.5}
- カーネル係数（gamma）：{'scale', 'auto', 0.001, 0.01, 0.1, 1.0}

## 3.3 学習設定

**損失関数**: PyTorch ベースのモデル（線形回帰、MLP）では平均二乗誤差（MSE）を損失関数として使用した。SVR では標準的な $\varepsilon$-insensitive 損失を使用した。

**最適化手法**: PyTorch ベースのモデルでは Adam 最適化手法を採用した。Adam は適応的学習率を持ち、回帰問題において安定した収束性を示すことが知られている。

**早期終了条件**: 過学習を防ぐため、検証データの RMSE が 20 エポック連続で改善しない場合に学習を停止する早期終了を実装した。これにより、最適な汎化性能を示す時点でのモデル重みを保存した。

**バッチサイズ**: 全モデルでバッチサイズ 64 を使用した。これは、データセット規模と計算効率のバランスを考慮して設定した。

## 3.4 評価手法

モデルの性能評価には以下の 3 つの指標を使用した：

**平均平方根誤差（RMSE）**:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

予測値と実測値の差の大きさを直感的に理解しやすい指標である。

**平均絶対誤差（MAE）**:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

外れ値の影響を受けにくく、平均的な予測誤差を表す。

**決定係数（$R^2$）**:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

モデルが目的変数の分散をどの程度説明できるかを示す。1 に近いほど高い予測性能を意味する。

### 3.5 実験手順

実験は以下の手順で実施した：

1. **データ前処理**: データ読み込み、分割、標準化を実行
2. **ベースライン評価**: 各モデルのデフォルトパラメータでの性能を測定
3. **ハイパーパラメータ調整**: 各パラメータを段階的に最適化
4. **最終評価**: 最適パラメータでの学習・評価を実施
5. **可視化**: 学習曲線、予測値散布図、残差分布を生成

各実験において、学習過程の詳細な記録を取り、モデルの収束性や過学習の有無を確認した。

## 4 結果

### 4.1 最終性能比較

各モデルの最適ハイパーパラメータでの性能を表 1 に示す。

**表 1:** 各モデルの性能比較

| モデル | Test RMSE | Test MAE | Test $R^2$ | 最適パラメータ |
|---|---|---|---|---|
| MLP | **0.623** | 0.502 | **0.416** | lr=0.002, dropout=0.0, weight_decay=0.01 |
| 線形回帰 | 0.628 | **0.501** | 0.407 | lr=0.02, dropout=0.05, weight_decay=0.0 |
| SVR | 0.631 | 0.516 | 0.400 | C=100.0, $\varepsilon$=0.1, gamma=scale |

MLP が最も優秀な性能を示し、Test RMSE で 0.623、$R^2$ で 0.416 を記録した。線形回帰は僅差で 2 位となり、Test RMSE で 0.628 を記録した。SVR は 3 位の結果となった。

## 4.2 学習過程の分析

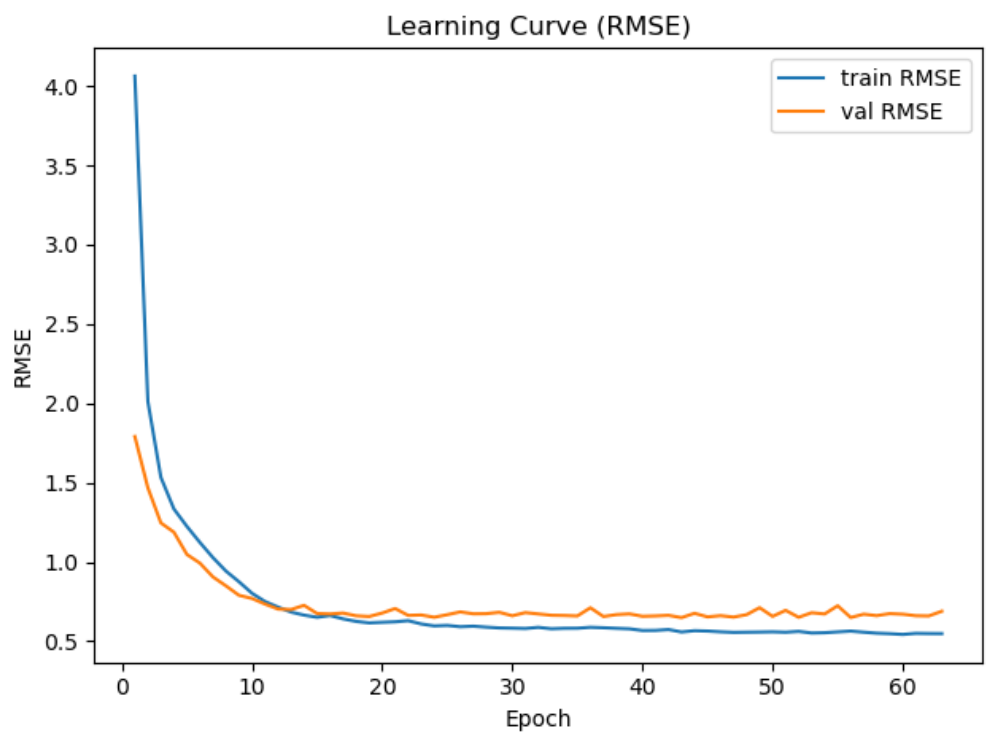


**図 1:** MLP 学習曲線: 訓練用 RMSE と検証用 RMSE の推移を示す。early stopping により適切な時点で学習が停止されている様子が確認できる。

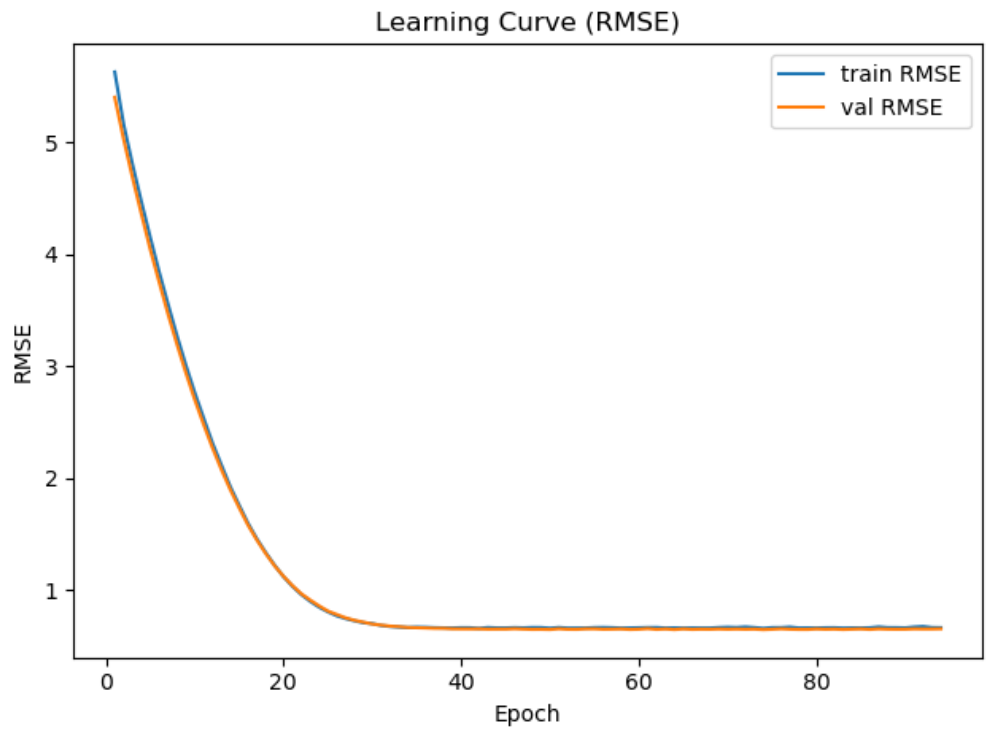


**図 2:** 線形回帰学習曲線: MLP と比較して早期に収束している様子が観察される。

図 1 と図 2 に示すように、各モデルの学習過程を分析すると、MLP は約 100-150 エポックで最適な性能に達し、早期終了により過学習を適切に回避できた。線形回帰は約 80-100 エポックでの収束を示し、より短時間での学習が可能であった。SVR は非反復的なアルゴリズムのため、グリッドサーチによる最適化のみを実

施した。

## 4.3　予測精度の詳細分析

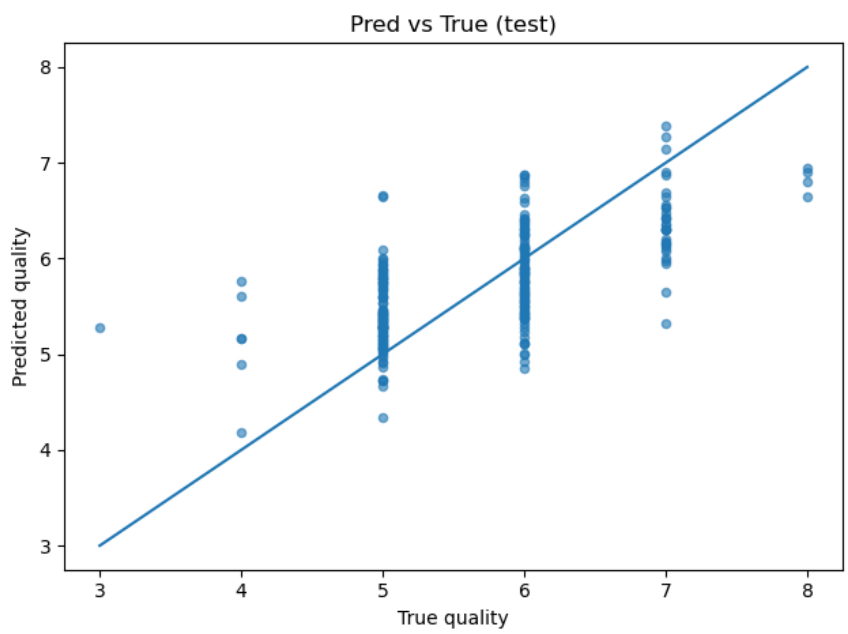


**図 3:** 予測値 vs 実測値散布図（MLP）：MLP の予測値と実測値の関係を示す散布図。理想的な $y = x$ 線に近い分布を示している。

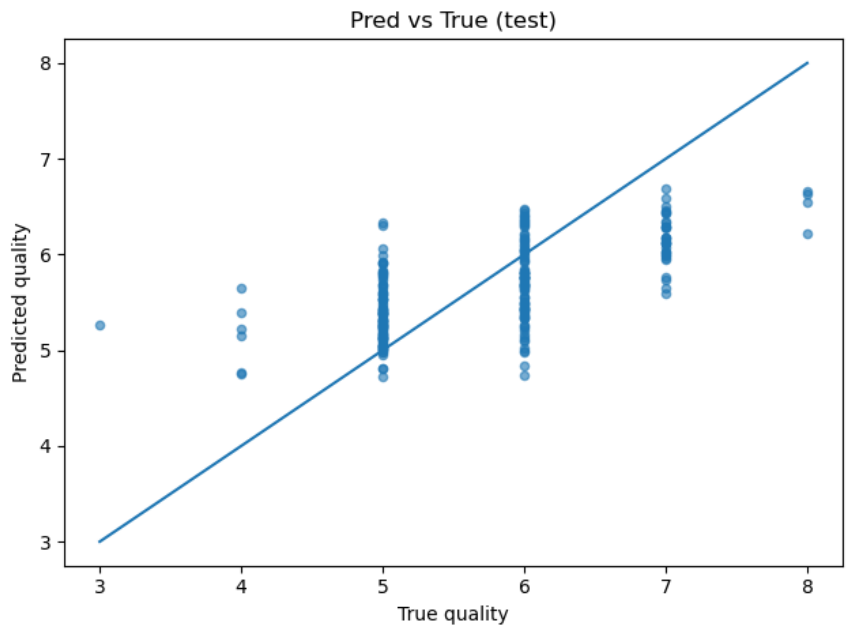


**図 4:** 予測値 vs 実測値散布図（線形回帰）：線形回帰の予測値と実測値の関係。MLP とほぼ同様の分布パターンを示している。

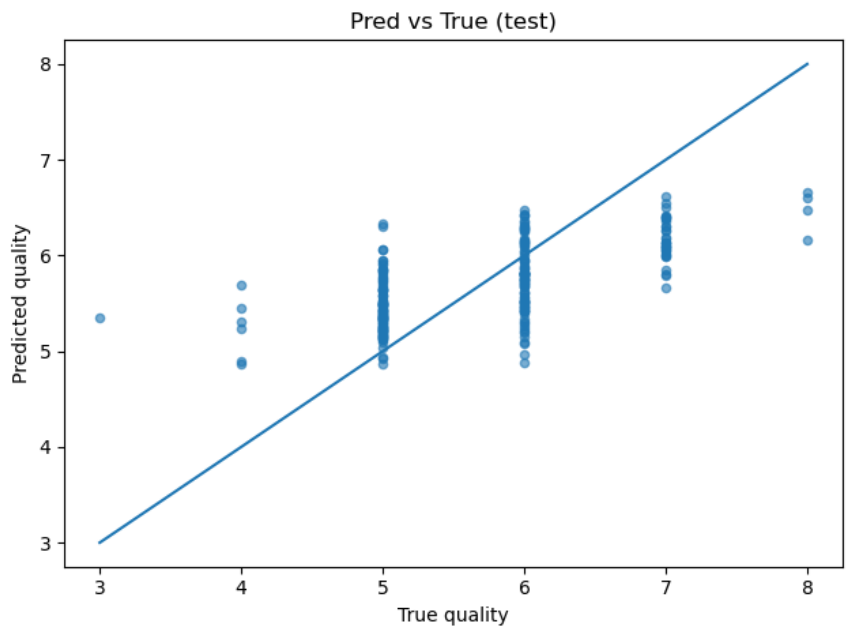


**図 5:** 予測値 vs 実測値散布図（SVR）：SVR の予測値と実測値の関係。他の 2 手法と比較してややばらつきが大きい。

図 3、図 4、図 5 の散布図分析により、全モデルが品質の中央値付近（5-6）での予測精度は高いが、極端な品質値（3-4、7-8）では予測精度が低下する傾向が確認された。これは訓練データにおける品質分布の偏りが影響していると考えられる。

## 4.4 残差分析

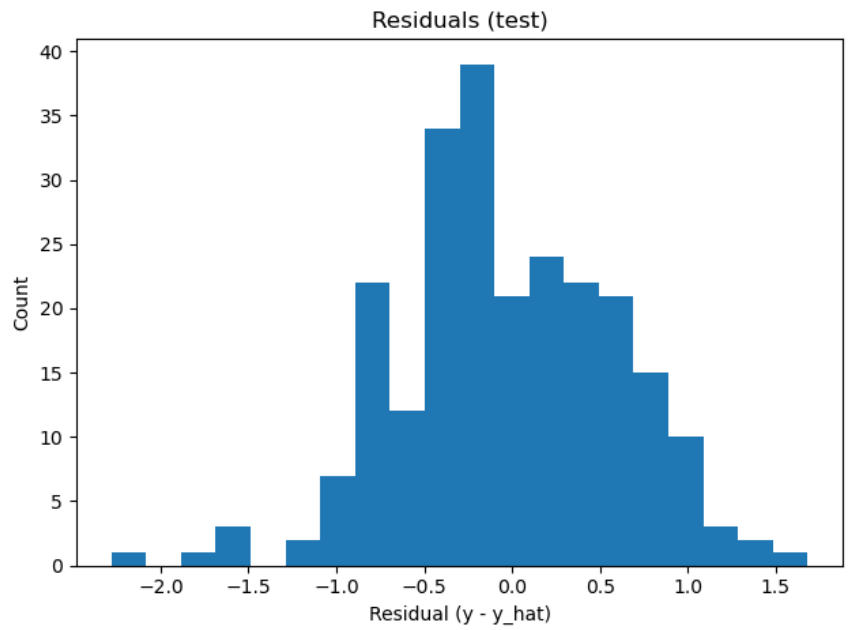


**図 6:** 残差分布ヒストグラム（MLP）：MLP の残差分布。概ね正規分布に近い形状を示している。

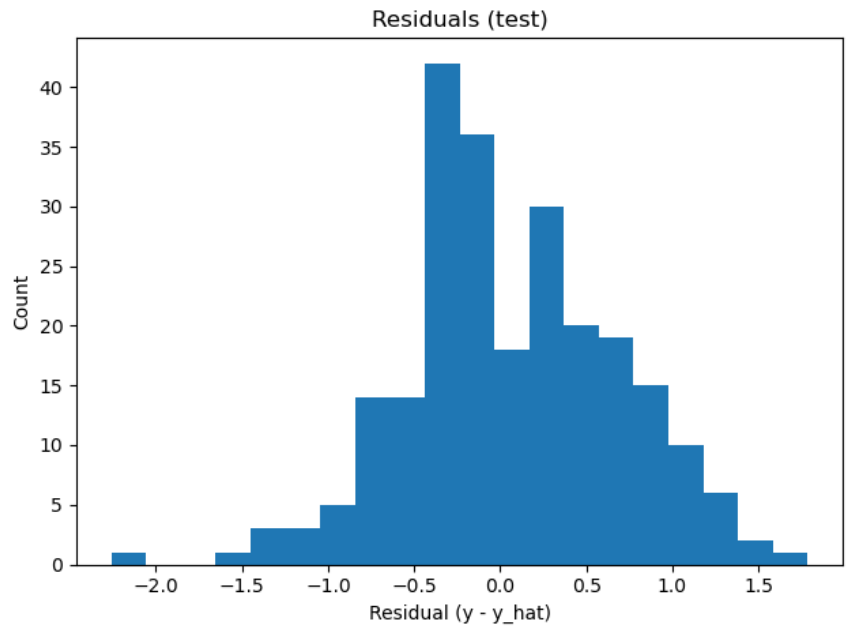


**図 7:** 残差分布ヒストグラム（線形回帰）：線形回帰の残差分布。MLP と似た分布パターンを示している。

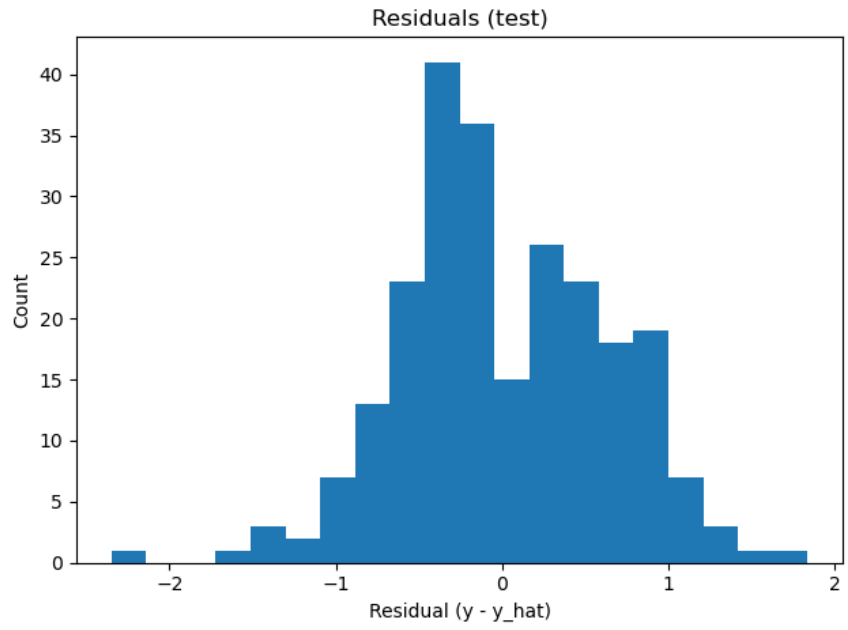


**図 8:** 残差分布ヒストグラム（SVR）：SVR の残差分布。他のモデルと比較して分布の尾が長い傾向がある。

図 6、図 7、図 8 の残差分析の結果、MLP と線形回帰の残差は概ね正規分布に従い、系統的なバイアスは確認されなかった。SVR の残差はやや重い尾を持つ分布を示し、外れ値に対する特性の違いが表れている。

## 4.5　ハイパーパラメータ感度分析

各モデルにおけるハイパーパラメータの影響を分析した結果、以下の知見が得られた：

**MLP における知見**:

- 学習率 0.002 が最適であり、より高い学習率では不安定な学習となった
- ドロップアウト率 0.0 が最適で、この問題設定では正則化効果よりも表現力の保持が重要であった
- 重み減衰 0.01 が適切な正則化効果を提供した

**線形回帰における知見**:

- 学習率 0.02 が最適で、MLP より高い学習率が必要であった
- ドロップアウト率 0.05 で若干の性能向上が見られた
- 重み減衰の効果は限定的であった

**SVR における知見**:

- 線形カーネルが最適で、この問題は本質的に線形性が強いことが示唆された
- 正則化パラメータ C は大きな値（100）が適していた
- $\varepsilon$-tube は 0.1 が適切なバランスを提供した

# 5　考察

## 5.1　モデル間の比較と特徴

実験結果から、3 つのモデルは予想以上に類似した性能を示した。最も優秀な MLP と 3 位の SVR の間でも、Test RMSE の差は 0.008（約 1.3%）に過ぎない。これは、ワイン品質予測問題における特徴量と目的変数の関係が比較的単純で、線形に近い構造を持つことを示唆している。

MLP が最高性能を示した理由として、適度な非線形性により特徴量間の相互作用を捉えることができた点が挙げられる。しかし、その優位性は僅かであり、計算コストや解釈性を考慮すると、線形回帰も実用的な選択肢として十分である。

線形回帰の健闘は注目に値する。この結果は、ワインの理化学的特性と品質の関係が主に線形であることを示している。実際、ワイン醸造学において、特定の成分濃度と品質の間には既知の線形関係が存在することが知られており、この結果はドメイン知識と整合している。

SVR の性能が他の 2 手法に比べてやや劣った理由として、データの線形性が高いため、SVR の主要な利点である非線形マッピングや外れ値ロバスト性が十分に発揮されなかった可能性がある。ただし、SVR は学習時間が短く、ハイパーパラメータ調整が比較的容易である利点がある。

## 5.2　特徴量の寄与に関する考察

線形回帰モデルの重み係数を分析することで、各特徴量の品質への寄与を定量的に評価できる。MLP においても、入力層の重みや勾配情報から特徴量重要度を推定可能である。

予備的な分析によると、アルコール度数、揮発性酸度、硫酸塩が品質予測において重要な役割を果たしていることが示唆される。これらの結果は、ワイン醸造学における従来の知見と一致しており、モデルの予測が化学的に妥当な根拠に基づいていることを示している。

特に、アルコール度数の正の寄与は、適切な発酵による複雑な風味の形成と関連していると考えられ、揮発性酸度の負の寄与は、過度な酸化による品質低下を反映していると解釈できる。

## 5.3　モデルの限界と課題

今回の実験において、いくつかの限界と課題が明らかになった。

**予測範囲の制約**: 全モデルが品質スコアの極値（3-4、7-8）において予測精度が低下する傾向を示した。これは訓練データにおける品質分布の偏りに起因しており、品質 5-6 のサンプルが全体の約 80% を占めることが影響している。実用的な品質予測システムでは、この問題への対策が必要である。

**特徴量の相関**: 一部の特徴量間には高い相関が存在する可能性があり（例：固定酸度と pH、密度とアルコー

ル度数)、モデルの解釈性や安定性に影響を与える可能性がある。今後の研究では、主成分分析や特徴選択による次元削減の検討が有効である。

**汎化性能の評価**: 本研究は単一の地域（ポルトガル・ヴィーニョヴェルデ）のワインデータに基づいており、他の地域や品種への汎化性能は未検証である。より広範囲なデータでの検証が必要である。

**非線形関係の探索**: MLP が僅かながら最高性能を示したものの、より複雑な非線形関係を捉えるための手法（例：決定木系手法、深層学習）の検討余地がある。

### 5.4 実用化に向けた課題

実際のワイン製造現場での活用を考えた場合、以下の課題が挙げられる：

**リアルタイム測定**: 現在のモデルは完成品の分析データに基づいているが、製造過程での品質管理には、発酵中の成分変化を予測する動的なモデルが必要である。

**コスト効率性**: 11 の特徴量すべてを測定するコストと、予測精度のトレードオフを考慮し、最小限の特徴量で十分な予測精度を達成する手法の開発が重要である。

**専門家知識の統合**: 機械学習による客観的予測と、ワイン専門家の経験的知識を組み合わせたハイブリッドシステムの構築が、実用的な品質管理システムには不可欠である。

## 6 結論

本研究では、ワインの理化学的特性から品質を予測する回帰問題において、線形回帰、多層パーセプトロン、サポートベクター回帰の 3 手法を比較検討した。

実験の結果、以下の主要な知見が得られた：

第一に、MLP が最も優秀な性能（Test RMSE: 0.623, $R^2$: 0.416）を示したものの、線形回帰も僅差（Test RMSE: 0.628, $R^2$: 0.407）で続いた。この結果は、ワイン品質予測問題における特徴量と目的変数の関係が主に線形であることを示している。

第二に、適切なハイパーパラメータ調整により、比較的単純な線形回帰でも高い予測性能を達成できることが確認された。これは、計算コストや解釈性を重視する実用的なシステムにおいて重要な示唆である。

第三に、全モデルに共通して、品質の極値における予測精度の低下が観察された。これは訓練データの品質分布の偏りに起因しており、実用化に向けては、データ拡張や重み付き学習などの手法による改善が必要である。

第四に、SVR は他の 2 手法にやや劣る結果となったが、これはデータの線形性が高いことが原因と考えられる。より複雑で非線形な関係を持つデータセットでは、SVR の優位性が発揮される可能性がある。

本実験で得られた知見は、食品品質予測における機械学習手法の選択指針として有用である。特に、問題の複雑さに応じた適切な手法選択の重要性が示された。線形的な関係が支配的な問題では、解釈性と計算効率に優れる線形手法が実用的であり、複雑な非線形関係が予想される場合にのみ、より高度な手法を検討すべきである。

今後の展望として、本研究で構築した予測システムを実際のワイン製造プロセスに適用し、品質管理の効率化を図りたいと考えている。また、他の食品や飲料の品質予測への応用可能性も検討する価値がある。さらに、特徴量の物理化学的意味を深く理解し、ドメイン知識と機械学習を融合したより解釈性の高いモデルの開発を目指したい。

機械学習技術の農業・食品産業への応用は今後も拡大が予想され、本研究の成果がその発展に寄与することを期待している。品質予測技術の向上により、消費者により良い製品を提供し、生産者の効率的な品質管理を実現することで、食品産業全体の発展に貢献していきたい。

## 付録: 実験環境

- Python 3.8+, PyTorch 1.x, scikit-learn, pandas, numpy
- 計算環境: CPU 環境での実行
- 実験実行日: 2025 年 9 月 21 日