Weld Quality Analysis and Prediction

Machine Learning Project Report

By: Ikram Firadous, Paul Guimbert, Hugues Du Moulinet d'Hardemare, Colin Frisch

1. Data Preprocessing

1.1 Dataset Overview

The welddb dataset contains:
- **1,652 welding samples** from various steel welding processes
- **44 features** spanning multiple categories:
- Chemical composition (12 features): Carbon, Silicon, Manganese, Sulfur, Phosphorus, etc.
- Welding parameters (6 features): Current, Voltage, Heat Input, Temperature, etc.
- Mechanical properties (9 features): Tensile Strength, Toughness, Elongation, Hardness, etc.
- Microstructure properties (5 features): Ferrite phases, Martensite, etc.
- Process variables (3 categorical): AC/DC, Electrode Polarity, Weld Type

1.2 Data Cleaning Strategy

Challenge: Significant missing data and ambiguous values (e.g., "<0.002", ">500")

Solutions implemented:
1. **Ambiguous value handling**:

- Values with "<" → multiplied by 0.5 (half the detection limit)
- Values with ">" → multiplied by 1.5 (1.5× the upper limit)

1. **Missing data strategy**:
   - Removed columns with >50% missing values (22 columns eliminated)
   - Removed rows with >30% missing values
   - Retained **31 features** for full dataset analysis
2. **Outlier treatment**:
   - Z-score method (threshold = 3)
   - Marked 681 outliers as NaN
   - Applied **KNN imputation** (k=5) to fill missing values
3. **Categorical encoding**:
   - Created **12 dummy variables** for categorical features
   - Avoided multicollinearity by dropping first category

Result: Clean dataset with 1,652 complete samples and 43 features (31 numeric + 12 categorical)

1.3 Feature Engineering

Nine engineered features were created based on metallurgical principles:

Impurity Features:
- `impurities_index` = Sulfur + Phosphorus
- `sulphur_phosphorus_ratio` = Sulfur / Phosphorus

Electrical Features:
- `power_input_kw` = (Current × Voltage) / 1000
- `power_efficiency` = Heat Input / (Current × Voltage)
- `current_density_proxy` = Current / Voltage

Carbon/Manganese Features:
- `hardenability_index` = Carbon + (Manganese / 6) *(Carbon Equivalent)*
- `carbon_manganese_ratio` = Carbon / Manganese

Thermal Features:
- `cooling_rate_proxy` = Heat Input / (Interpass Temp + 273)
- `thermal_cycle_intensity` = Heat Input / (Interpass Temp + 1)

These features capture complex physical relationships that simple linear models might miss.

1.4 Dimensionality Reduction (PCA)

Principal Component Analysis was applied to reduce dimensionality while preserving information:

- **11 principal components** selected (optimal based on scree plot)
- **88.19% variance explained** by these 11 components
- Component breakdown:
- PC1 (27.81%): Dominated by electrical parameters (power, current, voltage)
- PC2 (12.67%): Chemical composition (manganese, silicon)
- PC3-PC11 (47.71%): Mixed thermal, chemical, and process parameters

Key findings from PCA loadings:
- PC1 strongly correlates with engineered features: `power_input_kw` (0.94), `current_density_proxy` (0.85), `cooling_rate_proxy` (0.88)
- PC2 captures chemical composition: `manganese` (0.74), `hardenability_index` (0.64)

---

2. Regression Analysis: Predicting Weld Quality

2.1 Target Variable Definition

A composite **quality score** was created from four mechanical properties:

Formula:

```
Quality Score = 0.33 × (UTS_normalized) + 0.33 × (Toughness_normalized) + 0.34 × [(Elongation_no
```

Where:
- **UTS**: Ultimate Tensile Strength (MPa) - measures maximum stress before failure
- **Toughness**: Charpy Impact Toughness (J) - measures shock absorption
- **Elongation**: Elongation (%) - measures ductility/stretchability
- **Reduction**: Reduction of Area (%) - measures ductility/deformation capacity

Rationale: This weighted score balances strength (33%), toughness (33%), and ductility (34%), representing overall weld quality.

2.2 Dataset Reduction

Challenge: Many samples lacked sufficient target values for quality score calculation.

Solution:
- Retained only samples with **≥2 of 4 target values**
- Applied KNN imputation to complete missing target values
- **Final regression dataset**: 720 samples (43.6% of original data)

Quality score statistics:
- Mean: 0.585
- Median: 0.631
- Std Dev: 0.098

2.3 Model Comparison

Three models were evaluated using 5-fold cross-validation:

Model	Cross-Val R^2	Test R^2	Notes
Ridge Regression	0.544 ± 0.055	-	Baseline linear model
Random Forest	0.646 ± 0.043	0.703	Good performance, robust
Gradient Boosting	0.653 ± 0.060	0.727	Best traditional ensemble
XGBoost	**-**	**0.747**	**Best overall performance**

2.4 Hyperparameter Optimization

XGBoost - Best Configuration:

```
{
    'n_estimators': 300,
    'learning_rate': 0.05,
    'max_depth': 3,
    'subsample': 0.8,
    'colsample_bytree': 0.7
}
```

Performance Metrics (Test Set):
- R^2 Score: 0.747
- Mean Squared Error: 0.00290
- Mean Absolute Error: 0.0357

This means the model explains **74.7% of variance** in weld quality, which is excellent for this complex domain.

2.5 Feature Importance for Regression

Top predictive features (from Gradient Boosting analysis):

1. **Principal Components** (PC1-PC11): Capture complex interactions
2. **Categorical Variables**: Weld type significantly impacts quality
3. **Original Features** (when analyzed):
   - Electrical parameters dominate
   - Chemical impurities (S, P) strongly influence quality
   - Thermal cycle parameters moderately important

---

3. Clustering Analysis: Identifying Welding Patterns

3.1 Clustering Approach

Data: PCA-transformed features (11 components) from full dataset (1,652 samples)

Methods Evaluated:

Method	N Clusters	Silhouette Score	Davies-Bouldin	Calinski-Harabasz
K-Means	**6**	**0.206**	**1.550**	**353.7**
Hierarchical	6	0.184	1.625	334.9
Spectral	6	0.236	1.515	171.9
Mini-Batch K-Means	6	0.188	1.657	341.5
DBSCAN	5	0.443*	0.579*	228.6*

*DBSCAN excluded 42% of data as noise, limiting practical utility.

Selection Criteria: Multi-metric composite score (40% Silhouette + 30% Davies-Bouldin + 30% Calinski-Harabasz)

Winner: **K-Means with K=6** (Composite Score: 0.769)

3.2 Optimal Number of Clusters

K=6 was selected based on:
- **Elbow method**: Inertia curve flattens after K=6
- **Silhouette analysis**: K=6 shows good balance (0.206)
- **Physical interpretation**: Six clusters represent distinct welding regimes

3.3 Cluster Characteristics

Cluster 0 (649 samples, 39.3%) - *"Standard High-Manganese Welds"*

- **Largest cluster**, representing typical welding practices
- **High**: Manganese content (1.49% vs 1.20% overall)
- **Low**: Current (171A vs 262A), heat input, power
- **Quality**: Average toughness (85.4J), low test temperature (-47.8°C)
- **Interpretation**: Low-current, manganese-rich steels for low-temperature applications

Cluster 1 (168 samples, 10.2%) - *"High-Efficiency, High-Impurity Welds"*

- **High**: Power efficiency, phosphorus (0.015% vs 0.010%), thermal cycle intensity
- **Moderate**: Current (228A), power (5.8kW)
- **Quality**: Lower toughness (72.3J), moderate test temperature (-31.6°C)
- **Interpretation**: Energy-efficient processes, but impurities reduce toughness

Cluster 2 (146 samples, 8.8%) - ***"Vanadium-Rich, Heat-Treated Welds"***

Cluster 3 (343 samples, 20.8%) - ***"Low-Manganese, Low-Hardenability Welds"***

Cluster 4 (101 samples, 6.1%) - ***"High Carbon-Ratio, Premium Welds"***

Cluster 5 (245 samples, 14.8%) - ***"High-Power, High-Current Welds"***

3.4 Key Discriminating Features (ANOVA F-test)

Features that **best distinguish clusters** (ranked by F-statistic):

1. **Current** (F=6817) - Most important differentiator
2. **Power input** (F=5398)
3. **Current density** (F=2833)
4. **Heat input** (F=1579)
5. **Cooling rate** (F=1078)

> *Insight: Electrical parameters dominate cluster separation, followed by chemical composition and thermal treatment.*

3.5 Cluster Quality Analysis

Within-cluster homogeneity (lower std dev = more homogeneous):

- **Most homogeneous**: Cluster 0 (0.70) and Cluster 3 (0.97)
- **Least homogeneous**: Cluster 1 (7.28) and Cluster 5 (7.00)

Most similar clusters: Cluster 0 ↔ Cluster 3 (distance: 2.90)
Most different clusters: Cluster 2 ↔ Cluster 5 (distance: 7.87)

---

4. Key Findings and Insights

4.1 Critical Success Factors for Weld Quality

1. **Electrical Parameters are Paramount**
   - Current, voltage, and power input are the strongest predictors and cluster differentiators
   - Optimal balance varies by material and application

2. **Impurities Significantly Reduce Quality**
   - Sulfur and phosphorus (impurities_index) strongly correlate with reduced toughness
   - Cluster 1 (high impurities) shows 15% lower toughness than average

3. **Engineered Features Add Value**
   - Features like `power_efficiency`, `cooling_rate_proxy`, and `hardenability_index` improve both prediction and interpretation
   - Validate the importance of domain knowledge in feature engineering

4. **Trade-offs Exist**
   - Cluster 2 (high vanadium, heat-treated) sacrifices toughness for strength
   - High current/power doesn't always guarantee best quality (Cluster 5)

4.2 Practical Recommendations

For High-Quality Welds (inspired by Cluster 4):
- Optimize carbon/manganese ratio
- Apply appropriate post-weld heat treatment
- Control current density carefully
- Target test for specific temperature ranges

For Low-Temperature Applications (Clusters 0, 3):
- Prioritize manganese content
- Use lower current settings
- Minimize impurities (S, P)
- Expect good toughness at -45°C or lower

For High-Strength Applications (Cluster 2):
- Consider vanadium additions
- Implement post-weld heat treatment (PWHT)
- Accept trade-off in impact toughness
- Suitable for ambient temperature service

Conclusions

This comprehensive analysis of 1,652 steel welds demonstrates the power of machine learning in understanding complex manufacturing processes:

1. **Predictive Success**: XGBoost achieved 74.7% R^2 in predicting weld quality, enabling:
   - Quality prediction before physical testing
   - Parameter optimization for target quality
   - Cost reduction through fewer failed welds

2. **Pattern Discovery**: K-Means clustering revealed 6 distinct welding regimes:
   - Each cluster represents a coherent strategy (e.g., high-power vs. low-current)
   - Clear trade-offs between strength, toughness, and ductility
   - Practical guidance for process selection

3. **Critical Factors Identified**:
   - **Electrical parameters** (current, voltage, power) dominate both prediction and clustering
   - **Chemical impurities** (S, P) critically impact quality
   - **Heat treatment** (PWHT) enables high-strength applications
   - **Engineered features** capture complex physical relationships

4. **Actionable Insights**:
   - Cluster 4 characteristics (high C/Mn ratio, optimized heat treatment) produce best quality
   - Managing impurities is essential for high-toughness applications
   - Different applications require different welding strategies (clusters)