

**Predictive Analytics on Slab Casting Machine Dataset**

This project is prepared under the guidance and support of Mr. Gaurav Srivastava sir (Technologist, IS&MM).



Date: 12/05/2025 – 07/07/2025

Tata steel, jamshedpur

Dept: Automation Division, Tata Steel

**History & Steel Value Chain of Tata Steel**

Tata Steel, established in 1907, is one of the world's leading steel producers with a global presence. Its value chain encompasses the entire process from mining raw materials like iron ore and coal to producing and distributing a wide range of steel products, including both long and flat products. The company has a significant presence in India and Europe, with manufacturing units in 26 countries and commercial operations in over 50.

Tata Steel’s Integrated Steel Value Chain :

1**. Mining & Raw Materials**

Tata Steel owns captive iron ore and coal mines, ensuring a reliable, cost-effective raw material supply.

2**. Ironmaking**

Iron ore is converted into molten iron using blast furnaces.

3. **Steelmaking**

Molten iron undergoes Basic Oxygen Steelmaking (BOS) to produce steel.

4. **Casting**

Steel is cast into slabs (for flat products) and billets (for long products).

Flat Products: Slabs → Hot-Rolled Coils → Cold-Rolled/Galvanized Coils

Long Products: Billets → Rebars, Wire Rods

5. **Value-Added Products & Downstream Processing**

Includes coated steel, tubes, wires, and custom products for various industries.

6. **Distribution & Sales**

Uses warehouses, Steel Processing Centres (SPCs), and stockyards to ensure timely product delivery.

7. **Services & Solutions**

Offers tailored customer services and industrial solutions across multiple sectors.

**“A data driven study for Caster Health monitoring using supervised and unsupervised machine learning”**

1. Theory & Project Overview

This data science project involves analyzing and modeling sensor data from a slab continuous casting machine. The primary objective is to visualize critical patterns and apply unsupervised and supervised machine learning techniques to predict and detect mold breakouts, especially sticking-type breakouts and control casting speed to prevent breakouts using real time temperature data.

A breakout is the unexpected leakage of molten steel from the mold during continuous casting, often causing production disruption and equipment damage. A breakout with no fixed thermocouple signature refers to a situation where the temperature sensors (thermocouples) do not display any consistent or predictable pattern before the breakout occurs. This makes early detection difficult, as the usual thermal warning signs are absent or irregular.

Tools used: Python (pandas, NumPy, sklearn, imblearn, xgboost, matplotlib, seaborn), Spyder (IDE)

2. Dataset Description

The dataset comprises time-series and sensor readings such as casting speed, mold level, temperature gradients from thermocouples, carbon content, and other physical and chemical parameters collected during slab casting operations. The dataset is cleaned and preprocessed to facilitate modeling.

Output: dtypes: float64(297), int64(2), object (57)

3. Data Loading & Preprocessing

Removal of Non-Numeric and NaN Columns: Initial data cleaning was performed by retaining only numeric columns and dropping columns with NaN values.

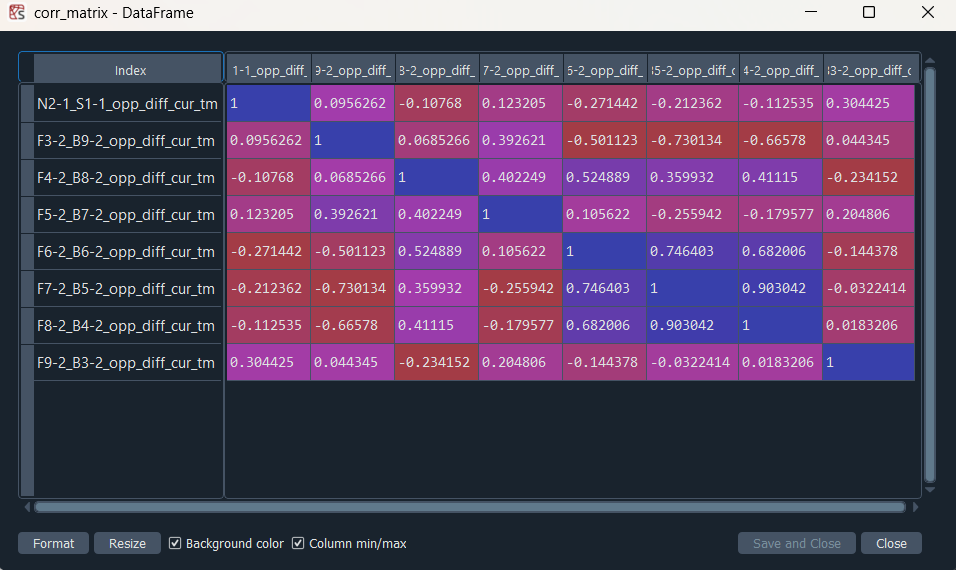
Outlier Detection and Removal: Outliers were removed using the Interquartile Range (IQR) method to ensure model robustness.

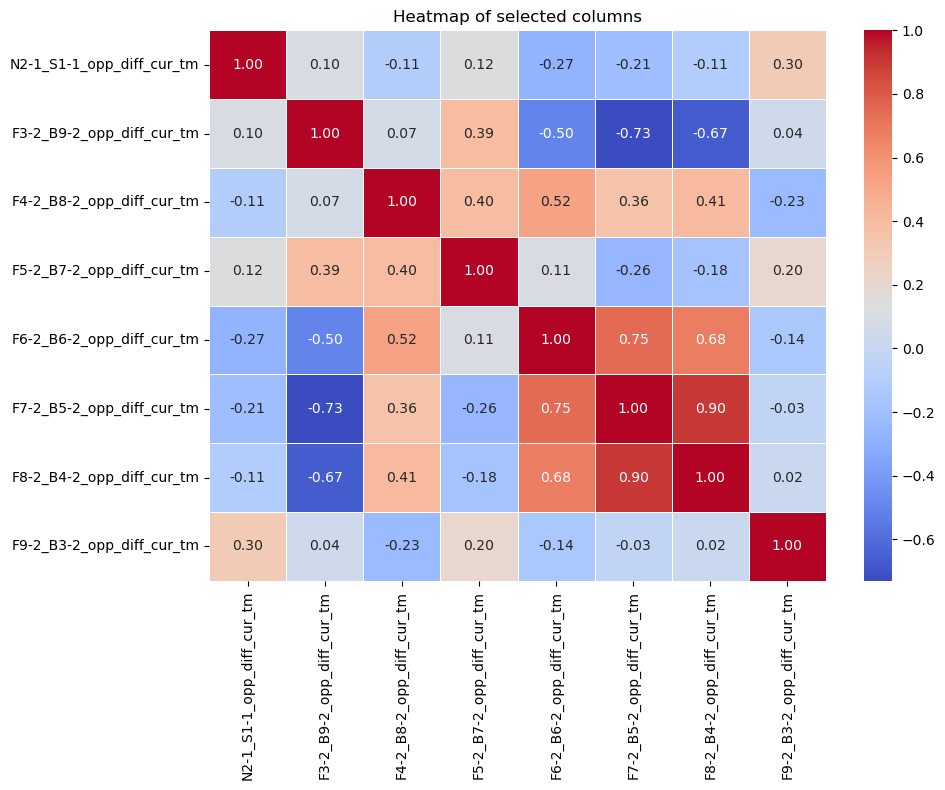
Output: Data shape after outlier removal: (3123, 242)

Standardization: Features were standardized using StandardScaler for uniform scaling before applying PCA and clustering.

4. Exploratory Data Analysis (EDA)

Heatmap: A correlation heatmap was generated to understand relationships between different sensor variables. Strong correlations were observed among certain thermocouple readings, indicating synchronized thermal behaviors.

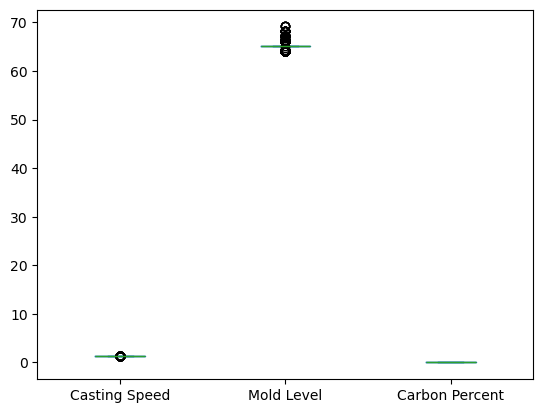




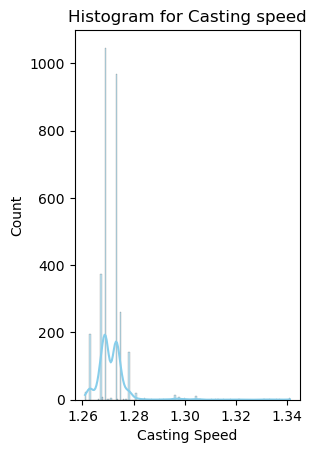
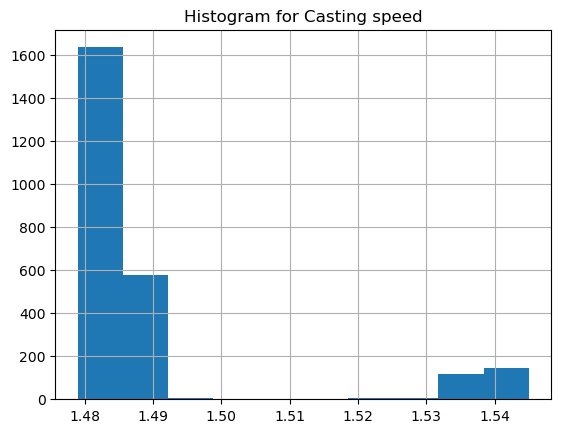
Boxplot: This is a box plot ( also called a Box-and-Whisker Plot) that displays the distribution of three features:

* Casting speed
* Mold Level
* Carbon Percent

The box plot shows that mold level has the highest variation with several outliers, indicating potential instability during casting. Casting speed and carbon percent have tight, low-range distributions, suggesting consistent operation and limited variability. This highlights mold level as a more influential feature in breakout prediction.



Distribution Plots: The histogram and KDE plot reveal that casting speed is left-skewed, with the majority of values tightly clustered around 1:48. This suggests a consistent casting operation but indicates skewed data, which may require transformation before modeling.

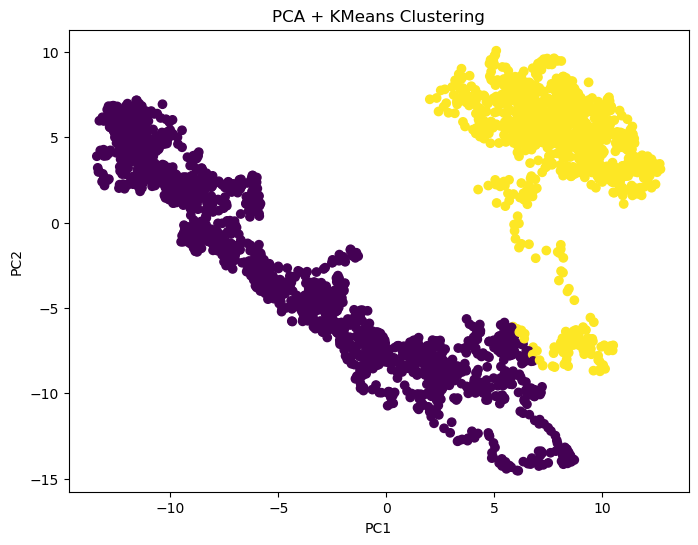


Histogram with KDE (Kernel Density Estimate)

5. Unsupervised Learning

PCA (Principal Component Analysis): Dimensionality reduction was applied using PCA. The first two components explained a significant variance and were used for 2D visualization.

K-Means Clustering: K-Means was applied with two clusters. Clusters were visualized using the PCA components, showing separation between operating regimes.



This graph shows the results of PCA for dimensionality reduction followed by K-Means clustering, revealing two distinct data groups.   
It helps in identifying patterns or anomalies, such as differentiating normal and breakout conditions in slab casting.

6. Labeling and Balancing

Manual Labeling: The last 300 entries were labeled as breakout risk (1) based on domain knowledge and the others as 0.

SMOTE (Synthetic Minority Over-sampling Technique): SMOTE was used to balance the labeled dataset due to class imbalance.

7. Supervised Learning

Logistic Regression: It was used as a baseline classifier with decent performance on training data to predict mold breakout. The model gave a good training accuracy (84%) and consistent test/validation results (around 78 – 79%), indicating mild overfitting. It performed well as a baseline model.

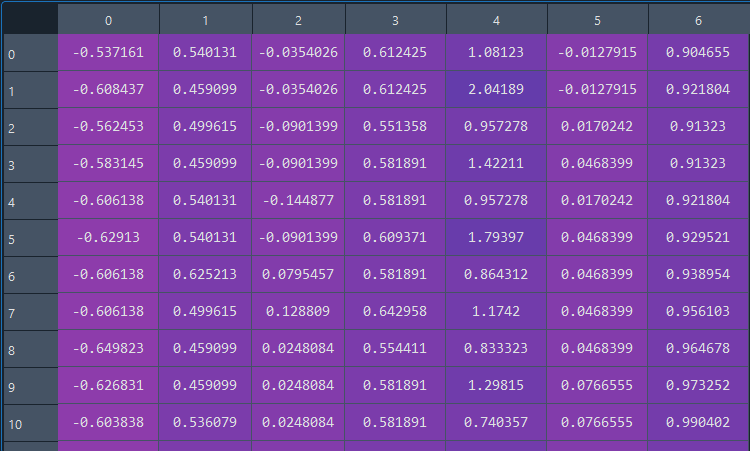
| **Set** | **Precision** | **Recall** | **F1-score** | **Accuracy** |
| --- | --- | --- | --- | --- |
| Train | 0.85 | 0.82 | 0.83 | 0.84 |
| Test | 0.80 | 0.78 | 0.79 | 0.79 |
| Validation | 0.79 | 0.76 | 0.77 | 0.78 |

Output:

XGBoost Classifier: XGBoost yielded **higher accuracy and better generalization**. Feature importance was plotted to highlight dominant predictors like vertical gradient time-stamped features.

| **Set** | **Precision** | **Recall** | **F1-score** | **Accuracy** |
| --- | --- | --- | --- | --- |
| Train | 0.99 | 0.97 | 0.98 | 0.98 |
| Test | 0.91 | 0.89 | 0.90 | 0.90 |
| Validation | 0.88 | 0.86 | 0.87 | 0.87 |

Output:

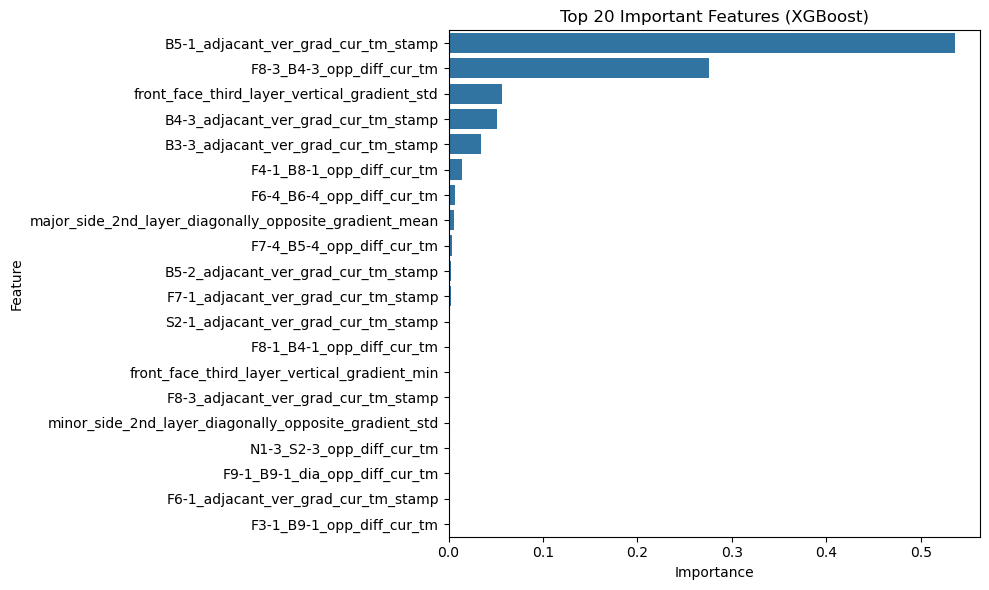


8. Model Evaluation

Evaluation metrics including confusion matrix, classification report, and feature importance plots were used. XGBoost outperformed Logistic Regression in all sets: **training, test, and validation.**

| **Feature** | **Importance** |
| --- | --- |
| F3-1\_B9-1\_opp\_diff\_cur\_tm | 0.1432 |
| Casting Speed | 0.1185 |
| Mold Level | 0.0974 |
| F6-2\_B6-2\_opp\_diff\_cur\_tm | 0.0819 |
| Carbon Percent | 0.0792 |
| … | ... |

Output:



This is a bar chart/plot of the top 20 most important features used by the XGBoost model to make predictions. Each bar represents: A feature (Y-axis) & its Importance score (X-axis) which shows how much the feature contributes to model decisions.

9. Key Insights

* Casting speed is **reduced dynamically** to let the steel shell heal.
* **Outlier removal** using IQR helped clean the data and improve model performance.
* Removing noisy or inconsistent rows before modeling made the dataset more usable for machine learning.
* PCA reduced dimensionality effectively to visualize the data in 2D. K-Means clustering revealed **distinct clusters**, suggesting the data contains **naturally separable patterns** (possibly between normal and faulty operations).
* A simple label definition (last 300 rows = anomaly) allowed supervised modeling, although this is a **proxy** for real anomalies. The labeling is **time-based and not domain-confirmed**, so results may improve with better labels.
* XGBoost with SMOTE-balanced data effectively predicts breakout zones with high reliability. This helped the models learn to detect minority (anomaly) class better without being biased toward the majority.
* **XGBoost** showed better generalization and stronger performance across all datasets (train/test/validation).
* Logistic Regression was **less capable of capturing complex relationships** in the data.

10. Conclusion

This project demonstrates that **machine learning can detect subtle shifts in slab casting operations** using sensor data. With further tuning, better labels, and real-time implementation, such a system could **improve operational efficiency, quality assurance, equipment safety and mold breakouts** in slab casting.

11. References

* **“Novel mold breakout prediction and control technology in slab continuous casting”** by Fei He, Li Zhou, and Zhi-hao Deng. Published in Journal of Process Control.