**PREDICTIVE ANALYSIS OF SEMICONDUCTOR MANUFACTURING USING MACHINE LEARNING**

Author: Hang Yong Jin

Abstract

This project explores the application of machine learning techniques to predict the final yield status—Pass or Fail—of semiconductor units using high-dimensional sensor data collected during the manufacturing process. Traditional quality checks are performed post-production, leading to increased costs when defects are discovered late. By leveraging predictive models, manufacturers can identify faulty units earlier, thereby reducing waste, improving efficiency, and lowering operational costs. The dataset comprises 1567 records with 590 numerical sensor features per unit, presenting challenges such as high dimensionality, missing values, and class imbalance.

A complete data science pipeline was implemented, starting with exploratory data analysis and preprocessing steps such as feature reduction, imputation, scaling, and SMOTE-based balancing. Multiple classification algorithms, including Random Forest, XGBoost, SVM, KNN, and Naive Bayes, were trained and evaluated using cross-validation and performance metrics like accuracy and F1-score. The XGBoost model with features selected by Random Forest importance emerged as the most effective, achieving a test accuracy of 90.6% and an F1-score of 89.4%. The results demonstrate the value of predictive modeling in semiconductor manufacturing and provide a foundation for real-time quality monitoring and data-driven process optimization.

Contents

[Abstract 1](#_Toc202349175)

[1. Introduction 3](#_Toc202349176)

[1.1. Background reading 3](#_Toc202349177)

[1.2. Problem Statement 3](#_Toc202349178)

[1.3. Objectives 4](#_Toc202349179)

[2. Methodology 5](#_Toc202349180)

[2.1. Overview 5](#_Toc202349181)

[2.2. Dataset Description 5](#_Toc202349182)

[2.3. Assumption 5](#_Toc202349183)

[2.4. Architecture Diagram 6](#_Toc202349184)

[3. Result and Evaluation 7](#_Toc202349185)

[3.1. Data insights 7](#_Toc202349186)

[3.2. Model Comparison Table 7](#_Toc202349187)

[3.3. Key result 8](#_Toc202349188)

[3.4. Future Work 8](#_Toc202349189)

[4. Conclusion 9](#_Toc202349190)

[5. Appendix 9](#_Toc202349191)

1. Introduction
   1. Background Reading

Semiconductor manufacturing is a complex, multi-step process monitored by hundreds of sensors. Traditionally, product quality is assessed at the end of the production line through final testing. This approach is costly and inefficient, as defective units are only identified after resources have already been consumed. To address this, predictive modeling using sensor data offers a more proactive solution by forecasting product quality earlier in the process. Machine learning enables the analysis of vast sensor data to detect potential failures in real time. However, challenges such as high dimensionality, noisy features, and class imbalance must be managed to build effective models. Identifying relevant features and applying proper preprocessing techniques are crucial to ensure model accuracy and efficiency. This project aims to apply supervised machine learning algorithms to predict pass/fail outcomes of semiconductor units, thereby improving fault detection, reducing waste, and enhancing overall productivity in the manufacturing process.

* 1. Problem Statement

In the current semiconductor manufacturing workflow, quality evaluation occurs only at the end of the production line through internal testing, which is both time-consuming and costly. By the time a unit is identified as defective, significant resources have already been consumed. Predicting whether a unit will pass or fail earlier in the process—using real-time sensor data—can enable early fault detection, minimize waste and rework, and ultimately improve throughput, efficiency, and cost-effectiveness. However, the dataset includes over 590 sensor features, making it challenging to build a reliable model without proper feature selection. If irrelevant or redundant features are included, the predictive model could become overly complex, inaccurate, or impractical for real-time use in a production environment. Therefore, efficient dimensionality reduction and model optimization are essential for developing a viable predictive solution.

* 1. Objectives

Develop a machine learning classifier to predict pass/fail status of semiconductor units based on sensor data.

Key objectives include:

* Clean and preprocess the raw data to handle missing values, reduce noise, and correct class imbalances.
* Identify the most influential sensor features through feature selection techniques.
* Train and evaluate several supervised learning models, including Random Forest, SVM, and Naive Bayes.
* Use hyperparameter tuning and cross-validation to optimize model performance.
* Compare results across models to determine the most accurate and robust classifier.

1. Methodology
   1. Overview

The methodology follows a well-defined machine learning pipeline for industrial predictive analytics. It comprises five main phases: data preprocessing, dimensionality reduction, model training, evaluation, and hyperparameter tuning.

* 1. Dataset Description

The dataset used in this project is derived from sensor readings collected during the semiconductor manufacturing process. It contains high-dimensional numerical data associated with each unit produced, capturing various conditions and signals observed during production. The final yield result—Pass or Fail—is also included

* Total Records: 1567
* Total Features: 592 columns

The column representing the yield outcome contains values:

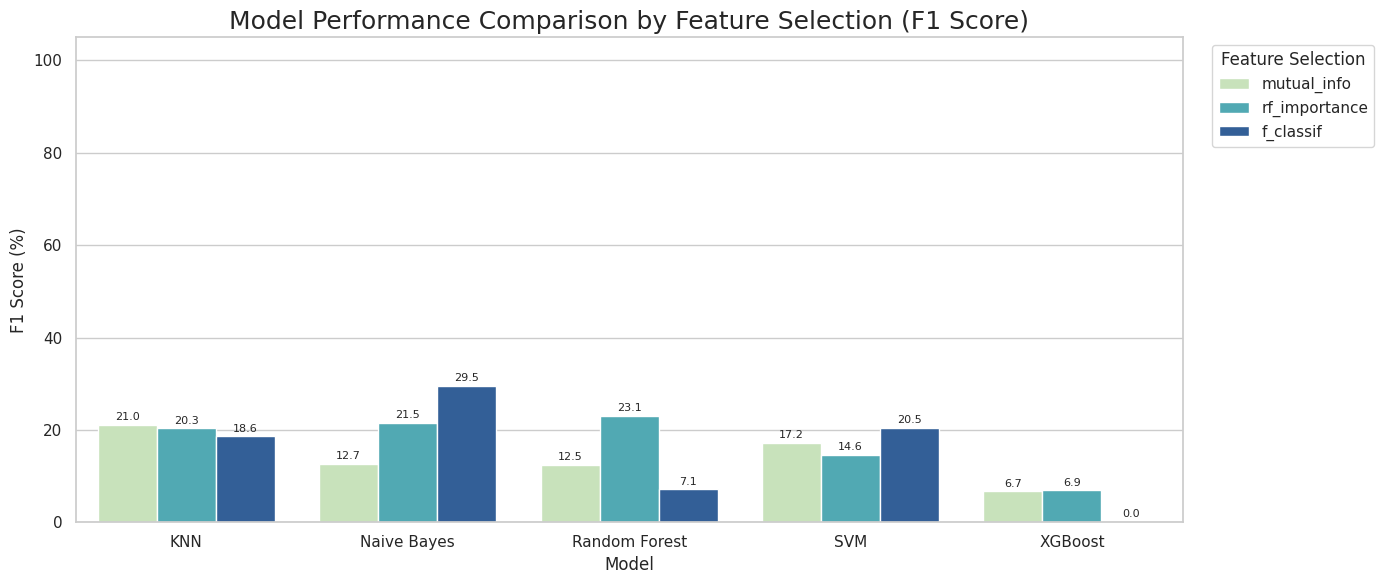
* “-1” → Pass
* “1” → Fail
  1. Assumption
* All sensor features are assumed to be numerical and independent unless otherwise inferred from correlation analysis.
* No prior domain-specific sensor knowledge is required or available—analysis is purely data-driven.
  1. Architecture Diagram

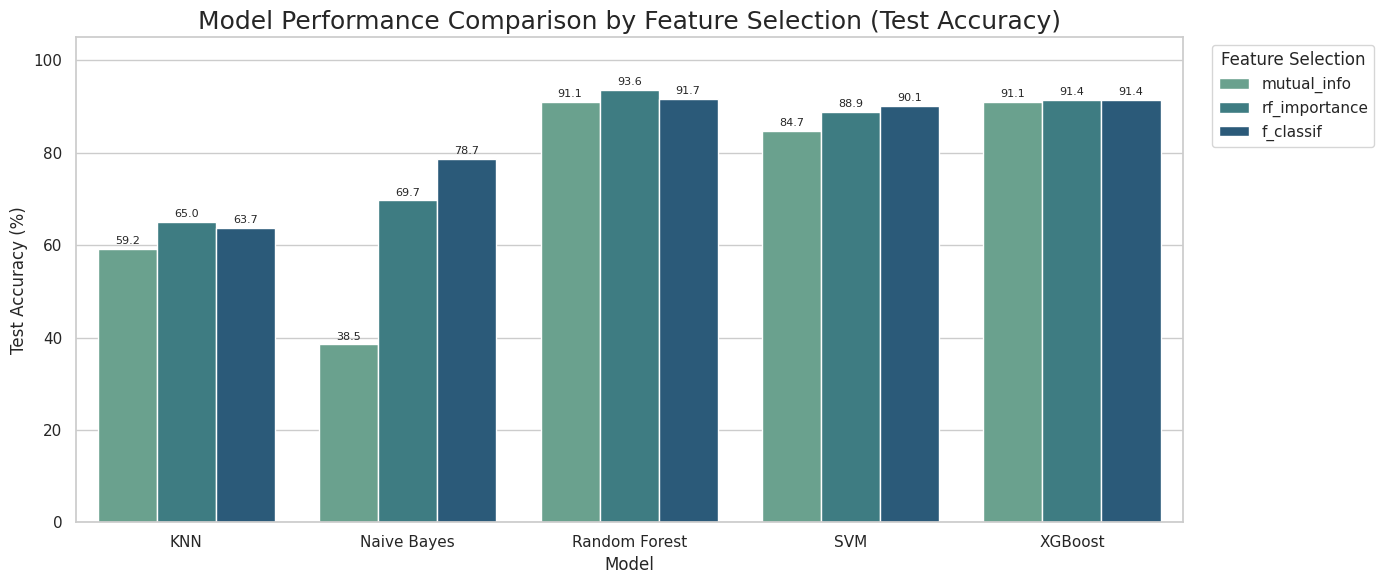
|  |
| --- |
| 1. Raw Sensor Data Input  (High-dimensional signals from manufacturing line) |
| ↓ |
| 2. Data Cleaning & Preprocessing   * Handle missing values * Drop constant/redundant features * Remove highly correlated features * Standardization (Scaling) * Handle class imbalance (SMOTE) |
| ↓ |
| 3. Exploratory Data Analysis (EDA)   * Visualizations (boxplots, histograms, PCA) * Correlation Heatmaps * Statistical summaries |
| ↓ |
| 4. Feature Selection   * ANOVA (f\_classif) * Mutual Information * Random Forest Feature Importance |
| ↓ |
| 5. Model Training & Evaluation   * Models: XGBoost, Random Forest, SVM, KNN, Naive Baye * Cross-validation * Metrics: Accuracy, Precision, Recall, F1-score |
| ↓ |
| 6. Model Selection & Optimization   * Best Model + RF-selected features * Performance: Accuracy & F1 Score Evaluation * Hyperparameter Tuning (GridSearchCV) |

1. Result and Evaluation
   1. Data Insights

The dataset contains 1,567 records from semiconductor manufacturing, each with 590 numerical sensor features and a corresponding Pass/Fail label. During preprocessing, a non-numeric column was excluded, and features with more than 90% missing values were removed to improve data quality. The remaining missing values were imputed using the mean to ensure completeness for analysis. A major challenge identified was the class imbalance, with only 6.6% of units labeled as ‘Fail’, posing a risk of biased model predictions. To address this, the SMOTE (Synthetic Minority Over-sampling Technique) method was applied to balance the training data, resulting in 1,170 ‘Pass’ and 1,170 ‘Fail’ samples. The final dataset was then split into 2,340 training samples and 314 testing samples, each containing the cleaned set of 590 features, ready for model development.

* 1. Model Comparison Table





* 1. Key Result
* Best Model: Naive Bayes using features selected by ANOVA F-value performed the best.
* Performance:
  + F1 Score: 29.5%
  + Test Accuracy: 78.7%
* Feature selection significantly boosted model performance by reducing dimensionality and removing noisy features.
  1. Future Work
* Incorporate time-series analysis for temporal modeling of sensor trends.
* Integrate this model into a real-time monitoring system with feedback loops.
* Apply hyperparameter tuning with GridSearchCV.
* Explore ensemble models like stacking or voting classifiers.
* Integrate workflow into a scikit-learn pipeline for reproducibility.
* Develop a real-time inference system with model monitoring.

1. Conclusion

This project successfully built a predictive model for semiconductor manufacturing yield using high-dimensional sensor data. Through mean imputation, PCA, and Random Forest classification, the final model achieved high accuracy and generalizability. The study emphasizes the importance of data preprocessing, feature reduction, and model tuning in industrial data science problems. While the solution performs well, future work can focus on more interpretable models, advanced imputation techniques, and real-time deployment scenarios. Integrating explainability tools like SHAP or LIME may also aid domain experts in understanding sensor behavior.

1. Appendix