Final Project Report
Semiconductor Yield Prediction

1. Introduction

Semiconductor manufacturing involves complex processes where minor variations in sensor readings can impact yield. Yield prediction aims to classify whether a manufactured wafer will pass or fail based on process and sensor data. This project builds a machine learning pipeline to predict semiconductor yield using signal measurements.

2. Problem Statement

Given a dataset containing multiple sensor signal readings from semiconductor production lines, the objective is to predict the yield outcome (Pass/Fail). The solution should support early defect detection and help reduce production losses.

3. Dataset Description

The dataset used in this project consists of signal measurements recorded during the manufacturing process. Each row represents a production sample, and each column represents a signal feature. The target variable indicates whether the sample passed quality checks (yield = 1) or failed (yield = 0).

4. Data Preprocessing

Key preprocessing steps performed:
- Handling missing values by imputing or removing incomplete rows.
- Removing constant or near-constant features.
- Feature scaling using StandardScaler to normalize sensor readings.
- Train-test split to evaluate model performance fairly.

5. Exploratory Data Analysis (EDA)

EDA was performed to understand feature distributions and relationships:
- Checked class imbalance between pass and fail samples.
- Analyzed feature correlations to identify redundancy.
- Visualized sensor signal distributions for anomalies and outliers.

6. Model Development

Several machine learning models were trained and evaluated:
- Logistic Regression (baseline)
- Random Forest Classifier
- Gradient Boosting / XGBoost (if applicable)
- Support Vector Machine (optional)

Hyperparameter tuning was applied to improve performance and reduce overfitting.

7. Model Evaluation

The trained models were evaluated using:
- Accuracy
- Precision, Recall, and F1-score
- Confusion Matrix
- ROC-AUC Score

The best-performing model was selected based on balanced metrics, especially recall for detecting failed yield samples.

8. Results & Findings

Key findings from the analysis:
- Certain sensor signals were strong indicators of yield failures.
- Ensemble models (Random Forest / Gradient Boosting) typically performed better than linear models.
- Class imbalance can reduce detection of failures; using class weights or resampling can improve recall.

9. Conclusion

This project demonstrates how machine learning can assist semiconductor yield prediction by analyzing sensor signal data. The final model provides early detection of potential failures, improving decision-making and reducing manufacturing costs.

10. Future Work

Potential improvements:
- Apply advanced feature selection and dimensionality reduction (PCA).
- Use deep learning models for time-series sensor data.
- Implement real-time deployment with monitoring dashboards.
- Improve handling of imbalanced datasets using SMOTE or anomaly detection methods.