Predictive Fleet Maintenance Using Machine Learning

KRUNAL SONI

About Scania & Vision

Driving the Future with Predictive Maintenance

🚛 **About Scania**

Scania is a **global leader in heavy vehicle manufacturing**, recognized for its **performance, safety, and sustainability**.

Today, Scania is transforming from a vehicle manufacturer into a **data-driven transport solutions company**.

💡 **Vision & Strategy**

At Scania's **Fleet Management & Data Analytics Division**, our mission is to build **intelligent, connected, and predictive systems** that keep vehicles on the road longer and safer.

⚙️ **Key Capabilities**

- 📡 **Real-time telematics** from connected trucks
- 🧠 **AI-driven predictive maintenance** to prevent costly failures
- 💰 **Maximized uptime** and operational efficiency
- 📊 **Data-informed decisions** for global fleet operators

Business Problem

When Downtime Costs Millions

Scania's system data provides an opportunity to detect early warning signs of component failure. By predicting breakdowns **before they happen**, Scania can prevent costly repairs, avoid delivery delays, and improve fleet reliability.

Why this matters

🚚 Unexpected breakdowns disrupt delivery schedules.

💸 Reactive repairs drive up maintenance costs.

⏱️ Downtime reduces fleet efficiency and impacts customer satisfaction.

Business Impact

Reactive Maintenance (Before)	Predictive Maintenance (After)
🔧 Repairs only after breakdowns	🧠 Issues detected *before* failure
🚛 Frequent unplanned downtime	⏱️ **Reduced downtime**, higher fleet availability
💸 Costly emergency repairs	💰 **Planned maintenance**, lower costs
🧯 Sudden part failures	⚙️ **Data-driven part replacement** before failure
📅 Fixed-interval servicing (not optimized)	📊 **Condition-based servicing** using sensor data
📦 Poor spare parts planning	📦 **Optimized inventory**, minimal waste
⏳ Delivery delays & route disruptions	🚚 **On-time deliveries**, smoother operations
😟 Loss of customer trust	🤝 **Customer confidence** through reliability
⚠️ Risk to driver safety	🛡️ **Improved vehicle & driver safety**
😞 Damage to brand reputation	🌟 **Stronger brand image & trust**

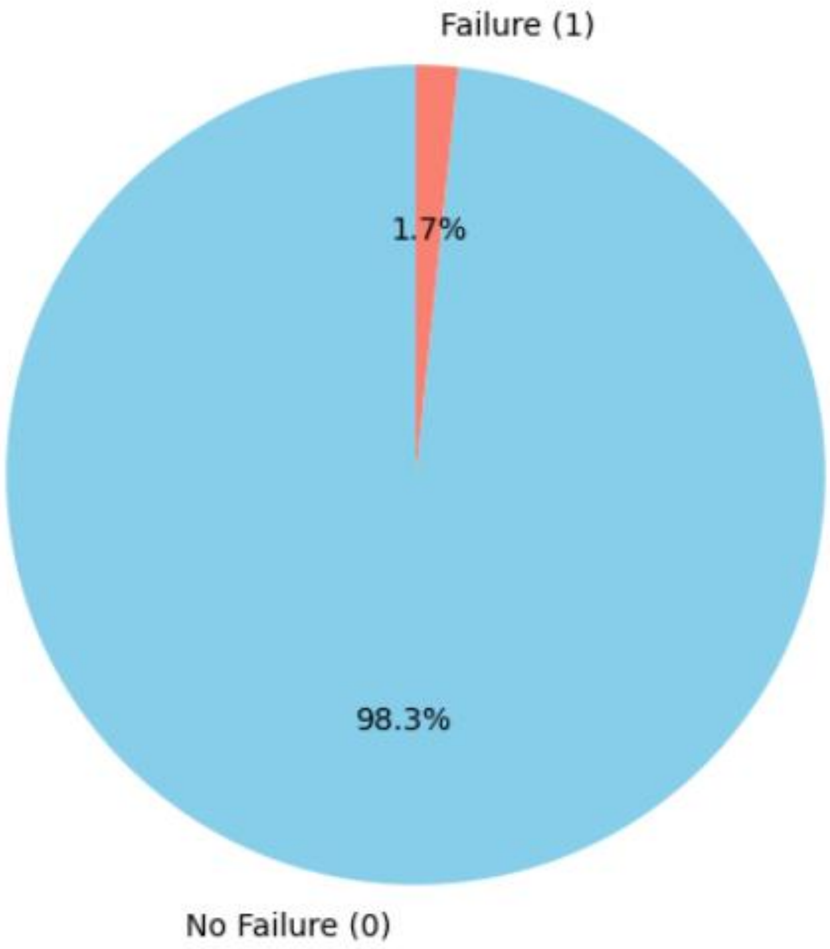
Dataset Information

- **170 Sensor Features**
  (Numerical readings from vehicle sensors)

- **2-Class Classification Problem**
  - **Class 0**: Normal (Non-failure)
  - **Class 1**: Failure (Fault detected)

- **Train Dataset**
  - **60,000 Total Samples**
  - **1,000 Failure Samples** *(~1.67%)*
  - **Imbalanced**: Majority class dominates

- **Test Dataset**
  - **16,000 Total Samples**
  - **375 Failure Samples** *(~2.34%)*



Class Distribution in Scania Dataset

Preprocessing Pipeline

Handle Missing Values

- Remove features with high null values, Impute with median

Drop Correlated Features

Remove highly correlated (|r| > 0.9), Reduce multicollinearity

Feature Scaling

Normalize features by standardizing scale for model stability

Balancing Data

- Oversampling of minority class for balance distribution

Machine Learning Approaches

Logistic Regression

Interpretable linear model

Fast and simple to deploy

Used as baseline classifier

Gaussian Naive Bayes

Probabilistic, assumes Gaussian features

Effective for high-dimensional numeric data

Fast training and inference

Summary

Both models were evaluated for predictive power and generalization. Chosen for **simplicity**, **speed**, and **suitability** for imbalanced fault detection tasks.

Model Training Process

Step-by-Step Supervised Learning Workflow

Split: Training, Validation, Test Sets

- Divide dataset to prevent overfitting

Logistic Regression

- A classification algorithm for binary outcomes

Gaussian Naïve Bayes

- Probabilistic Classifier based on Bayes' theorem

Evaluation Metrics

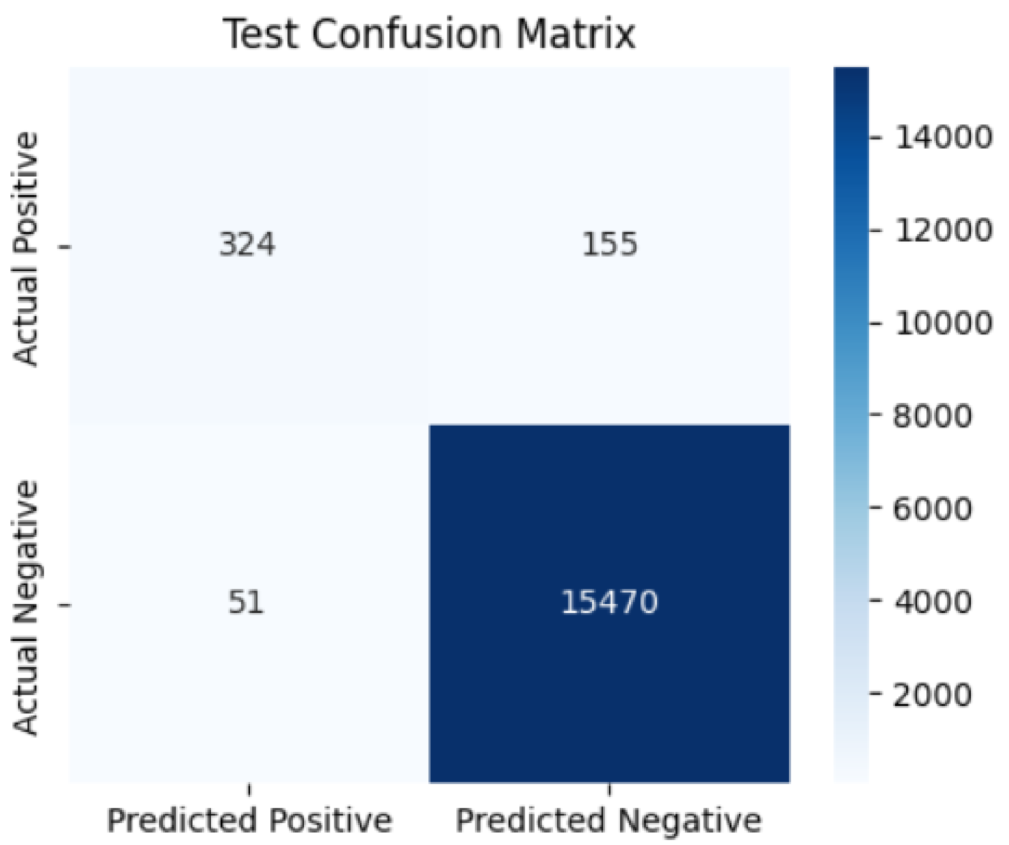
- Measures accuracy, precision, recall, F1-score

Confusion Matrix

- Visualize true vs predicted classifications

Visual Confusion Matrix

Term	Meaning	Real-World Interpretation
TP	Failure correctly predicted	You scheduled maintenance → avoided breakdown ✅
FP	Predicted failure, but truck was actually fine	You serviced unnecessarily → wasted maintenance cost ⚙️
FN	Predicted fine, but failure actually happened	Truck broke down unexpectedly → costly downtime 🚨
TN	Correctly predicted fine	Truck runs smoothly ✅



- **TP: 324** | **FP: 155**
- **FN: 51** | **TN: 15470**
- FN (Missed Failure) – High Risk – Low FN, High Recall
- FP (False Alarm) – Moderate Cost – Acceptable to some Extent

Model Performance Result

Focus Metric → Recall (Sensitivity)

- **Recall -** Measures how many *actual failures* - model correctly caught.
- High Recall = Fewer missed failures (low FN)
- Low Recall = You're missing real faults

Logistic Regression		Gaussian Naive Bayes	
TP: 324	FP: 155	TP: 343	FP: 528
FN: 51	TP: 15470	FN: 32	TN: 15097
Accuracy: 0.9871 | Precision: 0.6764 **Recall: 0.8640 | F1: 0.7588**		**Accuracy: 0.9650 | Precision: 0.3938** **Recall: 0.9147 | F1: 0.5506**	

Key Insight:

- Recall is critical → We prefer slightly more **false positives (FP)** rather than **missing actual failures** (**FN**).
- Logistic Regression outperformed Gaussian Naive Bayes by achieving a better balance between precision and recall, indicating more reliable fault detection with fewer false alarms.

Conclusion

📊 Key Outcomes

- **Comprehensive data preprocessing**: handled missing values, scaled features, and balanced classes.
- **Models built from scratch** — Logistic Regression & Gaussian Naive Bayes.
- **Best performance:** Logistic Regression with balanced **Precision (0.68)** and **Recall (0.86)**.
- **Naive Bayes strength:** High recall for early fault detection.

💡 Impact

⚙️ **Reduced unplanned breakdowns**	💰 **Lower maintenance costs**
🚛 Improved **fleet reliability**	🔒 Enhanced **driver safety**

🌍 Business Value

- Supports Scania’s vision of **proactive, data-driven maintenance**, enabling smarter logistics and sustainable operational excellence.

Thank you

QUESTIONS?