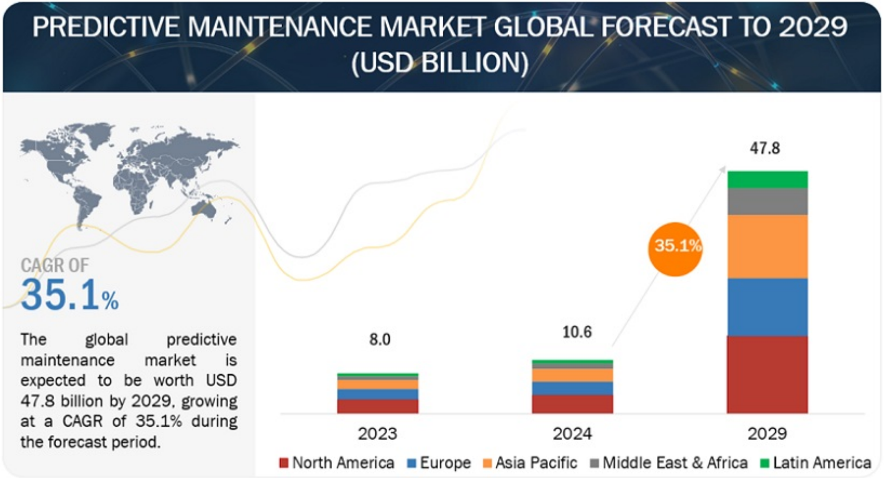
Title:	CRISP Case Study
Course:	Data Mining
Instructor:	Claudio Sartori
Master:	Data Science and Business Analytics
Master:	Artificial Intelligence and Innovation Management
Academic Year:	2024/2025

BOLOGNA BUSINESS SCHOOL

Alma Mater Studiorum Università di Bologna

Developed with the aid of ChatGPT – "The art and science of asking questions is the source of all knowledge" (*Thomas Berger*)

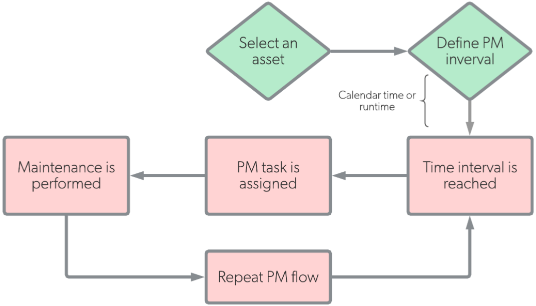


Source https://www.marketsandmarkets.com/

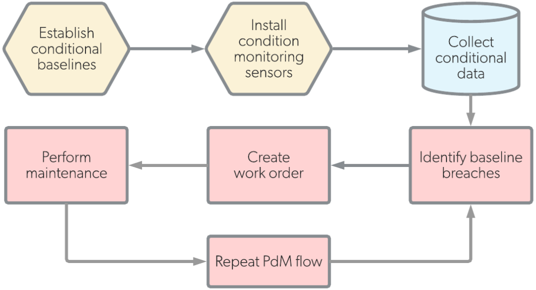
Data Mining for Predictive Maintenance in Industrial Environment

- Predictive maintenance is a *proactive* maintenance strategy that uses data analysis, machine learning, and real-time monitoring of equipment conditions to predict when a machine is likely to fail.
- This approach enables maintenance to be performed just before a failure occurs, minimizing unplanned downtime, reducing costs, and extending the lifespan of equipment
- Data mining techniques uncover patterns and insights from historical and real-time equipment data

Maintenance – Two alternative workflows



Preventive maintenance



Predictive maintenance

Reference: `https://ukeep.com`

Outline

1 Business Understanding 5

2 Data understanding 12

3 Data Preparation 18

4 Modelling 27

5 Evaluation 35

6 Deployment 44

7 Final Notes 53

8 57

Industrial context

The case study focuses on applying data mining techniques to predictive maintenance in a manufacturing environment, specifically targeting the maintenance of Computer Numerical Control (CNC) machines

Overview of CNC Machines

- Role in Manufacturing
  - Automated tools used to manufacture precision parts and components
  - Commonly employed in industries such as aerospace, automotive, electronics, and heavy machinery
- Importance
  - operate continuously under high precision requirements
  - Unplanned downtime can lead to significant losses in production, missed delivery deadlines, and increased operational costs
- Vulnerability
  - Key components like spindles, bearings, and motors are prone to wear and tear due to continuous use
  - Environmental factors (temperature, vibrations, . . .) exacerbate degradation

Challenges faced I

- Unexpected Failures
  - Machines often break down without warning, disrupting production schedules
  - Traditional reactive or scheduled maintenance methods fail to prevent such occurrences
- Cost Implications
  - Repairs during unplanned downtimes are expensive and involve replacement of costly components
  - Prolonged downtime impacts production efficiency, labor costs, and customer satisfaction

Challenges faced II

- Data Complexity
  - CNC machines generate massive amounts of operational data from embedded sensors
  - Extracting actionable insights from this data requires advanced techniques like data mining
- Maintenance Scheduling
  - Balancing machine utilization and preventive measures is challenging without accurate failure prediction
  - Over-maintenance wastes resources, while under-maintenance increases the risk of failures

Objective in the Case Study

- Minimize Unplanned Downtime
  - Predict failures before they occur to avoid production halts
- Optimize Maintenance
  - Move from reactive or scheduled maintenance to a predictive approach
- Reduce Costs
  - Prevent major breakdowns by addressing minor issues early
  - Improve maintenance team efficiency by prioritizing tasks based on risk
- Improve Productivity
  - Ensure machines are operational for the maximum possible time
  - Enhance production reliability, especially for tight schedules

Why CNC Machines are Ideal for Predictive Maintenance

- Rich sensor data
  - Equipped with advanced sensors monitoring temperature, vibration, motor current, and more
  - Continuous data generation allows for detailed analysis and pattern recognition
- Impact on production
  - CNC machines are often bottlenecks in production lines, so their uptime is critical
  - Reliable performance has a direct correlation with overall manufacturing output
- Scalability of Solutions
  - Predictive maintenance frameworks developed for CNC machines can be adapted to other critical industrial equipment

Outline

1. Business Understanding 5
2. Data understanding 12
3. Data Preparation 18
4. Modelling 27
5. Evaluation 35
6. Deployment 44
7. Final Notes 53
8. 57

Data Collection: Overview

- Data collection is critical for building an effective predictive maintenance system
- The case study relies on various data sources to monitor machine performance and predict failures

Data Sources I

- Sensor Data
  - Collected from embedded sensors on CNC machines
  - Examples of metrics:
    - Vibration levels
    - Temperature readings
    - Pressure levels
    - Motor current usage
  - Recorded at high frequency (e.g., every second) during machine operation
- Maintenance Logs
  - Historical data detailing:
    - Repair activities
    - Component replacements
    - Failure events

Data Sources II

- Operational Data
  - Includes:
    - Machine workload levels
    - Runtime hours
    - Environmental conditions (e.g., humidity, ambient temperature)
- Quality Control Reports
  - Tracks defect rates in manufactured parts
  - Serves as an indirect indicator of machine performance issues

Challenges in Data Collection

- Noisy Sensor Data
  - High-frequency data often contains noise due to environmental interference or faulty sensors
- Missing Values
  - Occasional connectivity issues result in gaps in data streams
- Imbalanced Dataset
  - Failures are rare compared to normal operation
  - Imbalance makes it harder for predictive models to detect failure patterns

Significance of Collected Data

- Holistic View
  - Combining sensor, maintenance, operational, and quality data provides a complete picture of machine health
- Key Insights
  - Sensor data identifies real-time anomalies
  - Historical logs provide trends and failure patterns
  - Quality control links machine performance to product defects
- Data-Driven Decisions
  - Enables accurate failure predictions and optimized maintenance scheduling

Outline

1 Business Understanding 5
2 Data understanding 12
3 Data Preparation 18
4 Modelling 27
5 Evaluation 35
6 Deployment 44
7 Final Notes 53
8 57

Data Preparation: Overview

- Preparing data for predictive maintenance in CNC machines involves cleaning, organizing, and enhancing data to extract actionable insights
- The process ensures raw sensor, maintenance, and operational data is suitable for machine learning and predictive analytics

Data Cleaning for CNC Machines

- Outlier Removal
  - Sensor data often contains abnormal readings caused by noise or transient events
  - Statistical methods like z-score analysis are used to identify and remove outliers
- Handling Missing Data
  - Causes of missing data:
    - Sensor malfunctions
    - Connectivity issues
  - Methods for imputation:
    - Mean/Median Imputation Suitable for stable, continuous variables
    - k-Nearest Neighbors (k-NN) Leverages similarity among instances for accurate estimation

Feature Selection for CNC Machines

- Key Sensor Features
  - Vibration statistics (e.g., mean, variance, skewness)
  - Temperature patterns (e.g., peak and trend analysis)
- Operational Features
  - Machine workload
  - Runtime and idle time metrics
- Aggregated Features
  - Cross-sensor interactions (e.g., vibration changes correlated with temperature spikes)

Feature Engineering for CNC Machines

- Sensor-Based Features
  - Extract key statistics:
    - Mean, standard deviation, and range of vibration signals
    - Temperature gradients over time
- Domain-Specific Features
  - Derived using knowledge of CNC machine operations:
    - Rate of spindle speed variation
    - Frequency of abnormal motor current spikes
- Cross-Feature Aggregation
  - Combine data from multiple sensors to uncover complex patterns
  - Example:
    - Correlating high temperature with rapid vibration changes to predict bearing wear

Normalization for CNC Machines

- Purpose
  - Ensure data from different sensors (e.g., temperature in °C, vibration in m/s^2) is on a comparable scale
- Methods
  - Min-Max Normalization
    - Scales data to a [0, 1] range
  - Z-Score Normalization
    - Standardizes data to have a mean of 0 and a standard deviation of 1

Dimensionality Reduction for CNC Data

- Need
  - High-frequency sensor data often results in high dimensionality
  - Reducing dimensions improves computational efficiency and removes redundant features
- Technique: Principal Component Analysis (PCA)
  - Captures the most critical information by transforming data into principal components
  - Retains significant variance while discarding noise

Final Prepared Dataset

- Integrated and Cleaned Data
  - Combines historical maintenance logs, operational data, and sensor data
- Key Features
  - Statistical metrics (e.g., mean, standard deviation)
  - Domain-specific insights (e.g., heat dissipation rate)
  - Aggregated cross-sensor indicators (e.g., combined vibration and temperature trends)
- Normalized and Reduced
  - Scaled and transformed data ready for predictive modeling

Significance of Data Preparation

- Improved Prediction Accuracy
  - Clean and enriched data leads to better model performance
- Operational Efficiency
  - Focused on relevant features, reducing unnecessary computational overhead
- Actionable Insights
  - Enables early detection of potential failures in CNC machines

Outline

1. Business Understanding 5
2. Data understanding 12
3. Data Preparation 18
4. Modelling 27
5. Evaluation 35
6. Deployment 44
7. Final Notes 53
8. 57

Modeling for Predictive Maintenance: Overview

- Predictive maintenance models aim to forecast failures or predict the remaining useful life (RUL) of CNC machines
- The approach combines machine learning techniques with domain-specific knowledge of CNC operations

Types of Predictive Models I

- Classification Models
  - Objective: Predict whether a failure will occur within a specified time frame
  - Example algorithms:
    - Logistic Regression
    - Support Vector Machines (SVM)
    - Random Forest
- Regression Models
  - Objective: Estimate the RUL of machine components
  - Example algorithms:
    - Linear Regression
    - Gradient Boosted Trees (e.g., XGBoost, LightGBM)

Types of Predictive Models II

- Anomaly Detection Models
  - Objective: Identify abnormal operating conditions indicating potential failure
  - Example algorithms:
    - Autoencoders
    - Isolation Forest
    - DBSCAN Clustering

Model Training Process

- Data Splitting
  - Dataset divided into training, validation, and test sets
  - Ensures robust performance evaluation
- Handling Class Imbalance
  - Failures are rare compared to normal operations
  - Techniques used:
    - Oversampling minority class using SMOTE (Synthetic Minority Oversampling Technique)
    - Undersampling majority class
- Cross-Validation
  - k-Fold Cross-Validation ensures model generalization

Evaluation Metrics

- Classification Metrics
  - Accuracy, Precision, Recall, and F1-Score for binary failure prediction
  - ROC-AUC for assessing overall performance
- Regression Metrics
  - Mean Absolute Error (MAE)
  - Root Mean Square Error (RMSE)
  - R^2 Score
- Anomaly Detection Metrics
  - Precision-Recall Curve for imbalanced datasets
  - Mean Squared Reconstruction Error for autoencoders

Advanced Techniques

- Deep Learning Models
  - Recurrent Neural Networks (RNNs)
    - Capture temporal patterns in sequential sensor data
  - Convolutional Neural Networks (CNNs)
    - Analyze sensor data as images (e.g., spectrograms of vibration signals)
- Hybrid Approaches
  - Combine traditional machine learning with deep learning for feature extraction and prediction
- Transfer Learning
  - Leverage pretrained models for specific failure scenarios

Deployment of Predictive Models

- Real-Time Integration
  - Models deployed on edge devices for real-time failure prediction
  - Data pipelines established for continuous sensor data monitoring
- Periodic Retraining
  - Models updated with new operational and failure data
  - Ensures adaptability to evolving machine conditions
- Integration with Maintenance Systems
  - Predictive outputs trigger automated maintenance scheduling
  - Reduces human intervention and response time

Outline

1. Business Understanding 5
2. Data understanding 12
3. Data Preparation 18
4. Modelling 27
5. Evaluation 35
6. Deployment 44
7. Final Notes 53
8. 57

Evaluation for Predictive Maintenance: Overview

- Evaluation ensures the predictive model's effectiveness and reliability in identifying failures or estimating Remaining Useful Life (RUL)
- It involves measuring performance against specific metrics tailored to the model's objectives

Evaluation Metrics: Classification Models

- Accuracy
  - Suitable for balanced datasets but less informative for imbalanced cases
- Precision
  - Focuses on the fraction of predicted failures that are correct
  - Important when false positives are costly (e.g., unnecessary maintenance)
- Recall
  - Measures the proportion of actual failures that are correctly predicted
  - Critical when missing a failure is unacceptable
- F1-Score
  - Balances false positives and false negatives
- ROC-AUC
  - Evaluates the trade-off between true positive and false positive rates
  - Suitable for comparing different classification models

Evaluation Metrics: Regression Models

- Mean Absolute Error (MAE)
  - Measures the average absolute difference between predicted and actual RUL
  - Easy to interpret and sensitive to large errors
- Root Mean Square Error (RMSE)
  - Penalizes large errors more heavily than MAE
  - Suitable when large deviations are particularly undesirable
- R^2 Score
  - Indicates the proportion of variance in RUL explained by the model
  - Higher values signify better model performance

Evaluation Metrics: Anomaly Detection Models

- Precision-Recall Curve
  - Evaluates performance in detecting rare failure events
  - Focuses on balancing false positives and true positives in imbalanced datasets
- Reconstruction Error
  - Used for models like autoencoders
  - Measures how well the model reconstructs normal behavior, flagging deviations as anomalies

Cross-Validation for CNC Machines

- Purpose
  - Ensures models generalize well to unseen data
- k-Fold Cross-Validation
  - Dataset is split into k subsets (folds)
  - Each fold is used as a test set while the others are used for training
  - Helps assess model stability and reliability
- Time-Based Validation
  - For sequential sensor data, ensures training data precedes test data
  - Prevents data leakage and ensures realistic evaluation

Interpretation of Results

- Threshold Tuning
  - Adjust decision thresholds based on evaluation metrics
  - Trade-offs:
    - Higher recall often reduces precision
    - Balance depends on operational priorities
- Root Cause Analysis
  - Evaluate feature importance to identify failure drivers
  - Helps optimize CNC machine operations
- Model Comparisons
  - Compare multiple models using consistent metrics and validation methods
  - Select the model with the best trade-off between accuracy, complexity, and interpretability

Challenges in Evaluation

- Imbalanced Datasets
  - Failure events are rare, leading to biased accuracy
  - Metrics like precision, recall, and F1-score are preferred
- Dynamic Conditions
  - Machine operating conditions vary over time
  - Continuous retraining and re-evaluation are required
- Complex Failure Patterns
  - Subtle anomalies may be missed by simple models
  - Advanced evaluation metrics (e.g., precision-recall curves) provide deeper insights

Significance of Evaluation

- Ensures Reliability
  - Models are tested for robustness under real-world scenarios
- Informs Deployment Decisions
  - Helps decide whether a model is ready for real-time integration
- Supports Continuous Improvement
  - Identifies weaknesses to guide model tuning and retraining

Outline

1 Business Understanding 5
2 Data understanding 12
3 Data Preparation 18
4 Modelling 27
5 Evaluation 35
6 **Deployment** 44
7 Final Notes 53
8 57

Deployment for Predictive Maintenance: Overview

- Deployment involves integrating predictive maintenance models into CNC machine workflows
- It ensures real-time failure prediction and supports proactive maintenance decisions
- Key steps include infrastructure setup, integration with existing systems, and model monitoring

Infrastructure Requirements

- Edge Computing
  - Deploy models on local devices near CNC machines
  - Reduces latency for real-time predictions
- Cloud Integration
  - Centralized storage and processing for large-scale data analytics
  - Supports periodic retraining and model updates
- Data Pipelines
  - Establish automated pipelines for continuous data collection, preprocessing, and prediction
  - Ensure data security and compliance with industry standards

Real-Time Model Deployment

- Predictive Models at the Edge
  - Models predict machine health based on live sensor data
  - Outputs are delivered to operators or maintenance systems in real time
- Integration with Machine Control Systems
  - Alerts generated by models trigger actions:
    - Maintenance scheduling
    - Emergency shutdown to prevent damage
- Latency Optimization
  - Ensure prediction speed meets real-time requirements
  - Use optimized algorithms and hardware accelerators

Model Monitoring and Updates

- Performance Monitoring
  - Continuously evaluate prediction accuracy in real-world conditions
  - Metrics to monitor:
    - False positives triggering unnecessary maintenance
    - Missed failures causing downtime
- Drift Detection
  - Identify changes in data distribution due to new operating conditions or equipment upgrades
  - Retrain models periodically to maintain accuracy
- Feedback Loops
  - Incorporate operator feedback and maintenance outcomes to refine models

Integration with Maintenance Systems

- Automated Maintenance Triggers
  - Predictive models send alerts to maintenance management systems
  - Systems schedule maintenance based on severity and priority
- Downtime Minimization
  - Predictions align maintenance with planned downtimes
  - Reduces unexpected halts in production
- User-Friendly Dashboards
  - Visualize real-time machine health and predictions
  - Provide actionable insights to operators and engineers

Scalability and Adaptability

- Scalable Solutions
  - Models are designed to handle increasing numbers of CNC machines and sensors
  - Cloud platforms support horizontal scaling
- Adaptability to New Machines
  - Transfer learning enables rapid adaptation to new CNC models
  - Fine-tune existing models with minimal retraining
- Customizable Pipelines
  - Allow for easy addition or modification of sensors and features
  - Accommodates changes in machine configurations

Challenges in Deployment

- Data Security
  - Ensure compliance with industrial data protection regulations
  - Implement encryption and secure access controls
- Model Reliability
  - Validate models under varied operating conditions
  - Handle edge cases effectively
- Cost of Infrastructure
  - Balance the trade-off between edge and cloud resources
  - Optimize investments in hardware and computational resources

Expected Benefits after Deployment

- Reduced Downtime
  - Predictive alerts prevent unexpected machine failures
- Cost Savings
  - Minimizes unnecessary maintenance and part replacements
- Enhanced Efficiency
  - Enables operators to focus on critical tasks, improving overall productivity

Outline

1 Business Understanding 5
2 Data understanding 12
3 Data Preparation 18
4 Modelling 27
5 Evaluation 35
6 Deployment 44
7 Final Notes 53
8 57

Impact (rough estimate)

- Cost Savings
  - Reduced downtime by 25%
  - Maintenance costs decreased by 15%
- Improved Productivity
  - Production reliability increased by 20%
- Employee Efficiency
  - Focus shifted to high-priority tasks instead of routine inspections

Conclusion

- Data mining techniques proved effective for predictive maintenance in CNC machines
- Combined feature engineering, machine learning models, and real-time monitoring reduced costs and improved efficiency
- Approach is scalable to other industrial environments (e.g., oil refineries, logistics)

Future Work

- Integration with Digital Twins
  - Simulate machine behavior for more robust predictions
- Enhanced Models
  - Incorporate Reinforcement Learning for adaptive maintenance
- Scalability
  - Deploy solutions across diverse facilities

"The only way to discover the limits of the possible is to go beyond them into the impossible."

Arthur C. Clarke

Bibliography

- Shearer, C. (2000).
  The CRISP-DM model: The new blueprint for data mining.
  Journal of Data Warehousing, 5:13–22.

- Wirth, R. and Hipp, J. (2000).
  CRISP-DM: Towards a standard process model for data mining.
  Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.