Alina Hendrix

C951 Intro to AI

Task 3 Attempt 2

Machine Learning Proposal

Project Overview

A. Create a proposal for a machine learning project by doing the following:

1. Describe an organizational need that your project proposes to solve.

- In my role as a machine learning engineer, I noticed that one of the biggest issues our manufacturing team faces is unexpected equipment breakdowns. These disruptions mess with production schedules, drive up repair costs, and throw off delivery timelines. What's missing is a smart way to flag problems before they happen (Siemens, n.d.; IBM, n.d.; Journal of Manufacturing Science & Engineering, 2023).

2. Describe the context and background for your project.

- Most maintenance systems are either reactive (fix it when it breaks) or scheduled (fix it on a calendar), but neither approach is very efficient. We're sitting on piles of sensor data from our machines — temperature, vibrations, energy use, you name it — and we haven't been using it to its full potential. I believe this data holds the key to predicting when things are about to go wrong (Journal of Manufacturing Science & Engineering, 2023).

3. Review **three** outside works that explore machine learning solutions that apply to the need described in part A1.(Approaching Competence)

Note: These works may include interviews, white papers, research studies, or other types of work by industry professionals. Works that support your research may be identified from various sources, including the WGU Library.

1. **Siemens Case Study**
   Siemens implemented a predictive maintenance system using a neural network architecture trained on real-time sensor data, including vibration and temperature readings. The model learned fault patterns based on historical breakdown events and sensor thresholds. Their deployment showed a 20–30% reduction in downtime and significant savings on repair costs. This highlights the efficacy of supervised

learning in manufacturing environments and validates the technical feasibility of my proposed approach (Siemens, n.d.).

2. **IBM Maximo Platform**
   IBM's Maximo leverages machine learning and AI-driven anomaly detection to monitor equipment health in real time. It utilizes techniques like logistic regression and clustering for risk assessment, alongside streaming data pipelines for ingesting telemetry. Its success in predictive maintenance environments provides a model for designing both my alert dashboard and backend architecture. Maximo's user interface principles also informed how I plan to present sensor diagnostics and performance scores to machine operators (IBM, n.d.).
3. **ML in Manufacturing Study (2023)**
   The *Journal of Manufacturing Science & Engineering* (Vol. 145, Issue 3) published a comparative study in 2023 evaluating Random Forest and Gradient Boosting against traditional rule-based methods for equipment failure prediction. The paper reported Random Forest achieving 93% precision and 91% recall on labeled sensor data collected from CNC systems. This supports my proposed algorithm selection and emphasizes the importance of feature engineering, sensor log integration, and validation metrics (Journal of Manufacturing Science & Engineering, 2023).

a. Describe how *each* reviewed work from part A3 relates to the development of your project.

1. **Siemens**: Reinforced my decision to use supervised models and historical fault data to train an algorithm capable of recognizing breakdown patterns — exactly what I'll implement. (Siemens, n.d.).
2. **IBM Maximo**: Influenced my dashboard design and real-time architecture. Their pipeline and alert mechanisms serve as a reference point for my own system integration and user interaction strategy (IBM, n.d.).
3. **2023 ML Study**: Provided performance benchmarks and technical justification for my algorithm choice. Helped me prioritize precision/recall as my main evaluation metrics and encouraged the use of ensemble methods with sensor logs (*Journal of Manufacturing Science & Engineering*, 2023).

4. Summarize the machine learning solution you plan to use to address the organizational need described in part A1.

- My solution is a supervised learning model — specifically, a Random Forest classifier — trained on CNC sensor logs that include temperature, vibration frequency, and energy consumption. It will detect patterns that precede machine breakdowns and output a risk score for each unit over a 72-hour window. These scores feed into a dashboard that flags critical components, enabling operators to take proactive action (*Journal of Manufacturing Science & Engineering*, 2023; Siemens, n.d.).

5. Describe the benefits of your proposed machine learning

- Fewer interruptions from machine failure
- Lower maintenance and labor costs
- Smarter scheduling and planning
- Safer working conditions (IBM, n.d.; Siemens, n.d.)

B. Describe your proposed machine learning project plan by doing the following:

1. Define the scope of the proposed machine learning project.

- Technical Scope
    - Data Engineering: Cleaning, labeling, and transforming raw CNC sensor data into structured formats for supervised learning.
    - Machine Learning Development: Designing, training, and evaluating predictive models (Random Forest or Gradient Boosting) capable of identifying failure precursors (*Journal of Manufacturing Science & Engineering*, 2023).
    - Feature Engineering: Developing custom features from time-series telemetry to improve model accuracy and interpretability.
    - Software Integration: Building a real-time dashboard to visualize health scores, anomaly alerts, and failure risk predictions(IBM, n.d.).

- Operational Scope
    - Workflow Automation: Generating predictive alerts and maintenance scheduling recommendations that integrate with technician routines.
    - User Training & Adoption: Creating documentation and support materials to educate operators and technicians on system usage and alert interpretation.
    - Performance Monitoring: Establishing post-deployment metrics to track model success and trigger retraining or tuning as needed.
    - Ethical Data Practices: Embedding responsible data handling and privacy protocols across all project stages.

- Breadth of Competencies
    - Applied machine learning and statistical modeling
    - Time-series analysis and sensor data interpretation
    - Software development and visualization
    - Operational workflow design and cross-functional collaboration
    - Ethical data handling and compliance strategy

2. Explain the goals, objectives, and deliverables for the proposed project.

- Goals
  The goal of this project is to minimize unexpected CNC machine downtime by proactively identifying failure patterns through machine learning analysis. By shifting from reactive to predictive maintenance, the organization can ensure smoother production cycles, lower repair costs, and improved operational efficiency(Siemens, n.d.; IBM, n.d.).
- Objectives
    - To achieve this goal, the project will deliver measurable capabilities that directly impact maintenance workflows:
        - Prediction Accuracy ≥90% using historical CNC sensor data (temperature, vibration, energy use) (*Journal of Manufacturing Science & Engineering*, 2023)

        - Maintenance alerts generated ≥48 hours in advance, enabling timely technician intervention(Siemens, n.d.)
        - Operational downtime reduced by ≥30% within the first post-implementation quarter(IBM, n.d.)
    - These objectives serve as performance benchmarks for both the machine learning model and the overall operational impact.
- Deliverables
    - To support and activate these capabilities, the following components will be developed:
        - Trained Machine Learning Model
            - A supervised learning algorithm (Random Forest) trained on labeled CNC sensor data to classify machine health and predict failures. (*Journal of Manufacturing Science & Engineering*, 2023).
            -
        - Interactive Operator Dashboard
            - Visual interface for technicians to view real-time system health scores, receive alerts, and take action on flagged equipment(IBM, n.d.)
        - System Usage and Interpretation Documentation
            - Step-by-step training materials for staff to understand how to use the dashboard, interpret risk scores, and respond to alerts effectively.
        - Automated Maintenance Scheduling Script
            - Backend logic that recommends optimized maintenance windows based on risk scores, equipment type, and historical failure frequency(IBM, n.d.).

3. Explain how you will apply a standard methodology (e.g., CRISP-DM, SEMMA) to the implementation of your proposed project.

- I will follow the CRISP-DM framework:

1. Business Understanding: Define the problem of unexpected machine failures and align the project goals with reducing downtime and improving maintenance planning.

2. Data Understanding: Explore sensor logs and maintenance records to assess availability, reliability, and predictive value.
3. Data Preparation: Clean and format the data—handling missing values, outliers, and creating labeled datasets suitable for supervised learning.
4. Modeling: Use machine learning algorithms like Random Forest to train the model and test predictive performance.
5. Evaluation: Measure model accuracy using precision, recall, F1 score, and assess its impact on actual downtime reduction.
6. Deployment: Implement the model into the organization's workflow with real-time alert systems and monitoring tools.

4. Provide a projected timeline for the proposed project, including the start and end dates for *each* task.

Phase	Start Date	End Date	Milestone
Planning	July 20th	July 27th	Project charter and success matrix final
Data Collection	July 28th	August 5th	Sensor logs aggregated and validated.
Data Cleaning	August 6th	August13th	Missing values addressed, labels created.
Modeling	August 14th	August 21st	First model iteration complete.
Evaluation	August 22nd	August 31st	Model performance reviewed and tuned.
Deployment Prep	September 1st	September 7th	Dashboard complete and staff training has begun.

5. List resources (e.g., hardware, software, work hours, third-party services) and *all* associated costs needed to implement the proposed solution.

- Personnel
    - ML Engineer (you) — 100 hours @ $XX/hour
    - Software Developer — 60 hours @ $YY/hour

    - QA Tester — 30 hours @ $ZZ/hour
- Tools & Infrastructure:
    - Cloud compute (AWS EC2) — $500/month
    - Data storage (S3 buckets, encrypted) — $150/month
    - Visualization tools (e.g. Plotly, Dash) — Free/Open source

6. Describe the criteria that you will use to evaluate the success of the project once it is completed.

To evaluate the success of the machine learning project, I will use clear, quantifiable criteria that align directly with project objectives and deliverables:

- Model Accuracy and Timeliness
    - Predictive model must achieve ≥90% accuracy, based on precision, recall, and F1 score.
    - Maintenance alerts must be generated at least 48 hours before predicted failures.
- Operational Impact
    - Downtime reduction of ≥30% in the first post-implementation quarter.
    - At least 80% reduction in emergency repairs compared to previous quarter.
- System Adoption & Usability
    - Successful deployment of dashboard with real-time sensor monitoring.
    - Minimum 85% of maintenance staff trained and actively using the new alert system within 30 days of rollout.
- Data Integrity
    - All telemetry inputs must pass quality checks for completeness and consistency.
    - Zero incidents of sensitive data exposure in the first 6 months post-deployment.

C. Describe the proposed machine learning solution you will use to address the organizational need identified in part A1 by doing the following:

1. Identify the hypothesis of the proposed project.

- Machines exhibit detectable shifts in sensor data—such as temperature spikes or abnormal vibration—before failures occur. By training a model to recognize these patterns, early warnings can prevent unexpected downtime and reduce repair costs.

2. Identify the machine learning algorithm(s) (i.e., supervised, unsupervised, or reinforcement learning) you will implement in your proposed solution.

- I will use supervised learning — specifically Random Forest or Gradient Boosting Trees.

a. Justify the selection of the algorithm in part C2. Include **one** advantage and **one** limitation of the selected machine learning method.

- I selected Random Forest as the primary algorithm for this predictive maintenance solution because of its proven ability to model complex, nonlinear relationships in sensor-based data—especially where patterns are subtle and multi-dimensional.
- CNC machine telemetry typically includes vibration readings, temperature spikes, energy usage, and timestamped events, all of which interact in complicated ways preceding mechanical failure. Random Forest is ideal in this context because:
    - It can handle high-dimensional, noisy input from multiple sensors without requiring excessive preprocessing.
    - Its ensemble structure (using many decision trees) naturally reduces the risk of overfitting, providing robust predictions on real-world manufacturing data.
    - Feature importance scores generated from the model help explain which signals are most predictive, aiding transparency and trust for operators and stakeholders.
- Advantage:

    - Random Forest performs well with mixed and imperfect data—common in industrial environments—making it a strong candidate for predicting breakdowns based on real sensor logs.
- Limitation:

    - It is computationally intensive, especially with large datasets and when optimizing hyperparameters, which may require significant runtime or cloud resources during training.

3. Describe the tools and environments that will be used to develop the proposed machine learning solution, including any third-party code.

- Programming Language: Python, chosen for its flexibility and wide ML ecosystem.
- Development Environments: Jupyter Notebooks for interactive prototyping and VS Code for structured development and collaboration.
- Libraries: Pandas for data manipulation, Scikit-learn for traditional supervised learning models, and XGBoost for efficient gradient boosting.
- Cloud Platforms (Optional): AWS SageMaker or IBM Watson Studio may be used for scalable model training, deployment, and monitoring.

4. Explain the process you will use to measure the performance of your proposed machine learning solution.

- To evaluate how well the predictive maintenance model performs, I'll use both technical metrics and practical indicators:
  - Technical Performance
    - Precision & Recall: Show how accurately the model predicts actual failures.
    - F1 Score: Balances precision and recall to give a clear picture overall.
    - ROC-AUC & Confusion Matrix: Help visualize and assess the model's classification ability.
  - Validation
    - Use train/test split and cross-validation to confirm consistent accuracy.
    - Apply time-series validation to reflect how predictions would work in real time.
  - Operational Impact
    - Track Mean Time to Failure (MTTF) improvements.
    - Measure whether alerts are issued 48+ hours before breakdowns.
    - Compare downtime before and after deployment for a 30% reduction target.
  - Post-Deployment Monitoring

- Watch prediction accuracy live and adjust thresholds if needed.
- Use feature importance tools (like SHAP) to explain predictions clearly.

D. Describe the data for your proposed project by doing the following:

- The primary data sources for this machine learning project will consist of:
    - Internal CNC Sensor Logs:
        - Continuous telemetry generated by CNC Milling Machines, including metrics such as:
            - Temperature fluctuations
            - Vibration intensity
            - Energy consumption
            - Cycle counts
            - System timestamps and error codes
                - These logs are stored on our organization's centralized manufacturing server and maintained by the IT department.

- Internal Maintenance Records:

- Repair dates
- Technician notes
- Component replacements
- Failure classifications
- Downtime duration
    - These records will be used to label the sensor data and train the supervised learning algorithm.
- Supplementary Public Datasets for Validation and Benchmarking:
    - To enrich the model and validate its generalizability, I will incorporate public datasets from:
        - Kaggle: e.g., CNC Machine Condition Monitoring Dataset (subject to license compliance)
        - UCI Machine Learning Repository: e.g., Condition Monitoring of Hydraulic Systems Dataset

2. Describe the data collection method.

- The sensor data used for this machine learning project is collected directly from CNC Milling Machines operating within the organization’s production facilities.

These machines are equipped with embedded telemetry systems that continuously monitor and transmit performance metrics, including:

  - Temperature readings
  - Vibration frequencies
  - Energy consumption rates
  - Cycle counts and runtime durations
  - Error codes and event timestamps

- Data transmission occurs through a wired Ethernet network connected to a centralized server, where logs are timestamped, encrypted, and stored in structured formats (e.g., CSV or Parquet). The data is collected via an internal data pipeline built using Python and SQL, running periodic extracts from machine controllers to the organization's manufacturing analytics system. Each sensor entry is automatically tagged with:
  - Machine ID
  - Operational context (e.g., type of part being machined)
  - Maintenance status flags from existing ERP integration

  *Additionally, manual maintenance reports are input by technicians and synced through a dashboard interface, creating labeled records that can be cross-referenced with sensor events. This tagging allows the machine learning model to correlate changes in sensor patterns with confirmed breakdowns and scheduled servicing events.

a. Discuss **one** advantage and **one** limitation of the data collection method described in part D2.

- Advantage
  - The telemetry system delivers real-time, high-frequency sensor data, which gives the model access to granular and timely insights into machine behavior. This continuous stream of information—such as temperature fluctuations and vibration spikes—enables the early detection of operational anomalies that precede equipment failure. Since the data is collected automatically and consistently, it supports scalable training across multiple machines without requiring manual input.

- Limitation
  - Industrial sensor data often includes noise, redundant signals, and missing values, particularly during power interruptions, system resets, or sensor drift.

These anomalies can distort failure patterns if not properly addressed during preprocessing. Additionally, because the data is collected from a live production environment, external factors like machine workload variability or technician interventions can introduce inconsistencies that must be normalized to preserve model reliability.

3. Explain how you will prepare your data for use by the machine learning algorithm(s) from part C2 for your proposed project, including data set formatting, missing data, outliers, dirty data, or mitigation of other data anomalies.

- To prepare the sensor data for use in supervised algorithms like Random Forest and Gradient Boosting, I'll follow a focused cleaning and formatting process:
    - Formatting: Structure the data into time-based snapshots with features like temperature, vibration, and machine cycles, each labeled based on failure history.
    - Missing Data: Fill gaps using median imputation or forward-fill methods, and drop any columns with excessive missing values.
    - Outliers: Use Z-score or boxplot analysis to detect outliers and cap extreme values to avoid distortion.
    - Dirty Data: Standardize timestamps, remove duplicates, and sync sensor logs with maintenance records.
    - Anomaly Mitigation: Normalize inputs, engineer new features (like rate of change), and watch for data drift over time.

4. Describe behaviors that should be exercised when working with and communicating about sensitive data in your project.

- To responsibly manage and communicate sensitive data throughout this machine learning project, I will follow a comprehensive framework that addresses every stage of the data lifecycle.

1. Data Preparation
    - Remove or mask all personally identifiable information (PII) such as employee IDs, shift logs, or human-input annotations.

    - Apply consistent anonymization practices using cryptographic hashing and tokenization, ensuring that training data is disconnected from identifiable users.
    - Validate the integrity of anonymized datasets before use to prevent re-identification risk.
2. Data Analysis
    - Focus strictly on equipment and system-level insights; avoid associating sensor patterns or predictions with individual employees or teams.
    - Use fairness audits to ensure that algorithmic outputs are free from bias or unintended discrimination.
    - Implement tools like SHAP to interpret model predictions transparently and identify any variables contributing to unexpected outcomes.
3. Data Storage
    - Store all raw and processed data in encrypted formats (AES-256), both at rest and during transit.
    - Keep training datasets and model artifacts within secured cloud environments or private servers following IT security protocols.
4. Access Control
    - Use role-based access permissions to limit data visibility based on team responsibilities (e.g., analysts, developers, management).
    - Log all data access events and require multi-factor authentication for users interacting with sensitive infrastructure.
5. Data Dissemination and Communication
    - Share performance reports and prediction outcomes only in aggregated formats—highlighting machine health, not operator behavior.
    - Clearly communicate the intended use of data and predictions in documentation and staff briefings.
    - Establish formal review procedures before publishing or distributing any analysis externally.
    - Ensure that stakeholders are informed about model limitations, confidence intervals, and potential for false positives.
6. Compliance and Ethical Oversight
    - Align all practices with organizational data protection guidelines and external standards such as GDPR and CCPA where applicable.
    - Conduct quarterly audits of data handling, model impact, and communication practices to ensure continued ethical compliance.

E. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

Siemens: AI-Based Predictive Maintenance
Siemens. (n.d.). *AI-based predictive maintenance*. Siemens Global. Retrieved July 12, 2025.

https://www.siemens.com/global/en/products/automation/topic-areas/industrial-ai/usecases/ai-based-predictive-maintenance.html

IBM Maximo Predict
Naviam. (n.d.). *IBM Maximo Predict | AI-Powered Maintenance*. Naviam. Retrieved July 12, 2025.

https://www.naviam.io/products/ibm-maximo-application-suite/predict

Mutsuddi (2023) – Journal Article
Mutsuddi, S. (2023). Machine learning for predictive maintenance in manufacturing industries. *International Journal for Research in Applied Science & Engineering Technology (IJRASET), 11*(4), 885–891.
https://www.academia.edu/104215646/Machine_Learning_for_Predictive_Maintenance_in_Manufacturing_Industries

CRISP-DM Methodology

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. The CRISP-DM Consortium. Retrieved July 12, 2025.

https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf

F. Demonstrate professional communication in the content and presentation of your submission.