PREDICTIVE MAINTENANCE

By Tan Bao Lin

A Data Mining Case Study

Introduction and Data Preprocessing

- **Predictive maintenance minimizes downtime, reduces costs, improves safety, and boosts productivity by preventing machine failures.**
- **This dataset has 10000 rows and 14 columns.**

Outlier removal via Z-score: 168 rows removed → 9832 rows remaining

Features: air temperature [K], process temperature [K], rotational speed [rpm], torque [Nm], tool wear [min] → Remove Target → Remove Type, UDI and Product ID
→6 columns → Discretization of continuous values
→ 12 columns

Data Distribution: No Failure: 9488, Heat Dissipated Failure: 109, Overstrain Failure: 95, Power Failure: 78, Random Failures: 44, Tool Wear Failure: 18 → SMOTE
→ 45612 rows

Failure Type: No Failure: 0, Heat Dissipated Failure: 1, Overstrain Failure: 2, Power Failure: 3, Random Failures: 4, Tool Wear Failure: 5

Machine Learning Models

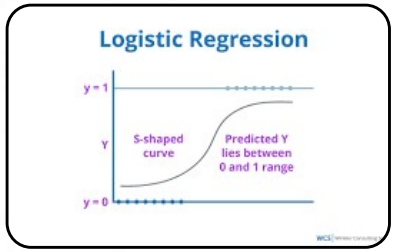
- Logistic Regression
- K-Nearest Neighbor
- Support Vector Machine
- Decision Tree

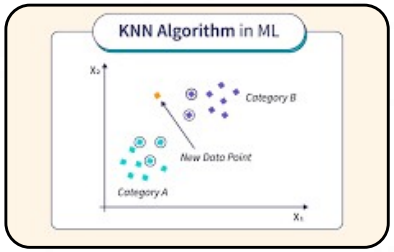
These models were found commonly in AzureML

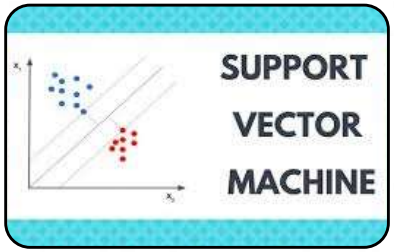
Common ML Algorithm

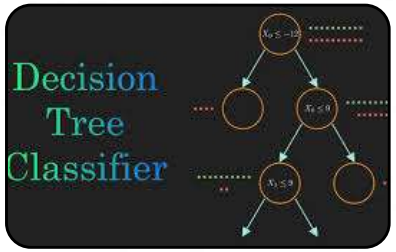
- **Splitting** → 80% train; 20% test
- **Training** → This is where the difference lies
- **Validation** → 10-Fold cross validation, with StratifiedKFold to help maintain class distribution in each fold
- **Evaluation** → Accuracy, standard deviation, confusion matrix, ROC curve, precision recall curve, learning curve

ML Model Differences









Linear

- Fits mathematical equation to the data
- Not flexible for complex, nonlinear patterns

Most Hardworking

Nonlinear

- Stores the entire dataset and evaluates during prediction
- Scales poorly with large datasets

Laziest

Nonlinear

- Maximizes margin between data points and decision boundary
- Takes long time to run due to complexity

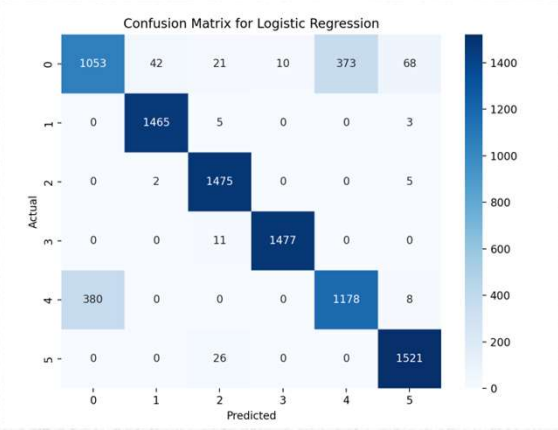
Hardworking

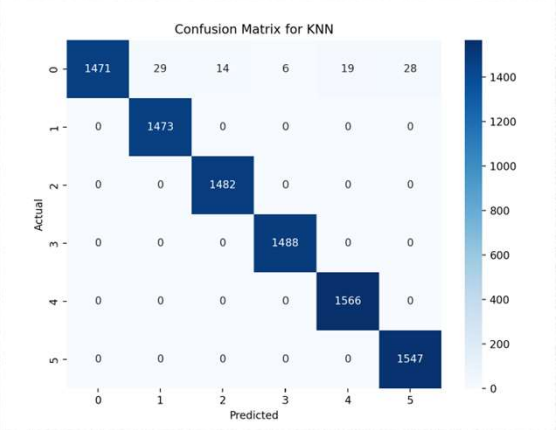
Nonlinear

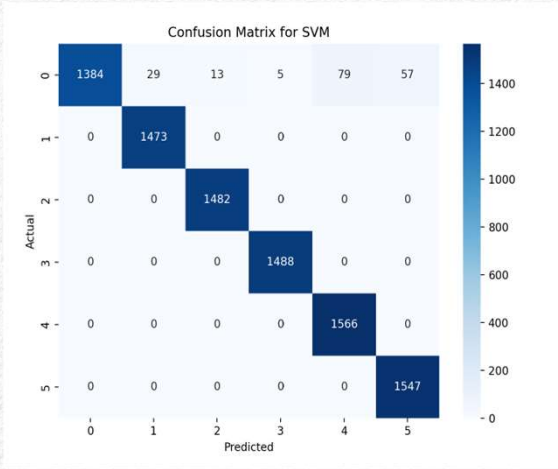
- Build hierarchical tree based on feature splits
- Risk of overfitting

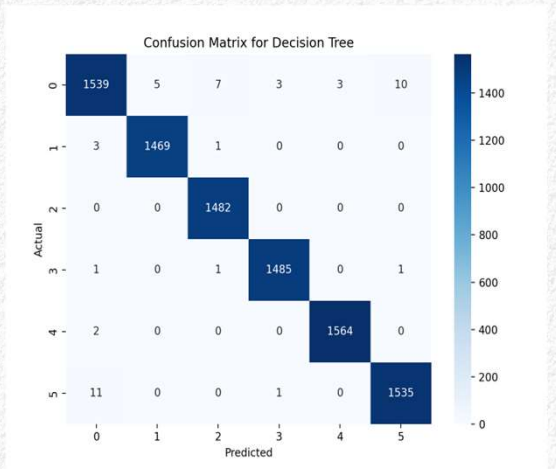
Lazy

Results – Train and Test









Accuracy

- Logistic Regression 90.04%
- KNN 98.81%
- SVM 97.92%
- Decision Tree 100%

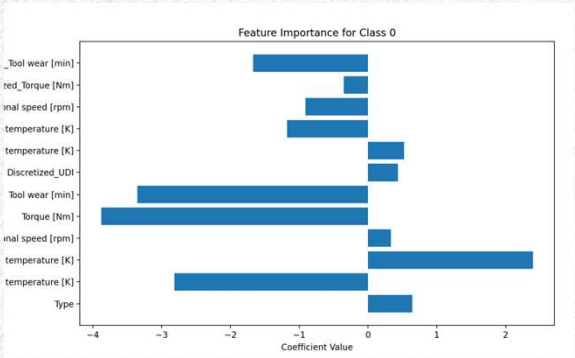
Time taken

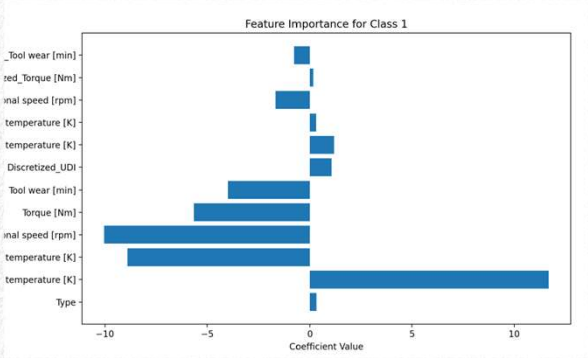
- Logistic Regression 11s
- KNN 2s
- SVM 3 min 26 s
- Decision Tree 2s

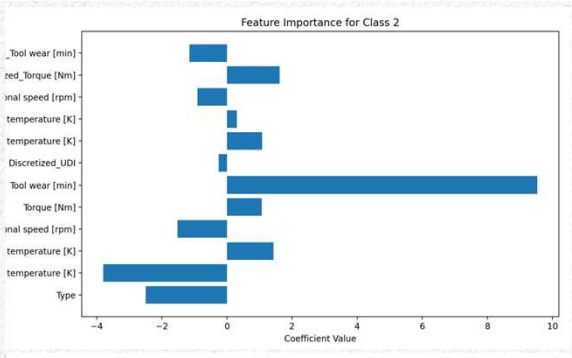
ROC Score Class 0

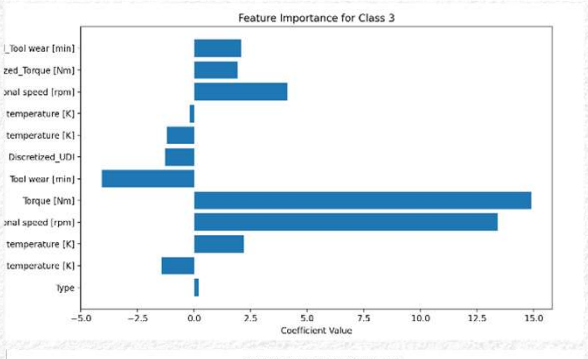
- Logistic Regression 0.94
- KNN 0.99
- SVM 0.99
- Decision Tree 0.99

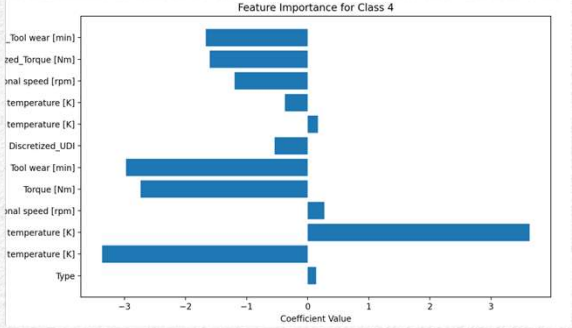
Results - Logistic Regression

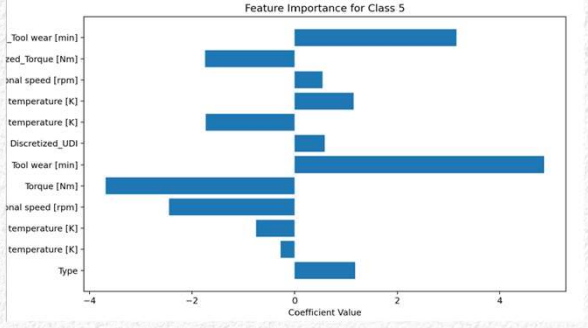












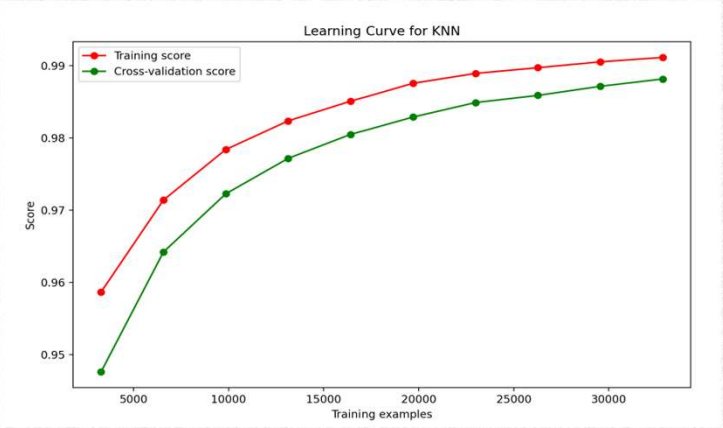
Class 0 (No Failure)

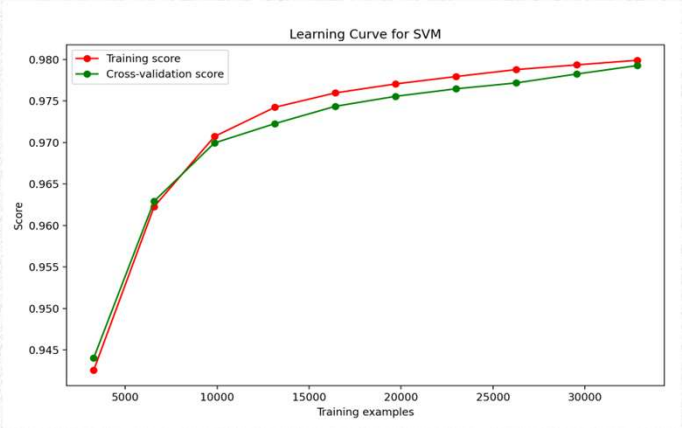
- Precision 0.7522
- Recall 0.6615
- All data seem significant, hard to determine which feature

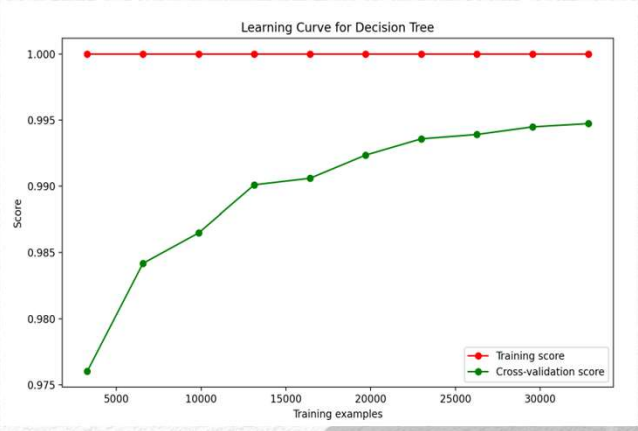
Class 4 (Random Failure)

- Precision 0.7595
- Recall 0.7742
- No correlation due to the failure cause being unknown

Results - Learning Curve







- kNN show consistent gap between training score and cross-validation score, which indicates less noise at the environment
- The high range between 0.95 to 0.99 and similarity of pattern between training score and cross validation score shows a suitable choice of k is used
- Decision Tree show signs of overfitting
- Despite overfitting, the cross-validation score for Decision Tree is still highest (0.975 to 0.995)
- SVM shows very close training score to cross-validation score, indicating well-generalized data and high adaptability to new data

Results – Prediction

Sample	Logistic Regression	KNN	SVM	Decision Tree
sample1.csv: 5 No Failure 1 for each Failure Type	No Failure: 5/5 Failure: 5/5	No Failure: 4/5 Failure: 5/5	No Failure: 4/5 Failure: 5/5	No Failure: 5/5 Failure: 5/5
sample2.csv: 2 for each Failure Type	Heat Dissipation Failure: 2/2 Power Failure: 2/2 Overstrain Failure: 2/2 Random Failures: 2/2 Tool Wear Failure: 2/2	Heat Dissipation Failure: 2/2 Power Failure: 2/2 Overstrain Failure: 2/2 Random Failures: 2/2 Tool Wear Failure: 2/2	Heat Dissipation Failure: 2/2 Power Failure: 2/2 Overstrain Failure: 2/2 Random Failures: 2/2 Tool Wear Failure: 2/2	Heat Dissipation Failure: 2/2 Power Failure: 2/2 Overstrain Failure: 2/2 Random Failures: 2/2 Tool Wear Failure: 2/2
sample3.csv: 10 No Failure	No Failure: 9/10	No Failure: 9/10	No Failure: 9/10	No Failure: 10/10

Based on prediction results, dataset size and effort required, Decision Tree is the suitable model for this case study.

THANK YOU