NORTON UNIVERSITY

GRADUATE SCHOOL

FINAL PROJECT DELIVERABLES

Machine Learning-Based Demand Prediction

for Factory Production Planning of Shampoo, Body Wash, and Laundry Detergent

(Regression)

Course: Machine Learning, **Group:** Group 16
Student: SOUT RAHIM
Lecturer: SEK SOCHEAT
Prepared: August 2026

Final report prepared in journal-style format and aligned with the course final-assignment guideline.

Data Status and Academic Integrity Disclosure

Important: No authorized Medtherm historical dataset was provided with the uploaded proposal. To avoid presenting invented factory evidence as real, this report uses a clearly labeled synthetic demonstration dataset, following the proposal backup plan. The model results below demonstrate the complete ML workflow only and are not claims about actual Medtherm demand. If real records become available, replace the synthetic CSV, rerun the supplied workflow, and update the Results, Discussion, and Abstract.

The synthetic dataset contains the same kinds of variables proposed at midterm - product type, month/season, previous sales, stock, price, promotion, customer orders, previous production, and next-period demand - and is used only to make the final methodology, code flow, evaluation, figures, and production-output format fully reproducible.

Final Deliverables Summary

Deliverable	Included evidence
Final project paper	This document; journal-style structure with methodology, results, discussion, conclusion, references, figures, and tables.
Dataset	Synthetic demonstration CSV with 177 monthly product records from Feb 2021 to Dec 2025; clearly labeled as non-real factory data.
Source code / reproducibility	Python/scikit-learn workflow for preprocessing, model training, evaluation, feature importance, and production recommendation.
Evaluation evidence	MAE, RMSE, R2, actual-vs-predicted visualization, model comparison, and permutation feature importance.
System output demonstration	January 2026 scenario showing predicted demand and recommended production quantities for the three products.
Appendix material	Data dictionary, model configuration, execution steps, sample rows, and submission checklist.

Table of Contents

Data Status and Academic Integrity Disclosure 2
Final Deliverables Summary 2
Table of Contents 3
Abstract 4
1. Introduction 4
1.1 Background and Motivation 4
1.2 Problem Statement 5
1.3 Project Objectives 5
1.4 Research Questions 5
1.5 Target Users and Practical Contribution 5
2. Literature Review / Related Work 5
2.1 Machine Learning for Demand Forecasting 5
2.2 Regression Models Used in This Project 6
2.3 Evaluation and Reproducibility 6
2.4 Relationship to the Proposed Study 6
3. Materials and Methods / Methodology 6
3.1 Project Scope 6
3.2 Dataset Source and Status 6
3.3 Synthetic Dataset Construction 7
3.4 Features and Target Variable 7
3.5 Preprocessing and Leakage Prevention 7
3.6 Train/Test Strategy 7
3.7 Machine Learning Algorithms 8
3.8 Evaluation Metrics 8
3.9 Production Recommendation Rule 8
3.10 Software Tools 8
4. Results 9
4.1 Dataset and Split Summary 9
4.2 Model Performance Comparison 9
4.3 Actual vs. Predicted Demand 10
4.4 Test-Period Demand Trends 11
4.5 Feature Importance 12

4.6 Illustrative January 2026 Production Output 12
5. Discussion 13
5.1 Answer to Research Question 1 13
5.2 Answer to Research Question 2 13
5.3 Answer to Research Question 3 13
5.4 Practical Interpretation 13
5.5 Limitations 14
6. Conclusion and Future Work 14
7. Final Submission Checklist 14
References 15
Appendix A. Reproducibility and Execution Steps 15
Appendix B. Sample Dataset Rows 15
Appendix C. Model Configuration and Output Files 16

Abstract

Accurate demand prediction can support factory production planning by reducing the risk of overproduction, stockouts, and inefficient inventory decisions. This project develops a supervised machine learning regression workflow for shampoo, body wash, and laundry detergent. Because real factory records were not supplied, a synthetic demonstration dataset was used transparently to implement and test the proposed pipeline without claiming real factory performance. The dataset contains 177 monthly product records and includes product type, month and season, previous sales, stock quantity, price, promotion, customer order quantity, and previous production quantity. Three regression algorithms - Linear Regression, Decision Tree Regressor, and Random Forest Regressor - were trained using a time-aware train/test split. Performance was evaluated using mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R2). On the synthetic test set, Linear Regression achieved the best result with MAE 62.5, RMSE 81.2, and R2 0.913. The workflow also produces feature-importance evidence and an illustrative production recommendation based on predicted demand, current stock, and a 10% safety-stock rule. The project demonstrates an end-to-end, reproducible framework that can be rerun with authorized real factory data for operational use.

Keywords: Machine Learning, Demand Prediction, Regression, Production Planning, Random Forest, Linear Regression

1. Introduction

1.1 Background and Motivation

Production planning for fast-moving daily-use products requires a balance between having enough finished goods to satisfy demand and avoiding unnecessary inventory. Shampoo, body wash, and laundry detergent can exhibit changing demand because of seasonality, price, promotions, current stock, and customer orders. Manual planning based mainly on experience or simple averages may not use all available information consistently.

Machine learning offers a structured way to learn relationships between historical business variables and future demand. In this project, demand prediction is formulated as a supervised regression problem in which the target is the quantity expected for the next planning period. The prediction can then be converted into an actionable production recommendation.

1.2 Problem Statement

The factory planning problem is to estimate future demand for three product categories before production decisions are finalized. Inaccurate estimates may lead to overproduction, which increases storage pressure and ties up working capital, or underproduction, which may cause stockouts and delayed customer fulfillment. The proposed solution trains regression models on structured product, sales, stock, pricing, promotion, and order information and compares their predictive performance.

1.3 Project Objectives

- Build a supervised regression model to predict product demand for the next period.
- Compare Linear Regression, Decision Tree Regressor, and Random Forest Regressor using common regression metrics.
- Identify the most useful predictive features for production-planning decisions.
- Produce a clear demand forecast and an illustrative recommended production quantity for each product category.
- Create a reproducible workflow that can later be rerun with authorized real factory records.

1.4 Research Questions

- RQ1: Can available product and business features predict future demand quantity?
- RQ2: Which of the three selected regression algorithms produces the lowest prediction error?
- RQ3: Which input features contribute most to prediction performance and planning interpretation?

1.5 Target Users and Practical Contribution

The intended users are factory managers, production teams, sales/order staff, and inventory personnel. The contribution is not only a numerical forecast; it is a repeatable decision-support process linking data preparation, model comparison, transparent evaluation, and an interpretable production-output table.

2. Literature Review / Related Work

2.1 Machine Learning for Demand Forecasting

Demand forecasting is a central supply-chain task because production, inventory, and fulfillment decisions depend on expectations about future customer needs. Carbonneau, Laframboise, and Vahidov [2] examined machine-learning approaches for supply-chain demand forecasting and showed the relevance of data-driven models for complex demand patterns. Their study supports the general motivation for comparing multiple predictive approaches rather than relying on a single forecasting rule.

2.2 Regression Models Used in This Project

Linear Regression provides a simple baseline and is useful when demand responds approximately linearly to variables such as recent sales, price, and order quantity. Decision trees can capture nonlinear rules and interactions but may overfit small datasets when tree complexity is not controlled. Random Forest, introduced by Breiman [3], averages predictions from many randomized trees and is widely used to improve stability and predictive performance for tabular data. The project therefore compares a transparent linear baseline with two tree-based nonlinear alternatives.

2.3 Evaluation and Reproducibility

Forecast accuracy should be measured with more than one metric because different metrics emphasize different error characteristics. MAE summarizes the average absolute error, while RMSE gives greater weight to larger errors. R2 indicates the proportion of target variance explained by the model. Hyndman and Koehler [4] discuss the importance of appropriate forecast-accuracy measures. The implementation uses scikit-learn, a widely used Python machine-learning library described by Pedregosa et al. [5].

2.4 Relationship to the Proposed Study

The midterm proposal identified demand history, price, promotion, stock, orders, product type, and season as candidate predictors and proposed comparison of three regression algorithms. The final methodology follows that design while adding explicit time-aware splitting, leakage prevention, model-agnostic permutation importance, and a documented rule for converting predicted demand into an illustrative production recommendation.

3. Materials and Methods / Methodology

3.1 Project Scope

The project scope covers three product categories: shampoo, body wash, and laundry detergent. The prediction target is monthly demand quantity. The design is intentionally limited to tabular regression so that the workflow remains understandable, manageable, and suitable for a course project.

3.2 Dataset Source and Status

The preferred source remains authorized historical Medtherm factory records. However, those records were not included in the materials available for this final report. Therefore, the implementation uses the midterm backup plan: a structured synthetic dataset with the same proposed columns. The synthetic data spans February 2021 through December 2025 and contains 177 observations (59 forecast months for each of three products).

3.3 Synthetic Dataset Construction

The demonstration dataset was generated with deterministic random seeds so that the same data can be reproduced. Product-specific demand levels, seasonality, a modest time trend, promotion effects, price sensitivity, recent demand, stock variation, customer-order signals, and random noise were combined to create plausible but artificial monthly values. This design is suitable for testing the workflow, but it must not be interpreted as evidence of actual factory demand behavior.

3.4 Features and Target Variable

Field	Type	Example / Unit	Purpose
product_type	Categorical	Shampoo, Body Wash, Laundry Detergent	Separates product-specific demand patterns.
month	Numeric / calendar	1-12	Captures monthly seasonality.
season	Categorical	Dry-Q1, Hot-Q2, Wet-Q3, Peak-Q4	Adds a broader seasonal representation.
previous_sales	Numeric	Previous month units	Represents recent demand history.
stock_quantity	Numeric	Units in stock	Represents inventory available before production.
price	Numeric	Unit price	Allows demand sensitivity to pricing to be learned.
promotion	Binary	0 or 1	Indicates whether a promotion is planned.
customer_order_quantity	Numeric	Confirmed/expected order units	Provides a near-term business demand signal.
previous_production_quantity	Numeric	Previous production units	Adds recent production context.
target_demand	Numeric target	Current forecast-month units	Quantity to be predicted.

Table 1. Features and target variable used in the demonstration dataset.

3.5 Preprocessing and Leakage Prevention

- Rows are sorted by forecast month so that training data precedes test data in time.
- Categorical variables are imputed if needed and one-hot encoded.
- Numeric variables are median-imputed if needed and standardized inside the training pipeline.
- Preprocessing is fitted on training data only, preventing information from the test period from leaking into training.
- All candidate models use the same training and test periods for fair comparison.

3.6 Train/Test Strategy

A time-aware holdout split was used instead of a random row split. The earliest 80% of forecast months were used for model training and the latest 20% for final testing. This produced 141 training observations and 36 test observations. The test set therefore represents later periods that were not used during fitting, which more closely matches a real forecasting scenario.

3.7 Machine Learning Algorithms

Algorithm	Configuration	Reason for inclusion
Linear Regression	Baseline linear model	High interpretability; appropriate when relationships are approximately linear.
Decision Tree Regressor	max_depth=5; min_samples_leaf=4	Captures nonlinear rules while limiting tree complexity.
Random Forest Regressor	400 trees; max_depth=8; min_samples_leaf=2	Ensemble model intended to improve stability and capture interactions.

Table 2. Candidate regression algorithms and demonstration configurations.

3.8 Evaluation Metrics

The models are compared using MAE, RMSE, and R2. Lower MAE and RMSE indicate smaller prediction errors. Higher R2 indicates that the model explains more of the variation in demand. The best model is selected using test-period performance, with RMSE used as the primary ranking metric and the other metrics used as supporting evidence.

3.9 Production Recommendation Rule

To demonstrate how a demand prediction can become a planning output, the following illustrative rule is applied: recommended production = max(0, predicted demand + 10% safety stock - current stock). This rule is deliberately simple and is not presented as an optimized factory policy. A real deployment should replace the fixed 10% safety factor with a value based on service-level targets, lead time, batch size, capacity, shelf life, and management policy.

3.10 Software Tools

Python is used as the main programming language. pandas and NumPy support data preparation, scikit-learn supports preprocessing, model training, and metrics, and matplotlib is used to generate figures. The workflow can be executed in Jupyter Notebook, Google Colab, or a local Python environment.

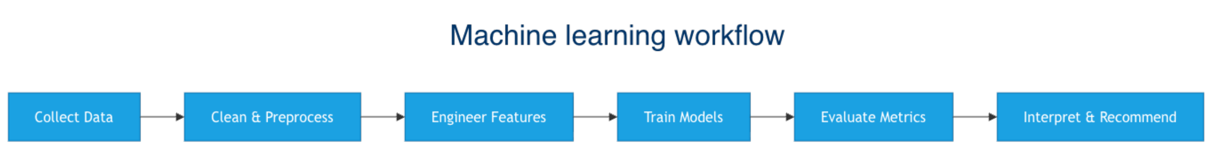

Figure 1. End-to-end machine learning workflow used in the project.

Figure 1 summarizes the reproducible sequence from data collection through interpretation. The same flow can be preserved when synthetic data is replaced with real factory records.

4. Results

4.1 Dataset and Split Summary

Item	Value
Total records	177
Products	3
Forecast period	Feb 2021 - Dec 2025
Training records	141
Test records	36
Target	Monthly demand quantity

Table 3. Demonstration dataset summary.

4.2 Model Performance Comparison

Model	MAE	RMSE	R2
Linear Regression	62.54	81.15	0.9133
Random Forest	80.71	102.59	0.8614
Decision Tree	108.53	141.72	0.7355

Table 4. Test-period performance of the three regression models on the synthetic demonstration dataset.

Linear Regression produced the lowest test error in this synthetic experiment, with MAE 62.54, RMSE 81.15, and R2 0.9133. Random Forest ranked second and Decision Tree ranked third. Because the dataset was generated with relatively smooth relationships, the strong performance of the linear baseline is reasonable and should not be assumed to carry over to real factory data.

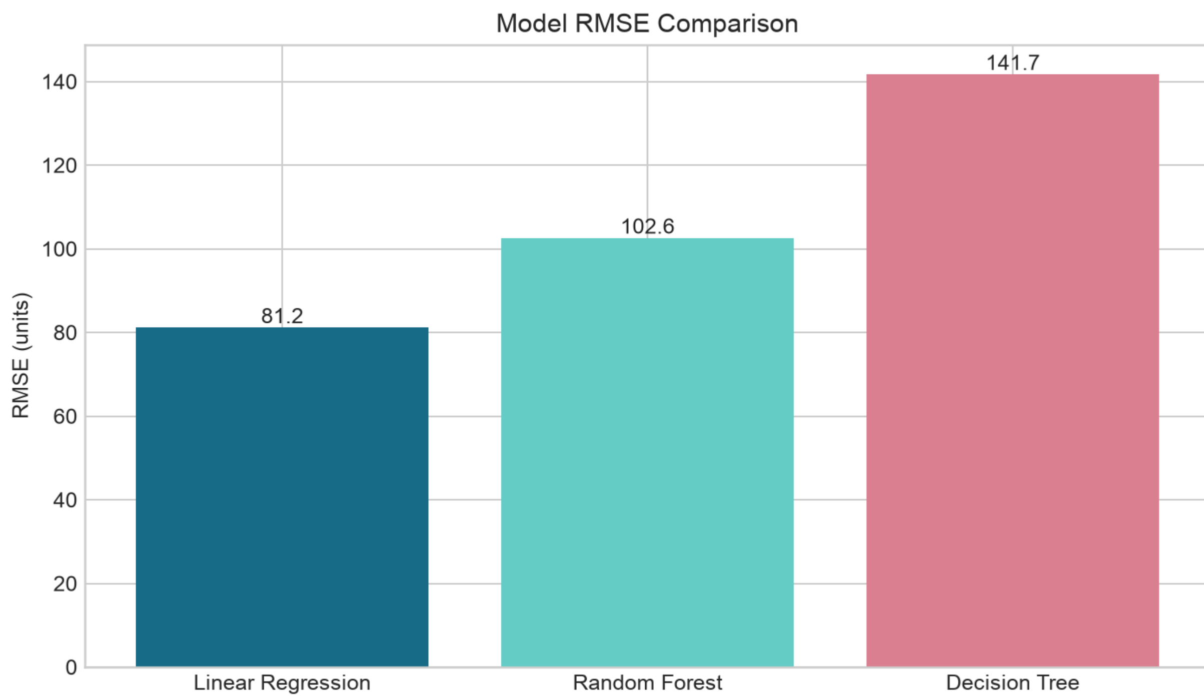

Figure 2. RMSE comparison across candidate regression models.

4.3 Actual vs. Predicted Demand

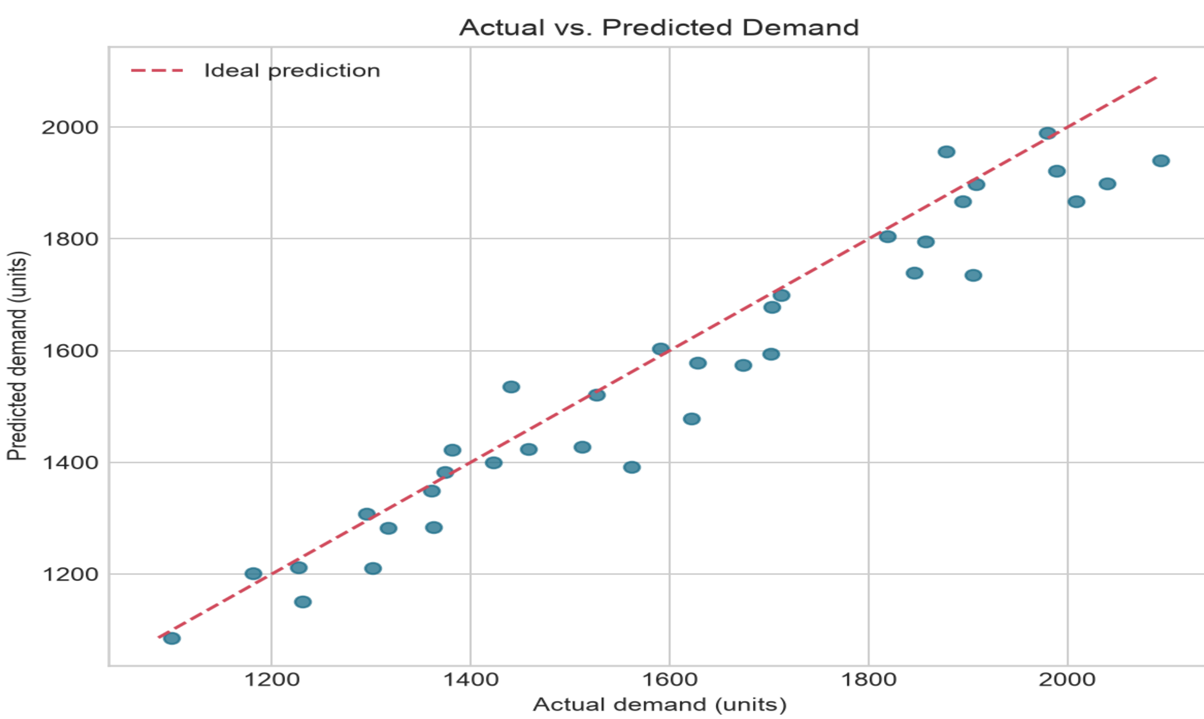

Figure 3. Actual versus predicted demand on the test set using Linear Regression.

Figure 3 shows that most predictions lie close to the ideal diagonal line on the synthetic holdout set. This supports the numerical metrics in Table 4, while still showing that prediction error remains present and should be included in production risk management.

4.4 Test-Period Demand Trends

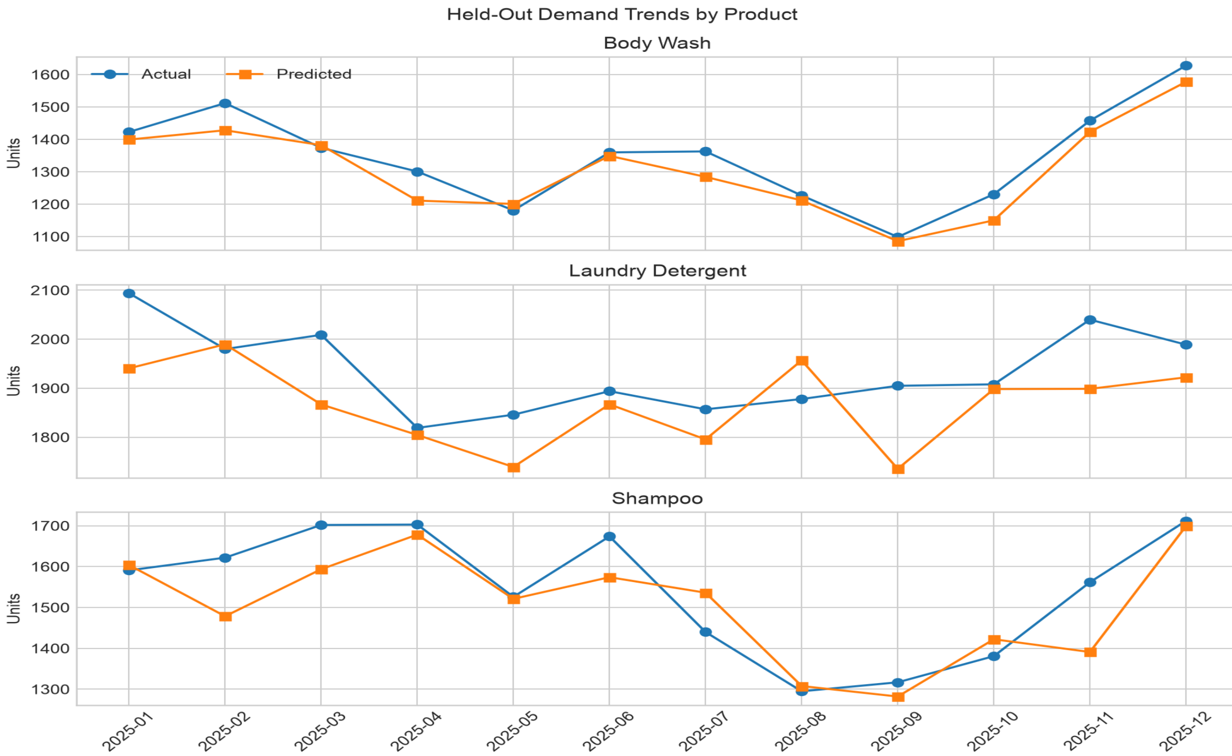

Figure 4. Actual and predicted monthly demand by product during the held-out test period.

Figure 4 provides a time-oriented view of the final holdout predictions. The model follows the broad product-level movement, although individual months still deviate because of noise and changing business inputs.

4.5 Feature Importance

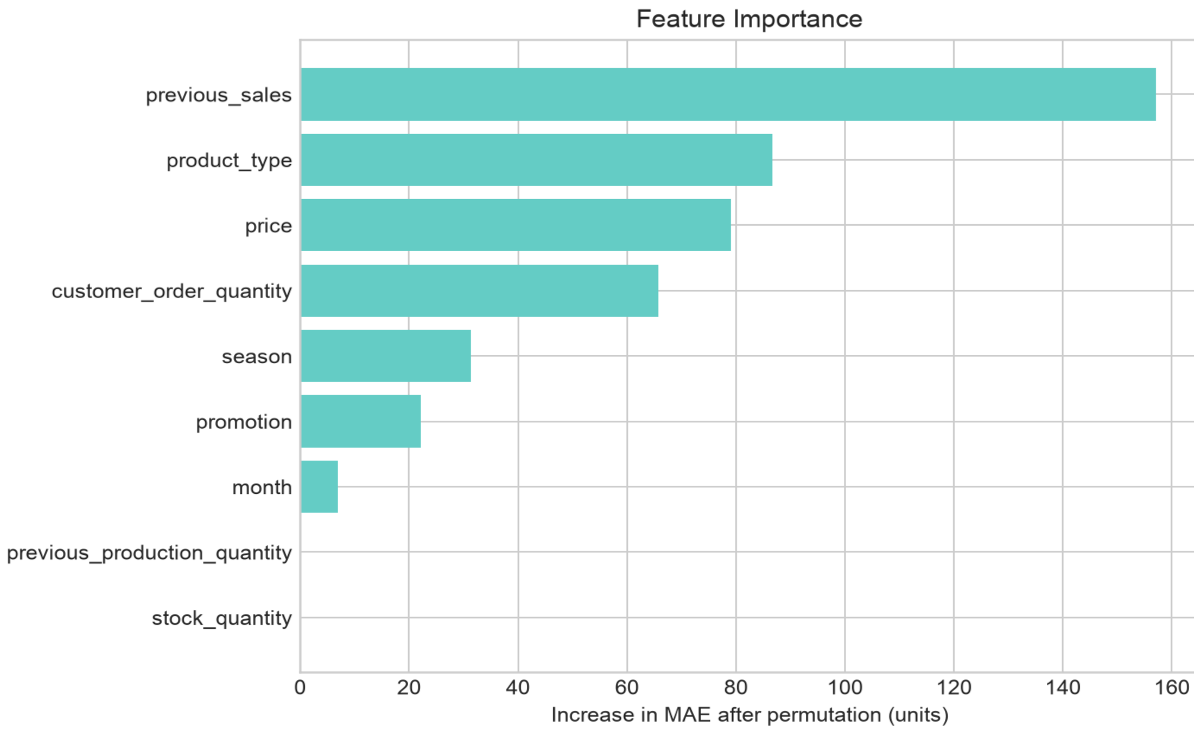

Figure 5. Permutation feature importance for the selected Linear Regression pipeline.

Feature	Permutation importance
previous_sales	157.15
product_type	86.73
price	79.02
customer_order_quantity	65.76
season	31.35
promotion	22.11

Table 5. Highest permutation-importance values on the synthetic test set.

Recent sales and customer-order quantity are the strongest signals in the synthetic data, followed by price and seasonal information. This result is consistent with the construction of the demonstration dataset and should be re-estimated from real data before drawing operational conclusions.

4.6 Illustrative January 2026 Production Output

Product	Predicted demand	Current stock	10% safety stock	Recommended production
Shampoo	1,653	353	165	1,465
Body Wash	1,167	101	117	1,183
Laundry Detergent	1,741	391	174	1,524

Table 6. Illustrative production-planning output for a January 2026 scenario.

Table 6 demonstrates the intended system output: the model forecast is combined with current stock and a simple safety-stock rule to produce a suggested production quantity. The input values for this scenario are synthetic planning values and must not be presented as actual Medtherm orders or inventory.

5. Discussion

5.1 Answer to Research Question 1

Within the synthetic demonstration, the available tabular features predicted demand with relatively low error, as shown by the best-model R2 of 0.913. Therefore, the implemented feature set is technically sufficient to support a regression workflow. This is a demonstration result only; RQ1 must be re-evaluated using genuine factory records before making a real operational claim.

5.2 Answer to Research Question 2

Linear Regression achieved the lowest RMSE on the held-out synthetic period. This finding shows why multiple algorithms should be evaluated rather than assuming that a more complex model will always perform better. On real factory data, nonlinear interactions, missing variables, promotions, holidays, product lifecycle changes, and data quality may change the ranking.

5.3 Answer to Research Question 3

Permutation importance indicates that previous sales and customer-order quantity provide the strongest predictive contribution in the synthetic experiment. Price and seasonal features also contribute. In practical planning, these variables are useful because they are understandable to managers and can often be recorded before the production decision.

5.4 Practical Interpretation

The main practical value is a structured decision-support process. Instead of choosing a production quantity only from experience or a simple average, planners can review predicted demand, recent sales, current stock, confirmed/expected orders, and model error together. The final decision should remain a management decision because production capacity, material availability, lead time, minimum batch size, and service-level requirements are outside the current model.

5.5 Limitations

- The dataset is synthetic and cannot establish real Medtherm forecasting accuracy.
- Only three product categories are modeled, so results cannot be generalized to the entire factory catalog.
- The sample size is small compared with industrial forecasting systems.
- External factors such as holidays, distributor behavior, competitor actions, weather, and marketing campaigns are not fully represented.
- The production recommendation uses a fixed 10% safety-stock rule rather than an optimized inventory policy.
- No formal hyperparameter-search procedure was used; the project emphasizes a clear baseline comparison and reproducible methodology.

6. Conclusion and Future Work

This final project implemented the proposed machine-learning demand-prediction workflow for shampoo, body wash, and laundry detergent. The system includes data preparation, leakage-aware preprocessing, three regression models, standardized evaluation metrics, model comparison, visual diagnostics, feature-importance analysis, and an illustrative production recommendation. On the synthetic demonstration dataset, Linear Regression produced the best holdout performance.

The most important next step is to obtain authorized real factory records and rerun the workflow without changing the evaluation discipline. Future work should expand the feature set to include holidays, lead time, channel/customer information, promotion intensity, and capacity constraints; evaluate time-series cross-validation; tune model hyperparameters systematically; estimate prediction intervals; and integrate the forecast with a production/inventory optimization rule. A simple user interface or dashboard could then present predicted demand, uncertainty, stock position, and recommended production for each product.

7. Final Submission Checklist

Requirement	Status / Evidence
Project includes at least one ML algorithm	Yes - three regression algorithms compared.
Code runs and produces results	Yes - reproducible Python workflow prepared.
Dataset or clear data reference	Yes - synthetic demonstration CSV, clearly labeled; replace with real authorized data if available.
Standard academic paper sections	Yes - abstract through conclusion and references.
Methodology/workflow figure	Yes - Figure 1.
Dataset/features table	Yes - Tables 1 and 3.
Evaluation comparison table	Yes - Table 4.
Graphs/charts	Yes - Figures 2 to 5.
Limitations and future work	Yes - Sections 5.5 and 6.
Consistent references	Yes - IEEE-style numbered references below.

Requirement	Status / Evidence
Presentation slides	Prepare separately for the final defense.
System demonstration	Run the Python workflow and show the generated model results and production table.

Table 7. Final project submission checklist aligned with the course guideline.

References

[1] S. Socheat, "Machine Learning Course: Midterm and Final Group Assignment Guideline," Norton University, Graduate School, 2026.

[2] R. Carbonneau, K. Laframboise, and R. Vahidov, "Application of machine learning techniques for supply chain demand forecasting," European Journal of Operational Research, vol. 184, no. 3, pp. 1140-1154, 2008, doi: 10.1016/j.ejor.2006.12.004.

[3] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001, doi: 10.1023/A:1010933404324.

[4] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," International Journal of Forecasting, vol. 22, no. 4, pp. 679-688, 2006, doi: 10.1016/j.ijforecast.2006.03.001.

[5] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[6] Group 16, "Machine Learning-Based Demand Prediction for Factory Production Planning of Shampoo, Body Wash, and Laundry Detergent," Midterm Assignment, Norton University, 2026.

Appendix A. Reproducibility and Execution Steps

1. Place the dataset CSV and Python script in the same project folder.
2. Install Python 3 with pandas, NumPy, scikit-learn, and matplotlib.
3. Run the script. It loads the dataset, applies the time-aware split, fits preprocessing only on the training data, trains all three models, and calculates MAE, RMSE, and R2.
4. The script saves model-comparison results, test predictions, feature importance, charts, and the January 2026 demonstration output.
5. If real factory data is supplied, preserve the same column names or update the feature mapping, then rerun the workflow and replace the synthetic results in this paper.

Appendix B. Sample Dataset Rows

Month	Product	Prev. sales	Stock	Price	Promo	Orders	Target
2021-02	Body Wash	1058	328	3.92	0	838	1112
2021-02	Laundry Detergent	1664	445	6.33	0	1141	1636
2021-02	Shampoo	1275	328	4.45	1	1133	1587

Month	Product	Prev. sales	Stock	Price	Promo	Orders	Target
2021-03	Body Wash	1112	139	3.83	0	825	1156
2021-03	Laundry Detergent	1636	426	6.22	1	1248	1727
2021-03	Shampoo	1587	417	4.37	1	1000	1746
2021-04	Body Wash	1156	297	3.82	0	811	1183
2021-04	Laundry Detergent	1727	427	6.22	0	1156	1697
2021-04	Shampoo	1746	416	4.73	1	1005	1727

Table A1. First nine rows of the synthetic demonstration dataset.

Appendix C. Model Configuration and Output Files

Output file	Purpose
synthetic_factory_demand_dataset.csv	Input demonstration dataset.
model_evaluation_results.csv	MAE, RMSE, and R2 for all three models.
test_predictions.csv	Held-out actual demand, predicted demand, and absolute error.
feature_importance.csv	Permutation importance values for model interpretation.
jan_2026_production_recommendation.csv	Illustrative next-period prediction and production recommendation.
model_rmse.png	Model comparison chart.
actual_vs_predicted.png	Prediction diagnostic chart.
demand_trends.png	Time-oriented test-period comparison.
feature_importance.png	Feature-importance chart.

Table A2. Reproducibility outputs generated by the project workflow.

Thank You.