DSK807: Applied Machine Learning

Exam Information

Exam 6-9 January 2026

The portfolio exam consists of the following elements:

1. a **group project** with a written short report (max. 3 pages) describing the group's project
2. a **shared group presentation** of the project with an oral group discussion
3. a **short individual oral exam** with questions on aspects of the implemented project (directly after the shared group presentation)

To achieve a passing grade overall, all elements must independently meet the objectives.
The assessment of element 1 will take place in conjunction with the completion of elements 2 and 3.
***Element 1 counts for 40%**, **element 2 counts for 30%**, and **element 3 counts for 30%**, in which a overall evaluation is applied.*

Exam 6-9 January 2026

- **Exam Duration:** 45 Minutes

  20 minutes presentation and group discussion + 5 minutes per group member.

- **Group Size: 3-4 Students** (Sign Up here)

- **Submission Deadline:** 3rd January 2026

- **Submission via Digital Examen**.

Project 1

Customer Churn Prediction in a Subscription-Based Service

Customer Churn Prediction in a Subscription-Based Service

Task: Tabular Data (Classification)

Dataset: Telco Customer Churn (Kaggle) **WA_Fn-UseC_-Telco-Customer-Churn.csv**

Short Description:

Telecommunication and subscription-based companies rely heavily on churn prediction models to retain customers. YOU will build an end-to-end ML pipeline to classify customers as “likely to churn” or “likely to stay,” using tabular data that combines demographic, behavioral, and usage features.

Task 1: Exploratory Data Analysis (EDA)

- Analyze dataset structure: distributions, outliers, skewness
- Examine correlations among numerical features
- Explore categorical feature cardinality
- Study churn vs. non-churn population imbalance
- Identify missing values and propose imputation strategies

Task 2: Shallow Learning Approaches

Train and evaluate at least three classical models:

- Random Forest, XGBoost / Gradient Boosting
  (others: Logistic Regression, LightGBM allowed as extras)
- Use Ensemble Learner

Expected steps:

- Tune hyperparameters through cross-validation
- Use stratification for splits
- Evaluate using accuracy, confusion matrix, F1

Task 3: Neural Network Models

Feed Forward Neural Networks

- Apply regularization: dropout, batch normalization, L1, L2, Early Stopping, ..etc
- Experiment with different optimizers, learning rate, activation functions,...etc
- Analyze training curves (loss/accuracy)

Autoencoder

- Train an Autoencoder for feature compression
- Feed the latent embedding into shallow ML / FFNN (use the AE for feature extraction)
- Compare performance vs raw features

Task 4: Model Comparison & Interpretability

- Compare performance across shallow models, FFNN, and AE

- Provide a detailed error analysis (misclassified users)

- Use **SHAP values** to explain:
  - Important features
  - Differences between models

- Conclude with:
  - What model is best for deployment?
  - Which features contribute most to churn?
  - How could performance be improved?

Project 2

Predicting House Prices Using Real-Estate Features

Predicting House Prices Using Real-Estate Features

Task: Tabular Data (Regression)

Dataset: Ames Housing Dataset (Kaggle) **AmesHousing.csv**

Short Description:

House price prediction is a classical ML task that combines structured real-estate and socio-economic data. YOU will build models that estimate housing value while analyzing nonlinear relationships, feature interactions, and encoding strategies.

Task 1: Exploratory Data Analysis (EDA)

- Visualize price distribution, skewness, and outliers
- Explore relationships: square footage, number of rooms, location features
- Investigate correlations using heatmaps
- Treat missing values (imputation strategies)
- Encode categorical variables using appropriate techniques

Task 2: Shallow Learning Approaches

Train and evaluate at least two classical models:

- Random Forest regressor, XGBoost / Gradient Boosting regressor
  (others: Linear Regression, LightGBM allowed as extras)
- Use Ensemble Learner

Expected steps:

- Tune hyperparameters through cross-validation
- Use stratification for splits
- Evaluate using RMSE and MAE
- Analysis of prediction error across price ranges

Task 3: Neural Network Models

Feed Forward Neural Networks

- Apply regularization: dropout, batch normalization, L1, L2, Early Stopping, ..etc
- Experiment with different optimizers, learning rate, activation functions,...etc
- Analyze training curves (loss/accuracy)

Autoencoder

- Train an Autoencoder on the full dataset
- Extract latent features
- Feed into RF/FFNN and compare performance

Task 4: Model Comparison & Interpretability

- Compare performance across shallow models, FFNN, and AE

- Discuss overfitting tendencies

- Use **SHAP values** to explain:
  - Important features
  - Differences between models

- Conclude with:
  - Which features drive price most?
  - Does any model capture neighborhood effects better?
  - Recommendations for improving accuracy (more features, geo-encoding, etc.)

Project 3

Plant Disease Recognition From Leaf Images

Plant Disease Recognition From Leaf Images

Task: Image Data (Classification)

Dataset: Plant Disease dataset (Kaggle)

Short Description:

Early identification of plant diseases is important for agriculture. Students will classify leaf images into healthy vs multiple disease categories and evaluate the effect of data augmentation, CNN depth, and transfer learning.

Task 1: Exploratory Data Analysis (EDA)

- Visualize sample images

- Analyze class distribution and imbalance

- Discuss resolution, image quality, color channels

- Prepare train/validation/test splits

Task 2: Neural Network Models

Train custom CNN models

- Multiple convolutional blocks,
- Experiment with different number of filters, kernel sizes, stride, paddings. activation function..etc
- Optimizer comparison: Adam vs SGD with different learning rates
- Dropout, batch normalization, Global Average pooling

Transfer Learning using Pre-Trained CNNs

- Test at least **two** pretrained architectures: ResNet18/34/50 , EfficientNet-B0/B1, MobileNetV3 DenseNet121
- Freeze vs fine-tune layers and compare

Task 3: Autoencoder and ViT

Autoencoder

- Train a Convolutional Autoencoder to reconstruct the leaf images
- Use the latent representation as features for:
  - Random Forest, FFNN, Logistic Regression

    Compare whether compressed representations capture meaningful texture patterns.

Vision Transformer (ViT) BONUS (optional, if you do it you will enhance your grade)

- Use a simple pretrained ViT (tiny or small variant)
- Fine-tune classification head
- Compare sample-efficiency with CNN models

Task 4: Model Comparison & Interpretability

- Compare CNN vs transfer learning vs AE/Vision Transformer
- Discuss overfitting and the role of augmentation
- Use SHAP (DeepSHAP) or Grad-CAM to interpret predictions

Provide insights:

- Which leaf regions influence predictions?
- What resolutions/models are optimal?