Project Report

Data Mining and Maching Learning

Group 42
Zarja Javh Dobernik; s1169080
Lan Stare; s1169977

January 5, 2026

1 Preliminaries (600 words max)

The goal of this project was to design a classification system that performs well in terms of AUC on an unlabeled test set. The training data consists of 4000 labeled examples evenly split between two classes, each described by 64 numerical attributes. The test set contains 50000 unlabeled examples with known class priors of 0.4 for class 1 and 0.6 for class 2.

As an initial step, we expected the data for missing values and anomalies. No missing values were detected, and all attributes were numerical and already scaled well, making extensive preprocessing unnecessary. Standardization was still applied for models sensitive to feature scale; logistic regression, neural networks, and support vector machines.

We evaluated five different classification approaches: logistic regression, decision tree classifier, random forest classifier, neural networks, and support vector machines (SVM). Each model was evaluated using stratified 5-fold cross-validation on the balanced training data, with AUC as the primary metric. We also used additional performance measures such as accuracy, precision, recall, and F1-score because we wanted to gain further insight into model behavior mostly to detect possible overfitting.

Logistic regression served us as a linear baseline. It achieved a mean cross-validated AUC of approximately 0.875, with moderate variance across folds. While its performance was stable and interpretable, it was outperformed by all the other classification techniques used which are more flexible non-linear models, indicating that the class boundary is likely non-linear.

Decision trees achieved a mean AUC of roughly 0.891. However, performance varied across folds, and trees showed sensitivity to training splits,

which suggests that they aren't robust. While decision trees are easy to interpret, they couldn't compare to ensemble methods.

Random forests delivered a substantial performance improvement, achieving a mean AUC of approximately 0.967 with low variance. The ensemble approach reduced overfitting and more effectively captured complex feature interactions. Training AUC reached 1.0, which raised mild concerns regarding overfitting. To confirm that the performance was not due to chance, we used permutation testing.

Neural networks also performed strongly, reaching a mean AUC of around 0.958. Despite this strong performance, they required more careful tuning and were less transparent in terms of decision-making, which complicated validation and interpretation. They were also more computationally expensive.

Support vector machines with an RBF kernel achieved the highest and most consistent performance. With optimized hyperparameters, the SVM obtained a mean cross-validated AUC of approximately 0.978. Although we could observe a small train–validation gap, permutation tests indicated that the learned structure was meaningful (not due to overfitting).

Based on the performance on these classification techniques, the SVM was selected as the final model due to its superior AUC performance, stability across folds, and robustness.

2 Classification Approach (600 words max)

As mentioned in the previous section, the final classification model selected for this project is a support vector machine (SVM) with a radial basis function (RBF) kernel. We decided on SVM due to its performance and its suitability for high-dimensional numerical data with potentially non-linear class boundaries.

All features are numerical and no missing values were present in the dataset, but we still used feature scaling since it's essential for SVM as the kernel function is sensitive to the magnitude of the inputs. Therefore, all attributes were standardized to zero mean and unit variance using a `StandardScaler`. To ensure correct preprocessing, scaling was embedded within a `Pipeline`, so that the scaler was fitted exclusively on training folds during cross-validation and then applied to validation data.

The SVM we implemented uses an RBF kernel, which introduces two key hyperparameters: the regularization parameter C, controlling the trade-off between margin maximization and classification error, and the kernel parameter y, determining the width of the Gaussian kernel. Since only these

two parameters directly affect the model's capacity, hyperparameter tuning was restricted to them. We performed a grid search over a small set of values for C and y, using stratified 5-fold cross-validation and the area under the ROC curve (AUC) as the optimization criterion.

To obtain an unbiased estimate of performance during tuning, the labeled dataset was first split into a training set (70%) and a validation set (30%), while preserving class balance through stratification. The grid search was conducted on the training portion only. The best-performing configuration was found to be C=10 with y=scale, achieving a mean cross-validated AUC of approximately 0.98. The resulting classifier demonstrated stable performance across folds, indicating robust generalization.

The final SVM was configured to output posterior class probabilities by enabling probabilistic calibration. These probabilities were used both for ROC/AUC evaluation and for generating the required submission scores. Since the competition metric is AUC, which depends solely on the ranking of samples, no hard classification threshold was applied. Instead, the predicted probability of belonging to class 2 was used as a continuous score to rank all test instances.

After hyperparameter selection, the optimized SVM pipeline was retrained on the complete labeled dataset to maximize the use of available information. The trained model was then applied to the unlabeled test set to produce one score per sample. These scores were written to a CSV file in the same order as the test data, with lower values corresponding to a higher likelihood of belonging to class 1, as required.

To assess model robustness and detect potential overfitting, several additional diagnostics were performed. The difference between training and validation AUC was examined, revealing a small but acceptable gap, consistent with mild model complexity rather than severe overfitting. Furthermore, a label permutation test was conducted by randomly shuffling class labels and re-evaluating performance. As expected, this resulted in AUC values close to 0.5, confirming that the model's strong performance on the original data reflects genuine structure rather than chance correlations.

Overall, the SVM-based approach provides a strong balance between flexibility, robustness, and ranking performance, making it well suited for the objectives of this classification task.

3 Performance Estimation (600 words max)

The estimated test performance is your estimate

4 Other Matters (250 words max)