Laboratory assignment

Component 4

Authors: Patricia Moga, Ciobanu Sergiu-Tudor
Group: 6

January 13, 2026

1 Related work summary

Various research papers were gathered in order to collate and discuss multiple machine learning results performed on the same dataset. The information for this literature survey was taken mainly from studies based on the Airline Passenger Satisfaction Dataset, publicly sourced from Kaggle. The goal is to evaluate existing performance benchmarks and see how algorithms relevant to our project, specifically K-Nearest Neighbors (KNN) for classification and K-Means Clustering for segmentation, performed on this data.

The study conducted by Ashwika et al. [ADH20] focused on evaluating and comparing the performance of five traditional supervised Machine Learning algorithms for predicting passenger satisfaction. The research included a practical application by implementing a prediction interface using Python's Flask microframework, allowing users to input data and forecast the satisfaction outcome. The dataset required essential preprocessing steps, including removing missing values, dropping non-contributory features, and applying label encoding to all categorical variables. Crucially, the SMOTE (Synthetic Minority Over-sampling Technique) resampling method was applied to address the class imbalance, specifically counteracting the majority of "Neutral/Dissatisfied" passenger records in the dataset. The evaluation compared predictive capabilities across several models, including Random Forest (RF) and K-Nearest Neighbors (KNN). The Random Forest Classifier emerged as the most accurate prediction model, significantly outperforming the other supervised ML algorithms with an accuracy of 0.9471. The K-Nearest Neighbor (KNN) model also delivered a high-performance profile, securing third place with an accuracy of 0.9091, a precision of 0.92, a recall of 0.86, and an F1-score of 0.89, demonstrating a strong balance between performance metrics in classifying passenger satisfaction.

Research by Nurdina et al. [NP23] studied the classification capabilities of the Naive Bayes and K-Nearest Neighbors (K-NN) algorithms for the prediction of passenger satisfaction and the evaluation of service quality. The methodology was constrained, using a subset of only 10 features of the original 25 columns, focusing on: id, gender, type of customer, age, type of travel, class, flight distance, wi-fi service in flight, departure/Arrival time and target label (satisfaction). Training and testing were conducted using a smaller sample of records (not more than 26,000) drawn exclusively from the original testing dataset, with records containing missing or invalid data being explicitly excluded. The entire process was run through the RapidMiner Studio application. The final evaluation showed Naive Bayes achieved a significantly higher accuracy of 0.8448 compared to K-NN's 0.6538. Specifically, Naive Bayes recorded better precision (0.8225), although K-NN managed a slightly higher recall (0.7433). The conclusion emphasized that while Naive Bayes was superior for prediction based on accuracy, results should be examined cautiously due to the limited number of attributes employed and the specific evaluation method.

The research by Huliyad et al. [Hul21] proposed an extensive evaluation system involving several classification models, using a large initial dataset of approximately 130,000 entries. Their methodology heavily focused on data preprocessing and feature engineering to optimize model performance. Key structural changes included removing redundant columns (like the ID), transforming the target Satisfaction class, and critically, removing both Delay features (Departure Delay In Minutes and Arrival Delay In Minutes). Furthermore, the Class feature was simplified by merging 'Eco Plus' into the 'Economy' category. The authors adopted a strict data cleaning procedure, removing records where service ratings (originally 1-5) were evaluated as 0, which they considered 'unanswered' and detrimental to performance, resulting in a reduced and cleaner dataset of about 70,000 entries.The training approach utilized 5-fold cross-validation, where 80% of the data was held for training and 20% for testing. GridSearchCV was employed to systematically find the optimal hyperparameters for all models, including K-Nearest Neighbors (K-NN), where the best performance was achieved with k=7. The Random Forest (RF) model was ultimately identified as the best performer based on all evaluation metrics, achieving a high AUC of 0.99, Precision of 0.97, and Recall of 0.94. The K-NN model also performed exceptionally well due to the stringent data cleaning, securing high scores with an Accuracy of 0.9774, Precision of 0.9529, and Recall of 0.9035.

The research by Kevser [Şah22] focused specifically on unsupervised classification, employing clustering algorithms to segment the airline passenger dataset based on feature similarity. The study utilized only the entries from the original testing dataset.The authors employed a slightly unique approach of conducting the experiment on the airline passenger satisfaction dataset, namely by segmenting the entries in certain groups based on similarity. The study was based on two clustering algorithms, K-Means and DBSCAN, using internal evaluation methods, such as Silhouette Coefficient, Calinski-Harabasz Index, and Davies-Bouldin Index, to appreciate performance. The Elbow Method was used to determine the optimal number of clusters for K-Means, which was set at K=8. K-Means secured slightly better results than DBSCAN, achieving a Silhouette Coefficient of 0.145, a Calinski-Harabasz Index of 2854.193, and a Davies-Bouldin Index of 2.078. This analysis provides a direct benchmark for the unsupervised component of the current project.

References

[ADH20] Ashwika, Dishali G K, and Hemalatha N. Airline passenger satisfaction prediction using machine learning algorithms. *Redshi Arch*, 1, July 2020.

[Hul21] Khodijah Hulliyah. Predicting airline passenger satisfaction with classification algorithms. *IJIIS : Int. J. Inform. Inform. Systems.*, 4(1):82–94, March 2021.

[NP23] Annisa Nurdina and Audita Bella Intan Puspita. Naive bayes and KNN for airline passenger satisfaction classification: Comparative analysis. *J. Inf. Syst. Explor. Res.*, 1(2), July 2023.

[Şah22] Kevser Şahinbaş. Performance comparison of K-Means and DBSCAN methods for airline customer segmentation. *Black Sea Journal of Engineering and Science*, 5(4):158–165, October 2022.