Assignment 3: Data Analytics

188.429 Business Intelligence 2025WS

Student A: Emina Skrijelj (12410074)
Student B: Azra Sisic (12024721)

December 15, 2025

Considering the server was down most of the time in the last few days before the deadline, we have decided to create a manual report which contains all the same information copied from our jupyter notebook template. Alongside the report we have also crated a separate notebook that contains only the code that is needed for solving this exercise (whithout the knowledge graph creation code and the automatic report creation code).

1 Business Understanding

The following questions have been answered together by both student A and B.

- **Define and describe the data source and a scenario in which a business analytics task based on the data set you identified should be solved**

  The dataset consists of 20,000 samples of handwritten English capital letters, where each instance represents a scanned image that has been converted into 16 numeric features describing shape, stroke patterns, and pixel distributions. These features allow the letters to be analyzed without storing the original images.

- **Business Objectives**

  A scenario for using this data is an automated mail-sorting system in a postal company, where handwritten characters on envelopes must be recognized quickly and accurately so that routing and delivery decisions can be made without manual intervention.

  The main business objective is to support automated mail processing by developing a model that can recognize handwritten letters without human involvement. By improving the speed and accuracy of character recognition, the company aims to reduce manual sorting costs, decrease processing time per mail item, and minimize sorting errors that lead to delays.

- **Business Success Criteria** Business success will be achieved if the final system consistently recognizes handwritten letters with high accuracy and operates fast enough for large-scale mail processing. A useful threshold for business success is achieving at least around 90previously unseen samples, reducing manual sorting effort by a measurable amount, and making sure that the automated system does not introduce operational risks like frequent misclassification of visually similar letters. The model must be reliable enough to support continuous deployment in a logistics workflow.

- **Data Mining Goals** The goal is to build a predictive model that can classify each input into one of 26 letter categories based on the extracted numerical features. This includes selecting a suitable machine learning algorithm, performing preprocessing, tuning key hyperparameters, and evaluating the model using appropriate metrics for multiclass classification.

- **Data Mining Success Criteria** The goal is to have a trained model that reaches strong classification performance on a test set, ideally approaching or exceeding 90while maintaining balanced performance across all classes. The process should demonstrate good improvement through hyperparameter tuning and produce reproducible results documented in the provenance graph. Success means that the model meets the technical requirements needed to fulfill the business objectives.

- **AI risk aspects** Some of the AI risks still exist even though the dataset does not involve personally sensitive attributes. The system may misclassify visually similar letters, which could lead to operational errors in automated sorting. The dataset might represent only a limited range of handwriting styles, making the model less robust when applied to real-world data that varies in pen type, writing habits, or scanning quality. Also the system must be monitored for drift and provide fallback options so that uncertain predictions do not automatically trigger incorrect routing decisions. Transparency about model limitations is important to guarantee safe deployment.

2 Data Understanding - Student A

The Letter Recognition dataset was loaded from a CSV file containing 20,000 instances of handwritten English capital letters. Each instance consists of 16 numerical features derived from scanned character images, describing geometric and pixel-based properties of the characters. The target variable represents the letter class from A to Z.

- **Outlier Detection:** Potential outliers were identified using a simple z-score-based approach applied to all numeric feature columns. Values with an absolute z-score greater than 2.2 were flagged as outliers. This threshold is intentionally chosen for demonstration purposes and exploratory data understanding, rather than as a statistically strict criterion. After inspecting the outlier report, the identified extreme values are considered in the context of handwritten character data. No immediate removal is performed at this stage, as such values may represent valid but rare writing styles. A final decision on handling outliers will be taken during the data preparation phase.

- **Skewness for all numerical features:** Skewness was computed for all numerical feature attributes of the Letter Recognition dataset.

  The features represent engineered, image-derived descriptors such as bounding box geometry, pixel distributions, and edge-related measures. Several features exhibit mild positive or negative skewness, which is expected due to the bounded and discrete nature of the underlying values (e.g., integer ranges from 0 to 15).

  No extreme skewness values were observed that would indicate data quality issues. The observed skewness patterns are typical for handcrafted feature representations derived from scanned character images.

- **Missing values:** Checking the dataset for missing values by counting null entries per column. The resulting report summarizes the number of missing values for each feature. The missing values report was inspected. No missing values were detected in any of the dataset columns. Therefore, no data cleaning or imputation is required at this stage.

- **Feature scale activity:** Identifying differences in feature scales by computing minimum, maximum, mean, and standard deviation values for each numeric feature. The observed minimum and maximum values fall within the expected bounded range of the dataset, indicating that the feature values are plausible and no invalid measurements are present. After inspecting the feature scale report, it was observed that although all numeric features share the same bounded range, their standard deviations differ substantially. To prevent features with larger variance from dominating the learning process, the decision was made to apply feature scaling during the Data Preparation phase.

- **Visual exploration:** Visual exploration of the Letter Recognition dataset shows that the target variable (letter classes A–Z) is approximately evenly distributed, indicating no severe class imbalance or underrepresented categories.

  The histogram of the representative numeric feature 'xbox' shows a bounded distribution with values concentrated within the expected range, confirming plausibility of the image-derived measurements and absence of extreme values.

  These visualizations support the suitability of the dataset for multi-class classification and motivate the application of feature scaling in the data preparation phase.

2.1 2.e

Ethical Sensitivity: The Letter Recognition dataset does not contain any personal, demographic, or otherwise ethically sensitive attributes. All instances represent scanned handwritten characters without any linkage to identifiable individuals. Therefore, the dataset poses minimal ethical risk with respect to privacy, discrimination, or unfair treatment of individuals.

Minority Classes and Class Balance: The target variable consists of 26 capital letters (A–Z). Visual inspection of the class distribution indicates that the letter classes are approximately evenly represented, and no severe class imbalance or underrepresented categories were identified. As a result, no specific over-sampling or under-sampling strategies are required.

Potential Bias Considerations: Although no explicit sensitive attributes are present, implicit bias may still arise from the limited diversity of handwriting styles captured in the dataset. For example, the dataset may underrepresent certain writing habits, pen types, or writing conditions. This could affect generalization performance when the model is applied to handwriting data collected under different conditions.

Evaluation Implications: Model evaluation should focus on balanced multi-class metrics (e.g., macro-averaged accuracy or F1-score) to ensure consistent performance across all letter classes.

2.2 2.f

Potential Risks and Additional Biases: Although the Letter Recognition dataset does not contain ethically sensitive or personal attributes, there do exist potential sources of bias and data quality risks. The dataset may not capture the full diversity of real-world handwriting styles, since it is unclear how many writers contributed to the data or whether variations in age, cultural background, writing habits, or motor skills are adequately represented.

Secondly, the dataset only contains English capital letters written using the Latin alphabet, which is quite limiting. A model trained on this data could perform poorly on lowercase letters, non-Latin text etc.

Another potential risk lies in the feature extraction process. Since the dataset contains precomputed numeric features rather than raw images, any bias or loss of information introduced during feature extraction is inherited by downstream models and cannot be corrected at later stages.

Questions for External Experts:

To better assess potential bias and data quality issues, the following questions would need to be clarified by dataset curators or domain experts: - How many distinct writers contributed to the dataset? - Under what conditions were the handwritten characters collected and scanned? - Are certain handwriting styles or character shapes overrepresented? - What preprocessing or normalization steps were applied during feature extraction? - Is the dataset considered representative of modern, real-world handwriting use cases?

3 Data Preparation - Student B

- **Column name normalization**

  Column name normalization was applied by removing leading and trailing whitespace from all attribute names. This step was done because of inconsistencies that have been noticed in the raw dataset, where some feature names contained trailing spaces (e.g., 'xbox '), which led to access errors during data exploration.

  The normalization improves robustness, readability, and reproducibility of subsequent preprocessing and analysis steps, without altering the underlying data values.

- **Separation of feature and target variables**

  The dataset was separated into a feature matrix that contains all numeric descriptors and a target vector that contains the letter labels. This will ensure a clear distinction between the input variables and the prediction target. This was decided in the Data Understanding phase and was done to prepare the data for steps like different transformations and modeling that are to come in this exercise.

- **Encoding**

  The categorical target variable representing handwritten letters (A–Z) has been encodued into numeric class labels using label encoding. This encoding assigns a unique integer value to each letter class and preserves a one-to-one correspondence between the original labels and their encoded representation. This transformation was chosen in the previous phase as a prefect step required to make the target variable compatible with machine learning algorithms in subsequent steps.

- **Scaling**

  Feature scaling was applied using standardization (z-score normalization) to ensure that all numeric attributes contribute equally during model training. Although all features are bounded within a similar range, in the Data Understanding phase we have seen some differences in feature variance. This was done with the idea to improve numerical stability and performance of classification algorithms that will be implemented in the following steps of the exercise. A side-by-side boxplot comparison before and after scaling was created to visually confirm that feature variance was aligned across all attributes. The visualization demonstrates that scaling affects magnitude but preserves the original distributional structure of the data.

Describe other pre-processing steps considered but not applied due to which reason

There are several steps that are usually part of preprocessing but we decided against using them. The argumentation for them is:

1. Outlier removal - considering that our dataset did not have any outliers we did not have the need for performing any steps that are related to outlier handling.
2. Missing value imputation - we have also explored if our dataset had any missing values, and since it did not have any,there was no need to take any steps regarding handling missing values.
3. Feature removal was something that we did consider and discuss, but all 16 numeric attributes represent distinct geometric orpixel-based properties and they seem to all be informative and important for the classification task.
4. Binning or discretization of numeric features was not applied. The features are alreadydiscrete and bounded, so there was no need for these steps.
5. Alternative scaling methods such as min–max scaling were considered. Standardization (zero mean, unit variance) was selected instead, as it is more suitable for models sensitive to feature variance.

Analyze options and potential for derived attributes The potential for derived attributes was evaluated for the Letter Recognition dataset.

Possible derived attributes that we have considered include:

1. Ratios or combinations of existing geometric features (e.g., width-to-height ratio, edge density ratios).
2. Polynomial feature expansions to capture non-linear interactions between attributes.
3. Aggregate measures combining multiple edge or pixel-related features.

None of these options we are done because:

- The existing attributes are already compact, low-dimensional, and semantically meaningful.
- Additional derived features could introduce redundancy and multicollinearity.
- The primary modeling goal is evaluating classification performance using the
- original standardized feature representation.

Analyze options for additional external data sources

Several potential external data sources and additional attributes could in theory be useful. Writer-related metadata (e.g., age group, handedness, or writing style) could help assess whether the model performs consistently across different handwriting styles and detect possible biases, which are some of the issues we spoke about earlier.

Some additional information about data acquisition conditions, such as scan resolution or noise level, could improve robustness analysis, since the dataset contains only clean features only. It could mean a lot for the following steps to know the process of getting these clean features insted of just the end resulta

Using an additional dataset or more of them (e.g., EMNIST or similar benchmarks) could be used for external validation to test generalization.