RMIT University
School of Computing Technologies
Practical Data Science with Python
Assignment 2: Data Modelling and Presentation

Title: Data Modelling and Presentation of the Avila Bible Case

Author: Lance Belen
Student ID: S3944846
Affiliations: RMIT University
Email: S3944846@student.rmit.edu.au

Date of Report: 26/05/2024

I certify that this is all my own original work. If I took any parts from elsewhere, then they were non-essential parts of the assignment, and they are clearly attributed in my submission. I will show I agree to this honor code by typing "Yes": *Yes*.

TABLE OF CONTENTS

ABSTRACT / EXECUTIVE SUMMARY

The Avila Bible case consists of a dataset that has been created from 800 images of the Abila Bible. The Avila Bible is a giant Latin copy of the whole Bible from the 12th century (Stefano, Fontanella, Maniaci, Freca, 2018). Manuscript studies showcase different methods and techniques of analysis that is applicable in different areas of the industry (Herman, Ransom, 2023). Here, this is utilised for the purpose of exploring the world of computing technology through a core step in data science. The Avila Bible Case is the study of standardised handwriting and book typology that dives into the analysis of basic layout features, page organisation, and exploitation of available space which provides significant findings that can be used to develop a predictive system that will identify the scribes who worked together to write the Avila Bible. This report will go through the data science process that I have implemented using the Avila Bible Case dataset to focus on data modelling which is a core step in the data science process. This is intended for the gaining of practical knowledge with the steps of data modelling, presentation and automation. The data is first retrieved and prepared for exploration through data preprocessing. Then, it is modelled through two supervised machine-learning methods that fall under the category of Classification. I used two different methods for modelling, namely the K-Nearest Neighbour method and the Decision Tree method. Finally, the evaluation of these models is done through K-Fold Cross Validation to confirm and compare the effectiveness of the two models in predicting/identifying the scribes given a set of trained typological features/data against the test set.

INTRODUCTION

The core step of the data science process is data modelling. This allows the communication or representation of connections between data points and structure (IBM, 2024). Through this, a system can be created to predict an outcome by providing a set of trained data of features that correspond to actual results. In this case, the scribes are the "results" of 10 typological features that can be used by the model to analyse and learn the traits of each scribe's handwriting and distinguish each of their characteristics in terms of the medieval scripting of the Avila Bible. With the two Classification models, built through the K-Nearest Neighbour method and the Decision Tree method, I was able to reach the goal of the project, which is to develop a scribe recognition system that, to an extent, accurately predicts the scribe given a set of test data.

Before data modelling, I first performed data retrieval and preprocessing which did not require much more cleaning as the dataset, as per Stefano, Fontanella, Maniaci, and Freca, in 2018, the authors of the report "Reliable writer identification in medieval manuscripts through page layout features: The "Avila" Bible case", has already been normalized by using the Z-normalization method. Given that the normalisation of data is for the removal of a possible unorganised nature and redundancy to allow a standardised dataset, I still handled the outliers and checked for missing values to serve as a second layer of data preparation to reach a closer chance of predictive accuracy for the model.

Following preprocessing, I explored the data to further provide an identification of possible inaccuracies or errors and hence, handle these again accordingly. However, the data exploration is mainly to observe how the features in the dataset are depicted and the characteristics or nature that each holds, as well as the relationships between pairs of selected features to reach a plausible hypothesis for the target variable. The features have been separated into three categories; Set 1 is for highlighting chronological and/or typological differences; related to the geometrical properties of the page, set 2 is mainly concerned with the exploitation of each column of the written area, and set 3 provides the characteristics of the scribe in terms of text distribution per row. That being said, the exploration of each of the features is separated into the same three categories as well to individually address the observations that provide signification information accordingly.

The specific categorization of the features:

Set 1 - Intercolumnar distance, Upper margin, Lower Margin
Set 2 - Exploitation, Row number, Modular ratio, Interlinear spacing
Set 3 - Weight, Peak number, Modular ratio/Interlinear spacing

Finally, as mentioned already, data modelling is the last task of this assignment which I performed through the methods K-Nearest Neighbour and Decision Tree, both of which are a Classification method of modelling. The data set has actually already been separated into two sets, the train and test set, so I just re-established these sets of data after the necessary preprocessing and exploration. Training and evaluating the model are the two main tasks associated with data modelling, and are intended to achieve the goal of creating the predictive/recognition system and to verify the accuracy and effectiveness of the model.

METHODOLOGY

This section outlines the systematic approach taken to complete the assignment. The key tasks are: retrieving and preparing the data, data exploration, and data modelling.

1. Retrieving and Preparing the Data

1.1. Data Retrieval

The dataset for this project was sourced from the study by Stefano, Fontanella, Maniaci, and Freca regarding the Avila Bible case in 2018. Their report is entitled “Reliable writer identification in medieval manuscripts through page layout features: The ‘Avila’ Bible case”, and the data retrieved includes 10 typological features and the target variable ‘Class’, which is the scribe. Lastly, the dataset has been normalised using the Z-normalisation method and evenly split into train and test sets. These sets came as a .txt file, so I converted it to a .csv file before importing it into the notebook file through the Python library, Pandas.

1.2 Data Preprocessing

Although the dataset was normalised using the Z-normalisation method, additional preprocessing steps were performed to ensure data quality. First, I checked for any outliers using the IQR. Upon identifying, I reviewed how these should be handled and made the changes accordingly by either removing the outliers or not to prevent skewing the analysis. I also checked the data types of the variables, but since the dataset is already numerical, no further encoding was necessary. Lastly, the normalisation already ensured the data was standardised to ensure consistency in the data range, which provides assistance in having an effective subsequent modelling process.

2. Data Exploration

Data exploration was performed to understand the characteristics and relationships within the dataset. The steps that were taken involved:

1. **Descriptive statistics** - summary statistics (including, but not limited to mean and standard deviation) were computed to understand the distribution of each feature.
2. **Visualisation** - plots were created using the matplotlib and seaborn libraries. Specifically, histograms, heatmaps, regression plots, and line plots were used for the depiction of the characteristics of each individual feature and the relationships in between.
3. **Feature categorisation** - the features were categorised according to how the study was conducted for the case as well:
    - Set 1: Highlighting the chronological and typological differences related to the geometrical properties of a page.
    - Set 2: Focuses on the exploitation of each column of the written area.
    - Set 3: Characterization of the scribe in terms of how they distributed the text per row.
4. **Correlation Analysis** - relationships between features were analysed to identify possible multicollinearity and inform feature selection for modelling.

3. Data Modelling

Both the K-Nearest Neighbor and Decision Tree models are non-parametric models, hence, they are not characterised by a “finite set of parameters and a predetermined functional form” (DeepAI). Their flexibility allows techniques such as Grid Search to be utilized to find the best combination of hyperparameters for the two algorithms.

3.1 Model Preparation

- **Data Splitting**: The dataset was already split into training and test sets to ensure robust evaluation. It was closely evenly split with a 50:50 ratio. I just re-established these sets after preprocessing in the first step.
- **Model Selection**: The K-Nearest Neighbor and Decision Tree algorithms were chosen for their simplicity and interpretability in Classification modelling.

3.2 Model Training

1. **K-Nearest Neighbor**: The number of neighbours, weights (whether uniform or distance), and the value of p (1 = Manhattan distance, 2 = Euclidean distance) were optimised through grid search to balance the bias-variance trade-off.
2. **Decision Tree**: Similarly, the model parameters, maximum depth and criterion, were tuned using grid search to prevent overfitting.

3.3 Model Evaluation

- **Performance Metrics**: The models were evaluated using precision, recall, f1-score, and accuracy.
- **Cross-validation**: Through K-Fold Cross-validation, we've ensured the robustness of the models, avoided overfitting, and evaluated the performance by "folding" or dividing the dataset into K subnets.
- **Model Comparison**: The performance of the KNN model and the Decision Tree model are compared to identify the more effective approach for scribe recognition. This comparison will be included in the Discussion section of this report.

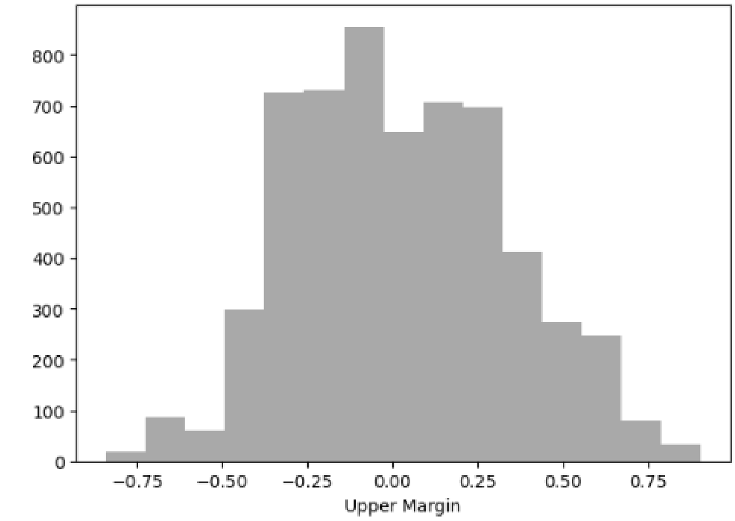
3.4 Model Selection

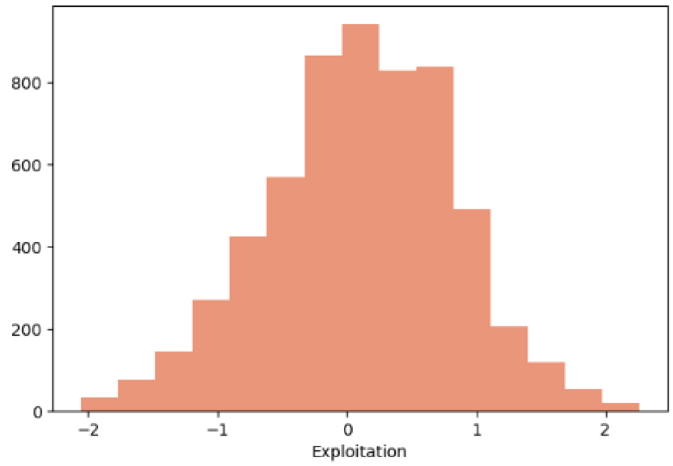
- Based on the evaluation metrics, a conclusion is reached regarding the model that demonstrates the highest accuracy and generalizability as a selection and recommendation of the final predictive model for scribe recognition.

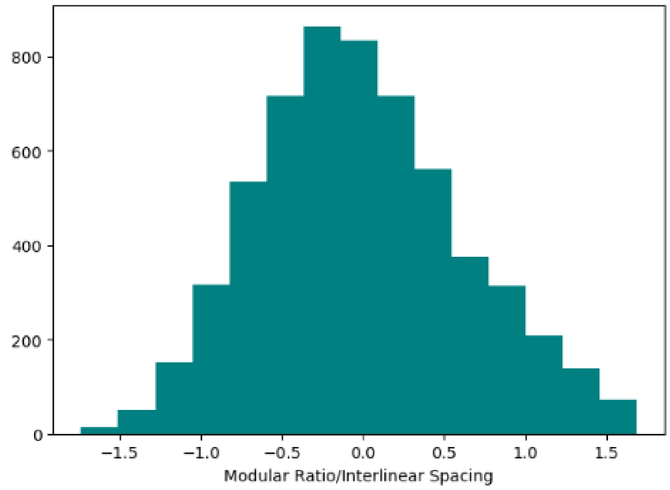
RESULTS

Exploration by Single Feature

As mentioned, the data set has already been normalized for data standardisation. This is observed in the histograms for the plots of each column, and as an example, I've added three plots below, one for a column in each of the three categories in the dataset.







This is evident in the skew of the features being close to 0 as well, hence, we confirm that the data has been standardized.

Exploration by Pairs of Features

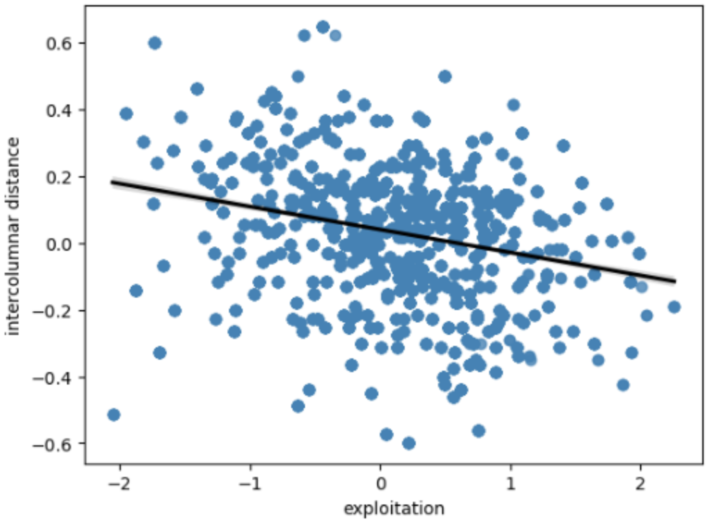
From the plots and statistics that were calculated, the pairs of features that are deemed to show representative and/or significant information are:

a. **Intercolumnar distance and Exploitation**

<u>Hypothesis</u>

- If the exploitation of the column increases, then the intercolumnar distance decreases

From this observation and looking at the correlation between the two features, I have identified that the exploitation has a negative correlation with the intercolumnar distance, which is reasonable. The more that the space in the column is exploited, the less intercolumnar distance there is.

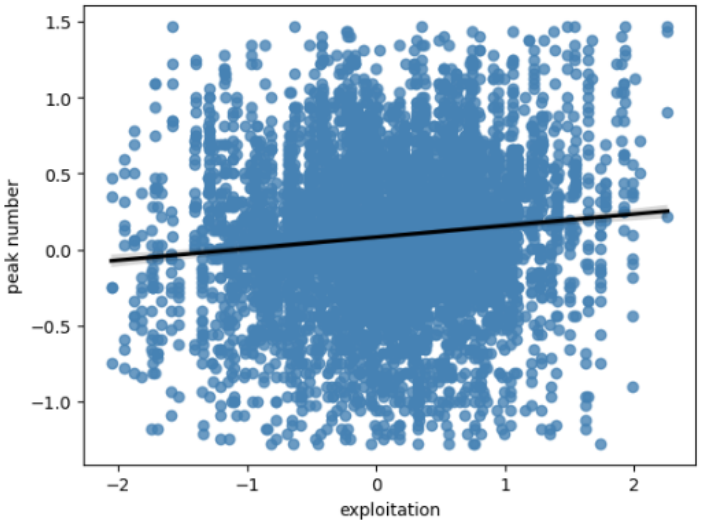


b. Peak number and Exploitation

Hypothesis

- If the peak number increases, then the exploitation also increases.

The peak number is an estimate of the number of characters in a row, which is obtained by counting the number of peaks in the pixel projection histogram on the horizontal axis for the given row. Hence, the observation of the positive correlation between the features is justified. As the number of peaks increases, the exploited area also increases as expected.

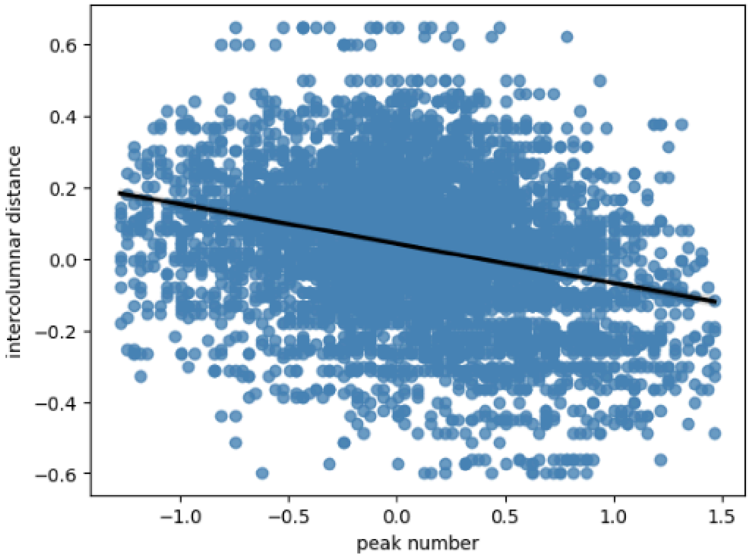


c. Intercolumnar distance and Peak number

Hypothesis

- If the peak number of the row increases, then the intercolumnar distance decreases.

Similar to intercolumnar distance and exploitation, these two features have a negative correlation with one another. The intercolumnar distance is calculated by the distance of pixels between the "centre zones" of two columns. The "centre zone" is obtained through estimation of the peak of the rows, hence, the observation for this pair is reasonable. As the peak number increases, the distance between the “centre zones” decreases as expected because the peak number is associated with how the text is distributed on the given column of the page.

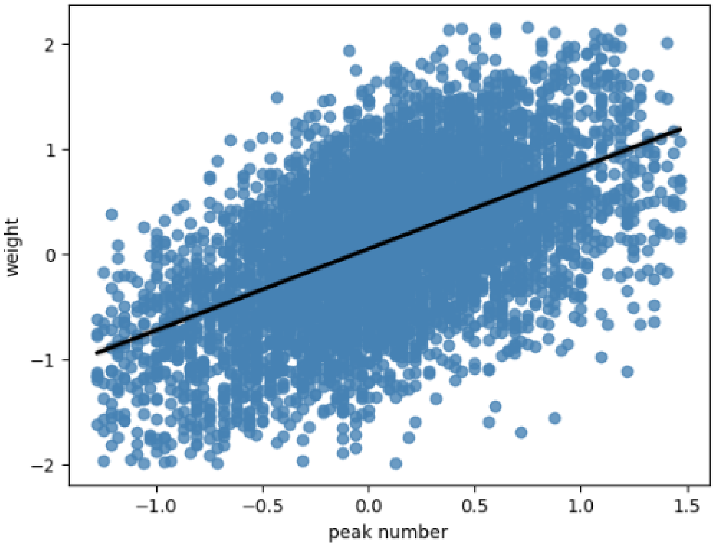


d. Weight and Peak number

Hypothesis

- If the peak number increases, then the weight increases as well.

This pair has the most significant observation among all of the pairs investigated. The weight measures how much a row is filled with ink, similar to exploitation. Hence, the justification of the positive correlation for the features. As the peak number increases, the weight increases as expected because the row contains more ink, and hence, more peaks.

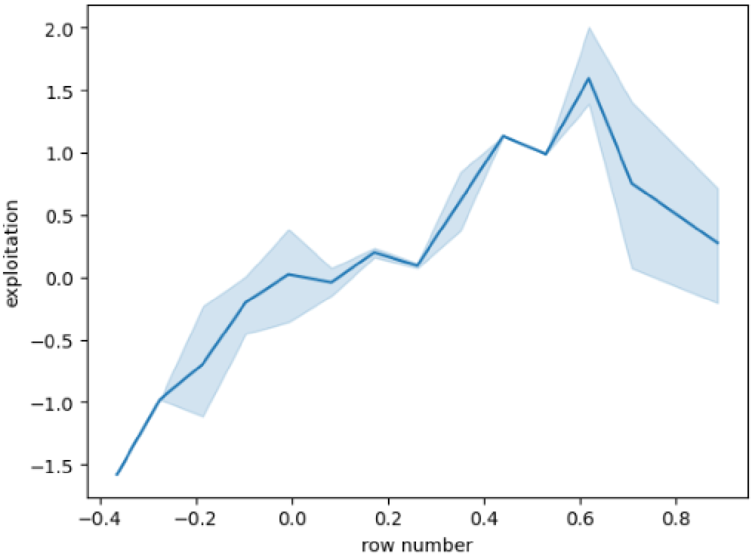


e. Exploitation and Row number

Hypothesis

- If the row number increases, then the exploitation increases as well.

The exploitation and row number has a positive correlation with one another as well. This one could be a coincidence since the row number does not necessarily provide information that characterises the scribe. However, it still provides the observation that as the row number increases, the exploitation increases as well, and this can still provide significant information. From this observation, we can also hypothesize that the scribes who wrote the higher page numbers of the Avila Bible tended to exploit the pages more.



Data Modelling

Comparing the KNN model and the Decision Tree Model, there is a noticeable difference in the statistical metrics, with the latter being more favourable than the former.

Metric	KNN	Decision Tree
Precision (macro avg)	0.87	0.92
Recall (macro avg)	0.73	0.97
F1-score (macro avg)	0.78	0.93
Accuracy	0.82	0.99

The results of K-Fold Cross-Validation performed on each of the models also further justify that the latter proves to be more powerful and effective in identifying the scribe.

Metric	KNN	Decision Tree
Fold 1	0.80	0.97
Fold 2	0.82	0.98

Fold 3	0.79	0.98
Fold 4	0.77	0.99
Fold 4	0.80	0.96

DISCUSSION

Data Exploration

Through statistics and visualisation performed, it can be identified that the most significant contributing factors to scribe distinction are the features that correlate to the exploitation of the page, i.e. the exploitation, weight, peak number, and intercolumnar distance. The normalised dataset also helped maintain consistency and standardised the data across the system, leading to great robustness and a more effective model. This can be proven as the features that do not correlate to the page's exploitation tend to have no correlation with other features at all, such as the upper margin and lower margin. This checks out with the fact that these two features belong to set 1, which is mainly for highlighting chronological and/or typological differences in relation to the geometrical properties of the whole page, not the individualism of the scribes.

Data Modelling

The parameters of the KNN model and the Decision Tree model were selected through the Grid Search hyperparameter optimization technique. This helped me identify which parameters to choose for training the model as it goes through the possible combinations of parameters for the specified machine learning model and provides the best combination that will achieve the best performance. Specifically, through this, I selected the parameters for the KNN model to be n_neighbors=4, weights=distance, metric=minkowski, and p=1. Whereas for the Decision Tree model, I selected criterion=entropy and max depth=18.

As mentioned, these two methods are non-parametric methods, and hence, have the flexibility to select the parameters required for best optimisation. While there may be parameters for the two models that I selected before performing the Grid Search technique or parameters that weren't included at all, it has still been identified that these combinations resulted in the most optimised model accordingly. That is, the parameters selected were identified to be a factor for the performance of the model and the parameters that weren't selected had little to no effect on the performance.

From the two models, as seen in the results, the Decision Tree model proved to be dominant on all of the metrics – precision, recall, f1-score, and accuracy. This is backed up by further evaluation through the K-Fold Cross-Validation technique where the Decision Tree model resulted in scores of at least 0.96 and at most 0.99, in comparison to a minimum of 0.77 and a maximum of 0.82 score for KNN.

CONCLUSION

The Decision Tree model demonstrated superior performance in scribe recognition, achieving an accuracy of 0.99, significantly higher than the KNN model's 0.82. Data exploration revealed that the exploitation of each column in the written area was the most significant contributing factor in distinguishing scribes. Additionally, features correlated to column exploitation further enhanced the model's ability to accurately identify scribes. This high accuracy and reliability suggest that the Decision Tree model is highly effective for distinguishing scribes based on typological features from the Avila Bible, providing a robust tool for historical manuscript analysis. Future work should explore ensemble methods and larger datasets to further enhance model performance and applicability.

REFERENCES

Herman, N., Ransom, L.. 2023. *Manuscript Studies*. University of Pennsylvania Press Journals. https://www.pennpress.org/journals/journal/manuscript-studies/.

Joseph, R. 2018. *Grid Search for model tuning*. Towards Data Science. https://towardsdatascience.com/grid-search-for-model-tuning-3319b259367e

Non-Parametric Model. DeepAI. https://deepai.org/machine-learning-glossary-and-terms/non-parametric-model

Stefano, C., Fontanella, F., Maniaci, M., Freca, A.. 2018. *Avila*. UCI Machine Learning Repository. https://doi.org/10.24432/C5K02X.

Stefano, C., Fontanella, F., Maniaci, M., Freca, A.. 2018. *Reliable writer identification in medieval manuscripts through page layout features: The "Avila" Bible case*. Elsevier. Engineering Applications of Artificial Intelligence 72 (2018) 99–110.

What is data modeling? IBM. https://www.ibm.com/topics/data-modeling.