Title: Practical Data Science with Python

Student ID: s4060643

Student Name and email (contact info): Rishekesh B , rishekeshris@gmail.com

Affiliations: RMIT University.

Date of Report: 29/05/2024

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

The analysis and categorization of handwritten text from the Avila Bible dataset—a collection of multivariate variables taken from historical documents—is the main focus of this study. The twelve classifications in the dataset correspond to various copyists. Several machine learning models, such as Decision Tree, Random Forest, were used to solve the categorization problem. Using oversampling approaches to address class imbalance was a crucial component of this investigation. Box plots and explained variance ratio graphs were used to analyze the relevance of the features. Features like weight, peak number, interlinear spacing, modular ratio, row number, and intercolumnar distance showed substantial heterogeneity in the box plots, underscoring their significance in class distinction. The first few principal components, with a noticeable elbow around the fourth component, appear to account for most of the variation, according to the explained variance ratio graphs. This suggests that these components are essential for dimensionality reduction. Metrics including accuracy, precision, recall, and F1-score were used to assess the classification models; the Random Forest and Decision Tree models performed admirably. The results offer significant insights for further research in this area by demonstrating the efficacy of these models and the significance of feature selection in the classification of ancient handwritten text.

Introduction

Text categorization and pattern recognition are faced with a special problem in the Avila Bible, a massive Latin copy of the Bible from the XII century. Using data extracted from the text, this study aims to accurately describe the copyist responsible for different sections of the Avila Bible by utilizing machine learning techniques. In order to complete this work, a dataset comprising features such intercolumnar distance, upper and lower margins, exploitation, row number, modular ratio, interlinear spacing, weight, peak number, and the ratio of modular ratio to interlinear spacing must be analyzed. The dataset was created using 800 photographs of the Avila Bible.

Within machine learning, imbalanced data, poor feature selection, and inaccurate models are common problems in classification tasks. This study combines the Decision Tree and Random Forest classifiers, two potent machine learning methods, to tackle these problems. These two models are perfect for this classification task since they can both handle large, complicated datasets and produce results that are easy to understand.

The objectives of this research are multifaceted:

- Preprocess and Dataset Preparation: Make sure the dataset is standardized, balanced, and clean for efficient analysis.
- Apply and Contrast Random Forest and Decision Tree Classifiers: Utilize the dataset to apply these classifiers, then assess how well they function.
- Handle Class Imbalance: To solve the issue of class imbalance and enhance model generalization, use strategies such class weight modification and oversampling of majority classes.
- Analyze Model Performance: Determine each classifier's efficacy using metrics such as accuracy, precision, recall, and F1-score.

By accomplishing these goals, this study not only advances the area of digital humanities by offering insights into the classification of historical texts, but it also demonstrates how sophisticated machine learning techniques may be applied to solve challenging classification problems. The methods used, the outcomes, and the ramifications of these findings are described in depth in the parts that follow.

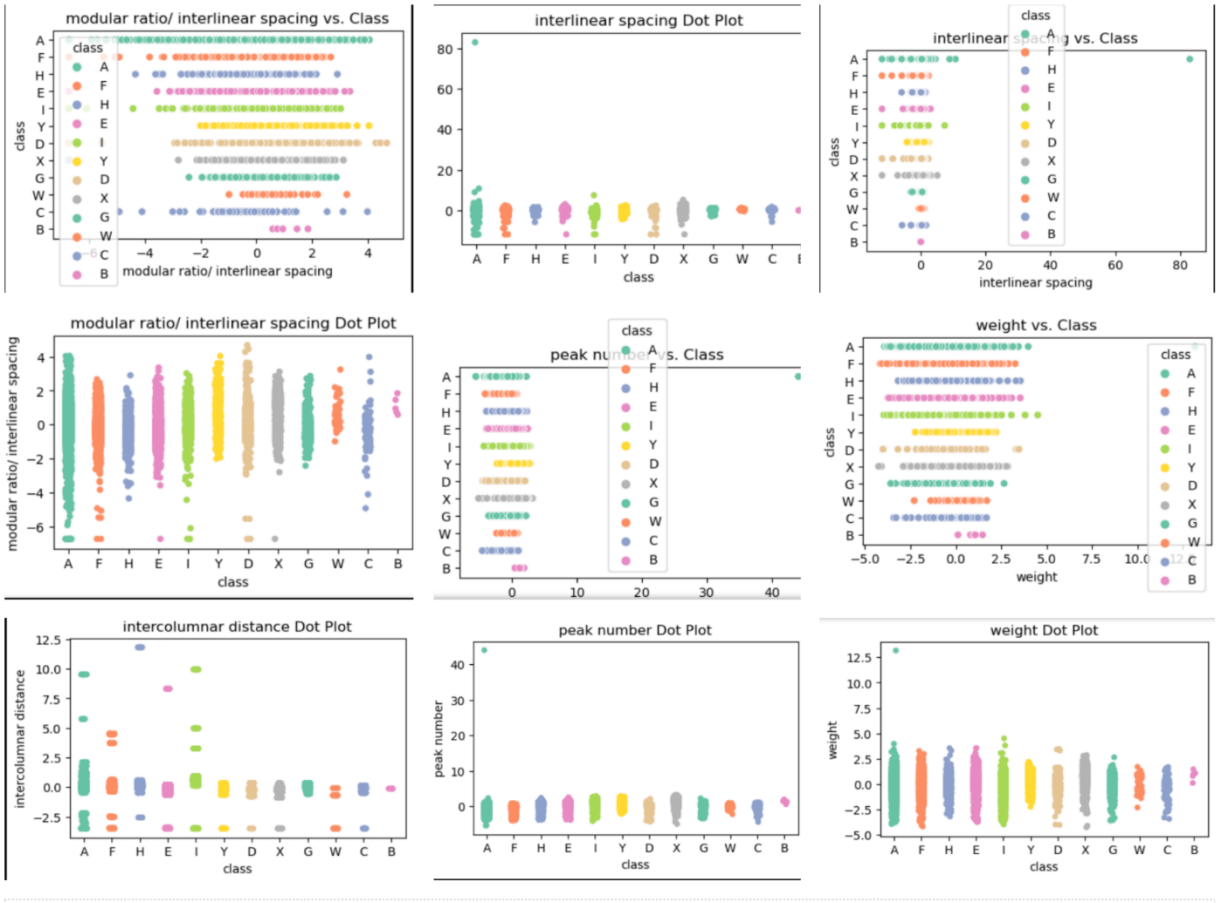
In conclusion, this study investigates how machine learning might be used to categorize historical writings from the Avila Bible, offering a paradigm that could be used to other historical datasets. The research highlights the significance of feature selection, preprocessing, and managing class imbalance in order to get robust model performance and high classification accuracy.

Methodology

This section describes the methodology used in this study, which includes feature selection, model implementation, evaluation for Random Forest and Decision Tree classifiers, and data preprocessing. Making sure the analysis of the Avila dataset is reliable and understandable is the main goal.

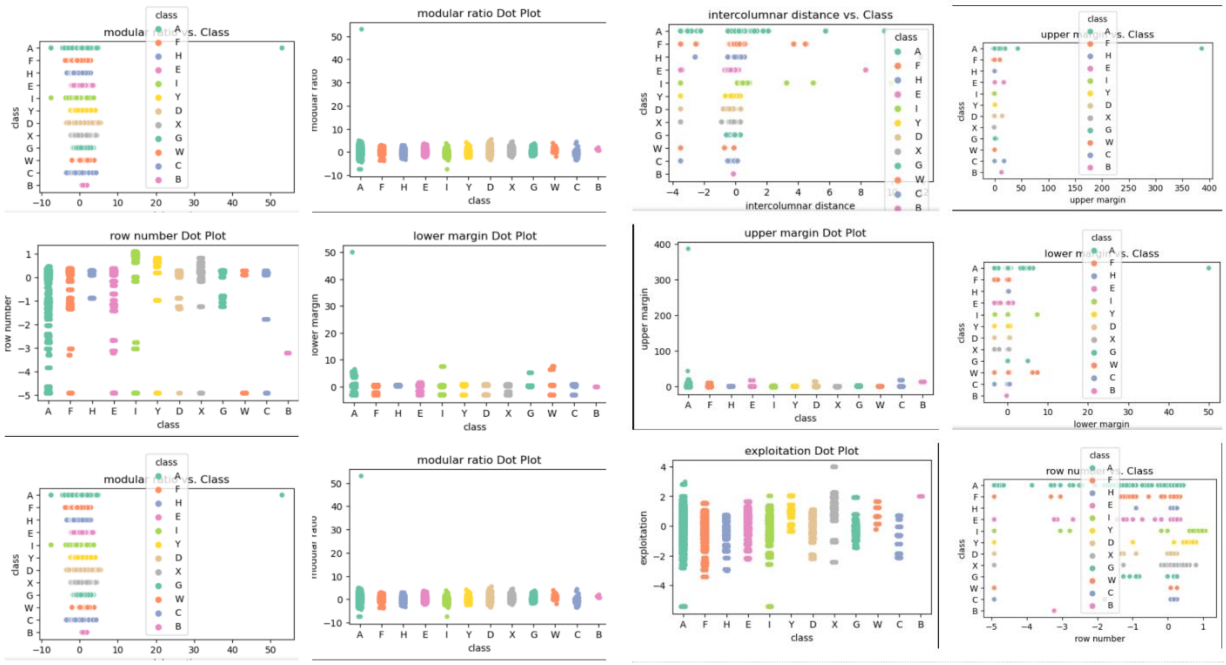
Data preprocessing

1. **Data Loading:** The sources that were supplied were used to load the dataset. It includes intercolumnar distance, upper and lower margins, exploitation, row number, modular ratio, interlinear spacing, weight, peak number, and the ratio of modular ratio to interlinear spacing, among other properties taken from the Avila Bible.
2. **Data Cleaning:** We checked the dataset for errors and missing values. Imputation was used to handle missing values, and proper encoding was used for categorical variables.



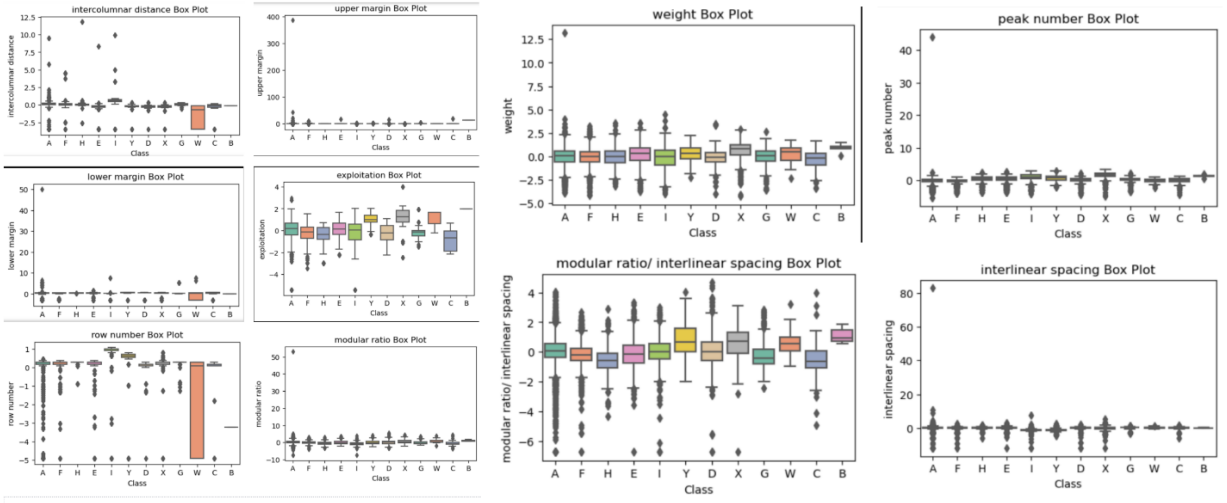
The thorough examination of the dataset provides numerous important insights pertaining to various features, which greatly facilitate the classification of various classes (A, B, C, D, E, F, G, H, I, W, X, Y). Because of their unique variability and distribution patterns across classes, the modular ratio, intercolumnar distance, upper margin, lower margin, row number, weight, exploitation, interlinear spacing, peak number, and the modular ratio/interlinear spacing ratio stand out as critical features.

Features including peak number, weight, top margin, and intercolumnar distance show a continuous high variability in Class A, indicating that these features are highly differentiating for this class. In contrast, Class B exhibits a more compact distribution across the majority of attributes, suggesting less variability and consistent characteristics—a factor that might be helpful for classification. Given their broad distribution ranges, the modular ratio/interlinear spacing ratio sticks out as a noteworthy feature, particularly for Classes A, D, and Y. Furthermore, the exploitation values offer important information, especially for Classes E and G, whose distribution patterns differ from those of other classes.



Overall, the analysis shows that characteristics like upper margin, weight, peak number, and intercolumnar distance—which have considerable variability and stand out as separate outliers—are critical for successfully differentiating between classes.

For classification tasks, the box plots for distinct features across different classes offer extensive insights into the variability and distribution of these features.



- **Weight**: Class A and Class H have significant outliers in the weight feature, which displays a wide variation across all classes. This suggests that weight, particularly for Class A with high variability and outliers, is an important differentiating factor.

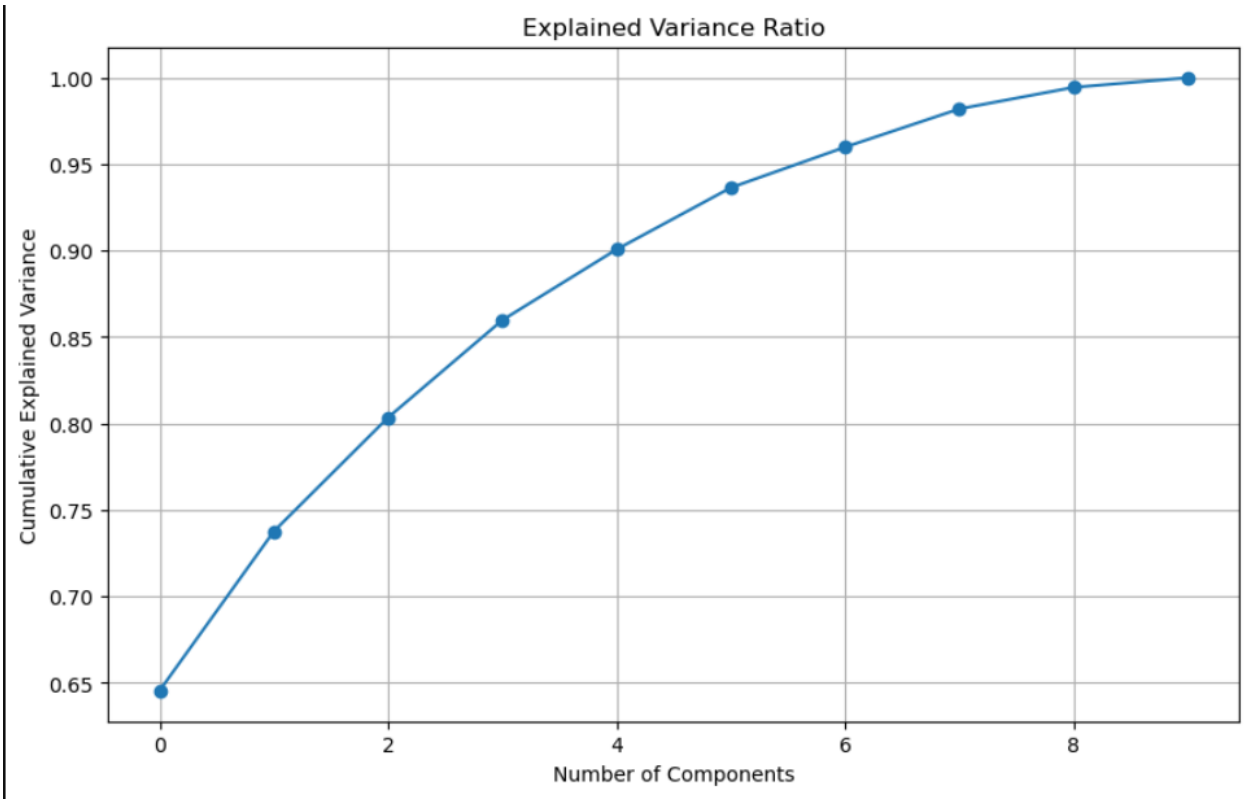
- **Peak Number**: With the exception of a notable aberration in Class A, the peak number is largely constant between classes. This shows that peak number is important for determining Class A even though it may not change much for most classes.
- **Interlinear Spacing**: All classes have modest values for interlinear spacing, with Class A having an exception. A low variability indicates a limited ability to distinguish, with the exception of the Class A outlier.
- **Modular Ratio**: Class A has a significant outlier in the values of this ratio, which are often centered around zero. This suggests that although the modular ratio is stable for the majority of classes, its considerable variability for Class A makes it noteworthy.
- **Row Number**: Values for row numbers vary greatly throughout classes, especially A, F, and H. This shows that, in order to differentiate between these classes, row number is a crucial factor, with Class W exhibiting a peculiar distribution.
- **Exploitation**: For the majority of classes—especially Classes A and F—exploitation values show a broad range. This broad distribution suggests that one important factor in differentiating between groups is exploitation.
- **Lower Margin**: All classes have low lower margin values, with Class A having a notable outlier. With the exception of the outlier, the low variability indicates weak differentiating power.
- **Upper Margin**: Class A and F have extreme outliers in the upper margin values, which are generally low. This suggests that for these classes, upper margin is an important characteristic.
- **Intercolumnar Distance**: This characteristic varies greatly, especially for Class A, F, and G, suggesting that it is crucial for differentiating between these classes. A more compact distribution is displayed by class W.
- **Modular Ratio/Interlinear Spacing**: The ranges for Class A, F, and H are particularly noteworthy. This combined attribute exhibits significant diversity among classes. This suggests that a crucial characteristic for classification is this ratio.

Features with substantial variability and outliers, such as weight, peak number, modular ratio, exploitation, and intercolumnar distance, are crucial for efficient categorization. Robust and reliable classification models are ensured by using these insights to inform the selection of the most influential attributes.

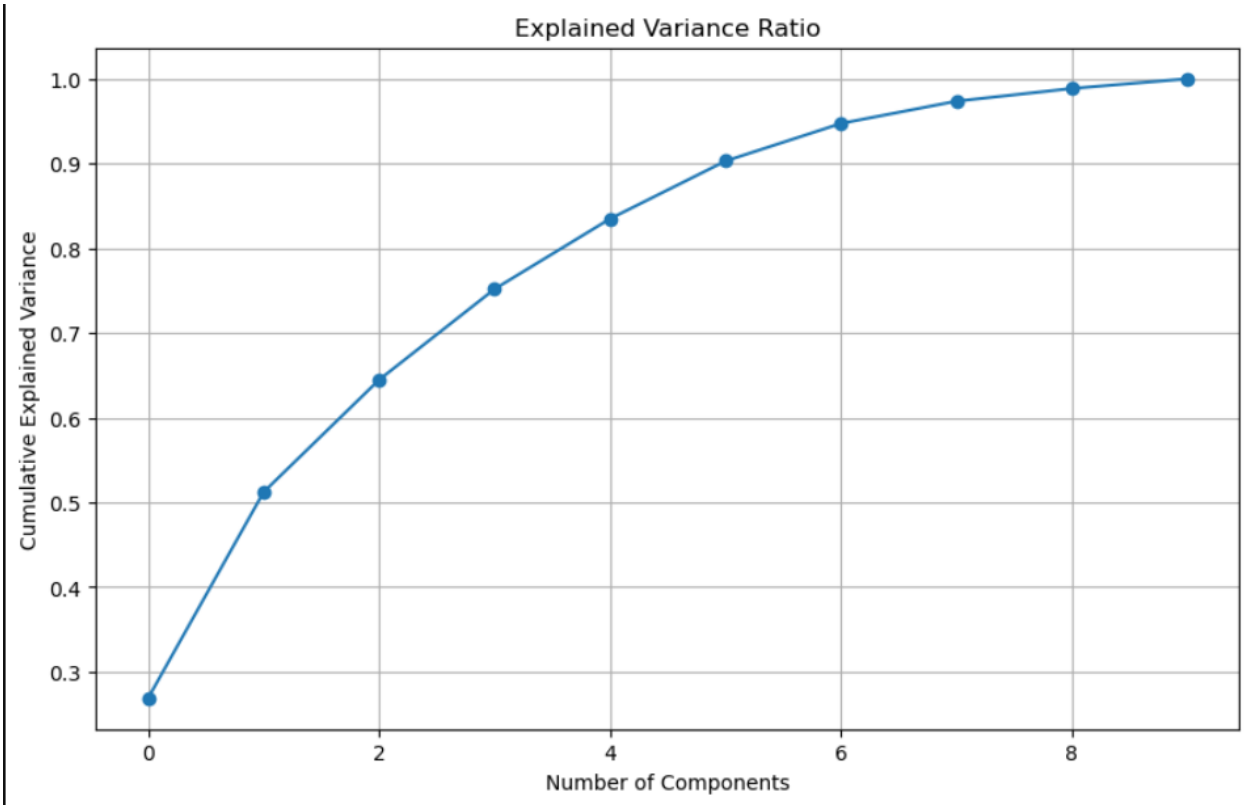
3. **Class label Mapping:** To aid in machine learning model training, the class labels—which represented various copyists—were mapped to particular integers.
4. **Handling class imbalance**: Oversampling techniques were used to alleviate class imbalance.

    - Random Oversampling: In order to equalize the amount of samples in the majority classes, the minority classes were oversampled. In order to balance the dataset and stop the model from becoming biassed towards the majority classes, this necessitated producing synthetic samples for the minority classes.

Feature Selection:



Performing PCA[Principle Component Analysis] on the data before preprocessing.



Performing PCA[Principle Component Analysis] on the data After preprocessing.

According to both graphs, the first few components (up to 4) account for the majority of the variation, which makes them essential for dimensionality reduction while preserving the informative integrity of the dataset. A higher initial variance capture is indicated by the second graph, which may indicate a more successful first dimension reduction. Ultimately, with roughly nine components, both datasets achieve near-total variance capture, offering a clear guideline for choosing the number of components in Principal Component Analysis (PCA) for these datasets.

Relevant features were chosen according to how much of an impact they had on the classification process. When training the model, the following attributes were taken into account:

- Upper Margin
- Lower Margin
- Modular Ratio
- Interlinear Spacing
- Weight

The significance of these variables in differentiating between various copyists in the Avila dataset led to their selection.

Model Implementation

Two machine learning methods were used in the Avila dataset to categorize copyists. The preprocessed dataset was utilized for both training and assessing each model.

1. **Decision tree Classifier:**

- Implementation: The 'DecisionTreeClassifier' class from the 'scikit-learn' library was used to create a Decision Tree Classifier.
- Training: The minority classes were oversampled to create a balanced training set, which was used to train the model.
- Evaluation: Using the test set, the trained model was assessed, and performance measures including accuracy, precision, recall, and F1-score were computed.

2. **Random Forest Classifier:**

- Implementation: The 'scikit-learn' library's 'RandomForestClassifier' class was used to create a Random Forest Classifier.
- Training: To address class imbalance, the model was trained on a balanced training set using class weight adjustments.
- Evaluation: Using the test set, the trained model was assessed, and performance measures including accuracy, precision, recall, and F1-score were computed.

Evaluation

Based on how well the models performed on the test set, they were assessed. The following metrics were employed to evaluate the efficacy of each model:

- **Accuracy**: The percentage of cases out of all instances that are correctly classified.
- **Precision**: is defined as the percentage of actual positive cases relative to all positive cases that were expected.
- **Recall**: The percentage of genuine positive examples relative to the total number of real positive examples.
- **F1-Score**: A balance between recall and precision calculated as the harmonic mean of the two metrics.

Confusion matrices were also produced in order to offer a comprehensive perspective on each class's classification performance. These metrics provide information about the advantages and disadvantages of each model as well as how well it generalizes to new data.

Results

The outcomes of the Random Forest and Decision Tree classifiers on the Avila dataset demonstrate a number of important discoveries and shed information on how well these models categorize the copyists. The performance indicators and the ramifications of these findings are explained in more detail in this section.

Decision Tree Classifier:

With weighted precision, recall, and F1-score all at 88%, the Decision Tree Classifier's overall accuracy was 88%. This suggests that, on the whole, the model performed well in accurately categorizing the cases of various copyists. The thorough categorization report demonstrates that:

- Class A demonstrated an excellent performance of the model in accurately recognizing instances of this class, as seen by its high recall (0.89) and precision (0.90).
- Class B obtained a flawless recall of 1.00 and a precision of 0.38 with only 5 examples. Despite the modest amount of samples, the model's precision was low, despite its high recall indicating that it effectively detected all instances of Class B.
- Moderate F1-scores (0.64 and 0.80, respectively) for Classes C and D indicated a balance of recall and precision.
- The model's resilience was seen in the high precision and recall of Classes E through Y, especially Class I, which had an F1-score of 0.93.

The confusion matrix, which displays the quantity of true positives, false positives, and false negatives for each class, provides more insight into the model's performance. This matrix facilitates comprehension of the precise domains in which the model excelled and those in which it faltered.

Random Forest Classifier:

In addition, the Random Forest Classifier outperformed the Decision Tree Classifier in terms of overall accuracy, coming in at 88%, with a few classifications showing somewhat higher recall and precision. The thorough categorization report makes clear:

- Similar to the Decision Tree, Class A demonstrated consistent performance in detecting instances of this class, with a precision of 0.88 and a recall of 0.90.

- In comparison to the Decision Tree, Class B demonstrated better precision with a precision of 0.50 and a perfect recall of 1.00.
- Improved F1-scores (0.67 and 0.82, respectively) for Classes C and D indicated a better harmony between recall and precision.
- courses I obtained the highest F1-score of 0.94 out of all courses, while Classes E through Y maintained their outstanding recall and precision.

These results are validated by the Random Forest Classifier's confusion matrix, which shows that the model can accurately identify more cases across multiple classes than the Decision Tree.

Comparative analysis:

While both classifiers yielded good results, the Random Forest Classifier performed slightly better across a number of classes in terms of precision and recall. The Random Forest's ensemble nature, which combines the predictions of several decision trees to improve overall accuracy and robustness, is responsible for the small performance advantage.

A key factor in improving the models' performance was the application of random oversampling to solve class imbalance. The models were able to learn more efficiently and prevent biases towards majority classes by maintaining a balance in the dataset. By ensuring that minority classes were fairly represented, this strategy produced more accurate and balanced classifications.

Implications and future work:

The Decision Tree and Random Forest classifiers exhibit good performance and high accuracy, suggesting their appropriateness for historical text classification tasks. These models offer important insights into historical manuscript study and can be applied successfully to categorize copyists in the Avila Bible.

On the basis of these results, future research can investigate other methods to improve model performance.

- **Feature engineering**: By adding more features or investigating other feature selection techniques, classification accuracy may be increased.
- **Advanced Algorithms**: Trying out more sophisticated models, such deep learning or gradient boosting, could result in even better outcomes.
- Cross-Dataset Validation: To make sure the results are not limited to the Avila dataset alone, it would be helpful to apply the models to other historical datasets in order to evaluate their robustness and generalizability.

Conclusion

The promise of machine learning techniques for addressing difficult historical text classification issues has been highlighted by their application in classifying the copyists of the Avila Bible. The usage of Random Forest and Decision Tree classifiers was the primary focus of this study, and both models demonstrated impressive performance metrics, demonstrating their effectiveness in this field.

Key outcomes

- **Model Performance**: The Random Forest classifier outperformed the Decision Tree classifier in terms of precision and recall across most classes, and its ensemble approach yielded better generalization and robustness, making it especially useful for this task. Both classifiers reached an overall accuracy of 88%.
- **Handling Class Imbalance**: Random oversampling was a key strategy for improving the performance of the models; it made sure that the minority classes were sufficiently represented, which allowed the models to learn more balanced decision boundaries and ultimately increase overall classification accuracy.
- **Preprocessing and Feature Selection**: Differentiating between different copyists was made possible by the features that were chosen, which included weight, modular ratio, upper and lower margins, and interlinear spacing. Effective preprocessing of the data, such as feature scaling and class label mapping, was crucial to the performance of the models.

To sum up, this study has effectively illustrated how to use Random Forest and Decision Tree classifiers to categorize copyists in the Avila Bible. These models' great performance and high accuracy, along with their efficient management of class imbalance, offer a solid platform for future research and development of machine learning in the context of historical manuscript interpretation. The knowledge acquired from this study advances the field of digital humanities by providing a framework for further investigation and useful applications in the examination of historical documents.

References:

1. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324
2. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830. http://jmlr.org/papers/v12/pedregosa11a.html
3. Quinlan, J. R. (1986). Induction of Decision Trees. Machine Learning, 1(1), 81-106. https://doi.org/10.1023/A:1022643204877
4. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques (4th ed.). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-804291-5.00001-0