# Jheronimus Academy of Data Science
# Group Project
# JBP041-B-6

‘Predictive Pollen-based Biome Modelling Using Machine Learning’

(Replication Paper)

Paulene Gueco, Camila Matoba, Austėja Vaitkutė
Dr. Dario Di Nucci, Dr. Gemma Catolino
Submitted in part for the PM Data Science & Entrepreneurship
21/22

# Abstract

The purpose of this study was to replicate [1] and examine the efficacy of the supervised classification models, a) parametric: Linear Discriminant Analysis, Logistic Regression, Naives Bayes, b) non-parametric: K-Nearest Neighbours, Decision Trees, Random Forest, and Support Vector Machines, against the conventional biomisation method for the task of biome prediction using high-dimensional pollen datasets. We take this a step further through scaling the data and applying SMOTE analysis to substantiate imbalanced datasets that potentially confine the performance of the classification models. In all instances, the Random Forest classifier boasts superiority, both in terms of ranking the highest in all the performance metrics utilised to evaluate the models and in forecasting four out of five biomes to a high degree of accuracy. The results of this paper have special implications for probabilistic reconstruction of past biomes and vegetation histories from fossil pollen records that sport the potential to enrich our comprehension of both spatial and temporal distributions. In addition, our method could be employed to help regulate disparate reconstructions realised through other methods.

**Keywords:** Pollen datasets, Pollen recognition, Fossil pollen, Vegetation dynamics, Vegetation reconstructions, Climate change, Southern Africa, Late Pleistocene, Holocene, Paleoenvironment, Objective classifications, Machine Learning.

# Nomenclature

| | |
|---|---|
| $m$ | Number of different classes |
| $TP$ | True Positive |
| $FP$ | False Positive |
| $TN$ | True Negative |
| $FN$ | False Negative |
| $MDA$ | Mean Decrease Accuracy |
| *LDA* | Linear Discriminant Analysis |
| *LR* | Linear Regression |
| *NB* | Naïve Bayes |
| *KNN* | K-Nearest Neighbours |
| *CDT* | Classification Decision Trees |
| *RF* | Random Forest |
| *SVM* | Support Vector Machines |

# Table of Contents

# 1. Introduction

## 1.1 Background

Proxy data such as pollen has been widely utilised to presuppose preceding environmental conditions. The analysis of fossil pollen, in particular, has played a prominent role in contributing to our understanding of vegetation shifts [2, 3] and quantitative climate fluctuations [4]. In the last few decades, the numerically appraised relationships between modern pollen assemblages and corresponding variables of interest have allowed for a good degree of accuracy in reconstructing paleoenvironments, and furthermore, distinguished modeling of pollen-vegetation-climate correlations. As such, any valuable approximations of prior environments are heavily dependent on colossal and reliable calibration sets [5].

The Biomisation Technique is the principal method for predicting and reconstructing biomes from pollen data. Numerous paleoecologists across the globe have applied it to fossil pollen sequences to simulate shifts in the apportioning of past biomes. Through implementing a plant functional types (PFTs) approach that surmises a plant's form and function are concomitant [6, 7], it reduces a biome's floral intricacies to a limited number of representative taxa. In essence, it delegates pollen taxa to one or more PFTs, and these PFTs are combined to define biomes following a specific set of protocols, resulting in two matrices. Binary matrix multiplication is then conducted to the matrices to produce the final biome-taxon matrix.

Notwithstanding the efficacy of the Biomisation Technique, a major caveat is that it employs only a subset of pollen taxa for its predictions. Through stipulating only a scarce number of taxa to characterise biomes, intricate interrelations and interplay within contributing factors may be undervalued and/or overlooked. Such debarment of data is crucial, particularly with regard to fossil pollen assemblages, as it very likely implies the loss of certain information that ultimately finalises to inordinately simplistic interpretations. For this reason, there is an imperative need for employing more "complete" datasets to better enhance the interpretability and results of pollen-based paleoenvironmental reconstructions.

The remaining sections of this paper are organised as follows: Section II describes the reference work. Section III reports the overall methodology. Section IV presents the results together with its analysis. Section V is an overview of related works that pose similar interrogatives and methodologies. Section VI provides concluding remarks. Section VII offers recommendations for further work.

## 1.2 Aims and Objectives

Pollen-based biome modelling using machine learning confers for more spatially described reconstructions, as the methodology is applied to both regional and local vegetation classification systems. Coupled with the recent advancements in computing and technology, the use of machine learning for pollen-based biome modelling not only alters our current comprehension of preceding biome distributions, but it also favours a more gradated view towards paleovegetation [11]. The main aim of this paper is to explore different machine learning classification methods for the prediction of biomes using pollen datasets, with the added improvement of data scaling and SMOTE. Four main objectives were decided upon for this study:

1) To acquire the modern African pollen dataset utilised by the authors of [1] and subsequently select and exclude the appropriate data.
2) To construct the various machine learning classification models as per the original paper and replicate its results.
3) To modify the previously skewed modern African pollen dataset through data scaling and SMOTE, and thereafter reconstruct the classification models and acquire its results.
4) To identify, using performance metrics, the highest performing classification model in the instance where data scaling and SMOTE were applied.

## 1.3 Project Planning

A Gantt Chart, which can be found in the Appendix , was created to assess how long the project would take and determine the resources needed, and more importantly, the order in which various tasks and milestones were to be completed. The key milestones in the Gantt Chart are as follows:

[03/02/2021] First Project Revision Session
[03/30/2021] Second Project Revision Session
[04/02/2021] Methodology Summary
[04/25/2021] Confirmation of Results
[05/11/2021] Project Report and Evaluation

Each of these milestones were treated as a gated conclusion of each section of the study. That is, satisfactory conclusions had to be drawn before advancing forward. For this project, it was unanimously agreed that there are no major ethical considerations applicable.

## 1.4 Threats to Validity

Principally, this paper has overall no threats to validity. The methodology used to verify such threats were based on "'Bad smells' in software analytics papers, Information and Software Technology" by Tim Menzies and Martin Shepperd. As stated by Wolpert, D.H., Macready, W.G. (1997), in "No Free Lunch Theorems for Optimization", "for any algorithm, any elevated performance over one class of problems is offset by performance over another class". Thus, it is of interest to study a wide range of different models to achieve the best results that were also properly tuned. Furthermore, there is an extensive research on related work, such as the state-of-the-art models. There is also a special regard in reproducibility: the tuning parameters were stated in the paper and both the dataset and the code were made available on GitHub. The author states that K. Gajewski provided them the modern African pollen dataset, which was made available as a csv file at: https://github.com/Betelgesse/PredictiveBiomeModelling, titled at the platform as OlsenVeg.csv. The data was loaded into the model using:

```
import os
import tarfile
from six.moves import urllib

#Variables
file_path = os.path.join(".")
file_name = "OlsenVeg.csv"
file_url ="https://raw.githubusercontent.com/octokami/PredictiveBiomeModelling/master/OlsenVeg.csv"

#Import
def fetch_file_data(file_url, file_path):
os.makedirs(file_path, exist_ok=True)
csv_path = os.path.join(file_path, file_name)
urllib.request.urlretrieve(file_url, csv_path)
fetch_file_data(file_url, file_path)
```

The only modification done to the original code was the use of DataFrame.values() instead of As_matrix, since it was deprecated since version 0.23.0 pandas.DataFrame.as_matrix — pandas 0.25.1 documentation in the following line:

pollen_matrix = pollen_only.values. Other than that, experiments using scaling and SMOTE were added as it was stated in the original paper that the database was very imbalanced. The only problem that could be present would be the 11th: "Not exploring simplicity". However, due to the intrinsic characteristics of pollen data, reducing the amount of taxa present could affect each of the biomes differently.

## 1.5 Our Hypothesis

On account of the high reproducibility of [1], the group believes that biome prediction from pollen data through the use of supervised classification models will result to the same classifier yielding the best results, i.e., the Random Forest classifier will predict four out of five biomes correctly. However, the "No Free Lunch" theorem propounds that all optimisation algorithms behave effectively when their performances are averaged across a viable set of problems. For predictive modeling problems like regression and classification, there exists no single best optimisation algorithm due to the close nature of optimising and searching. Despite our general enthusiasm in the proposed application of data scaling and SMOTE to accommodate for more potential sources of loss and error in dealing with unbalanced datasets, it will not compensate for biomes where there are fewer total pollen taxa present and where there are high sampling noises at the biome level, specifically the Flooded Grasslands and Savannas (FGS) biome.

# 2. Reference Work

This paper, in conjunction with the authors of [1], assesses the use of machine learning methods for biome prediction given complete sets of pollen taxonomic data. Considering that pollen-based biome modeling is a method that predicates a given biome will yield its complex characteristic patterns, biome modeling in this instance proves to be a laborious endeavour. Not only are pollen datasets high-dimensional in nature [10], but the potential interdependencies and interactions amongst the taxa are also inexpedient to unfold, thereby making them even more demanding to evaluate [10]. The evolution of computational power in terms of speed and processing in the fields of machine learning-- specifically those that focus on classification and prediction approaches-- increase the feasibility of reducing the loss of certain information through using whole multivariate datasets. As such, machine learning methods have important advantages, i.e., it retains a biome's complexity and proffers more impartial distribution.

## 2.1 Machine Learning

Supervised classification is a branch of machine learning and is a type of learning process that involves the use of training data that are considered paradigmatic of each surficial unit to be classified and is used to approximate a mapping function that sorts the data into multiple categories, chief of which are *binary* and *multi-class* classification [12]. Binary differs from multi-class classification in that it predicts one of two discrete values, whereas the latter deals with a particular value from a set of discrete values. In the context of pollen-based biome modelling, the prediction of terrestrial against marine surroundings for a pollen assemblage is a prime exemplar of binary classification, whilst the prediction of multiple biomes from pollen data is of multi-class classification. Supervised classification operates on datasets consisting of a list of *examples*, each of which contain a *set of features* and a target *label*. For the purposes of this paper, the samples represent the examples, the set of features are denoted by the individual pollen taxon abundances, and the labels pertain to the assignments of biomes that typify the target values being predicted.

A standard practice in machine learning is to gauge an algorithm by splitting the dataset into *training* and *testing sets*. The training set is utilised in the training phase so as to determine the set of parameter values that minimise a certain cost function in relation to the entire dataset; the testing set is reserved to appraise how the model will perform in the face of unforeseen data. *Model fitting* refers to the process at which the algorithm learns-- by virtue of the training set-- of the optimal parameters

distinct to a particular model. The accurate prediction of biome labels from the training data are largely dependent on the configuration of the variables; the *hyper-parameters* are separate and distinctive for each model.

Optimal hyper-parameters are determined through the process of cross-validation, whereby the data is divided into '$k$' folds, of approximately equal size, and ensuring that each fold is used as a testing set at some point. The $k$ value is any integer from 1 to 10 but must be carefully chosen to reflect the sample data. A poor choice of $k$ would lead to high *variance* and/or high *bias*. Commonly, the choice of $k$ is usually $k = 5$ or $k = 10$, as these values have been shown empirically to yield test error rate estimates that neither suffer from excessively high variance nor excessively high bias [13]. Variance measures the extent to which a classifier's prediction generated by a learning algorithm deviates from each training sample [17]. It originates from the amalgamation of a model's predictive power and sampling error. Models that are able to gain a more thorough understanding about the relationships between features and labels typically have a high variance and are considered unstable. Bias, on the other hand, is a byproduct of a given model's presumptions about the distribution of data. A high bias equates to a model with overly strong assumptions.

## 2.2 Performance Metrics

As one of the main objectives of this project is to replicate the original paper and compare the results, the chosen performance metrics were the same as the ones on the paper. These metrics are accuracy, f1_macro, f1_micro, f1_weighted, kappa, precision_macro, precision_micro, precision_weighted, recall_macro, recall_micro, recall_weighted.

The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative. The best value is 1 and the worst value is 0. The recall is intuitively the ability of the classifier to find all the positive samples. The best value is 1 and the worst value is 0.

Upon subjecting a model to the test sets to establish its predictive performance, results are stored in a *confusion matrix*. It imparts a quantitative precis on the correct and inapposite classifications done by the model. Fundamentally speaking, in the case of binary classification, a confusion matrix of size *m x m* cognated with a classifier depicts both the forecasted and existing classification [14]. Where m is the number of different classes, for a confusion matrix with $m = 2$, various evaluation methods such as *accuracy*, *recall* (true positive rate), *precision* (positive predicted rate) and *error* (Type 1 and Type 2) may be employed to ascertain how well the model performs on formerly unapprehended data.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \qquad (1)$$

$$Recall = \frac{(TP)}{(TP + FN)} \qquad (2)$$

$$Precision = \frac{(TP)}{(TP + FP)} \qquad (3)$$

$$Error = \frac{(FP + FN)}{(TP + TN + FP + FN)} \qquad (4)$$

Where,

$TP$ = True Positive ,
$FP$ = False Positive ,
$TN$ = True Negative,
$FN$ = False Negative,

True Positives and True Negatives both constitute the correct classifications, whilst False Positives and False Negatives are known as Type 1 and Type 2 errors, respectively. To determine the number of positively classified examples that are apropos, in the particular context of biome prediction, recall and precision are utilised to calculate the $F1$statistic, the harmonic mean of both metrics. Meanwhile, *Cohen's Kappa* aims to remedy the extent of agreement by discounting the bits of the tally that may not be entirely attributed to chance; it is a measure of equitable models' performances considering imbalances in class distributions [15, 16]. To calculate Kappa, the confusion matrix is converted into probabilities and the kappa is found through dividing the difference between the overall proportion of observed agreement and the overall expected value. In addition to what precedes, *features importances* for classification models are crucial. The *Mean Decrease in Accuracy (MDA)* estimates the degree of a model's test accuracy being reduced by randomly traversing the values of a given feature. More to the point, the weightage of each individual pollen taxa to the gross prediction accuracy of a model is verified by the MDA. Where an MDA value is low or a 0, it implies that a particular feature did not contribute significantly or was not involved in the prediction. As per contra, high MDA values suggest that a model's predictive performance was very much contingent on that feature.

## 2.2 Materials

For the training of the models, [1] uses an accumulation of circulated modern pollen data [20] from Africa which they then delegate to biome types by virtue of the world terrestrial ecosystem classification [18,19]. An exclusion of 73 modern pollen samples from a total of 1198 was necessary following a scarcity of coordinates and unbefitting context, i.e., marine. The resulting dataset contains 1125 biome samples, expressed in terms of 119 pollen predictors, that represent nine biomes and showcase variety across the spectrum: surface, lakes, rivers, traps, middens, and ice.

## 2.3 Description of Machine Learning Methods

The models considered by [1] represent parametric, non-parametric, and semi-parametric supervised classification methods. However, the latter method which deals with neural networks was excluded as it goes beyond the purview of JBP041-B-6.

The supervised classification models were selected following the premise of suitability in the context of ecological and paleoecological applications [22], but more to the point, their predictive capabilities in the presence of multivariate and high dimensional pollen data, coupled with their multi-class classification abilities in terms of forecasting more than two biome classes. The models are then further demarcated based on respective rules-- linear and non-linear-- required to transform data and distinguish between classes of biomes. As the name implies, a linear classifier is characterised by linear decision-making thresholds, i.e., straight lines or planes, which are applied to discretise groups of data. On the contrary, a non-linear classifier assumes any configuration, i.e., yes or no questions, non-linear shapes. By and large, these rules stipulate how to allocate a given modern pollen assemblage to a biome type. Nonetheless, oftentimes the nature of original data makes it difficult for any such model to correctly separate into discrete categories and therefore require data transformation to advance towards clearer distinctions amidst various classes.

### 2.3.1 Parametric Classification

Parametric classification is a type of learning process whereby huge amounts of data is not a prerequisite for the model to learn the mapping function, and the learnt mapping function has a recognised form with a fixed set of parameters [23]. For linear types of classification methods, parametric models are considered high bias [24] when the model assumptions do not correlate

with the actual data distribution. Hence, their predictive abilities to discern complex data patterns and configurations are purposely confined to avoid instances of high bias. Linear parametric classification models, such as Linear Discriminant Analysis and Logistic Regression, are traditionally compatible with basic types of problems.

- **Linear Discriminant Analysis (LDA):**
  Linear discriminant analysis detects linear amalgams of pollen features that best describe the separation of classes of data into groups and locates the epicentre of all pollen data, given a multi-class classification problem. The Mahalanobis metric [26] measures the distance between the central point and the points central to each biome. Straight lines are used to divide groups of biomes, taking into account the minimised scatter for each biome type and also the maximised distance between each biome type and the centre point [23]. LDA compartmentalises a new unlabelled example to a particular biome type through calculating distances from the biome categories.

- **Linear Regression (LR):**
  Linear Regression determines a biome class through computing a constant bias term and also the weighted sum of pollen abundances, which in this case are our inputs [27]. The weighted sum is transformed into a probability, with values ranging between 0 and 1, and the outcome of the logistic function establishes a linear decision boundary that guides the separation and apportioning of observations to biome classes, according to which side of the line they come to pass. Moreover, the concept of multinomial linear regression (mLR) enables LR to be universalised, i.e., given a multi-class classification problem, the probability of a sample pertaining to a certain biome class may be estimated by a single mLR model that is trained for all the biome classes.

- **Naïve Bayes (NB):**
  A Naïve Bayes model lessens the complexity of highly dimensional datasets through supposing conditional independence amongst predictor variables [29]. Throughout the course of the training phase, the fraction of biome classes present in the training dataset is computed alongside the probability of each pollen taxon that is contingent on the biome class. In the testing phase, the pollen features for which the values are continuous are converted into likelihood tables. The Bayes equation [28] then measures the probability for each biome class, where the class garnering the highest probability dictates the likelihood of the final prediction for the unlabelled instance.

### 2.3.2 Non-Parametric Classification

In contradistinction to parametric classification, non-parametric learning methods hold fewer assumptions in relation to the underlying functions and are also more computationally intensive. Where parametric models do not require massive amounts of data to learn the mapping function, a superior classification from a non-parametric model directly correlates with the amount of data available, i.e. the more data, the better the predictive capacity. In addition, the high degree of stochasticity intrinsic to highly complex non-parametric models [23] result in the interpretability of the outputs being all the more arduous.

- **Support Vector Machines (SVM):**
  Support Vector Machines are classifiers that map a vector of predictors into a new, higher dimensional feature plane through either linear and non-linear kernel functions [33]. These kernel functions determine the similarity between points in the original plane that corresponds to the points in the new plane and a maximum margin separator (or the widest street approach) then separates the ensuing groups [34]. In the kernel space, a hyperplane is constructed such that the margin of the decision boundary between the two nearest neighbouring points of each biome type is maximised. Similar to the LR in parametric classification, an unlabelled example is classified according to the side of the decision boundary they come to land, but in the kernel space.

- **K-Nearest Neighbours (KNN):**
  A K-Nearest Neighbour model accumulates all the samples from the training set at the time of training. By virtue of a similarity function, where an unlabelled example is encountered at the prediction stage, a KNN classifier explores for a number of 'K' nearest instances most analogous to the example. Once the labels for the nearest instances are captured, a majority vote rule [30] then determines the class most prevalent that is encompassed by the nearest neighbours. The biome label for the formerly unlabelled example is assigned to that class.

- **Classification Decision Trees (CDT):**
  Classification decision trees are based on the idea of an ordinary tree structure that is composed of roots and leaves, nodes and branches. In this context, a node symbolises a certain attribute, whilst branches symbolise a range of values that act as partition points for the given attribute [31]. A CDT is constructed starting from the root and its branches are represented by segments that connect the nodes, which are generally drawn from left to right. In the simplest sense, it classifies data through posing yes-or-no questions on features triggering a hierarchical decision process that yields an outcome.

  In the training phase, pollen data is partitioned into the best pollen taxon, where the daughter nodes are expected to preserve maximum heterogeneity between and within themselves. This iterative process continues indefinitely until the remainder subsets of pollen data are classified and the tree leaves embody a majority of a biome type. One caveat is that because of this procedure, we obtain an unnecessarily complex tree and a function that is too closely fitted to the training data, thereby hindering the effective capabilities of the model during testing. Thus, it can be stated that CDT's have a low bias and high variance. However, this is compensated for through the process of cross-validation.

- **Random Forest (RF):**
  A Random Forest is one of the best-performing non-parametric classifiers. It is an assemblage of individual decision trees germinated in a manner akin to CDT, but with the addition of two randomisation steps in the learning process that addresses the issue of having a low bias and high variance vis-a-vis CDTs [32]. Here, pollen data is randomly subsampled, and the nodes are also partitioned via a random subsample of pollen features from the original dataset, resulting in the RF being a more robust model than its stand-alone counterpart. When the RF encounters a new unlabeled example, it is passed through all the individual trees in the forest. Each tree then distributes a prediction of biome types and the predictions are averaged across all trees. As per the majority vote rule, the biome type with the highest probability leads to the final prediction for the unlabeled example.

## 2.4 Model Training

After an adequate description of the various supervised learning classification models as above, [1] analysed these aforementioned models in Python 2.7.12, through the aid of packages like scikit-learn, numpy and pandas [35, 36, 37]. The preprocessing of data involved, 1) scaling abundant pollen in the range of zero and one, 2) eradicating biomes that were underrepresented, e.g., below 10 sites, 3) eradicating rare pollen taxa in such a manner that only the taxa above 3% in at least one site was retained. Thereafter, the dataset was split into a training and testing set, following a 9:1 ratio.

The models were optimised using 50 iterations of random search so as to obtain their corresponding hyperparameters [38] and also using a k-fold cross-validation, where k = 10. The classification models were subsequently fitted to the training set using the derived hyperparameters, giving a total of 500 fits for 10 folds for 50 iterations. For ease of reference, the list of hyperparameters as determined by the authors of [1] which we later apply in our methodology and, in some cases, modify, has been attached below. See **Fig.1**.

To examine the predictive efficacy of the models, the reserved testing set was utilised and the results were evaluated using performance metrics found in Section 2.2 (accuracy, precision and recall, kappa statistic, F1). The classifier that scored the

highest on these metrics was deemed superior and was also evaluated further through calculating the accuracy metrics on individual biome predictions for the testing set. Furthermore, for each of the models, variable importances were also calculated using the MDA to portray the impact of individual pollen taxa on its predictive capabilities. Lastly, precision, recall and F1 are recalculated from the confusion matrix to compare the acquired results from the models to the traditional biomisation method.

# 3. Methodology

All Python codes for this section can be found at https://github.com/octokami/Introduction-to-Machine-Learning hosted on GitHub.

| | Our best hyperparameters | Paper's best hyperparameters | Argument Description |
|---|---|---|---|
| **LR** | C': 926.300878513349 | **973.7555188** | Inverse of regularization strength |
| | class_weight': 'balanced' | **None** | Weights associated with classes |
| | fit_intercept': True | **FALSE** | Specifies if a constant should be added to the decision function |
| | max_iter': 10000 | | Added to enforce conversion |
| | multi_class': 'multinomial' | multi_class': 'multinomial' | Class type; either 'one-versus-rest' or 'multinomial' |
| | solver': 'lbfgs' | solver': 'lbfgs' | Algorithm to use in the optimization problem |
| **RF** | class_weight': 'balanced_subsample' | class_weight': 'balanced_subsample' | Weights associated with classes |
| | criterion': 'entropy' | criterion': 'entropy' | Function measuring the quality of a split |
| | max_features': 'sqrt' | max_features': 'sqrt' | Number of features to consider when looking for the best split |
| | min_samples_split': 0.007066305219717406 | **0.007066305** | Minimum number of samples required to split an internal node |
| | n_estimators': 98 | n_estimators': 98 | Number of trees in the forest |
| **NN** | activation': 'relu' | activation': 'relu' | Activation function for the hidden layer |
| | alpha': 0.017436642900499146 | 0.0174366429 | Regularization term |
| | batch_size': 32 | batch_size': 32 | Size of minibatches for stochastic optimizers |
| | hidden_layer_sizes': 200 | hidden_layer_sizes': 200 | The n-th element representing the number of neurons in the n-th hidden layer |
| | learning_rate': 'adaptive' | learning_rate': 'adaptive' | Learning rate schedule for weight updates |
| | learning_rate_init': 0.0001 | learning_rate_init': 0.0001 | The initial learning rate used |

| | | | |
|---|---|---|---|
| | max_iter': 123 | max_iter': 123 | Maximum number of iterations |
| | solver': 'adam' | solver': 'adam' | Solver for weight optimization |
| **LDA** | n_components': 3 | n_components': 3 | Number of components for dimensionality reduction |
| | solver': 'svd' | solver': 'svd' | Solver to use |
| **NB** | alpha': 0.0007787658410143283 | 0.9737555188 | Smoothing parameter |
| | fit_prior': False | **TRUE** | Whether to learn class prior probabilities or not |
| | class_prior | None | Prior probabilities of the classes |
| **KNN** | algorithm': 'kd_tree' | brute | Algorithm used to compute the nearest neighbors |
| | n_neighbors': 1 | **6** | Number of neighbors to use |
| | p': 1 | p': 1 | Power parameter for the Minkowski metric |
| | weights': 'uniform' | **distance** | Weight function used in prediction |
| **CDT** | class_weight': None | class_weight': None | Weights associated with classes |
| | criterion': 'entropy' | criterion': 'entropy' | Function measuring the quality of a split |
| | max_features': 'sqrt' | max_features': 'sqrt' | Number of features to consider when looking for the best split |
| | min_samples_split': 0.0007787658410143283 | **0.031313293** | Minimum number of samples required to split internal node |
| | splitter': 'random' | splitter': 'random' | Strategy used to choose the split at each node |
| **SVM** | C': 83.24526408004218 | 21.23491107 | Penalty parameter C of the error term |
| | degree': 2 | **1** | Degree of the polynomial kernel function |
| | gamma': 0.7797658410143283 | **617.4825096** | Kernel coefficient |
| | kernel': 'rbf' | **poly** | Kernel type to be used in the algorithm |
| | max_iter': 180 | | Added to enforce conversion |

**Table 1.** List of hyperparameters

## 3.1 Data Selection

As the database is relatively imbalanced, the biomes and taxa with less samples were excluded from the model. The threshold for taxa was the sum from all instances being over 3. For the Biomes, it was required to have at least 10 instances. According to said threshold, from the 119 taxas, 5 of them were removed: CHRISTIANA, CAMPYLOSTE, HUGONIA, DIDIEREACEAE, and CHROZOPHOR. Likewise, 4 biomes were excluded: Temperate Grasslands, Savannas, and Shrublands (8 instances), Mediterranean Forests, Woodlands, and Scrub (4 instances), Mangroves (3 instances), Tropical and Subtropical Dry Broadleaf Forests (1 instance).

The chosen threshold seems reasonable as most data seems to be between 7 and 355:

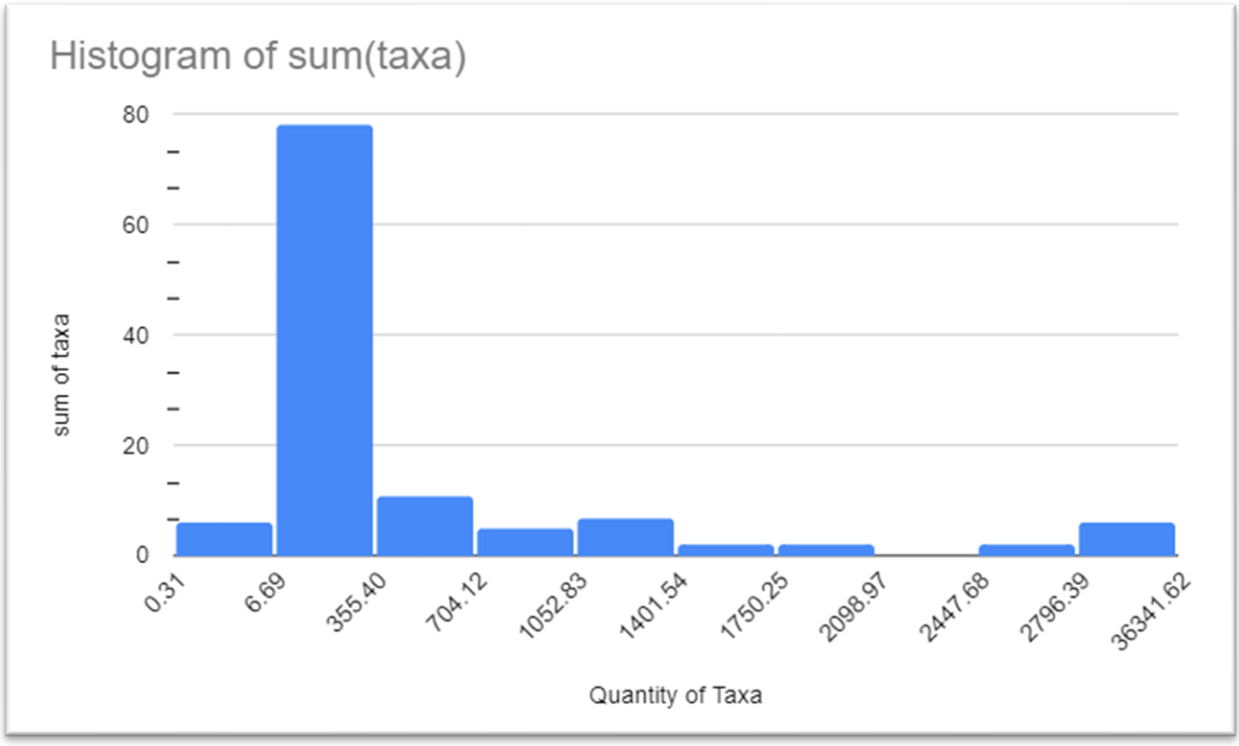



**Fig.1.** Histogram of sum(taxa)

The code to exclude taxa was:

```
def exclude_rare_taxa(x, threshold=3):
    ##Summing the whole column
    to_keep = (x > threshold).sum(axis=0) != 0
    return x[:, to_keep]
```

The code to exclude biomes was:

```
# Filter rows for biomes that occur in less than 10 sites
df = df.groupby("BIO_N").filter(lambda x: len(x) >= 10)
```

# 3.2 Data Scaling and SMOTE Implementation

SMOTE (Synthetic Minority Over-sampling Technique) and scaling were added as it was stated in the original paper that the database was highly imbalanced. An imbalanced dataset can greatly affect the performance of classification models for the underrepresented classes. As stated by Blagus, R., Lusa, L. (2013), in "SMOTE for high-dimensional class-imbalanced data.", "The bias is even larger for high-dimensional data, where the number of variables greatly exceeds the number of samples". For that reason, the group decided to apply this technique to better compare it with the others. The difference in the quantity of the samples was checked by using:

```
categories='BIO_N'
df[categories].value_counts()
```

Which outputs of the count of instances for each biome:

| | |
|---|---|
| Tropical and subtropical grasslands, savannas, and shrublands | 415 |
| Tropical and Subtropical Moist Broadleaf Forests | 314 |
| Deserts and Xeric Shrublands | 239 |
| Montane Grasslands and Shrublands | 120 |
| Flooded Grasslands and Savannas | 21 |
| Temperate Grasslands, Savannas, and Shrublands | 8 |
| Mediterranean Forests, Woodlands, and Scrub | 4 |
| Mangroves | 3 |
| Tropical and Subtropical Dry Broadleaf Forests | 1 |

Name: BIO_N, dtype: int64

After SMOTE, each biome had 373 samples, as can be checked with:

```
#Checking number of samples of each biome
unique_elements, counts_elements = np.unique(y_ovs, return_counts=True)
print("Frequency of unique values each Biome:")
print(np.asarray((unique_elements, counts_elements)))
```

Which outputs:

```
[[  0   1   2   3   4]
 [373 373 373 373 373]]
```

SMOTE is a technique that consists of synthetically creating the samples of the classes in the dataset which are lacking when compared to the majority class [21]. In this project, all underrepresented classes had samples synthetically created to equalize their number to the majority class training sample. As scaling is necessary when using smote and it should be done before applying SMOTE, a MinMaxScaler was used and fitted:

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler(copy=False)
x = scaler.fit_transform(x)
```

After scaling, the dataset was split and finally the SMOTE could be applied:

```
##SMOTE
#!pip install -U imbalanced-learn only needed to be used once
import seaborn as sns
from imblearn.over_sampling import SMOTE
# creating a dataset with SMOTE application
smt = SMOTE(random_state=seed)
x_ovs, y_ovs = smt.fit_resample(x, y)
```

## 3.3 Model Training

### 3.3.1 Splitting, Fitting, and Predicting

The data was split into 10% test and 90%, with the stratify=y method. That keeps the proportions within the separate parts close to the original proportions. The same procedure goes into the Cross validation, where the parameters were set as:

```
folds = StratifiedKFold(n_splits=n_folds, shuffle=True, random_state=seed)
```

According to the original author of the paper, StratifiedKFold for 10-folds data splitting was used due to the large class imbalances. The code is as follows:

```
x, x_test, y, y_test = train_test_split(x, y, test_size=0.1, random_state=seed, stratify=y)
```

The hyperparameters were optimized using a Randomized Search:

```
random_search = RandomizedSearchCV(
clf,
cv=folds, ##Cross_fold
verbose=1, #Controls the verbosity: the higher, the more messages.
n_jobs=4, #Parallel jobs
param_distributions=param_dist,
n_iter=n_iter_search,
random_state=seed)
```

The model was fit using the usual fit() from the RandomizedSearchCV.

# 4. Results and Analysis

The contribution of the 30 most vital pollen taxa to the overall prediction accuracy of each model is shown in **Fig.2,** where AMAR (Amaranthaceae) and EUPH (Euphorbiaceae) appear to be the most recurrent taxa chosen by the models to a fluctuating degree of significance.

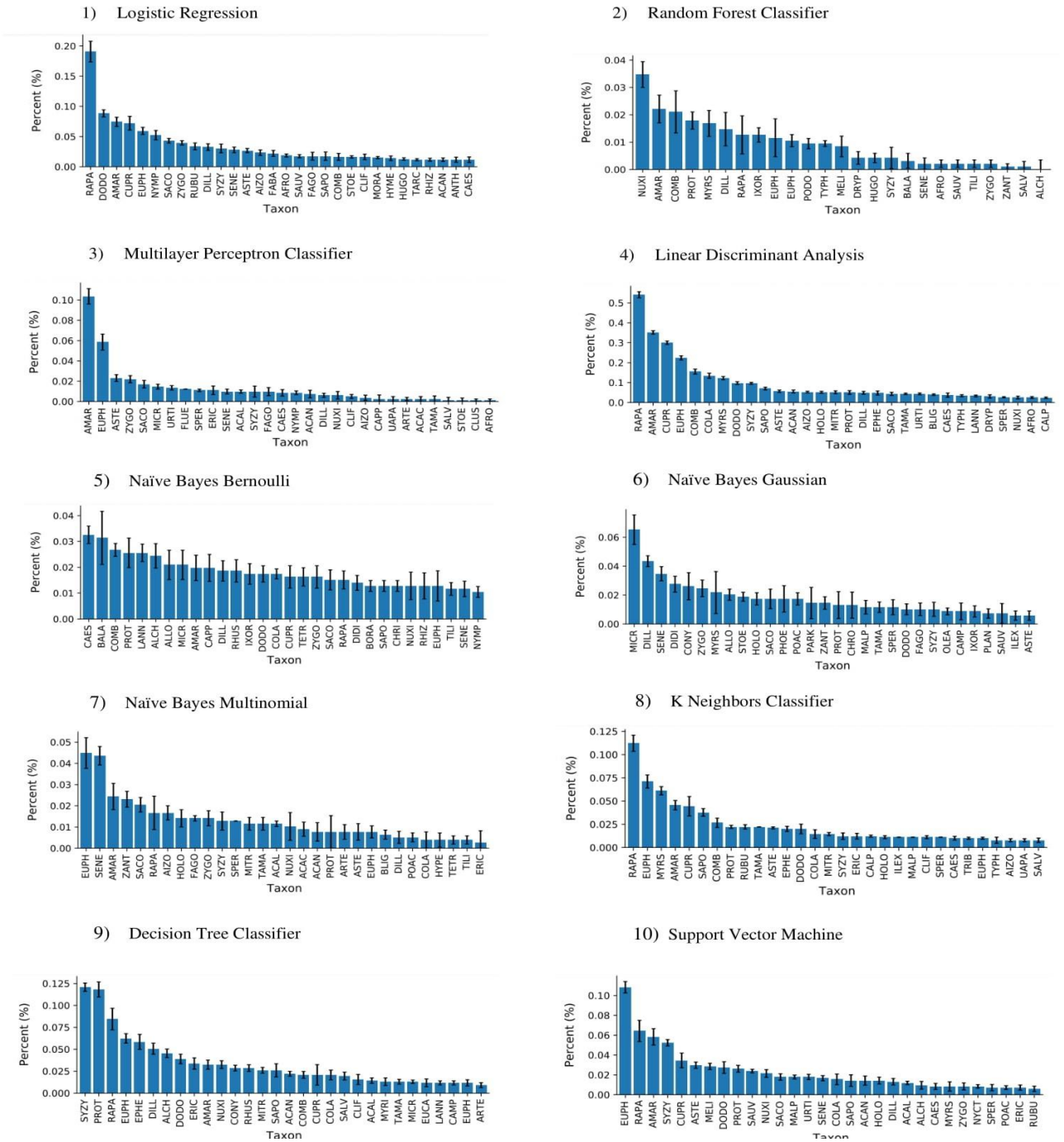



**Fig.2.** Mean Decrease in Accuracy for Machine Learning Classifiers

| SMOTE | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | accuracy | f1_macro | f1_micro | f1_weighted | kappa | precision_macro | precision_micro | precision_weighted | recall_macro | recall_micro | recall_weighted |
| RandomForestClassifier | 0.86 | 0.79 | 0.86 | 0.86 | 0.80 | 0.79 | 0.86 | 0.86 | 0.80 | 0.86 | 0.86 |
| KNeighborsClassifier | 0.81 | 0.74 | 0.81 | 0.81 | 0.74 | 0.71 | 0.81 | 0.82 | 0.77 | 0.81 | 0.81 |
| LogisticRegression | 0.77 | 0.69 | 0.77 | 0.79 | 0.70 | 0.68 | 0.77 | 0.81 | 0.75 | 0.77 | 0.77 |
| MLPClassifier | 0.74 | 0.65 | 0.74 | 0.77 | 0.66 | 0.67 | 0.74 | 0.84 | 0.74 | 0.74 | 0.74 |
| BernoulliNB | 0.77 | 0.73 | 0.77 | 0.77 | 0.70 | 0.72 | 0.77 | 0.81 | 0.77 | 0.77 | 0.77 |
| SVC | 0.76 | 0.70 | 0.76 | 0.76 | 0.67 | 0.68 | 0.76 | 0.76 | 0.73 | 0.76 | 0.76 |
| MultinomialNB | 0.70 | 0.61 | 0.70 | 0.74 | 0.61 | 0.63 | 0.70 | 0.81 | 0.70 | 0.70 | 0.70 |
| LinearDiscriminantAnalysis | 0.68 | 0.60 | 0.68 | 0.72 | 0.59 | 0.62 | 0.68 | 0.78 | 0.67 | 0.68 | 0.68 |
| DecisionTreeClassifier | 0.69 | 0.62 | 0.69 | 0.70 | 0.59 | 0.60 | 0.69 | 0.72 | 0.67 | 0.69 | 0.69 |
| GaussianNB | 0.62 | 0.57 | 0.62 | 0.62 | 0.50 | 0.59 | 0.62 | 0.69 | 0.64 | 0.62 | 0.62 |

| SMOTE - Calculated | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | accuracy | f1_macro | f1_micro | f1_weighted | kappa | precision_macro | precision_micro | precision_weighted | recall_macro | recall_micro | recall_weighted |
| SVC | 0.01 | 0.15 | 0.01 | 0.03 | 0.03 | 0.06 | 0.01 | 0.02 | 0.20 | 0.01 | 0.01 |
| KNeighborsClassifier | 0.01 | 0.08 | 0.01 | 0.02 | 0.02 | 0.05 | 0.01 | 0.03 | 0.11 | 0.01 | 0.01 |
| MLPClassifier | -0.03 | 0.04 | -0.03 | 0.01 | -0.01 | 0.05 | -0.03 | 0.09 | 0.12 | -0.03 | -0.03 |
| GaussianNB | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 | 0.00 | 0.01 | 0.01 |
| RandomForestClassifier | 0.00 | 0.11 | 0.00 | 0.01 | 0.00 | 0.11 | 0.00 | 0.01 | 0.11 | 0.00 | 0.00 |
| BernoulliNB | 0.01 | 0.04 | 0.01 | 0.00 | 0.01 | 0.05 | 0.01 | 0.01 | 0.02 | 0.01 | 0.01 |
| MultinomialNB | -0.04 | 0.02 | -0.04 | 0.00 | -0.02 | 0.04 | -0.04 | 0.07 | 0.09 | -0.04 | -0.04 |
| LogisticRegression | -0.05 | 0.01 | -0.05 | -0.04 | -0.06 | -0.01 | -0.05 | -0.01 | 0.08 | -0.05 | -0.05 |
| DecisionTreeClassifier | -0.06 | 0.01 | -0.06 | -0.05 | -0.07 | -0.03 | -0.06 | -0.04 | 0.07 | -0.06 | -0.06 |
| LinearDiscriminantAnalysis | -0.09 | -0.05 | -0.09 | -0.07 | -0.10 | -0.05 | -0.09 | -0.02 | 0.03 | -0.09 | -0.09 |

0.00

| SMOTE - Paper | accuracy | precision_weighted | f1_weighted | kappa |
|---|---|---|---|---|
| MLPClassifier | -0.03 | 0.17 | 0.00 | -0.11 |
| KNeighborsClassifier | 0.02 | 0.11 | 0.02 | -0.05 |
| BernoulliNB | -0.01 | 0.10 | -0.01 | -0.08 |
| LinearDiscriminantAnalysis | -0.09 | 0.09 | -0.05 | -0.18 |
| DecisionTreeClassifier | -0.07 | 0.06 | -0.06 | -0.17 |
| RandomForestClassifier | 0.00 | 0.06 | 0.00 | -0.06 |
| LogisticRegression | -0.05 | 0.07 | -0.03 | -0.12 |
| SVC | -0.01 | 0.09 | -0.01 | -0.10 |

**Fig.3** Results with SMOTE

As can be seen from **Fig.3**, the best performing model is the RF, scoring highest on all evaluation metrics and achieving an overall accuracy, precision and F1 score of 0.86. The KNN classifier follows closely behind, with an overall accuracy of 0.81, precision of 0.82, and F1 of 0.81, whilst the remaining models perform similarly to one another.

To verify the results achieved by [1], the hyperparameters presented by the paper were set in the models without using the Hyperparameter Search. The results achieved were very close to the expected, with an average over all models and metrics of +1.50%, as can be seen in **Fig.4**.

| Coded with the paramenters | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | accuracy | f1_macro | f1_micro | f1_weighted | kappa | precision_macro | precision_micro | precision_weighted | recall_macro | recall_micro | recall_weighted |
| RandomForestClassifier | 0.86 | 0.80 | 0.68 | 0.86 | 0.85 | 0.67 | 0.86 | 0.85 | 0.70 | 0.86 | 0.86 |
| MLPClassifier | 0.81 | 0.73 | 0.65 | 0.81 | 0.80 | 0.65 | 0.81 | 0.80 | 0.66 | 0.81 | 0.81 |
| LogisticRegression | 0.80 | 0.72 | 0.67 | 0.80 | 0.81 | 0.69 | 0.80 | 0.82 | 0.65 | 0.80 | 0.80 |
| BernoulliNB | 0.78 | 0.71 | 0.73 | 0.78 | 0.78 | 0.72 | 0.78 | 0.81 | 0.76 | 0.78 | 0.78 |
| LinearDiscriminantAnalysis | 0.77 | 0.69 | 0.65 | 0.77 | 0.79 | 0.67 | 0.77 | 0.80 | 0.64 | 0.77 | 0.77 |
| KNeighborsClassifier | 0.77 | 0.68 | 0.64 | 0.77 | 0.77 | 0.65 | 0.77 | 0.77 | 0.63 | 0.77 | 0.77 |
| DecisionTreeClassifier | 0.76 | 0.66 | 0.61 | 0.76 | 0.75 | 0.64 | 0.76 | 0.77 | 0.60 | 0.76 | 0.76 |
| MultinomialNB | 0.74 | 0.64 | 0.58 | 0.74 | 0.73 | 0.58 | 0.74 | 0.73 | 0.60 | 0.74 | 0.74 |
| SVC | 0.71 | 0.58 | 0.50 | 0.71 | 0.68 | 0.56 | 0.71 | 0.70 | 0.50 | 0.71 | 0.71 |
| GaussianNB | 0.61 | 0.49 | 0.56 | 0.61 | 0.61 | 0.58 | 0.61 | 0.69 | 0.63 | 0.61 | 0.61 |

| Coded - Paper | accuracy | precision_weighted | f1_weighted | kappa |
|---|---|---|---|---|
| LinearDiscriminantAnalysis | 0.00 | 0.11 | 0.00 | 0.02 |
| BernoulliNB | 0.00 | 0.10 | 0.00 | 0.00 |
| LogisticRegression | -0.02 | 0.08 | -0.02 | -0.01 |
| DecisionTreeClassifier | 0.00 | 0.11 | 0.00 | -0.01 |
| SVC | -0.06 | 0.03 | -0.06 | -0.09 |
| MLPClassifier | 0.04 | 0.13 | 0.04 | 0.03 |
| KNeighborsClassifier | -0.02 | 0.06 | -0.02 | -0.02 |
| RandomForestClassifier | 0.00 | 0.05 | 0.00 | -0.01 |

1.50%

**Fig.4** User-specified Parameters

In theory, several factors contribute to the reduction of accuracy and precision when it comes to the prediction of biomes. Sources of loss and error include but are not limited to: 1) pollen preservation correlating to the dispersal syndrome, where the largest source of deposited grains are procured from wind-pollinated vegetation [39]; 2) pollen preserves are accumulated from a wide array of sources, from water-logged and anaerobic environments (ideal) [40] to snow amd pack rat middens and hyena scat (less ideal) [41, 42]; 3) oxidation [43], microbial activity [44], wet-dry cycles and changes in pH [45] all contribute to the degradation of pollen grains differing in physical properties, so not all pollen are created equally; 4) the laboratory preparation of pollen samples also effectuate some loss of pollen grains [39]; and 5) the microscopic recognition of pollen grains as determined by any human is contingent upon their level of expertise and state of mind [46]. Notwithstanding, through supervised learning classification methods the group was able to successfully predict four out of five biomes in tandem with the authors of [1], and also achieve slightly better results following our implementation of SMOTE to the classification models.

As expected, biomes that were well-represented in the modern pollen dataset (Tropical and Subtropical Broadleaf Forests, Tropical and Subtropical Grasslands, Savannas and Shrublands, Xeric Shrublands, Montane Grasslands and Shrublands) were all predicted with a high level of accuracy and precision, as was the case with [1]. Meanwhile, our models, with the exception of the KNN classifier, also struggled to predict the Flooded Grasslands and Savannas biome, which again was expected due to the sampling noise both at the pollen and biome level, and also that there were fewer pollen taxa present in the Flooded Grasslands and Savannas as compared to the other four biomes at hand. Furthermore, as the authors of [1] have stated, "*Although pollen taxa specific to the FGS biome are present, including aquatics such as Typha and Nymphaea, their signal may be diluted by the cosmopolitan species."* It is also worth reiterating that datasets were split into a 9:1 training and testing ratio, which for the Flooded Grasslands and Savannas posed extra complications due to the inherent small size of both its training and testing samples, making it all the more difficult for the learning algorithms to map the relationships between the biome and pollen data. Going forward, in building a more superior and robust model, it is of paramount importance to have well-represented biomes in the pollen dataset. Where modern datasets are not readily accessible for some regions of the world, a strict machine learning approach is limited as model performances are reliant on these datasets for training. This is particularly crucial in palaeosciences for which reconstructions of past conditions are only as accurate as the empirical data available.

Nevertheless, in utilising esteemed parametric and non-parametric models, each emblematic of a specific set of assumptions, a machine learning approach to predict biomes from pollen datasets is well-substantiated, first and foremost because of the low signal-to-noise ratio intrinsic to the said datasets and the extensive computational resources required to analyse them. Although it is entirely plausible that the RF model in this instance does not fully epitomise valuable information present in the pollen data, it still incorporates more criteria for biome predictions from pollen sequences. Its non-linear property also establishes a strong validity as it maintains the best predictive performance amongst all the other models. That said, the LR model having attained a prediction accuracy of 0.77 and a precision of 0.81, denotes that linear assumptions in relation to the association between pollen and biome types are also valid. Furthermore, it was shown in both [1] and our results that even through simplified assumptions and disregarding a fraction of individual pollen features where only the use of presence/absence proxy data was had (NB model with Bernoulli), it is still possible to classify and predict biomes to a reasonably high degree of accuracy and precision.

# 5. Related Works

In pg.17, [1] mentions that the previous study, "Late Miocene vegetation and climate reconstruction based on pollen data from the Sofia Basin" attempts to identify potential pollen indicators for quantitative reconstruction of temperature and precipitation [47]. Here, a total of 109 pollen and spore samples were analysed by standard methods for disintegration of tertiary sediments, and taxonomic identifications of fossil pollen grains were taken purely from

pollen keys and atlases [48]. As opposed to supervised machine learning classification, they applied the Coexistence Approach Method (CAM), facilitated by the Paleoflora database, to palynological data to calculate four climatic parameters. Similar to an NB model with Bernoulli, CAM relies only on the presences/absences of taxa and not on their abundances and is therefore heavily dependent on the sampling size. Hence, although their results yielded a diverse flora and showcased three distinct stages of vegetation dynamics, it can be said that CAM would have been more robust to taphonomic filtering where the pollen assemblage was more diverse.

In another and perhaps more analogous and recent study, "Reconstructing past biome states using machine learning and modern pollen assemblages: A case study from Southern Africa", they applied a recently developed pollen-based vegetation classification utilising supervised machine learning to South African modern pollen assemblages [11]. More specifically, 211 surface pollen samples were trained using a RF algorithm in R version 4.6 and was optimised using the tuneRF function in the same package. Bootstrap aggregating was also done to subsample pollen data. Their model correctly classified pollen assemblages up to 95% for both sub-humid savanna and dry savanna, 91% for coastal forest and 90% for wet grasslands. These high success rates only reinforce the application of machine learning approaches for past vegetation dynamics. Unlike [47], the results are measures of vegetation type and not climatic factors like temperature, precipitation and moisture, which can only be arbitrarily determined from their ranges within a biome according to its definition.

# 6. Conclusion

In this project, we applied supervised machine learning classification approaches to modern pollen assemblages for biome prediction. Imbalanced nature of the original dataset lead to poor predictions of the underrepresented biome classes in [1]. To solve imbalanced classification SMOTE was applied and all biome classes were oversampled to mimic the number of instances in the majority class. This has been proven to significantly increase the overall accuracy of the prediction models signaling that data balancing was crucial for accurate more prediction. In the training and testing of various parametric and non-parametric models, the Random Forest proved to be the best performing classifier; however, we recognise that the predictive capabilities of our models are largely confined to the available dataset and its adequate representation of various biomes and vegetation types.

The results of this paper have special implications for probabilistic reconstruction of past biomes and vegetation histories from fossil pollen records that sport the potential to enrich our comprehension of both spatial and temporal distributions. In addition, our method could be employed to help regulate disparate reconstructions realised through other methods.

# 7. Future Work

As in any machine learning problem, the results could be greatly improved in the future if the dataset could be expanded. With emphasis in the biomes with very little sampling, such as The Flooded Grasslands and Savannas and the ones that were not even considered in the study due to its lack of data: Temperate Grasslands, Savannas, and Shrublands (8 instances), Mediterranean Forests, Woodlands, and Scrub (4 instances), Mangroves (3 instances), Tropical and Subtropical Dry Broadleaf Forests (1 instance).

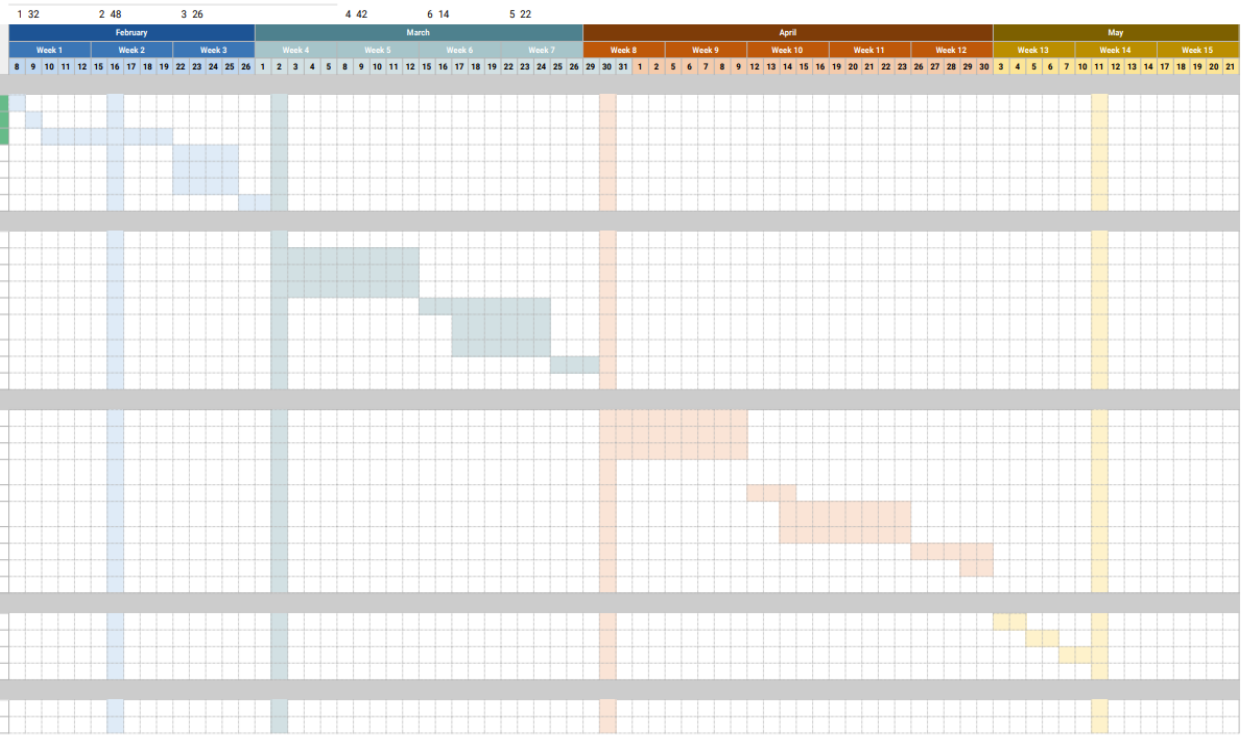
Biome modeling via machine learning could also be enhanced further when used conjointly with other methods such as traditional pollen analysis or pollen-based quantitative climate reconstructions.

# Appendix: Gantt Chart

## Machine Learning Project

| | |
|---|---|
| Title | Predictive pollen-based biome modeling using machine learning |
| Members | Austėja, Camila, Paulene |

| Month | TASK TITLE | Details | TASK OWNER | DUE DATE | PCT OF TASK COMPLETE |
|---|---|---|---|---|---|
| **2** | **February** | **Understanding** | | | |
| | **Choosing theme** | | All | **2021-02-16** | 100% |
| 2.1 | Business understanding | Reading the chosen paper and researching further references | All | 2021-02-09 | 100% |
| 2.2 | Data understanding | Understand the Database | All | | 100% |
| 2.3 | Data preparation | Searching more updated/comprehensive data | Austėja | | 0% |
| 2.4 | Methodology | Writing about the chosen part | Camila | | 0% |
| 2.5 | Introduction + Reference work | Writing about the chosen part | Paulene | | 0% |
| 2.6 | Making the 1st presentation ppt | | All | **2021-03-01** | 0% |
| **3** | **March** | **Parametric classification models** | | | |
| 3.1 | **Project revision session** | | | **2021-03-02** | 0% |
| 3.2 | Linear Discriminant Analysis | | Austėja | | 0% |
| 3.3 | Logistic Regression | | Camila | | 0% |
| 3.4 | Naïve Bayes | | Paulene | | 0% |
| | Testing with the updated database | | | | |
| 3.5 | Evaluation | A confusion matrix, accuracy, kappa statistic, F1, and weighted precision and recall for every model | All | | 0% |
| 3.6 | Deployment | Comparing the paper and our model | All | | 0% |
| 3.7 | Making the 2st presentation ppt | | | | 0% |
| 3.8 | **Project revision session** | | | **2021-03-30** | 0% |
| **4** | **April** | **Non-parametric classification models** | | | |
| 4.1 | K-Nearest Neighbors | | Austėja | | 0% |
| 4.2 | Classification Decision Tree | | Camila | | 0% |
| 4.3 | Random Forest | | Paulene | | 0% |
| 4.4 | Support Vector Machine | | **TBD: Whoever has more time** | | 0% |
| | Testing with the updated database | | All | | |
| 4.5 | Evaluation | A confusion matrix, accuracy, kappa statistic, F1, and weighted precision and recall for every model | All | | 0% |
| 4.6 | Deployment | Comparing the paper and our model | All | | 0% |
| 4.7 | Paper presentation | | All | | |
| 4.8 | Presentation preparation | | All | | |
| x | Neural Network | Will not be needed | - | | 0% |
| **5** | **May** | **Presentation** | | | |
| 4.1 | Presentation review | | All | | 0% |
| 4.2 | Paper review | | All | | 0% |
| 4.3 | **Project report** | | All | | 0% |
| 4.4 | **Project evaluation session** | | All | **2021-05-11** | 0% |
| **6** | **June** | | | | |
| 6.1 | Study for the exam | | All | | 0% |
| 6.2 | Test | | All | **2021-06-04** | 0% |




This Gantt chart was updated fortnightly to account for current progress and obstacles encountered. Some of the work formerly planned to be completed in March needed to be postponed until mid-April to allow for a more comprehensive literature review on the supervised classification models and learning the fundamentals of Python. This meant that the key focus from mid-April was writing a functional Python script, running the simulations and co-authoring our paper.

# References


[1] Sobol, M.K. and Finkelstein, S.A., 2018. Predictive pollen-based biome modeling using machine learning. *PloS one*, *13*(8), p.e0202214. https://doi.org/10.1371/journal.pone.0202214

[2] Trondman, A.K., Gaillard, M.J., Mazier, F., Sugita, S., Fyfe, R., Nielsen, A.B., Twiddle, C., Barratt, P., Birks, H.J.B., Bjune, A.E. and Björkman, L., 2015. Pollen-based quantitative reconstructions of Holocene regional vegetation cover (plant-functional types and land-cover types) in Europe suitable for climate modelling. *Global change biology*, *21*(2), pp.676-697.

[3] Davis, B.A., Brewer, S., Stevenson, A.C. and Guiot, J., 2003. The temperature of Europe during the Holocene reconstructed from pollen data. *Quaternary science reviews*, *22*(15-17), pp.1701-1716.

[4] Seppa¨ H, Birks HJB. July mean temperature and annual precipitation trends during the Holocene in the Fennoscandian tree-line area: pollen-based climate reconstructions. The Holocene. 2001; 11: 527–539

[5] Cao, X.Y., Herzschuh, U., Telford, R.J. and Ni, J., 2014. A modern pollen–climate dataset from China and Mongolia: Assessing its potential for climate reconstruction. *Review of palaeobotany and palynology*, *211*, pp.87-96.

[6] Prentice, C., Guiot, J., Huntley, B., Jolly, D. and Cheddadi, R., 1996. Reconstructing biomes from palaeoecological data: a general method and its application to European pollen data at 0 and 6 ka. *Climate Dynamics*, *12*(3), pp.185-194.

[7] Prentice IC, Cramer W, Harrison SP, Leemans R, Monserud RA, Solomon AM. A global biome model based on plant physiology and dominance, soil properties and climate. J Biogeogr. 1992; 19: 117– 134

[8] Dallmeyer, A., Claussen, M. and Brovkin, V., 2019. Harmonising plant functional type distributions for evaluating Earth system models. *Climate of the Past*, *15*(1), pp.335-366.

[9] Tim Menzies, Martin Shepperd, "Bad smells" in software analytics papers, Information and Software Technology, Volume 112, 2019, Pages 35-47, ISSN 0950-5849, https://doi.org/10.1016/j.infsof.2019.04.005

[10] Verleysen, M. and François, D., 2005, June. The curse of dimensionality in data mining and time series prediction. In *International work-conference on artificial neural networks* (pp. 758-770). Springer, Berlin, Heidelberg.

[11] Sobol, M.K., Scott, L. and Finkelstein, S.A., 2019. Reconstructing past biomes states using machine learning and modern pollen assemblages: A case study from Southern Africa. *Quaternary Science Reviews*, *212*, pp.1-17.

[12] Russell SJ, Norvig P. Artificial intelligence: a modern approach. Malaysia: Pearson Education Limited; 2016

[13] Casella G, Fienberg S, Olkin I. An Introduction to statistical learning with Applications in R. InSpringer Texts in Statistics 2013. Springer New York.

[14] Visa, S., Ramsay, B., Ralescu, A.L. and Van Der Knaap, E., 2011. Confusion Matrix-based Feature Selection. *MAICS*, *710*, pp.120-127.

[15] Ben-David, A., 2008. Comparison of classification accuracy using Cohen's Weighted Kappa. *Expert Systems with Applications*, *34*(2), pp.825-832.

[16] Powers, D.M., 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

[17] Brain, D. and Webb, G.I., 1999, December. On the effect of data set size on bias and variance in classification learning. In *Proceedings of the Fourth Australian Knowledge Acquisition Workshop, University of New South Wales* (pp. 117-128).

[18] Mack, R.N. and Lonsdale, W.M., 2001. Humans as global plant dispersers: getting more than we bargained for: current introductions of species for aesthetic purposes present the largest single challenge for predicting which plant immigrants will become future pests. *BioScience*, *51*(2), pp.95-102.

[19] Olson, D.M. and Dinerstein, E., 2002. The Global 200: Priority ecoregions for global conservation. *Annals of the Missouri Botanical garden*, pp.199-224.

[20] Gajewski, K., Lézine, A.M., Vincens, A., Delestan, A. and Sawada, M., 2002. Modern climate–vegetation–pollen relations in Africa and adjacent areas. *Quaternary Science Reviews*, *21*(14-15), pp.1611-1631.
[21] Blagus, R., Lusa, L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics 14, 106 (2013). https://doi.org/10.1186/1471-2105-14-106
[22] Hais, M., Komprdová, K., Ermakov, N. and Chytrý, M., 2015. Modelling the Last Glacial Maximum environments for a refugium of Pleistocene biota in the Russian Altai Mountains, Siberia. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *438*, pp.135-145.
[23] Murphy, K.P., 2012. *Machine learning: a probabilistic perspective*. MIT press.
[24] Hastie, T., Tibshirani, R. and Friedman, J., 2009. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.
[25] Simpson, G.L., 2012. Tracking Environmental Change Using Lake Sediments: Developments in Paleoenvironmental Research.
[26] Mahalanobis, P.C., 1936. On the generalized distance in statistics. National Institute of Science of India.
[27] Walker, S.H. and Duncan, D.B., 1967. Estimation of the probability of an event as a function of several independent variables. *Biometrika*, *54*(1-2), pp.167-179.
[28] Reynolds, O., 1886. IV. On the theory of lubrication and its application to Mr. Beauchamp tower's experiments, including an experimental determination of the viscosity of olive oil. *Philosophical transactions of the Royal Society of London*, (177), pp.157-234.
[29] Hastie, T., Tibshirani, R. and Friedman, J., 2009. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.
[30] Birks, H.J.B., Lotter, A.F., Juggins, S. and Smol, J.P. eds., 2012. *Tracking environmental change using lake sediments: data handling and numerical techniques* (Vol. 5). Springer Science & Business Media.
[31] Ali, J., Khan, R., Ahmad, N. and Maqsood, I., 2012. Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, *9*(5), p.272.
[32] Breiman, L., 1996. Bagging predictors. *Machine learning*, *24*(2), pp.123-140.
[33] Cortes, C. and Vapnik, V., 1995. Support-vector networks. *Machine learning*, *20*(3), pp.273-297.
[34] Vapnik, V., Golowich, S.E. and Smola, A., 1997. Support vector method for function approximation, regression estimation, and signal processing. *Advances in neural information processing systems*, pp.281-287.
[35] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, pp.2825-2830.
[36] Van Der Walt, S., Colbert, S.C. and Varoquaux, G., 2011. The NumPy array: a structure for efficient numerical computation. *Computing in science & engineering*, *13*(2), pp.22-30.
[37] McKinney, W., 2010, June. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51-56).
[38] Bergstra, J. and Bengio, Y., 2012. Random search for hyper-parameter optimization. *Journal of machine learning research*, *13*(2).
[39] Faegri, K., Kaland, P.E. and Krzywinski, K., 1989. *Textbook of pollen analysis* (No. Ed. 4). John Wiley & Sons Ltd..
[40] Moore, P.D., Webb, J.A. and Collison, M.E., 1991. *Pollen analysis*. Blackwell scientific publications.
[41] Bourgeois, J.C., Gajewski, K. and Koerner, R.M., 2001. Spatial patterns of pollen deposition in arctic snow. *Journal of Geophysical Research: Atmospheres*, *106*(D6), pp.5255-5265.
[42] Bourgeois, J.C., 2000. Seasonal and interannual pollen variability in snow layers of arctic ice caps. *Review of Palaeobotany and Palynology*, *108*(1-2), pp.17-36.
[43] Twiddle, C.L. and Bunting, M.J., 2010. Experimental investigations into the preservation of pollen grains: A pilot study of four pollen types. *Review of Palaeobotany and Palynology*, *162*(4), pp.621-630.
[44] Bryant VM, Holloway RG. Archaeological palynology. In: Jansonius CJ, M D., editors. Palynology: principles and applications. Dallas, TX.: American Association of Stratigraphic Palynologists Foundation; 1996. pp. 913–917.

[45] Dimbleby GW. Pollen Analysis of Terrestrial Soils. New Phytol. 1957; 56: 12–28. https://doi.org/10.1111/j.1469-8137.1957.tb07446.x
[46] Mander, L., Baker, S.J., Belcher, C.M., Haselhorst, D.S., Rodriguez, J., Thorn, J.L., Tiwari, S., Urrego, D.H., Wesseln, C.J. and Punyasena, S.W., 2014. Accuracy and consistency of grass pollen identification by human analysts using electron micrographs of surface ornamentation. *Applications in Plant Sciences*, *2*(8), p.1400031.
[47] Hristova, V. and Ivanov, D., 2014. Late Miocene vegetation and climate reconstruction based on pollen data from the Sofia Basin (West Bulgaria). *Palaeoworld*, *23*(3-4), pp.357-369.
[48] Faegri, K., Iversen, J. and Waterbolk, H.T., 1964. Textbook of pollen analysis.