Filling in the gaps in the NDFF

Student:

Debarupa Roy Choudhury	[2423340]	d.roychoudhury@students.uu.nl

Supervisors:

Ad Feelders	a.j.feelders@uu.nl
Menno Straatsma	m.w.straatsma@uu.nl

ADS Thesis

Thesis Applied data science [INFOMTADS], 2022-2023

Utrecht University: MSc Applied Data Science

02 July, 2023

Table of contents

1. Introduction

1.1. Motivation and context

We are doing this research in association with **TAUW**, which is an organization that is committed to sustainable development, environmental stewardship, and social responsibility. They aim to contribute to a more sustainable and resilient future by providing solutions that balance economic, environmental, and social interests.

TAUW uses the **NDFF** database to advise on measures pertaining to biodiversity and policy. Together with dedicated fieldwork that is complementary to these insights, they utilize this database to monitor biodiversity and provide limited insight into the current status of biodiversity.

The Dutch National Database on Flora and Fauna, or NDFF, has more than 20 million records spanning more than 40 years and practically all significant taxa found in the Netherlands.

The NDFF is a data warehouse that was built to house information about the distribution of plants and animals in the Netherlands. The collection includes observations from structured surveys as well as chance observations made by volunteers. The database as a result suffers from frequent observer effects, such as places with higher documentation than others.

The database specifically includes geospatial information related to species observations, species occurrences, and species-specific attributes such as abundance, habitat preferences, and conservation status. This information is usually collected through field surveys, monitoring programs, and citizen science contributions.

Researchers, conservationists, policymakers, and the general public can access the NDFF data to gain insights into the biodiversity of the Netherlands and support conservation and management decisions.

Throughout our research, we would predict missing data caused by the common observer effects that are often present in the NDFF.

We want to predict values of the occurrence of butterflies based on environmental characteristics and observations from the NDFF for areas where we don't have any information. Based on other areas that are better represented in the database, we would build our training set by combining lots of other data that describe the ecotope or habitat of that particular species.

A machine learning model that has been trained on the environmental traits of the more documented locations can be used to forecast the frequency of occurrences in the less well-recorded areas to fill in these "gaps" in the data. Data such as satellite imaging, land use data, soil data, distance to roads/build-up, etc. should be able to describe these qualities.

Following data preparation, a machine learning model can be trained on a subset of the data. Support vector machines, random forests, linear regression (with or without feature interaction), etc would all be contenders. Regularization techniques can be employed to enhance the model's performance because not all features will be of equal importance to it. A comparison between various models will be done in our research.

1.2. Research Question

We want to be able to make estimates for places where we don't have any observations, hence we're interested in forecasting a species' occurrences.

In places where we lack knowledge, we would like to make predictions about values such as the presence/absence of certain species. It’s better to create the training set by integrating additional data that speak to an ecotope or habitat of that specific species based on other places that are better represented in the NDFF such as vector data or aerial photography (Top10NL, Natura2000, etc.). Our main research question of this study is:

RQ: *How can we use machine learning to fill the gaps in the NDFF?*

To answer this question, two research areas have been formulated:

RQ1: *We will focus more on modeling the presence/absence of species rather than just predicting their count.*

RQ2: *Along with that, we will also focus on the features which help us to predict whether the species is observed or missing.*

Our research is distributed into the following stages:

A. The production of a "ready for analysis" data set with attributes of a location or area listed in each row together with the matching occurrence counts. We assembled data from many sources, choose the appropriate level of aggregation (such as area size), and included fabricated 0 counts.
B. Using machine learning algorithms such as linear regression, support vector regression, random forests, etc., analyze this data set to produce predictive models. Additionally, this section aims to provide light on the characteristics/features that are crucial for predicting the presence/absence of a species.
C. Reviewing the developed models.

1.3. Literature Overview

The goal of species distribution modeling (SDM) or predicting where a species is likely to be found, is to identify patterns of space (and occasionally time) in which a species occurs. Powerful machine learning methods have recently attracted a lot of attention as a solution to difficult ecological problems (Elith Jane Elith et al. – Presence-only and Presence-absence Data for Comparing Species Distribution Modeling Methods 70 et al. 2006).

When developing an SDM, a number of factors need to be taken into account, such as the suitability of the underlying model assumptions, the choice of modeling algorithms, the tuning of the model parameters and complexity, the selection of background data, and the availability of species data and environmental predictors (Araujo et al. 2019). Predictions may be greatly influenced by these factors.

Given the variety of possible methodologies, the selection of the modeling algorithm is frequently significant among these factors. There is continual interest in detecting broad trends in predictive performance across methodologies because the predictive success of different methods varies (Pearson et al. 2006; Thuiller et al. 2009).

Species-distribution modelers frequently use model averaging or ensemble modeling because it is thought to have a stronger predictive power and to be more trustworthy than single models (Araujo and New 2007, Marmion et al. 2009; Hao et al. 2019).

In biogeography, various types of generalized linear models (GLMs), classification and regression trees (CART), ecological niche factor analysis (ENFA), genetic algorithms (GA), point pattern analysis algorithms, maximum entropy-based techniques, and similar methods are preferred for spatial prediction. The kind of occurrence records we use—presence-only records, presence/absence records, counts, or real measurements of a species' attributes—determine the kind of model we can employ in significant part. (T. Hengl; H. Sierdsema)

In Ecological modeling, forecasting species abundance and predicting species presence or absence are two different goals. Estimating a species' population size or quantity in a certain location while taking environmental conditions and habitat features into account is the process of predicting species abundance. Its goal is to quantify a species' abundance, such as its population size or biomass, in order to gain knowledge of its population dynamics and geographic dispersion. A species' occurrence at a certain area or habitat is the focus of forecasting species presence/absence, on the other hand. Understanding species distribution patterns and guiding conservation efforts, these models evaluate the presence or absence of a species based on elements like environmental variables and habitat appropriateness.

It is less well-established how species traits affect the accuracy of abundance models as compared to that of presence/absence models. According to theory, it may be difficult to accurately predict abundance for some species, such as those with wide geographic ranges and dense populations (Chisholm and MullerLandau 2011, Peterson et al. 2011, Yaez-Arenas et al. 2014, Chu et al. 2016, Bowler et al. 2017, Yenni et al. 2017, Hallett et al. 2018). Contrarily, rare (low mean abundance) species with constrained niches frequently display more stable populations, and as a result, abundance might be more predictable (Yenni et al. 2017). A species distribution model's performance may also be impacted by the qualities of the data.

By being less regionally and environmentally biased, more samples generally enhance the performance of species distribution models, which should also enhance the performance of abundance models (Yaez-Arenas et al., 2014). These consequences haven't been examined, though.

The performance of species distribution models is frequently correlated with species and data properties. For conservation and management applications, specifically in relation to commonness and rarity, it is essential to establish how and why model performance varies for different species.

Common species frequently make the biggest contributions to ecosystem functioning in terms of their local and regional abundance (Genung et al. 2020). Low abundance and range-restricted species may be given priority for conservation since they are more likely to go extinct (Purvis et al. 2000; Ceballos et al. 2020) and may have special functions in ecosystems (Violle et al. 2017). In general, species with smaller ranges, less endemicity, and non-migratory behaviours perform better in species distribution models; performance is also favorably impacted by the number of observations (McPherson and Jetz 2007, Newbold et al. 2009, Chefaoui et al. 2011, Thuiller et al. 2019).

2. Data

2.1. Motivation for choice of target species

Our motivations behind the choice of the target species are listed below:

I. The "**Heideblauwtje**" species with a variety of traits are interesting for our research since they are also modeled by NDFF which can help us to validate our research in the future.
II. We chose the insect species based on the NDFF's thorough documentation, low dispersal, and lack of significant reliance on the presence of other kinds of animal species. Abiotic factors like vegetation characteristics, plant habitats, etc have a greater influence on these occurrences.

2.2. Data sources and pre-processing

We compared the land use classification found in open data sources (such as BRT TOP10NL, N2000, etc.) for the view landscapes and biotopes that are of importance for identifying biodiversity. The majority of these data sources were obtained at PDOK or nationaalgeoregister.

Our **Area of Interest (AOI)** or the geographic extent of our research was Amersfoort / RD New -- Netherlands - Holland – Dutch (EPSG:28992).

Some of the dataset's components were also created using satellite images that the planetary computer had downloaded in response to our AOI shapefile.

The general strategy for data preparation was:

1. Downloading the data
2. If necessary, reprojecting to epsg:28992
3. Clipping to the AOI's maximum extent
4. Using the CLC (Corine Land Cover dataset) 2018 as a model for rasterizing
5. Conversion to CSV

6. Combining all CSV data into a single final dataset.
7. Next, variable names were cleaned.

The NDFF was cleaned by removing all data older than 2012 and records whose area exceeded the size of a cell.

The **ndff_combined_dataset.csv** file contains the merged dataset. Also, there is the target raster named “rasterized_ndff_count.tiff” and a document “named lookup_merged_categories.txt” that defines the categories which are merged in the dataset.

Each record in the CSV file is an observation with information on how it was done. The observation may be a point or polygon. The butterfly observations are also added per year to the dataset, as well as there is a column containing the total.

Binary maps are added to datasets with only one category. Datasets with numerous categories are added as a stacked TIF, where each band represents a binary map for a single category.

The number of rows is 207,957 rows and 31 columns in the combined data set with 1243 non-zero records, where the grid cell size is 100x100m.

The **ndff_count_total** is the total abundance of observations in each row, corresponding to a cell (center). This represents the total of all observations made over the previous ten years (earlier data are not indicative of a species' current distribution).

Additionally, we got rid of the dataset's extremely long category names. More variable truncation was a possibility, but we ultimately decided not to consider it because it would make the dataset harder to understand. With regard to several of the categorical variables, we consolidated the categories with extremely low counts.

2.3. Selected data exploration

NDFF is built on observer records (point registrations). The number of observer logs is large in both urban and natural regions that are easily accessible to people, but it is low or non-existent in areas that are either difficult to access or not as well-known as natural areas. Additionally, there is a wide range in the number of observations made by each contributor. As a result, the dataset is biased because flora and fauna are only found in specific locations clustered together or close to the residences of few contributors. The quantity of observer logs also changes over time.

Thus, using the NDFF alone to study biodiversity will only provide extremely limited insights at this time. To build a model regarding the NDFF, further open geographic data will be needed like environmental elements like a biotope or a natural habitat. The model should become familiar with these environmental traits that might be related to the existence of an observation.

Considerable amounts of information are available, including the BRT, BGT, BAG, AHN, Natura2000, BRP (gewaspercelen), TOP10NL, soil maps, "Houtopstanden," as well as aerial or satellite imagery for research.

The following are environmental predictors of interest for our research analysis:

Input features:

I. AAN
II. bro_genese
III. bro_landform
IV. BRP_gewas
V. cbs_landuse
VI. clc2018
VII. ndvi_2022
VIII. natura_2000
IX. fysisch_geografische_regios
X. nationale_parken
XI. NOK_begrenzing
XII. NOK_beheer
XIII. NOK_planologische_ehs
XIV. NOK_verwervinginrichting
XV. SGM_ondergrond
XVI. Stiltegebieden
XVII. Wetlands

See **Table 1** for the description of different data sources(**features**).

Table 1: Description of the features used for the model.

Features	Description	Domain values
natura_2000	Natura 2000 is a geospatial dataset that combines the information about Natura 2000 protected areas with the PDOK infrastructure. It includes data on the boundaries and locations of Natura 2000 sites within the Netherlands, potentially accompanied by additional geographic information provided by PDOK.	**Integer values:** • 0 (Not protected area) • 1 (Protected area)

wetlands	Wetlands contain geospatial information about the boundaries and locations of wetland areas within the Netherlands. It may include details about different types of wetlands, such as their size, hydrological characteristics, and associated vegetation.	**Integer value:** • 0 (Not wetland area) • 1 (Wetland area)
nationale_parken	Nationale Parken includes geospatial information about the boundaries and locations of national parks within the Netherlands. It may provide details about the size, boundaries, and key features of each national park, such as notable landscapes, habitats, and recreational facilities.	**Integer values:** • 0 (Not National Park area) • 1 (National Park area)
fysisch_geografische_regios	Fysisch Geografische Regios contains geospatial information about the boundaries and locations of different physical geographic regions within the Netherlands. These regions could be classified based on various factors such as landform types, soil characteristics, or hydrological patterns.	**Object values:** This is a dataset on its own without categories being merged.
Basic Registration of Crop Plots (BRP): BRP_gewas	BRP_gewas specifically focuses on the crop or agricultural land use information associated with each registered parcel. It provides geospatial data that identifies the types of crops cultivated on different parcels of agricultural land in the Netherlands.	**Object values:** The categorical values that have a very low count, that is those that are very rare in the dataset, were merged together.
Agricultural Area Netherlands (AAN)	AAN includes data on the boundaries and characteristics of agricultural land parcels throughout the country. It may contain	**Integer values:** • 0 (Not agricultural area)

	information such as land use types, crop types, agricultural practices, parcel sizes, and ownership details.	- 1 (Agricultural area)
Normalized Difference Vegetation Index (NDVI): ndvi_2022	NDVI is a widely used vegetation index in remote sensing and satellite imagery analysis. It is calculated from the reflectance values of red and near-infrared (NIR) light wavelengths captured by satellite sensors. It is a valuable tool for assessing vegetation health, density, and biomass. It provides information on vegetation growth, changes, and spatial distribution over time.	**Float values: (-0.00009 to +0.5)** NDVI is mainly derived from the below formula: **NDVI = (NIR - Red) / (NIR + Red)**
cbs_landuse	CBS_landuse contains geospatial information about the land use categories and their spatial distribution across the country. It includes categories such as residential areas, agricultural land, forests, industrial zones, water bodies, and other land use types. The dataset also provides attributes such as area size, land use codes, and possibly additional information about land use patterns and trends.	**Object values:** The categorical values that have a very low count, that is those that are very rare in the dataset, were merged together. - Building site - Cemetery - Dry natural area - Other agricultural usage - Other inland water - Sports ground - Water with recreational usage - Wet natural area - Woodland
Nature measurement on map (NOK):	NOK typically includes geospatial information related to nature	**Integer values:** - 0

i. NOK_begrenzing ii. NOK_beheer iii. NOK_planologische_ehs iv. NOK_verwervinginrichting	conservation and management efforts. It contains data on various aspects, such as the boundaries of protected areas, habitat quality, biodiversity assessments, ecological connectivity, and species distributions. The dataset may also include information about specific nature management actions, restoration projects, and conservation targets.	• 1
stiltegebieden	Stiltegebieden contains geospatial information about the boundaries and locations of designated quiet areas throughout the country. These areas are typically chosen based on their natural characteristics, distance from major roads or urban centers, and absence of significant noise sources. Stiltegebieden is often established to protect and preserve natural soundscapes, promote a sense of calm and relaxation, and provide opportunities for solitude and reflection.	**Integer values:** • 0 (Not a quiet area) • 1 (Quiet area)
National Database of Flora and Fauna (NDFF): i. ndff_count_total ii. ndff_count_2012 iii. ndff_count_2013 iv. ndff_count_2014 v. ndff_count_2015 vi. ndff_count_2016 vii. ndff_count_2017 viii. ndff_count_2018 ix. ndff_count_2019 x. ndff_count_2020 xi. ndff_count_2021 xii. ndff_count_2022	NDFF is the national database that collects and manages data on the occurrence, distribution, and characteristics of various plant and animal species in the Netherlands. It serves as a central repository for biodiversity information and is maintained by various organizations, including government agencies, research institutions, and citizen science initiatives.	**Integer values:** **ndff_count_total***:* 0 to 3780 **ndff_count_2012*:*** 0 to 81 **ndff_count_2013:** 0 to 362 **ndff_count_2014*:*** 0 to 227 **ndff_count_2015:** 0 to 322 **ndff_count_2016*:*** 0 to 518 **ndff_count_2017*:*** 0 to 2001

		ndff_count_2018: 0 to 916 **ndff_count_2019:** 0 to 91 **ndff_count_2020:** 0 to 157 **ndff_count_2021:** 0 to 211 **ndff_count_2022:** 0 to 134
Corine Land Cover (CLC): clc2018	Corine Land Cover is a pan-European program that aims to classify and map land cover across the continent. It provides standardized information about different land cover categories, such as forests, agricultural areas, wetlands, urban areas, water bodies, and other land cover types. The dataset classifies land cover based on satellite imagery and other available data sources, allowing for the assessment of land use patterns, environmental changes, and habitat fragmentation.	**Object values:** This is a dataset on its own. • 112 - Discontinuous urban fabric • 121 - Industrial or commercial units • 124 – Airports • 142 - Sport and leisure facilities • 211 - Non-irrigated arable land • 231 – Pastures • 243 - Land principally occupied by agriculture with significant areas of natural vegetation • 311 - Broad-leaved forest • 312 - Coniferous forest • 313 - Mixed forest • 321 - Natural grasslands • 322 - Moors and heathland

		- 331 - Beaches - dunes – sands
Base Registration Subsoil (BRO): i. bro_genese ii. bro_landform	BRO aims to collect and manage data related to the geological, hydrological, and geotechnical properties of the subsurface. It includes information about soil composition, groundwater levels, geological formations, and other relevant subsoil features. It includes geospatial information such as borehole locations, soil profiles, lithological descriptions, hydrogeological parameters, and geotechnical data. It is collected through various methods, including drilling, sampling, and geophysical surveys.	**Object values:** The categorical values that have a very low count, that is those that are very rare in the dataset, were merged together.
Soil Map (SGM): SGM_ondergrond	Soil Map represents the distribution and characteristics of soils within a specific area or region. Soil maps typically provide information about soil types, soil properties (such as texture, organic matter content, pH, and nutrient levels), soil depth, and other relevant soil characteristics. These maps are created through field surveys, soil sampling, laboratory analysis, and spatial interpolation techniques.	**Object values:** The categorical values that have a very low count, that is those that are very rare in the dataset, were merged together.

Target Variable (Calculated Per Area):

ndff_count_total (Total species count for 10 years)

ndff_count_2012

ndff_count_2013

ndff_count_2014

ndff_count_2015

ndff_count_2016

ndff_count_2017

ndff_count_2018

ndff_count_2019

ndff_count_2020

ndff_count_2021

ndff_count_2022

See Table 2 in Appendix A, section 7.1 for an overview of the data.

3. Method

3.1. Selected methods description

In our research, we sampled pseudo-absences from locations where there are no observations, and using the selected model, we predict all the locations with zero and pseudo-absence observations on which we trained our model on.

An overview of all used methods can be found in Table 3 from Appendix B (section 7.2).

Our initial findings indicated that making predictions about species abundance levels holds limited potential. There is way too much noise in the NDFF count data, or there is not enough information in the environmental data to estimate abundance. We tried with linear regression model for that but it performed quite poorly (accuracies of +/- 10%). The accuracy might be improved significantly with a lot more data and careful tuning, but we think the improvements would only be slight and would not be feasible given our time and computational resource constraints. On the other hand, our initial findings for classification (presence/absence) indicate accuracy levels of about 90%. So, we made our move to the classification decision.

For pre-processing the data, we used one-hot encoding to convert categorical variables into numerical representations to use with models that primarily handle numerical data. Also, we used MinMaxScaler to scale/transform the input features to a given range.

We experimented with different machine learning models and evaluated their performance using k-fold cross-validation scores which helped us to measure how our model predicts out-of-sample.

We proceeded with classification models like the **Decision Tree classifier**, **Random Forest classifier, Extra Trees classifier, Support Vector Machine, K-Nearest Neighbors, and Naive Bayes** to analyze our research data. Our models helped us to conclude which features are of more relevance to our research using the Recursive Feature Elimination technique from Scikit Learn.

To perform hyperparameter tuning in the Random Forest Classifier, we used Grid Search with cross-validation to search through a range of hyperparameter values and find the optimal combination.

The best model was then evaluated based on Accuracy, F1-Score, Precision, Recall, training time, scoring time, and AUC-ROC and we made predictions on the test set using that model.

3.2. Motivation for Used Methods

For our research, we have experimented with various machine learning models. As mentioned in section 3.1, we focused on classification models to predict the presence/absence of species due to the nature of the data. Below are some important models used along with their motivation for our study.

Random Forest (RF) is an ensemble learning method that is effective for species presence/absence prediction due to its ability to handle complex interactions between predictor variables and capture non-linear relationships. RF can handle high-dimensional data, account for variable importance, and provide robust predictions. When building each tree, RF only picks a random subset at each split of the predictor variables. This results in decorrelated trees and lowers the final model's variance, both of which improve predictive performance (Hastie et al. 2009).

Decision Tree is a simple yet powerful algorithm for classification tasks. It partitions the predictor variables based on their values to create a tree-like model. Decision trees are intuitive to interpret, handle both numerical and categorical data (through one-hot encoding in Python), and can capture complex decision boundaries. It can manage correlations between predictor factors and species presence that are non-linear. They are capable of calculating the relative weights of various features (such as environmental variables) in the classification process. We can determine which characteristics have the biggest effects on the classification of species presence by looking at the tree structure. The key factors influencing species distributions can help us to prioritize data collection efforts and direct future ecological studies.

Extra trees classifier helps one to construct multiple decision trees using random feature subsets and random splits to improve diversity and reduce overfitting. It can handle high-dimensional data, provide feature importance measures, and perform well in the presence of noisy or irrelevant features.

We know that Logistic Regression models the relationship between predictor variables and the probability of species presence/absence using a logistic function. LR is computationally efficient, interpretable, and can handle both continuous and categorical predictors (through one-hot encoding in Python. However, it assumes a linear relationship between predictors and the log odds of presence/absence.

Complex linkages and significant predictors are adeptly captured by Random Forest and Extra Trees Classifier. While Logistic Regression calculates probability and manages several types of predictors, Decision Trees offer interpretability. Naive Bayes effectively manages high-dimensional data whereas KNN takes local context and spatial linkages into account.

We concentrated on the above modeling approaches that helped with model fitting and model selection when deciding which ones to include. By punishing and reducing regression coefficients, for example, regularization approaches (Friedman et al. 2010), hyperparameters tuning, and the Random elimination technique enhanced the predictive performance of models (James et al. 2013). These techniques also cause a significant decrease or removal of unimportant features.

4. Results and analysis

In Appendix D, section 7.4 (Figures 6, 7, 8, 9, and 10), we provide the descriptive analysis.

We considered the species which had a count equal to and greater than 1 as **presence** and all others as an **absence**.

In our research, we have combined **1243** records of presence and an equal amount of absence records randomly from the absence data to form a balanced dataset for our model. We have performed 5-fold cross-validation on this dataset with Decision Tree and Random Forest model. We found that the mean cross-validation score for Random Forest (0.82) is higher than the Decision Tree (0.76) model.

Also, we improved the performance of the RF model using data pre-processing, feature engineering, and hyperparameter tuning. We have performed hyperparameter tuning on our RF model with parameters as follows: 'n_estimators': [100, 200, 300], 'max_depth': [None, 5, 10], 'min_samples_split': [2, 5, 10], 'min_samples_leaf': [1, 2, 4], 'max_features': ['sqrt', 'log2'].

As a result, we have received optimal parameters ('max_depth': 10, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 300).

We have used 80% of this dataset as training data and the rest 20% as testing data. We have tried various classification models to evaluate and compare performance metrics. Overall, we found that the Random Forest model gives good results with parameters received through hyperparameter tuning.

Also, after successful model evaluation, we have predicted the presence of species for the whole absence data.

In our analysis, we found that the Random Forest (RF) models' accuracy was much higher compared to other machine learning models. It helped us to predict species presence/absence with about 91% accuracy.

Table 4 in Appendix C (section 7.3) provides an overview of our findings regarding all model's key performance metrics.

In Figure 1, we have filled the gaps and visualized the outcome for the complete dataset. It shows the possibility of species' presence predicted by our model.

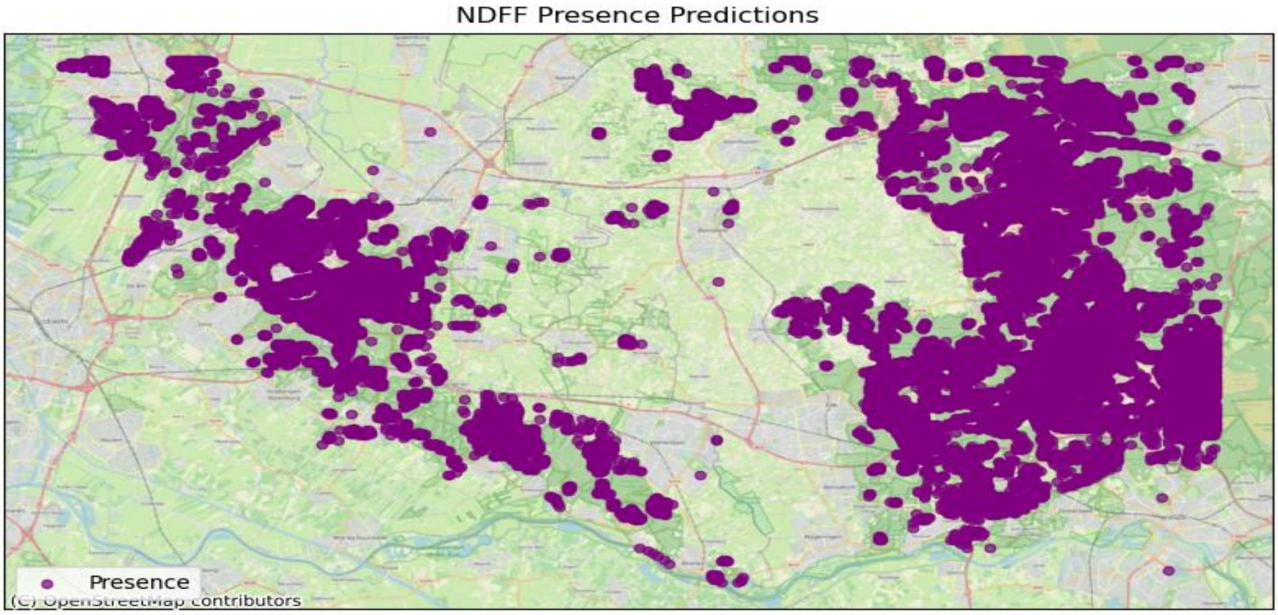

Figure 1: Visualization of the prediction of butterfly presence using Random Forest.

In Figure 2, we have shown the features which are more important to predict the presence of butterflies. Some feature names include actual features along with important values for that feature.

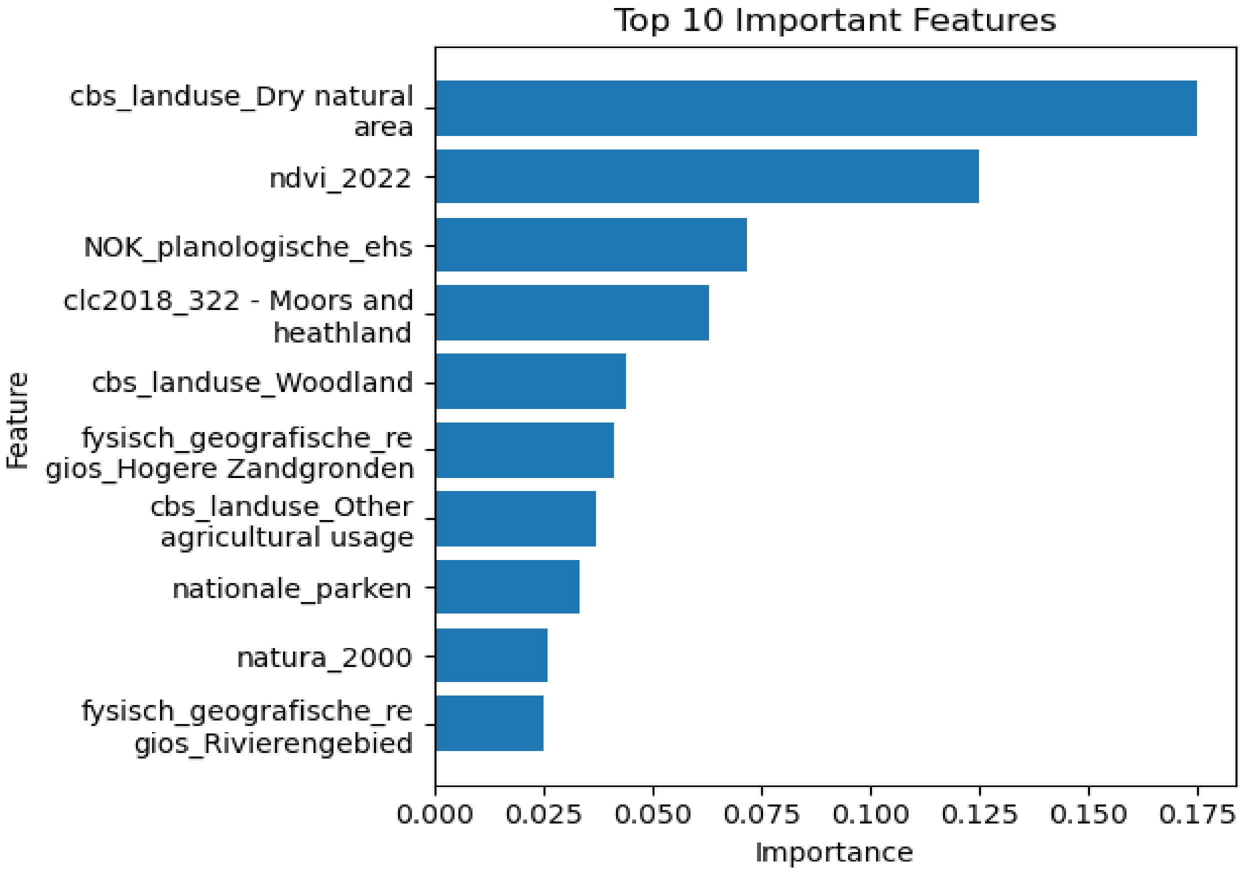

Figure 2: Barplot of top 10 features predicting butterfly presence/absence using Random Forest.

From Figure 3, we can see that the AUC score for the Random Forest model is better than other models.

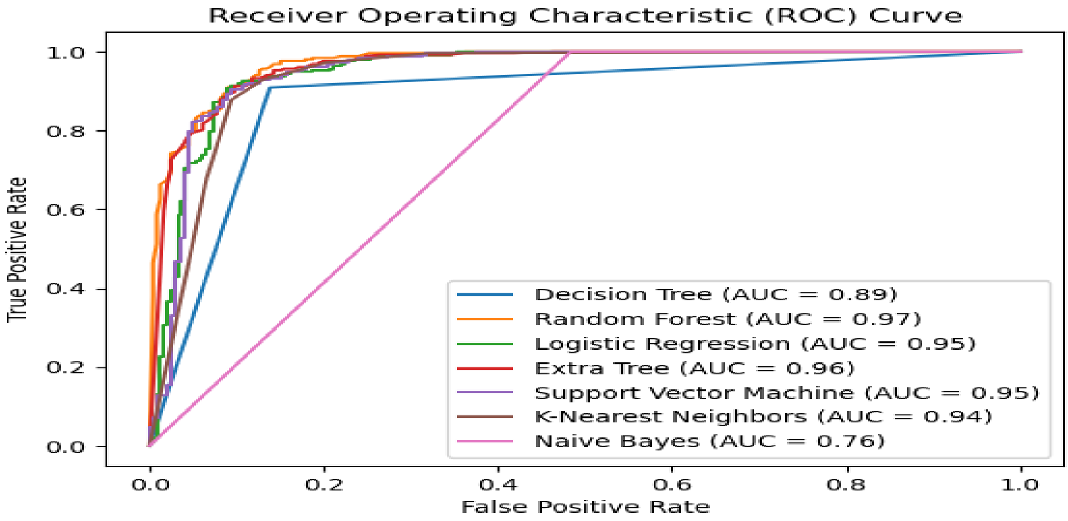

Figure 3: Plots for the ROC curve for each model with the AUC score displayed in the legend.

In Figure 4, we can see Support Vector Machine takes more time to train and score the model. On the other hand, Naive Bayes and Logistic Regression take less scoring time but give less accuracy. Also, Random Forest takes an average time to train along with less time to score and good accuracy.

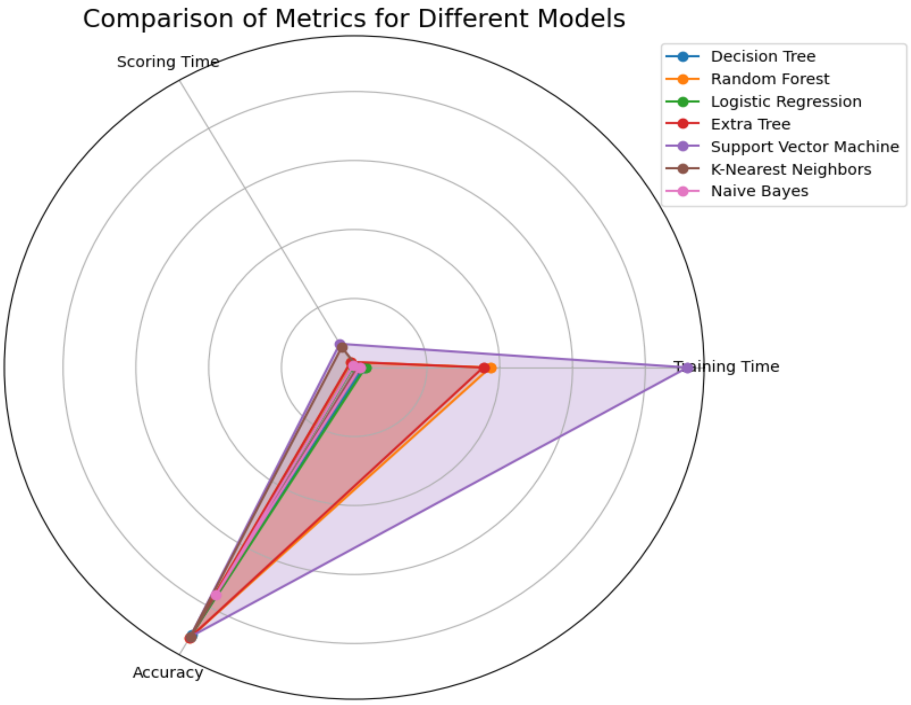

Figure 4: Radar Chart showing a comprehensive view of the performance of each model across multiple metrics.

5. Conclusion and discussion

5.1. Implications and Considerations for domain setting

A. After data analysis, we found that for the butterfly species, there have been no absence records in the NDFF during the past ten years across the entire Netherlands.
B. The features that can predict whether the species count is available for a certain location are also of much interest to our research.
C. We have considered different performance metrics of various models to find the best model and parameters for making good predictions.
D. As discussed with TAUW, we considered that older records are not relevant for our current distribution of butterflies.

5.2. Interpretation and Discussion

It is possible to estimate species geographic distributions using a variety of environmental factors. Although some modeling strategies generally outperform others, no one model is better than another in every circumstance. Autocorrelation, complex and nonlinear interactions, and changing spatial interaction are frequent features of natural systems. In these cases, nonparametric models frequently perform better than conventional parametric models (Evans and Cushman 2009).

Previous studies have shown that one of the main sources of variability in species presence/absence model performance is the structure of data (Fielding and Haworth 1995), especially the prevalence of species (Leathwick et al. 2006, Meynard and Quinn 2007, Syphard and Franklin 2009, Santika 2011, Madon et al. 2013) and the strength and shape of environmental gradients (Thuiller et al.2003, Austin et al. 2006, Santika and Hutchinson 2009, Hoffman et al. 2010, Santika 2011).

Here, we compare a variety of statistical and machine-learning methods frequently used for estimating species distributions. Our results highlight the significance of considering the peculiarities of presence-absence data when deciding how to implement different methods, in addition to providing particular conclusions concerning the effectiveness of alternative techniques. All models were fitted using the Python programming language.

The research shows good results with classification models for the presence/absence of species and the Random Forest classifier with hyperparameter tuning gave us optimal prediction with approx. 91 % accuracy.

Table 4 in Appendix C (section 7.3) provides an overview of our findings regarding the model's performance metrics.

5.3. Threats to validity

Some of the threats which our analysis may lead to are:

I. It's possible that there may be a discrepancy between the time at which a count was conducted and the time at which the feature values for that region were computed (for instance, if the area was once woodland but over that 10-year period houses were erected there). This is one of the aspects that was not considered in my analysis.

II. It can be challenging to successfully combine several data sources because Heideblauwtje observations may be gathered using a variety of different techniques. For instance, observations obtained by experts and volunteers have significantly different collection biases than the ones acquired through a well-designed scientific survey. It can be difficult to handle these biases in a rigorous, systematic manner, especially when dealing with big data sets made up of hundreds of diverse projects, each with its own unique sampling methods. It is frequently necessary to read up on the project's literature in order to comprehend the protocols employed for a particular data-gathering project inside a bigger repository. There aren't any readily available, defined definitions or methods for quantifying bias for many projects, though.

III. We know that observers are more likely to visit and report sightings in some areas than others (hereinafter referred to as "observer bias"), and presence-only data, which contain information on species existence but not absence, are subject to bias.

5.4. Future work

A. For our future research, we can focus on using deep learning algorithms where we can include features that are directly or indirectly influencing the habitat for butterflies as our input.

B. We can compare the predictions of our models against those of the NDFF-built models like ‘De Kansenkaart’ (models the likeliness of a species) for our future analysis, which may provide us with meaningful insights.

C. We can enhance our research using backward (or forward) stepwise regression to select only the features that contribute the most to predicting the observation count.

D. In order to confidently study biodiversity using the NDFF and external data sources, we can correct for spatio-temporal factors that can be found in the NDFF.

E. A critical area for model improvement is the incorporation of environmental variation at the appropriate spatiotemporal scale for a given species (Roslin et al. 2009; Ashcroft et al. 2014; Rebaudo et al. 2016), particularly for projections of future climate effects on species occurrence and abundance (Gillingham et al. 2012; Hannah et al. 2014; Maclean et al. 2015; Woods et al. 2015).

F. Fieldwork can be used to find an external source of validation in addition to the standard (spatial) evaluation techniques. The species and the locations to be selected will aid to determine this.

6. References

a. Species Distribution Modelling:

Valavi, R., G. Guillera-Arroita, J. J. Lahoz-Monfort, and J. Elith. 2022. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. Ecological Monographs 92(1): e01486. 10.1002/ecm.1486. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code.

Jane Elith1*, Catherine H. Graham2, Roozbeh Valavi1, Meinrad Abegg2, Caroline Bruce3, Simon Ferrier4, Andrew Ford5, Antoine Guisan6, Robert J. Hijmans7, Falk Huettmann, Lucia Lohmann9, Bette Loiselle10, Craig Moritz11, Jake Overton12, A. Townsend Peterson13, Steven Phillips14, Karen Richardson15, Stephen E. Williams16, Susan K. Wiser17, Thomas Wohlgemuth2, Niklaus E. Zimmermann2. PRESENCE-ONLY AND PRESENCE-ABSENCE DATA FOR COMPARING SPECIES DISTRIBUTION MODELING METHODS.

SARA BEERY∗ and ELIJAH COLE∗, California Institute of Technology JOSEPH PARKER, California Institute of Technology PIETRO PERONA, California Institute of Technology KEVIN WINNER, Yale University. Species Distribution Modeling for Machine Learning Practitioners: A Review.

SDM example in R, from URL: https://rpubs.com/mlibxmda/SDMPartOne

Spatio-temporal analysis of NDFF records: generating dynamic distribution maps of flora and fauna species (T. Hengl, H. Sierdsema), from URL: https://www.researchgate.net/profile/Tomislav-Hengl/publication/264855500_Spatio-temporal_analysis_of_NDFF_records_generating_dynamic_distribution_maps_of_flora_and_fauna_species/links/53fb3eb90cf2e3cbf566169e/Spatio-temporal-analysis-of-NDFF-records-generating-dynamic-distribution-maps-of-flora-and-fauna-species.pdf

Ecology, from URL: Habitat suitability modelling and niche theory (Alexandre H. Hirzel and Gwenaëlle Le Lay, doi: 10.1111/j.1365-2664.2008.01524.x0) and Uses and misuses of bioclimatic envelope modeling (MIGUEL B. ARAÚ JO AND A. TOWNSEND PETERSON, 2012)

b. Predicting species presence/absence:

Predicting abundance, from URL: https://land.copernicus.eu/pan-european/corine-land-cover

Norberg A., N. Abrego, F. G. Blanchet, F. R. Adler, B. J. Anderson, J. Anttila, M. B. Araujo, T. Dallas, D. Dunson, J. Elith, S. D. Foster, R. Fox, J. Franklin, W. Godsoe, A. Guisan, B. O'Hara, N. A. Hill, R. D. Holt, F. K. C. Hui, M. Husby, J. A. Kalas, A. Lehikoinen, M. Luoto, H. K. Mod, G. Newell, I. Renner, T. Roslin, J. Soininen, W. Thuiller, J. Vanhatalo, D. Warton, M. White, N. E. Zimmermann, D. Gravel, and O. Ovaskainen. 2019. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. Ecological Monographs 89(3):e01370. 10.1002/ecm.1370. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels.

c. SDM abundance modeling

Conor Waldock, Rick D. Stuart-Smith, Camille Albouy, William W. L. Cheung, Graham J. Edgar, David Mouillot, Jerry Tjiputra and Loïc Pellissier. A quantitative review of abundance-based species distribution models.

d. Modelling the observer bias

Warton DI, Renner IW, Ramp D (2013) Model-Based Control of Observer Bias for the Analysis of Presence-Only Data in Ecology. PLoS ONE 8(11): e79168. doi:10.1371/journal.pone.0079168. Model-Based Control of Observer Bias for the Analysis of Presence-Only Data in Ecology.

7. Appendix

7.1. Appendix A: Remarks on data

Table 2: Data description regarding the input features in our dataset.

Features	Categories Merged	Data type	Data source
cbs_landuse	Yes	object	https://www.pdok.nl/introductie/-/article/cbs-bestand-bodemgebruik
NOK_begrenzing NOK_beheer NOK_planologische_ehs NOK_verwervinginrichting	No	int64	https://www.pdok.nl/geo-services/-/article/natuurmeting-op-kaart-nok-
ndvi_2022	No	float64	https://www.pdok.nl/geo-services/-/article/luchtfoto-pdok
clc2018	No	object	https://land.copernicus.eu/pan-european/corine-land-cover/clc2018?tab=download
fysisch_geografische_regios	No	object	https://www.pdok.nl/introductie/-/article/fysisch-geografische-regio-s
natura_2000	No	int64	https://www.pdok.nl/geo-services/-/article/natura-2000

nationale_parken	No	int64	https://www.pdok.nl/geo-services/-/article/nationale-parken
SGM_ondergrond	Yes	object	https://www.pdok.nl/introductie/-/article/bro-bodemkaart-sgm-
We used Bro to create the two variables: **genese** and **landforms** bro_genese bro_landform	bro_geomorfologischekaart_genese -> bro_genese bro_geomorfologischekaart_landfrom -> bro_landform	object	https://www.pdok.nl/-/wms-service-voor-bro-geomorfologische-kaart
BRP_gewas	Yes	object	https://www.pdok.nl/introductie/-/article/basisregistratie-gewaspercelen-brp-
AAN	No	int64	https://www.pdok.nl/introductie/-/article/agrarisch-areaal-nederland-aan
Stiltegebieden	No	int64	https://www.pdok.nl/geo-services/-/article/stiltegebieden
Wetlands	No	int64	https://www.pdok.nl/geo-services/-/article/wetlands

7.2. Appendix B: Annotated scripts of method settings

Table 3: Overview of model implementation settings for predicting butterfly presence/absence and corresponding important features selection.

Machine Learning Model	Used techniques
Random Forest Classifier	● Load necessary packages and dataset ● Check the descriptive statistics of data and its shape ● Extract categorical and numeric features from the data ● One-hot encoded the categorical columns

	● Combine encoded categorical and numeric columns ● Convert non-zero values in the target column to 1 and store it in a new column ● Filter dataset based on an equal number of zero and non-zero values of the above column ● Extract the independent variables (X) and the dependent variable(Y) ● Split dataset into training set and test set ● Apply non-negative transformation using MinMaxScaler ● Create a pipeline with preprocessing, feature selection, and model ● Define the hyperparameters and their possible values ● Perform Grid Search with k-fold cross-validation ● Create a Random Forest Classifier with the tuned parameters ● Fit the model to the data ● Predict the response for the test dataset ● Evaluate the model, and check its accuracy, precision, and recall ● Visualize the model predictions using contextily, GeoDataFrame, and its CRS (epsg:28992)
Decision Tree Classifier	● Load necessary packages and dataset ● Check the descriptive statistics of data and its shape ● Extract categorical and numeric features from the data ● One-hot encoded the categorical columns ● Combine encoded categorical and numeric columns ● Convert non-zero values in the target column to 1 and store it in a new column ● Filter dataset based on an equal number of zero and non-zero values of the above column ● Extract the independent variables (X) and the dependent variable(Y) ● Split dataset into training set and test set ● Create a Decision Tree model object ● Train Decision Tree model using training data ● Predict the response for the test dataset ● Evaluated the model, and checked its accuracy, precision, and recall

	● Visualized the tree using the decision tree model
Linear Regression (Used for predicting butterfly occurrence count)	● Load necessary packages and dataset ● Check the descriptive statistics of data and its shape ● Extract categorical and numeric features from the data ● One-hot encoded the categorical columns ● Combine encoded categorical and numeric columns ● Convert non-zero values in the target column to 1 and store it in a new column ● Filter dataset based on an equal number of zero and non-zero values of the above column ● Extract the independent variables (X) and the dependent variable(Y) ● Perform linear regression on the extracted dataset ● Determined the coefficient of determination, R^2 ● Predicted the response
Extra Trees Classifier, KNN, Naïve Bayes, Logistic Regression, SVM	• Load necessary packages and dataset • Split dataset into training set and test set • Create a list of various classification models • Iterate through list of models to train them with training data • Test these models with test data to check the results • Calculate and compare performance metrics (accuracy, precision, recall, F1-score, and AUC-ROC) of all the models • Plot various graphs to compare the performance of these models
Features Selection	• Get important features from the model (Random Forest) using Recursive feature elimination (RFE) technique • Create a list of these feature names • Sort the features in descending order based on their importance • Visualize the top 20 important features

7.3. Appendix C: Full method exploration results

Figure 5 shows that the Random Forest model stands out based on performance metrics and gives good results.

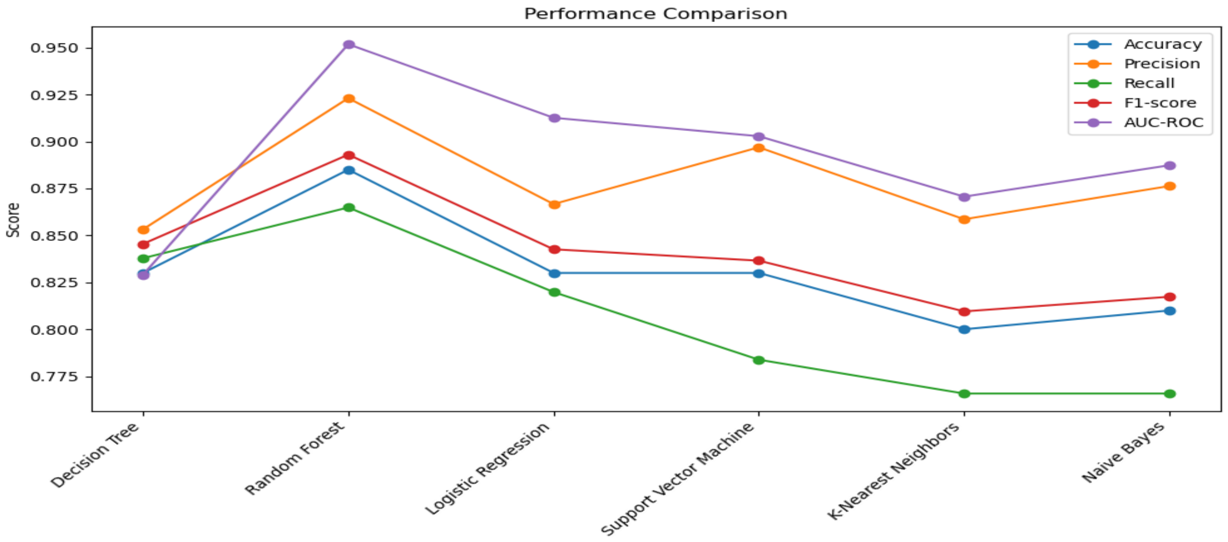

Figure 5: Parallel Coordinates plot showing different performance metrics for each model.

Table 4: A summary of performance metrics for implemented models.

Model	Performance
Random Forest Classifier	• RF accuracy: 0.91 • RF precision: 0.89 • RF recall: 0.93 • RF F1-score: 0.91 • RF AUC-ROC: 0.97 • Hyperparameter values: {'max_depth': 10, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 300}
Extra Tree Classifier	• Extra Tree accuracy: 0.90 • Extra Tree precision: 0.88 • Extra Tree recall: 0.94 • Extra Tree F1-score: 0.91 • Extra Tree AUC-ROC: 0.96
Logistic Regression	• LR accuracy: 0.90 • LR precision: 0.88 • LR recall: 0.93 • LR F1-score: 0.91 • LR AUC-ROC: 0.95
K-Nearest Neighbors	• K-NN accuracy: 0.90 • K-NN precision: 0.88 • K-NN recall: 0.93 • K-NN F1-score: 0.90

	- K-NN AUC-ROC: 0.94
Decision Tree Classifier	- Decision Tree accuracy: 0.90 - Decision Tree precision: 0.89 - Decision Tree recall: 0.92 - Decision Tree F1-score: 0.90 - Decision Tree AUC-ROC: 0.90
Naive Bayes	- Naive Bayes accuracy: 0.76 - Naive Bayes precision: 0.68 - Naive Bayes recall: 1.00 - Naive Bayes F1-score: 0.81 - Naive Bayes AUC-ROC: 0.76
Linear Regression (Used for predicting butterfly occurrence count)	- Coefficient of determination (R^2 value): 0.14

7.4. Appendix D: Descriptive Analysis

From Figure 6, we found that the total number of butterflies was relatively high in 2017 as compared to other years.

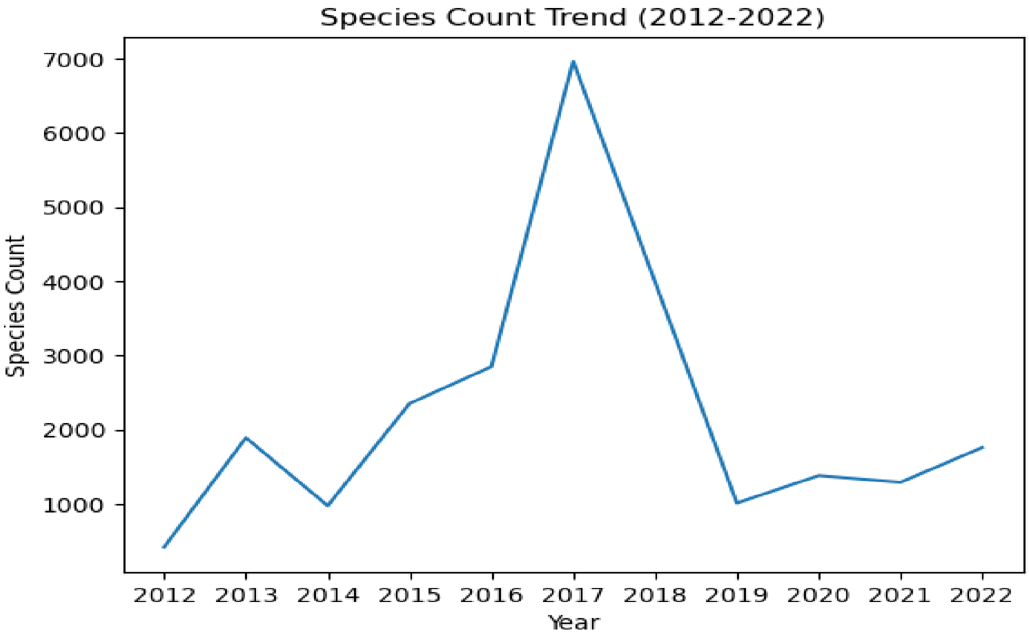

Figure 6: Lineplot of the trend of butterfly occurrences per year.

The following Figure 7, shows the presence of butterflies in various regions for each year.

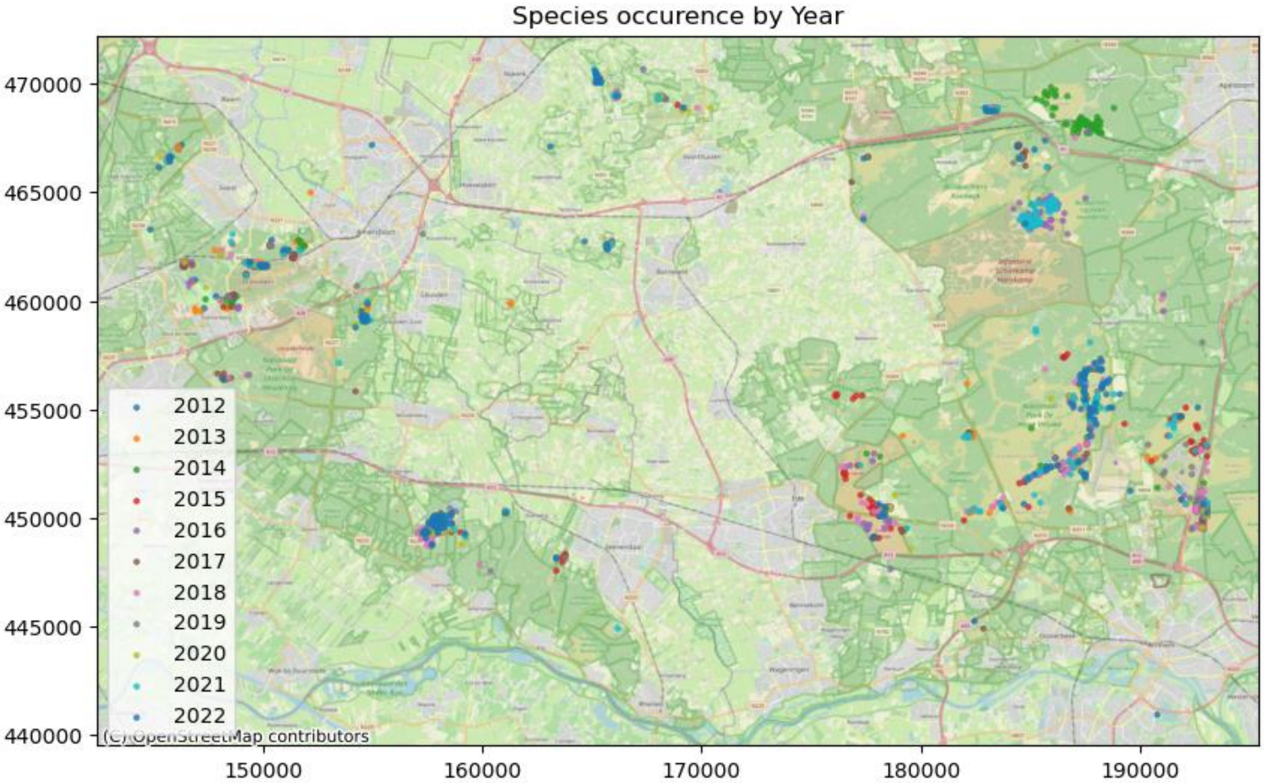

Figure 7: Visualization of butterflies' presence year wise.

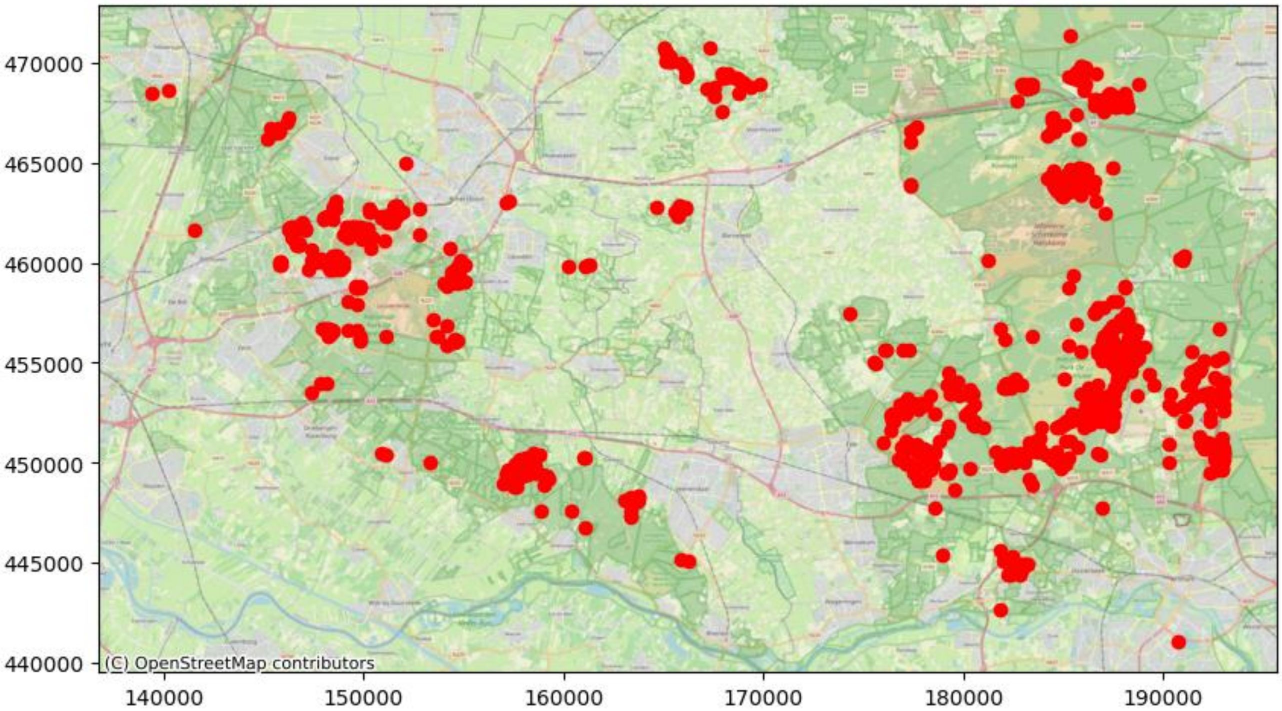

Figure 8: Visualization of butterflies' presence based on total ndff count.

The following heatmap provides a visual representation of the pairwise correlation between the features, where higher correlation values are indicated by brighter or darker colors. Positive correlations are shown in one color gradient (e.g., red), while negative correlations are shown in the opposite gradient (e.g., blue). The values in the heatmap can help identify relationships and patterns among the input features in the dataset.

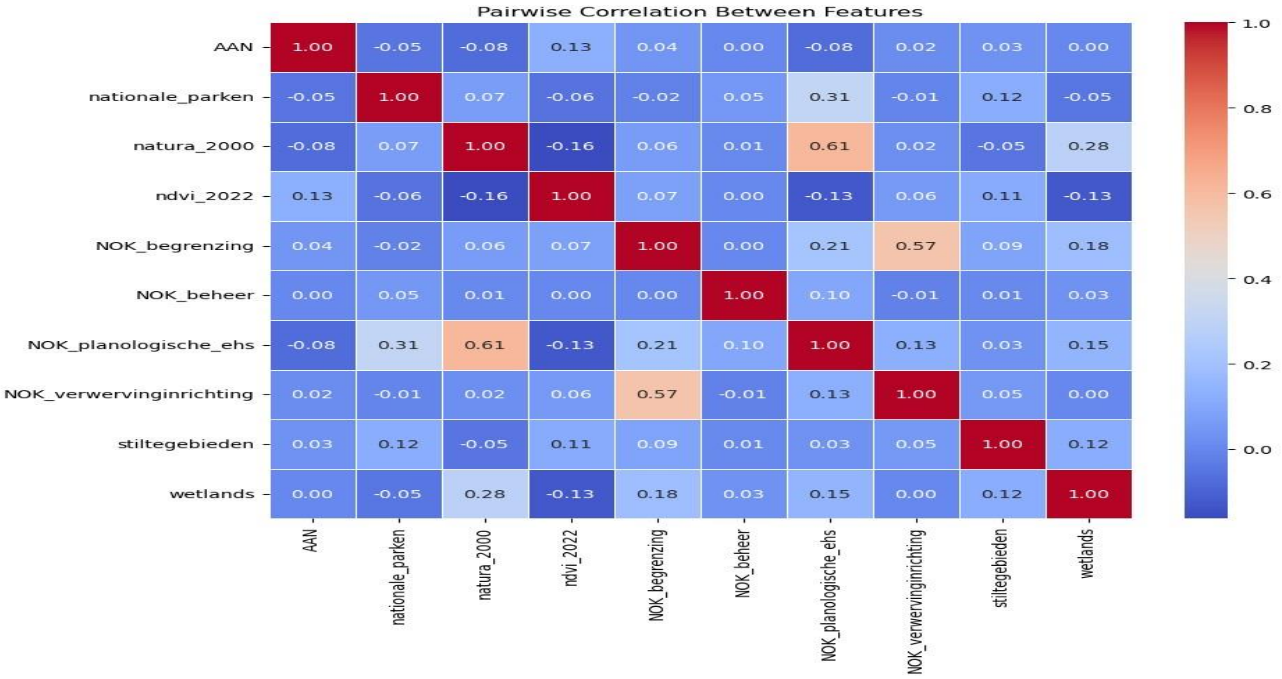

Figure 9: Heatmap of the pairwise correlation between input features.

In Figure 10, classification tree shows the decision flow based on important features.

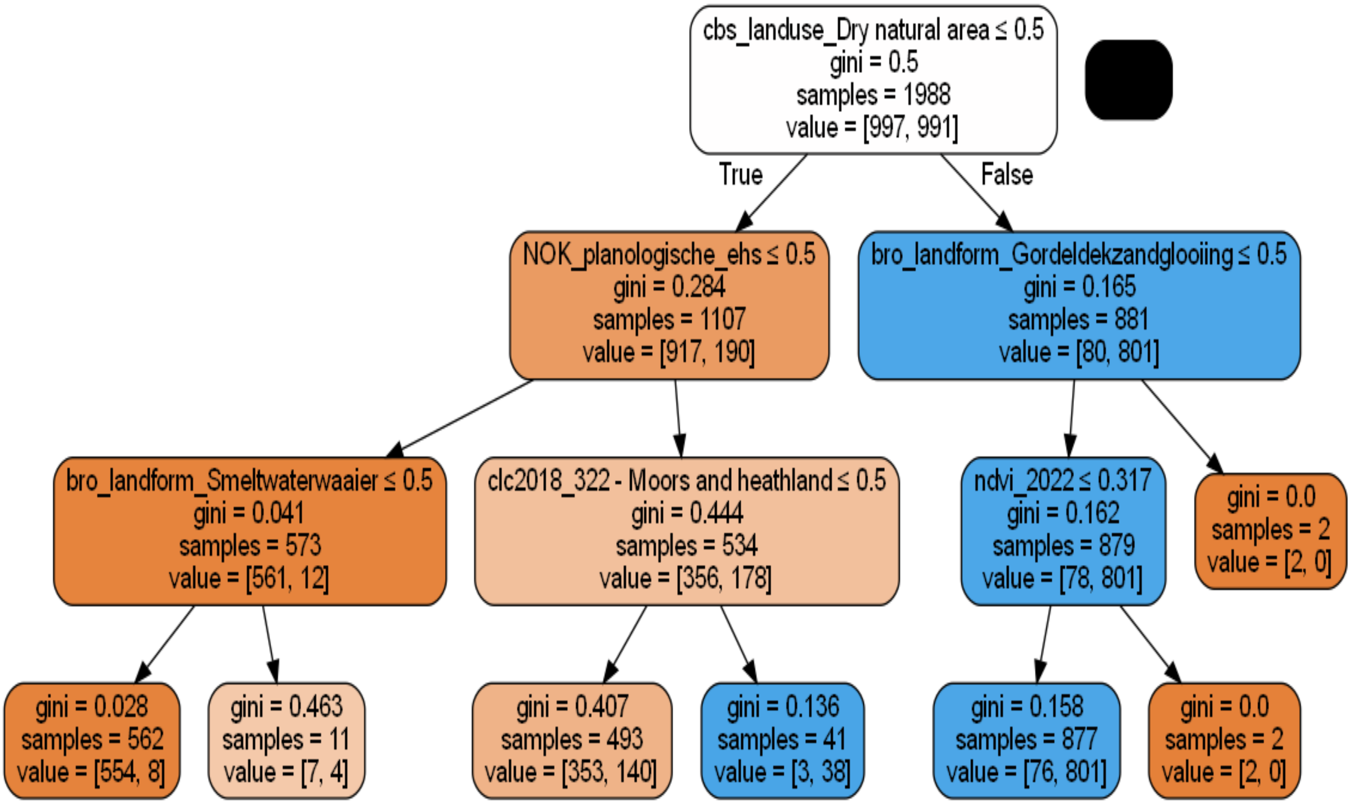

Figure 10: Decision Tree Model up to depth 3 for better readability.