PREDICTING THE PRESENCE OF GREY WOLVES IN NATURA 2000 SITES

A MACHINE LEARNING APPROACH

DAVID BLEIJLEVENS

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
BACHELOR OF SCIENCE IN COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE

DEPARTMENT OF
COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

STUDENT NUMBER

2024417

COMMITTEE

Dr. Eriko Fukuda
Dr. Travis Wiltshire

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 24, 2022

ACKNOWLEDGMENTS

I would hereby like to thank my supervisor, Dr. Eriko Fukuda, who has granted me the freedom to write a thesis about a topic that is of great interest to me. Besides that, she has given me valuable advice and made sure that thesis meetings were both informative and enjoyable.

PREDICTING THE PRESENCE OF GREY WOLVES IN NATURA 2000 SITES

A MACHINE LEARNING APPROACH

DAVID BLEIJLEVENS

Abstract

In recent years researchers have started to implement various machine learning approaches to achieve models that can predict species distributions with greater accuracy. To facilitate the use of machine learning techniques in the field of ecology; Beery, Cole, Parker, Perona, and Winner (2021) plea for the involvement of computer scientists in future research. In this study, an effort was made to investigate the extent to which different machine learning algorithms can predict the existence of wolves in Natura 2000 sites; natural areas that are part of a European Union-wide conservation strategy. Instead of collecting and combining multiple sources of data, this study made use of a pre-existing dataset, constituting information on Natura 2000 sites as reported by all Member States of the European Union. To achieve this, two previously established (*Random Forests & Support Vector Machines*), and one novel (*eXtreme Gradient Boosting*) algorithm were trained to correctly classify sites based on the presence or absence of wolves. To account for class imbalance, a combination of oversampling and undersampling methods was applied to the data. The results showed that eXtreme Gradient Boosting and Random Forest classifiers were able to make predictions at a low error rate, returning Matthews Correlation Coefficients (MCC) of .78 and .77 respectively. Besides, these models showed little to no change in performances after the class balancing methods were applied (+ 0.78% and - 1.54% on the MCC respectively).

1 INTRODUCTION

According to the Living Planet Index (WWF, 2020), global wildlife populations have decreased by 68% over the last 50 years. While Europe's biodiversity has seen the least decline, caution is still needed. Reintroduction projects have started for those species that were on the brink of extinction, such as the European bison (Lord, Wirebach, Tompkins, Bradshaw-Wilson,

& Shaffer, 2020). Other species such as the grey wolf have been able to expand their range naturally, thanks to European legislation for species protection and improved environmental laws.

At the centre of the European Union's (EU) biodiversity strategy for 2030 lie expansion plans for the Natura 2000 network (European Commission, 2020). The main objectives of this network of protected natural sites are an increase in biodiversity and the conservation of vulnerable protected species. The 27,031 Natura 2000 sites currently cover 18% of the EU's landmass and 9% of its marine territory (European Commission, n.d.). In 2015 the grey wolf, scientifically known as *Canis lupus*, made its reappearance in the Netherlands after a 150 year absence. Since then the wolf has officially settled, reproduced and spread to multiple locations in the Netherlands. The return of the largest living canine species has sparked a fierce public debate about its suitability in a country that is as densely populated as the Netherlands.

Ecologists and conservationists are delighted by the presence of wolves in Dutch nature, as the species plays an important role in the health of ecosystems (Linnell et al., 2005; Ritchie et al., 2012). On the other hand, those who oppose the wolf's westbound territorial expansion fear that the shortage of sizeable natural sites in the Netherlands will inevitably lead to the animal wandering into urbanized areas and threaten livestock. Hence, observing species distributions and understanding habitat preferences are vital not only for biodiversity protection, but also to ensure minimal conflict between protected species and humans.

Wildlife monitoring is challenging, especially for rare species. Current methods often include the placement of camera traps to observe wildlife or the use of GPS collars to track movements. These methods are not only expensive in resources and time, but can also interfere with wildlife (Berger-Wolf et al., 2017). Data collection has become a standard practice in most aspects of the world, and the field of ecology is no different. As national and international (governmental) organizations continue to collect data on natural sites, there is an ever-growing database that can be used for predictive modelling (Tuia et al., 2022). Making use of this readily available data to predict wildlife occurrences in new places is inexpensive and quick in comparison to traditional methods.

Ecological systems are difficult to model, due to the complex nature of interactions that take place in such a system (Beery et al., 2021). While a range of machine learning techniques have risen in popularity for this reason, collecting and combining data to account for all possible interacting variables remains challenging. Therefore, this study set out to predict the presence of grey wolves in a Natura 2000 site, based on the yearly-reported descriptive features of that specific natural area. As each Natura

2000 site has a sizeable amount of features, ranging from habitat type to possible pollutants to endangered species present on a Natura 2000 site, there lie possibilities to establish relationships between such variables and the presence of protected species such as the wolf. If accurate, these predictions could help conservation efforts and serve as an inexpensive and convenient tool to check the suitability of a natural site to host certain threatened animal species. Linear regression, Support Vector Machines, Random Forests, and eXtreme Gradient Boosting classifiers have been trained, tested, and compared on their performance. This resulted in the following research questions:

RQ1 *To what extent can a machine learning algorithm predict the existence of grey wolves in a Natura 2000 site, based on site-specific features?*

RQ2 *Which algorithm can best predict the existence of grey wolves in a Natura 2000 site?*

RQ3 *Which features best predict the existence of grey wolves in a Natura 2000 site?*

2 RELATED WORK

This section briefly reviews previous research on the topic of predicting the presence of animal species, and the associated use of different machine learning techniques. In addition, the use of such methods in the context of wolves is explored.

Species distribution models

The methods that combine data of species occurrences and environmental factors to predict spatial distributions are called species distribution models (SDMs). SDMs can be used to determine the places where an animal species currently exists (Jiménez-Valverde, Lobo, & Hortal, 2008) or where the conditions are right for potential existence (Soberón, 2010). SDMs can work as a presence-only model where only the presence of certain species are recorded, or as a presence-absence model, where the absence of said species in an area is also recorded (Beery et al., 2021). The latter are highly dependent on the quality of the available data and have to deal with some degree of uncertainty, as it is difficult to accurately claim the total absence of a species (Rocchini et al., 2011).

SDMs can be used for explanatory purposes such as finding drivers for species distribution, and for predictive purposes such as predicting species presence in a 'new' natural site (Elith & Leathwick, 2009). Historically SDMs are based on statistical methods such as generalized linear models

and logistic regression, where predictors are selected based on their perceived importance (Elith & Leathwick, 2009). While this methodology has served its purpose for many years, it is prone to biased predictor variables due to researchers over- or underestimating the importance of a variable or certain relationships (Rocchini et al., 2011), which are difficult to account for in complex ecological systems. To deal with these drawbacks, a number of machine learning algorithms have been introduced to the domain over the last two decades.

SDMs & machine learning techniques

Machine learning approaches are well suited to account for the complex relationship between species and environmental factors (Beery et al., 2021). While a multitude of models has been used in recent years, only some seem to be able to consistently outperform generalized linear models. One group of well-performing algorithms are support vector machines (SVM). Drake, Randin, and Guisan (2006) were one of the first to propose their use for SDMs when they found that they outperformed generalized linear models in predicting the presence of plant species in the Alps. Since then, SVM have regularly been successfully applied in SDMs (Pouteau, Meyer, & Stoll, 2011; Pouteau, Meyer, Taputuarai, & Stoll, 2012; Sadeghi, Zarkami, Sabetraftar, & Van Damme, 2012).

Another highly successful group of algorithms that are frequently used as SDMs are tree-based models. Their ability to handle different data types, deal with multicollinearity, and model complex nonlinear relations are some of reasons why they lend themselves well for ecological modelling. While there is a large variety of tree-based models, ensemble methods such as Random Forests (RF) seem to be the favourite among ecologists as their built-in bagging method results in models with stronger predictive performances (Elith, Leathwick, & Hastie, 2008). Fukuda, De Baets, Waegeman, Verwaeren, and Mouton (2013) conclude that both SVM and RF show superior performances compared to generalized linear models and other statistical methods, with RF performing marginally better than SVM. Furthermore, Sabat-Tomala, Raczko, and Zagajewski (2020) found that RF outperformed SVM when there is high homogeneity between classes, while SVM performed best on data where there is little uniformity between individual cases in a class. Though the majority of SDMs are still not based on RF, the ensemble model's popularity among ecologists is on the rise as multiple researchers advocate for making RF the gold standard among SDMs (Mi, Huettmann, Guo, Han, & Wen, 2017).

An ensemble algorithm that is similar to RF and rapidly gaining traction in the non-scientific community is *eXtreme Gradient Boosting* (XGBoost). Although there has been no implementation of XGBoost in SDM, other

scientific domains such as economics (Chang, Chang, & Wu, 2018) and medicine (Shi et al., 2019) have seen encouraging results with the use of XGBoost as a predictive model. This novel algorithm often performs better than more well-known ensemble models, opening the door to a successful implementation in SDM.

Grey wolf
While there is a lack of previous work on presence-absence SDMs for wolves in Europe, there are some studies that touch upon the subject. Linnell et al. (2005) claim that large predators such as the wolf are able to cope with changes in both habitat quality and the use of land in and around protected natural sites. This is reiterated by Cimatti et al. (2021) who suggest that - like brown bears and lynxes - wolves are highly adaptive in their habitat selection. On the other hand, Morales-González, Fernández-Gil, Quevedo, and Revilla (2022) found that individual wolves looking for new territories (i.e. dispersal) seem to avoid areas with high agricultural activity. Moreover, dispersal studies suggest that wolves prefer habitats that are similar to those where the individual previously resided (Sanz-Pérez et al., 2018).

3 METHOD

The following section outlines the methodological procedure of the study. Starting with a brief description of the dataset and explaining the preprocessing steps that were taken. Subsequently, the ML models are clarified and lastly, argumentation for the chosen evaluation metrics is given.

3.1 *Dataset*

This research made use of the Natura 2000 data — the European network of protected sites dataset, which was retrieved from the European Environment Agency (2022). The dataset is a compilation of the national databases that are submitted by the relevant authorities of each EU Member State for the year 2021. Each Member State submits their data yearly through a standardized data form. The dataset consists of 11 separate CSV files. The majority of files contain descriptive features of a Natura 2000 site itself, though the files SPECIES and OTHERSPECIES provide information on the protected species that are present on a site. SPECIES lists the species for which the corresponding Natura 2000 site is specifically designated, while OTHERSPECIES lists other protected or remarkable species that occur on a site. The files DESIGNATIONSTATUS, DIRECTIVESPECIES, MANAGEMENT and METADATA contain no information that was useful

for this project and were thus discarded. The remaining files consist of multiple descriptive features such as habitat type, biogeographic region, area size, geographical coordinates, and possible human-caused threats and pressures.

3.2 *Data cleaning*

The initial preprocessing steps have been conducted in R (R Core Team, 2022); version 4.1.3 by using the dplyr (Wickham, François, Henry, & Müller, 2022), SuperML (Saraswat, 2022), stringr (Wickham, 2022), & MICE (van Buuren & Groothuis-Oudshoorn, 2011) packages. In these first stages of data preparation, the aforementioned CSV files needed to be cleaned and combined to one cohesive dataset. In doing so, certain columns containing irrelevant information were removed.

As the BIOREGION and HABITATCLASS files contained multiple observations of the same unique sitecode – all the bioregions and habitat classes present in a single Natura 2000 site – only the observations with the highest percentage of coverage for a certain site were kept so that each unique site is referred to by their dominant bioregion and habitat class. This was done to avoid the negative impact including a large number of features can have on model performances (i.e. curse of dimensionality). Moreover, training on a limited number of features greatly reduces the runtime of ML algorithms.

For the IMPACT file, the information was again too fine-grained. To counter this, the impact codes were grouped according to their 'parent code' (e.g. *A*04 becomes *A*, where *A* stands for agriculture), as defined by Salafsky et al. (2008). This reduced the number of categories from 506 to 13. These were subsequently one-hot encoded to display the presence or absence of certain threats and pressures in or around a Natura 2000 site. To prevent the loss of possibly valuable information, the missing values in this file were mode-imputed.

After merging these files based on the key that is present in all files (i.e. the unique site code corresponding with a Natura 2000 site), a dummy encoded column to account for the presence / absence of wolves in a site was added. This is subsequently a means of ground truth labeling. As the SPECIES and OTHERSPECIES files contain information on the presence of all protected or noteworthy animal and plant species in a site, one-hot encoding all would not suffice (i.e. curse of dimensionality). Only the animal species that are reasonably expected to influence wolf presence were included in the final dataset. These species include the prey species that make up the diet of wolves in Europe (Newsome et al., 2016; Sin, Gazzola, Chiriac, & Rîșnoveanu, 2019), species known to interact with wolves (Gable,

Windels, Romanski, & Rosell, 2018), and the competing large predators that roam in parts of Europe; brown bear and lynx (concatenation of European and Iberian subspecies). As a means to limit information loss, a new column containing the number of distinct mammal species present in a site was constructed. The variables present in the final dataset are listed in Appendix A (page 27).

After merging the cleaned files into one final dataset, all sites that consist of > 90% marine territory were removed as wolves need land to live on. The remaining sites contained a total of 3391 missing values (distributed over the *AREAHA, LONGITUDE,* and *LATITUDE* variables), of which 161 occurred in sites with wolf presence. To avoid losing sites from the minority class the 161 missing values were imputed by predictive mean matching (PMM), making use of the *Multivariate Imputation by Chained Equations* (MICE) implementation by van Buuren and Groothuis-Oudshoorn (2011). The advantage of using PMM as a multiple imputation method is its robustness in dealing with non-linear data with many outliers (Kleinke, 2018). Moreover, compared to other imputation methods PMM introduces little bias in a model, especially when there are many features present to fit the prediction to (Landerman, Land, & Pieper, 1997; van der Palm, van der Ark, & Vermunt, 2016)

The sites that are located in countries with no wolf-presence whatsoever were removed from the dataset. As these sites can only default to 0 they offer little predictive value and can be seen as noise in the data that would limit model performances (e.g. a certain habitat type happens to occur often in wolf-less countries, leading to a model that unjustifiably sees that habitat type as a strong predictor for wolf presence). Moreover, the COUNTRY_CODE variable was removed as it could lead to biased models, where assumptions are made based on information that cannot be classed as a true feature of a natural site.

3.3 *Preprocessing*

Python (version 3.7.3) was used to build the ML models and evaluate their performance (packages/modules: Pandas, NumPy, imbalanced-learn, scikit-learn, Matplotlib, XGBoost, seaborn). After performing a 75:25 train-test split on the data, some further preprocessing steps were taken. These were deliberately implemented after the train-test split, as some of the chosen preprocessing methods could have lead to data leaking from the test set to the model. These methods were fitted solely to the training set, and thereafter both the training and test set were transformed. As the categorical variable SITETYPE consists of only three levels, dummy-encoding was possible without extending the dataset much further. In contrast,

BIOGEOGRAPHICREG and HABITATCODE are categorical variables with high cardinality, meaning that a different encoder was required. A target encoder (McGinnis, 2016) was used to encode the features to values in the 0-1 range. This is achieved by replacing the categorical values with the probability that a given category occurs for the target label (Micci-Barreca, 2001).

For the AREAHA variable a log transformation was applied to reduce the skewness. While log transform and transformations in general can lead to a degradation in model performance, and problems in interpreting feature importances (Feng, Wang, Lu, & Tu, 2013), it performed well on this particular data in practice.

Thereafter the continuous variables in the dataset were scaled to ensure all values would lie in the 0-1 range, as a large variability in range can lead to overfitting (Belkin, Hsu, Ma, & Mandal, 2019). This was achieved through an implementation of the MinMaxScaler, since this scaling method performs well for all chosen ML algorithms, as reported by Ahsan, Mahmud, Saha, Gupta, and Siddique (2021).

Subsequently, the models were trained on a semi-balanced dataset. This was done via a combination of oversampling and undersampling techniques, as this leads to less information loss than the use of a singular over- or undersampling technique (Batista, Prati, & Monard, 2004). *Synthetic Minority Oversampling TEchnique* (SMOTE) was used as it generally performs better than other oversampling techniques (Batista et al., 2004). Contrary to other oversampling methods that simply create duplicates of the minority class, SMOTE creates new minority-class data points based on characteristics of its K-nearest neighbors, thus creating plausible 'new' data (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). Here, K was set to 5 as proposed by Elreedy and Atiya (2019). Complementary to SMOTE, random undersampling was applied; resulting in the removal of a specified amount of randomly picked samples from the majority class.

After application of both methods, the class imbalance was reduced from 92:8 to 75:25, as seen in Table 1. While truly balanced classes would be preferable, it would come at the expense of either losing many informing instances from the majority class or generating a large number of synthetic minority class instances. Both approaches would be at risk of training models that do not translate well to authentic data (Sáez, Krawczyk, & Woźniak, 2016). Following the same reasoning, the test data has remained unaltered as the goal is to test model performances on data of existing Natura 2000 sites.

Table 1: Target label distribution for the training data

	n Positive class	n Negative class
Unbalanced	1,155	11,922
Balanced	2,384	7,224

3.4 *Models*

The following section outlines the four different classification algorithms used in this study and will give a general description of their workings. The algorithms were selected based on their performances in previous research, or their suitability and interpretability. A dummy classifier (often referred to as no-skill classifier) was chosen as the baseline model. Dummy classifiers consistently opt for the majority class without taking any other information into account, thus returning a score that is equal to the score one would achieve by guessing. This gives a fair impression on the performance of other models, as a classifier with true predictive powers should comfortably beat a model that blindly predicts the majority class.

Logistic Regression

Logistic regression classifiers are simple, robust and computationally efficient models. A logistic regression calculates the probability a certain feature belongs to the target class. By summing the probabilities for all features in the dataset, a decision boundary is defined. For each data point the model returns an output label (i.e. target class prediction), depending on which side of the decision boundary the data point falls.

While logistic regression has been a mainstay model since it was first proposed by Cox (1958), there are some drawbacks to take into consideration. It assumes a certain degree of linearity and has difficulty in dealing with outliers (Healy, 2006). As generalized linear models such as logistic regression classifiers have long been the standard in SDM, it is interesting to compare its performance to those of more novel approaches. As such, the logistic regression classifier will act as a baseline, complementary to the dummy classifier.

Support Vector Machines

SVM classifiers can model both linear and complex non-linear relationships. Contrary to generalized linear models such as the logistic regression, it does not fit all data points to create a decision boundary. Instead, SVM focus solely on the data points that can explain the largest proportion of variance. At the basis of SVM lies a decision function $f(x)$, where x is

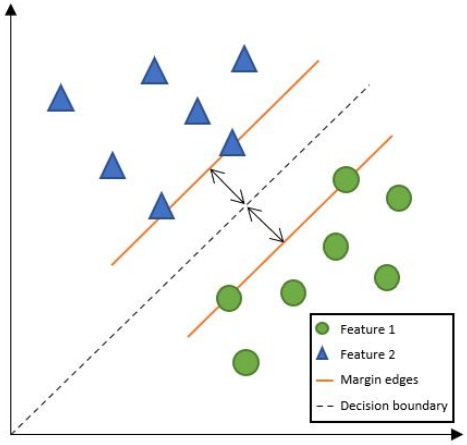

Figure 1: Linear SVM example

some data point. This decision function calculates the probability that x belongs to the positive (1) or negative (-1) class (Hearst, Dumais, Osuna, Platt, & Scholkopf, 1998).

The algorithm then tries to determine and maximize a buffer that segregates the two classes (i.e. the margin), with the decision boundary lying in the centre where ($f = 0$). The subset of data points (which are represented as vectors) that are used to determine the decision boundary are dubbed support vectors. In tandem, these support vectors are positioned on the plane such that $f = \pm 1$. All data points whose position lie at a greater distance are classified as a positive or negative class, while points that fall in the margin are treated as outliers (Hearst et al., 1998). This results in less data points 'crossing' the decision boundary and thus less misclassifications. For each feature in the data, support vectors are created and subsequently transformed to a N-dimensional space, where N is the number of features in the data. A simplified example of a linear SVM classifier is shown in Figure 1; where the dotted line represents the decision boundary, and the parallel lines indicate the edges of the margin. The points that fall on the coloured lines are the support vectors.

For linear SVM the decision boundary is represented as a straight line, but seeing as SVM can also be used for non-linear classifications the decision boundaries can be fitted as curved lines. The mathematical approach used to transform the support vectors to a N-dimensional representation and consequently determine the visual representation of the corresponding decision boundary is called the kernel function (Patle & Chouhan, 2013). This can either be a linear-, radial bias-, sigmoid-, or polynomial function. A grid search (as detailed in section 3.5) confirmed radial bias function (RBF) to be the most appropriate kernel.

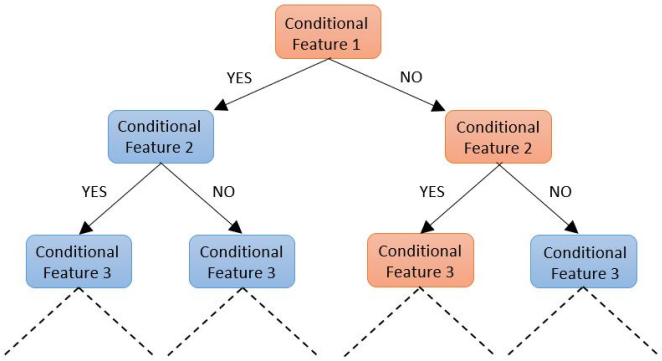

Figure 2: Decision tree example

Random Forests

RF are ensemble models: models that implement methods such that many weak learners unite to form a single strong learner. One such ensemble method is *bagging*, a contraction of the words bootstrap and aggregating (Bühlmann, 2012). A bootstrap sample is a subset of the original data, so to arrive from many weak learners to a single strong learner, bagging methods average the outcomes of multiple bootstraps.

In RF, the weak learners are decision trees: straightforward algorithms that evaluate a number of binary conditions to arrive at some conclusion. This is accomplished by randomly selecting features from the data, and testing whether the chosen instance meets the binary condition for said feature. Subsequently the root splits in two nodes, those nodes split and form new nodes, and so on. Figure 2 shows a simplified representation of a decision tree, where the orange path symbolizes the conditional checks that lead to a correct classification.

One major advantage of decision trees is their interpretability, as it is easy to understand why a decision tree returns a certain output. Conversely, a decision tree's simplicity is what makes them weak learners. Since the decision tree uses the same conditional checks on unseen data as were used during training, the accuracy score drops when this unseen data differs slightly from the training data. This proneness to overfitting can be avoided by combining the outcomes of many trees, each of which have been trained on different bootstraps (Oshiro, Perez, & Baranauskas, 2012). Consequently, this results in a strong learner that is robust enough to perform well on new data. While the risk of overfitting is already reduced through RF's built-in approach to bagging, it might be necessary to prune the trees to further improve the accuracy on the test set. Pruning means limiting the depth (and likewise the complexity) of the decision trees in the model, which can be achieved by tuning various hyperparameters.

eXtreme Gradient Boosting

XGBoost is an ensemble model founded on decision trees, similar to RF. However, they differ in the ensemble method that is used. Gradient boosted trees are highly effective models that, as the name suggests, incorporate a boosting method in their algorithm. Boosting lets a model run repeatedly on the full training data, where the training data is reweighted (hence *gradient*) after each run. The model focuses on the instances that were the most difficult to classify correctly, leading to a model that delivers even greater accuracy due to the shrinking misclassification rate (Schapire & Freund, 2013). Gradient boosted trees in itself are no novelty, but XGBoost differentiates in two ways primarily: only a portion of columns (i.e. the features) in the data are used per tree, and the algorithm allows for parallelization (Chen & Guestrin, 2016). This results in an accurate model that can handle large amounts of data at a relatively low time complexity O.

3.5 *Model Optimization*

To reduce the chance of overfitting the models, stratified 5-fold cross-validation was included in a grid search for hyperparameter tuning. The extensive grids included a wide range of parameters and corresponding values to find the optimal hyperparameter configurations for each model, yet remained in ranges that were within reason to limit computational complexity. This procedure was applied in the same manner to all four models, with the addition of a preliminary randomized grid search for the RF and XGBoost models. Ensemble models have a wide range of hyperparameters that can be tuned, resulting in an exhaustive grid search with enormous computational complexity (22,000 possible hyperparameter combinations).

Hence, a randomized search running 100 random hyperparameter combinations was implemented to establish a suitable but limited grid for the subsequent grid search. A complete overview of the grids and selected hyperparameters can be found in Appendix B (page 28).

3.6 *Evaluation Metrics*

To answer research questions 1 and 2, the models' performances had to be evaluated in ways that were most reflective of their true predictive capabilities. By default, classification models are evaluated on their accuracy to predict the correct target value. However, when evaluating model performance on imbalance data, accuracy scores can be misleading. Dummy classifiers would return 90 percent accuracy on $9:1$ imbalanced data,

without making any useful prediction whatsoever. Obtaining meaningful insights into model performance - irrespective of class imbalance - can be done by using a confusion matrix. A confusion matrix displays the number of correct predictions for each of the target labels as *true positive (TP)*, *false positive (FP) true negative (TN)*, and *false negative (FN)*. Where *true* means a correct prediction, and *positive* or *negative* corresponds with the target label (i.e. positive for sites that are ground truth labeled as having wolf presence in this study).

During cross validation and hyperparameter tuning, the scoring metric used to find the best performing model was the AUC score, which is the area that falls under the *receiver operating characteristic* (ROC) curve. ROC plots the true positive rate (i.e. *sensitivity* (1)) against the false positive rate for a range of decision thresholds. The false positive rate equals $1-$ the true negative rate (i.e. *specificity* (2)). The AUC score is the mean of all values on the ROC curve, meaning that an AUC score of 1.0 would indicate that the model is perfectly sensitive and specific (Fan, Upadhye, & Worster, 2006).

$$Sensitivity = \frac{TP}{TP + FN} \tag{1}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2}$$

To compare the different models the ROC curves were plotted. Additionally, the *Matthews Correlation Coefficient* (Equation 3) was used as a scoring metric, which Chicco and Jurman (2020) proposed as the most truthful scoring metric for unbalanced binary classification problems.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \tag{3}$$

RF and XGBoost have an integrated attribute that enables the assessment of the contribution each feature had in the construction of a decision tree. The built-in function computes the Mean Decrease in Impurity (MDI) over all trees in the model, which can be understood as the increase in a model's accuracy that is caused by some feature in the data. This comprehensible, albeit somewhat limited, method was used to answer RQ3.

4 RESULTS

In this section, classification performances of the different models will be presented. Subsequently, the results after the application of data balancing

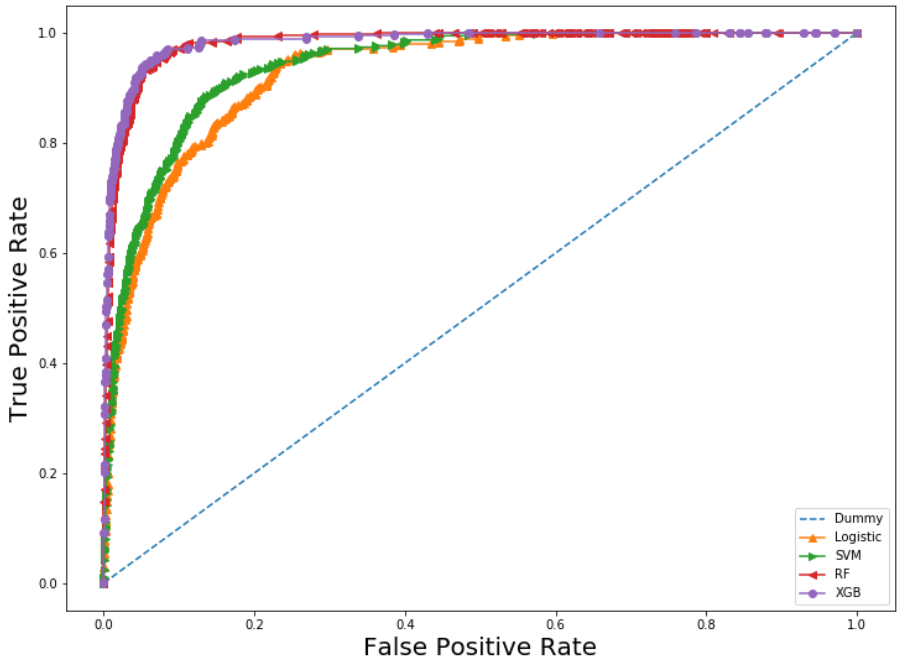

Figure 3: ROC curves on the original data.

methods will be put forward and compared to the results of the original data. Finally, the feature importances that were derived from the tree-based models will be presented.

Model evaluation

As specified in section 3.6, both a dummy classifier and a logistic regression classifier were used as baseline models since they serve different purposes. As seen in the ROC curves in Figure 3, all models were able to return higher AUC scores than a dummy classifier. Irrespective of class distribution, a dummy classifier's AUC score defaults to .50, as its $sensitivity = 0$ and $specificity = 1$. XGBoost and RF were the best performing models, and while their ROC curves are highly similar, XGBoost seems to perform slightly better. Besides, all classifiers show a better performance than the logistic regression, though the SVM is briefly surpassed by the baseline classifier at the .25 decision threshold.

Table 2 shows the results of the confusion matrices. This further clarifies classification performances, as they can be reviewed per target class. While overall the tree-based models performed best, they score lower than the SVM classifier in predicting the positive class. When reviewing the positive class results in isolation, RF is the only classifier that was not able to beat the logistic regression classifier. In contrast, RF was best able to predict the

Table 2: Confusion matrices for the original data.

		True	False
Dummy	Positive	0	0
	Negative	3,974	385
Logistic Regression	Positive	319	575
	Negative	3,399	66
SVM	Positive	337	520
	Negative	3,454	48
RF	Positive	306	85
	Negative	3,889	79
XGBoost	Positive	320	95
	Negative	3,879	65

true negative instances, closely followed by the XGBoost classifier. SVM and logistic regression classifiers were not able to correctly predict the negative class at a similar rate to the tree-based models, which explains the better overall performances of the tree-based models. Moreover, SVM did not beat the baseline classifier in predicting the negative class.

By default $MCC = 0$ for a dummy classifier, hence looking solely at the baseline score of the logistic regression classifier allows for a more intuitive and meaningful interpretation. Herein, the reported differences between the model performances become clearer, as the tree-based classifiers score significantly higher than the SVM classifier. All models returned higher scores than the baseline on *MCC* and specificity, while RF is the only classifier to record a lower sensitivity than the baseline classifier (.795 for RF, and .829 for logistic regression).

Model evaluation on balanced data

The application of SMOTE and random undersampling on the training data had some small effects on classification performances, as can be seen in Table 4. Interestingly, only the SVM classifier returns an *AUC* score that differs from the score on the original data, seeing a minor .01 decrease. When looking at Figure 4, the ROC curves seem to reflect something similar. Although the curves have slightly changed, XGBoost and RF lie at similar positions as before. Albeit marginal, the ROC curves of the logistic regression and SVM classifiers have noticeably changed shape to more rounded and flatter lines in the upper-left corner respectively.

This difference is explained in the confusion matrices (Table 3) as the number of *TP* predictions has decreased by 36 for the SVM. However, the

Table 3: Confusion matrices for the balanced data.

		True	False
Dummy	Positive	0	0
	Negative	3,974	385
Logistic Regression	Positive	318	571
	Negative	3,403	67
SVM	Positive	301	352
	Negative	3,622	84
RF	Positive	325	135
	Negative	3,839	60
XGBoost	Positive	335	112
	Negative	3,862	50

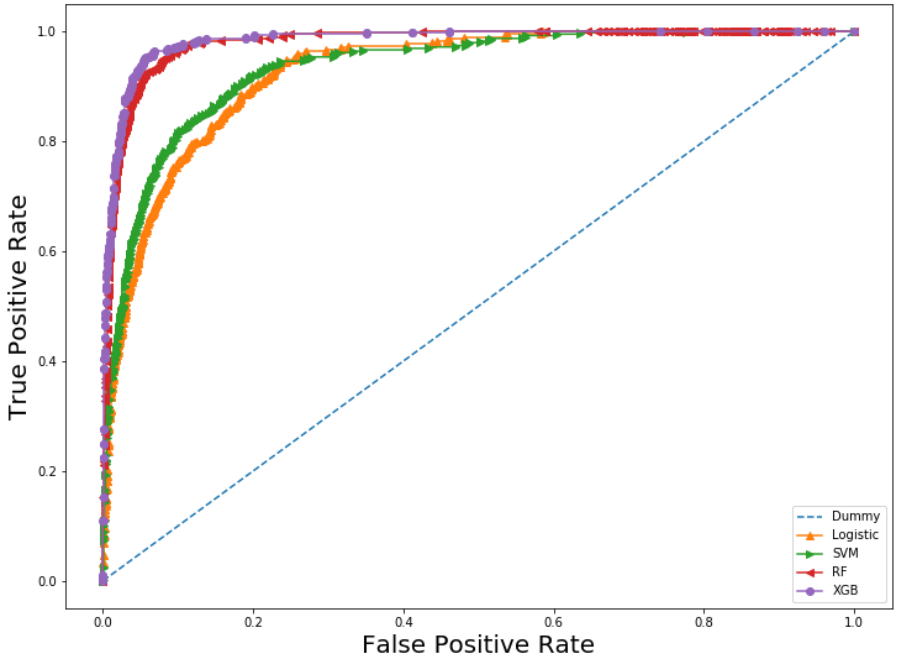

Figure 4: ROC curves on the balanced data.

classifier has drastically improved its *TN* predictions by 168. Moreover, the tree-based models have a higher number of *TP* predictions, yet slightly less *TN* predictions.

As shown in table 4 the logistic regression and SVM classifiers seem to be less sensitive, but more specific. Conversely, RF and XGBoost classifiers return higher sensitivity scores, and lower specificity scores on the balanced data than on the original data. This is in line with the contrasting effects seen in the confusion matrices, where the tree-based models act inverted to the SVM and baseline. Overall, none of the classifiers improves in both *TP* and *TN* by balancing the training data.

The MCC shows that the tree-based models have the strongest classification performance, as it takes class proportions best into account. In addition, the performance of the RF classifier seems to suffer from the data balancing methods, while SVM and XGBoost have benefited from it. All things considered, XGBoost returned the best scores, as it is both sensitive and specific, in addition to having high *AUC* and *MCC* scores.

Table 4: Model comparisons on original and balanced data, scored on *AUC*, sensitivity, specificity, and *MCC*.

	Data	AUC	Sensitivity	Specificity	MCC
Logistic Regression	Original	**.93**	**.83**	.86	.48
	Balanced	.93	.83	**.86**	.48
SVM	Original	**.94**	**.88**	.87	.53
	Balanced	.93	.78	**.91**	**.55**
RF	Original	.98	.80	**.98**	**.77**
	Balanced	.98	**.84**	.97	.75
XGBoost	Original	.98	.83	**.98**	.78
	Balanced	.98	**.87**	.97	**.79**

Feature importances

On the original data, RF returned mammal biodiversity, latitude, longitude, and area size as the most important features in building the decision trees respectively. XGBoost made use of different features, with the presence of brown bears being the most important, followed by mammal biodiversity, lynx presence, and chamois presence.

The RF classifier that was trained on the balanced data had similar feature importances, but this time area size carried more weight than longitude. XGBoost saw a rise in importance for the SITETYPE_A variable,

while the presence of chamois had little impact on the decision trees when trained on balanced data.

A complete overview of feature importances can be found in Appendix C (page 29).

5 DISCUSSION

This study set out to predict the presence of wolves in Natura 2000 sites. The models used to make predictions on the spatial distribution of animal species are referred to as SDMs (Jiménez-Valverde et al., 2008). While most SDMs are built by collecting and combining multiple sources of data, this study aimed at making predictions from a readily available extensive dataset that is updated yearly. In the last decade or so researchers have successfully implemented machine learning algorithms in SDMs, most notably SVM and RF models (Fukuda et al., 2013; Mi et al., 2017; Sabat-Tomala et al., 2020). The goal of this study was to utilize these two machine learning techniques, with the addition of a XGBoost classifier, and compare their performance as SDMs for wolves in Natura 2000 sites. More specifically, the study aimed to gain insight on the extent to which the specific features of a Natura 2000 site can be used by ML algorithms to predict the presence of wolves (RQ1). In addition, it sought to find the best-performing algorithm (RQ2), and the features that were most telling in making these predictions (RQ3).

To account for the class imbalance in the dataset, the machine learning models were trained once on the original data, and once more on data that was partially balanced.

The results show that all models, including the logistic regression that was used as a baseline model, could predict the presence of wolves to some extent. The *AUC* scores were comfortably higher than that of a dummy classifier that blindly predicts the negative class (i.e. the majority class), so it can be assumed that the models have a certain degree of predictive power. Moreover, the tree-based models were the best performing models by some margin. This becomes most apparent in the *MCC* scoring metric, which demonstrates that using the Matthews correlation coefficient is well suited for the evaluation of classifiers on data suffering from class imbalance, as suggested by Chicco and Jurman (2020).

It is noteworthy that the model that best predicted the positive class was the SVM classifier, as overall its performance was evidently worse than RF and XGBoost. This could be an indication that SVM is well able to classify the data, but struggles with variance in the data. In other words, its relatively poor performance in predicting the negative class might be

attributed to overfitting, if data in the test set somewhat varies from the data that was used to train the model.

While sensitivity and specificity scores for the SVM classifier are almost identical when trained on the original data, the tree-based models are clearly more specific than sensitive. Hence it could be argued that the ensemble methods that are incorporated in these models are the reason they perform better than SVM and logistic regression, as bagging and boosting are means to reduce overfitting. It could be that a number of features in the data are strong predictors for the absence of wolves, and due to their ensemble approaches RF and XGBoost are more likely to notice these features.

The data balancing methods had less impact on the classification performance of the models than anticipated. RF performed better on the original dataset, while SVM and XGBoost saw minor improvements after the application of balancing methods. Moreover, the impact differed per model, as an increase in sensitivity was accompanied by a decrease in specificity (or vice versa). The variability that is common in ecological data might be a reason for this, since the addition of new instances to the minority class could lead to the classifiers associating certain features with sites that have wolf presence, while there are no sites with such features in the test data. Conversely, removing features from the majority class might have led to information loss.

When the classification performances on both the original and balanced data are taken into consideration, it can be concluded that XGBoost was the best performing algorithm. This is in line with results from the small body of previous research on the models, as reported by Chang et al. (2018), and Shi et al. (2019). Nevertheless, RF performed only marginally worse than XGBoost, and was clearly able to beat both the baseline and the SVM classifier. Once more, this is in agreement with related works (Fukuda et al., 2013; Mi et al., 2017; Sabat-Tomala et al., 2020).

Features that seemed to benefit predictive performances were the presence of rivaling large predators, the number of distinct mammal species, and the presence of chamois. As brown bears and lynxes are often present in the same regions as wolves (e.g. dense forests in rural Romania) the inclusion of these features might have a disproportionate impact on the models performances. Moreover, it has to be noted that the data for most of the prey species is incomplete. The wolf's favorite prey species - wild boar and roe deer - are so common and widespread all over the European continent that only a small fraction of Natura 2000 sites opted to list their presence. Hence, prey species such as the chamois (a rare species only present in certain mountain regions) might carry more weight due to a lack of information on other prey species. Another important factor that

might have influenced model performances is the incompleteness of data on wolf presence. Since Member States could opt to withhold certain sensitive information to protect vulnerable species, countries such as the Netherlands are not listed as having any sites with wolves presence in the dataset. The inclusion of sites in countries where only a few Natura 2000 sites are ground truth labeled for wolf presence could have had some effect on classification performances.

Surprisingly, threats and pressures around a Natura 2000 site (impact codes), seem to have little importance. This is in agreement with the notion that large predators have little preferences in picking their habitats (Cimatti et al., 2021; Linnell et al., 2005), while not aligning with Morales-González et al. (2022) finding that wolves avoid regions with high amounts of agricultural activity. This could be accredited the preprocessing steps that were taken, as missing values imputation and simplification of numerous high cardinality features have inevitably led to information loss.

The impact that the area size of a Natura 2000 site has on model performances seems reasonable, as it can be expected that large predators need a relatively spacious territory. Mind that interpreting feature importances has to be done with some restraint. Features with high cardinality can be used for more tree splits than a feature with binary categories, and co-occurrence of certain features might impact their perceived importance (Ghorbani, Berenbaum, Ivgi, Dafna, & Zou, 2021).

Perhaps the most important limitations to this study is the lack of expert knowledge in the ecological domain, as Beery et al. (2021) suggest the cooperation between expert ecologists and computer scientists to model meaningful SDMs with machine learning techniques. Hence, certain important features might not have been missing in the data, or should not have been part of the model in the first place.

6 CONCLUSION

As this study's main research question aimed to explore the extent to which ML algorithms could predict the existence of wolves in Natura 2000 sites, the results can be seen as promising. Though assumptions about the way these results translate to practical uses have to be made with great care, it can be concluded that ML techniques are well suited to handle similar classification problems. XGBoost and RF were most able to correctly predict the existence of wolves in a Natura 2000 site, with the best performing classifiers obtaining Matthews correlation coefficients of .79 and .77 respectively. This provides supplementary evidence in support of using ensemble models as SDMs, as proposed by Elith et al. (2008); Mi et al. (2017). Concurring with results from previous studies, SVM was able to

make predictions more accurately than generalized linear models such as a logistic regression classifier (Drake et al., 2006; Fukuda et al., 2013), but was outperformed by RF and XGBoost (Sabat-Tomala et al., 2020). These outcomes simultaneously answer the second research question, which focused on finding the best performing ML algorithm.

Though the third research question could not be definitively answered, the tree-based models did provide some insight on the features that best predict the existence of wolves in Natura 2000 sites. These features can be categorized into two groups; one consisting of coarse descriptive features such as geolocation and area size of a site, and the other consisting of presence-absence data of certain (mainly large predator) species for a site.

The particularly encouraging results obtained by the XGBoost classifier suggest that this novel model can be successfully implemented as a SDM, and exploring the use of different ensemble methods in the ecological domain could benefit performances of future SDMs. Similar to XGBoost, there are a number of ensemble models that have not been used in SDMs to this date. These models have shown their potential in practical applications, as well as scientific research in a variety of domains. Combining sophisticated classifiers such as SVM with an ensemble model (i.e. stacking) could lead to better classification performances on both classes. Seeing as SVM displayed a specific ability to correctly predict the positive class, the model warrants a more in-depth analysis in the future.

REFERENCES

Ahsan, M. M., Mahmud, M., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of data scaling methods on machine learning algorithms and model performance. *Technologies*, *9*(3), 52.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, *6*(1), 20–29.

Beery, S., Cole, E., Parker, J., Perona, P., & Winner, K. (2021). Species distribution modeling for machine learning practitioners: A review. In *Acm sigcas conference on computing and sustainable societies* (pp. 329–348).

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, *116*(32), 15849–15854.

Berger-Wolf, T. Y., Rubenstein, D. I., Stewart, C. V., Holmberg, J. A., Parham, J., Menon, S., ... Joppa, L. (2017). Wildbook: Crowdsourcing, computer vision, and data science for conservation. *arXiv preprint arXiv:1710.08880*.

Bühlmann, P. (2012). Bagging, boosting and ensemble methods. In J. E. Gentle, W. K. Härdle, & Y. Mori (Eds.), *Handbook of computational statistics: Concepts and methods* (pp. 985–1022). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from https://doi.org/10.1007/978-3-642-21551-3_33 doi: 10.1007/978-3-642-21551-3_33

Chang, Y.-C., Chang, K.-H., & Wu, G.-J. (2018). Application of extreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Applied Soft Computing*, *73*, 914–920.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Chicco, D., & Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*(1), 1–13.

Cimatti, M., Ranc, N., Benítez-López, A., Maiorano, L., Boitani, L., Cagnacci, F., ... others (2021). Large carnivore expansion in europe is associated with human population density and land cover changes. *Diversity and Distributions*, *27*(4), 602–617.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, *20*(2), 215-

232. Retrieved from https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1958.tb00292.x doi: https://doi.org/10.1111/j.2517-6161.1958.tb00292.x

Drake, J. M., Randin, C., & Guisan, A. (2006). Modelling ecological niches with support vector machines. *Journal of applied ecology*, *43*(3), 424–432.

Elith, J., & Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annual review of ecology, evolution, and systematics*, *40*, 677–697.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of animal ecology*, *77*(4), 802–813.

Elreedy, D., & Atiya, A. F. (2019). A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance. *Information Sciences*, *505*, 32–64.

European Commission. (n.d.). *Natura2000.* Retrieved from https://ec.europa.eu/environment/nature/natura2000/index_en.htm

European Commission. (2020). *Biodiversity strategy for 2030.* Retrieved from https://ec.europa.eu/environment/strategy/biodiversity-strategy-2030_en

European Environment Agency. (2022). *Natura 2000 data – the european network of protected sites [data files].* Retrieved from https://www.eea.europa.eu/data-and-maps/data/natura-13

Fan, J., Upadhye, S., & Worster, A. (2006). Understanding receiver operating characteristic (roc) curves. *Canadian Journal of Emergency Medicine*, *8*(1), 19–20.

Feng, C., Wang, H., Lu, N., & Tu, X. M. (2013). Log transformation: application and interpretation in biomedical research. *Statistics in medicine*, *32*(2), 230–239.

Fukuda, S., De Baets, B., Waegeman, W., Verwaeren, J., & Mouton, A. M. (2013). Habitat prediction and knowledge extraction for spawning european grayling (thymallus thymallus l.) using a broad range of species distribution models. *Environmental modelling & software*, *47*, 1–6.

Gable, T. D., Windels, S. K., Romanski, M. C., & Rosell, F. (2018). The forgotten prey of an iconic predator: a review of interactions between grey wolves canis lupus and beavers castor spp. *Mammal review*, *48*(2), 123–138.

Ghorbani, A., Berenbaum, D., Ivgi, M., Dafna, Y., & Zou, J. Y. (2021). Beyond importance scores: Interpreting tabular ml by visualizing feature semantics. *Information*, *13*(1), 15.

Healy, L. M. (2006). Logistic regression: An overview. *Eastern Michighan College of Technology*.

Hearst, M., Dumais, S., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their Applications*, *13*(4), 18-28. doi: 10.1109/5254.708428

Jiménez-Valverde, A., Lobo, J. M., & Hortal, J. (2008). Not as good as they seem: the importance of concepts in species distribution modelling. *Diversity and distributions*, *14*(6), 885–890.

Kleinke, K. (2018). Multiple imputation by predictive mean matching when sample size is small. *Methodology*, *14*(1), 3-15. doi: 10.1027/1614-2241/a000141

Landerman, L. R., Land, K. C., & Pieper, C. F. (1997). An empirical evaluation of the predictive mean matching method for imputing missing values. *Sociological methods & research*, *26*(1), 3–33.

Linnell, J. D., Promberger, C., Boitani, L., Swenson, J. E., Breitenmoser, U., & Andersen, R. (2005). The linkage between conservation strategies for large carnivores and biodiversity: the view from the "half-full" forests of europe. *Large carnivores and the conservation of biodiversity*, 381–398.

Lord, C. M., Wirebach, K. P., Tompkins, J., Bradshaw-Wilson, C., & Shaffer, C. L. (2020). Reintroduction of the european bison (bison bonasus) in central-eastern europe: a case study. *International Journal of Geographical Information Science*, *34*(8), 1628–1647.

McGinnis, W. (2016). *Target encoder.* Retrieved from `https://contrib.scikit-learn.org/category_encoders/targetencoder.html`

Mi, C., Huettmann, F., Guo, Y., Han, X., & Wen, L. (2017). Why choose random forest to predict rare species distribution with few samples in large undersampled areas? three asian crane species models provide supporting evidence. *PeerJ*, *5*, e2849.

Micci-Barreca, D. (2001, jul). A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *SIGKDD Explor. Newsl.*, *3*(1), 27–32. Retrieved from `https://doi.org/10.1145/507533.507538` doi: 10.1145/507533.507538

Morales-González, A., Fernández-Gil, A., Quevedo, M., & Revilla, E. (2022). Patterns and determinants of dispersal in grey wolves (canis lupus). *Biological Reviews*, *97*(2), 466–480.

Newsome, T. M., Boitani, L., Chapron, G., Ciucci, P., Dickman, C. R., Dellinger, J. A., ... others (2016). Food habits of the world's grey wolves. *Mammal Review*, *46*(4), 255–269.

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How many trees in a random forest? In *International workshop on machine learning and data mining in pattern recognition* (pp. 154–168).

Patle, A., & Chouhan, D. S. (2013). Svm kernel functions for classification. In *2013 international conference on advances in technology and engineering*

(icate) (pp. 1–9).

Pouteau, R., Meyer, J.-Y., & Stoll, B. (2011). A svm-based model for predicting distribution of the invasive tree miconia calvescens in tropical rainforests. *Ecological modelling*, *222*(15), 2631–2641.

Pouteau, R., Meyer, J.-Y., Taputuarai, R., & Stoll, B. (2012). Support vector machines to map rare and endangered native plants in pacific islands forests. *Ecological Informatics*, *9*, 37–46.

R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Ritchie, E. G., Elmhagen, B., Glen, A. S., Letnic, M., Ludwig, G., & McDonald, R. A. (2012). Ecosystem restoration with teeth: what role for predators? *Trends in Ecology I& Evolution*, *27*(5), 265-271. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0169534712000067` doi: https://doi.org/10.1016/j.tree.2012.01.001

Rocchini, D., Hortal, J., Lengyel, S., Lobo, J. M., Jimenez-Valverde, A., Ricotta, C., ... Chiarucci, A. (2011). Accounting for uncertainty when mapping species distributions: the need for maps of ignorance. *Progress in Physical Geography*, *35*(2), 211–226.

Sabat-Tomala, A., Raczko, E., & Zagajewski, B. (2020). Comparison of support vector machine and random forest algorithms for invasive and expansive species classification using airborne hyperspectral data. *Remote Sensing*, *12*(3), 516.

Sadeghi, R., Zarkami, R., Sabetraftar, K., & Van Damme, P. (2012). Use of support vector machines (svms) to predict distribution of an invasive water fern azolla filiculoides (lam.) in anzali wetland, southern caspian sea, iran. *Ecological Modelling*, *244*, 117–126.

Sáez, J. A., Krawczyk, B., & Woźniak, M. (2016). Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. *Pattern Recognition*, *57*, 164–178.

Salafsky, N., Salzer, D., Stattersfield, A. J., Hilton-Taylor, C., Neugarten, R., Butchart, S. H., ... others (2008). A standard lexicon for biodiversity conservation: unified classifications of threats and actions. *Conservation Biology*, *22*(4), 897–911.

Sanz-Pérez, A., Ordiz, A., Sand, H., Swenson, J. E., Wabakken, P., Wikenros, C., ... Milleret, C. (2018). No place like home? a test of the natal habitat-biased dispersal hypothesis in scandinavian wolves. *Royal Society open science*, *5*(12), 181379.

Saraswat, M. (2022). superml: Build machine learning models like using python's scikit-learn library in r [Computer software manual]. Retrieved from `https://github.com/saraswatmks/superml` (R pack-

age version 0.5.5)

Schapire, R. E., & Freund, Y. (2013). Boosting: Foundations and algorithms. *Kybernetes*.

Shi, H., Wang, H., Huang, Y., Zhao, L., Qin, C., & Liu, C. (2019). A hierarchical method based on weighted extreme gradient boosting in ecg heartbeat classification. *Computer methods and programs in biomedicine*, *171*, 1–10.

Sin, T., Gazzola, A., Chiriac, S., & Rîșnoveanu, G. (2019). Wolf diet and prey selection in the south-eastern carpathian mountains, romania. *PloS one*, *14*(11), e0225424.

Soberón, J. M. (2010). Niche and area of distribution modeling: a population ecology perspective. *Ecography*, *33*(1), 159–167.

Tuia, D., Kellenberger, B., Beery, S., Costelloe, B. R., Zuffi, S., Risse, B., ... others (2022). Perspectives in machine learning for wildlife conservation. *Nature communications*, *13*(1), 1–15.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, *45*(3), 1–67. Retrieved from `https://www.jstatsoft.org/index.php/jss/article/view/v045i03` doi: 10.18637/jss.v045.i03

van der Palm, D. W., van der Ark, L. A., & Vermunt, J. K. (2016). A comparison of incomplete-data methods for categorical data. *Statistical methods in medical research*, *25*(2), 754–774.

Wickham, H. (2022). stringr: Simple, consistent wrappers for common string operations [Computer software manual]. (http://stringr.tidyverse.org, https://github.com/tidyverse/stringr)

Wickham, H., François, R., Henry, L., & Müller, K. (2022). dplyr: A grammar of data manipulation [Computer software manual]. (https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr)

WWF. (2020). *Living planet report 2020 - bending the curve of biodiversity loss.* Gland, CH. Retrieved from `https://livingplanet.panda.org/`

A DATASET VARIABLES

TPA in Table 5 stands for Threats, Pressures and Activities' near a Natura 2000 site. SPA stands for Special Protection Area (for bird species protection). Sites of Community Importance, SAC stands for Special Area of Conservation (for habitat protection).

Table 5: Overview of independent variables in the final dataset (post-cleaning)

Variable name	Description	Type
AREAHA	Area size in **ha**	contin. num.
MARINE_AREA_PERC.	percentage covered by marine territory	discrete num.
LATITUDE	geographical coordinate	discrete num.
LONGITUDE	geographical coordinate	discrete num.
BIOGEOGRAPHICREG	biogeographic region in which site falls	categorical
HABITATCODE	most prominent habitat type	categorical
IC_A	agriculture TPA	cat. (dummy)
IC_B	sylviculture, forestry TPA	cat. (dummy)
IC_C	mining, extraction TPA	cat. (dummy)
IC_D	transportation TPA	cat. (dummy)
IC_E	urbanisation TPA	cat. (dummy)
IC_F	hunting, fishing TPA	cat. (dummy)
IC_G	human intrusions TPA	cat. (dummy)
IC_H	pollution TPA	cat. (dummy)
IC_I	invasive species TPA	cat. (dummy)
IC_J	natural system modifications TPA	cat. (dummy)
IC_K	natural (a)biotic processes TPA	cat. (dummy)
IC_L	geological events TPA	cat. (dummy)
IC_M	climate change TPA	cat. (dummy)
hare	presence/absence of hare	cat. (dummy)
roe_deer	presence/absence of roe deer	cat. (dummy)
red_deer	presence/absence of red deer	cat. (dummy)
chamois	presence/absence of chamois	cat. (dummy)
boar	presence/absence of wild boar	cat. (dummy)
beaver	presence/absence of beaver	cat. (dummy)
reindeer	presence/absence of reindeer	cat. (dummy)
ibex	presence/absence of ibex	cat. (dummy)
brown_bear	presence/absence of brown bear	cat. (dummy)
lynx	presence/absence of lynx	cat. (dummy)
mam_biodiv	distinct number of mammal species	contin. num.
SITETYPE_A	SPA	cat. (one-hot)
SITETYPE_B	SCI, SAC	cat. (one-hot)
SITETYPE_C	both SPA & SCI, SAC	cat. (one-hot)

B HYPERPARAMETER SELECTION

Table 6: Complete selection of hyperparameters that were included in the grid search, and the resulting optimal hyperparameters

	Hyperparameters	
Models	Grid	Optimal
Logistic	*solvers*[newton-cg, lbfgs, liblinear, saga]	**saga**
	penalty[L2, L1, None]	**L1**
	C[100, 10, 1.0, 0.1, 0.01]	**1.0**
SVM	C[100, 10, 1.0, 0.1, 0.01]	**10**
	kernel[rbf, auto]	**rbf**
	gamma[scale, auto]	**auto**
RF	*n estimators*[100, 200, 300, ..., 1000]	**1000**
	max features[auto, sqrt, 0.2, 0.5]	**0.2**
	max depth[None, 10, 20, 30, ..., 100]	**60**
	min samples split[2, 5, 10, 25]	**5**
	min samples leaf[2, 5, 10, 25]	**2**
	bootstrap[True, False]	**False**
XGBoost	*booster*[gbtree, gblinear, dart]	**gbtree**
	eta[0.1, 0.2, 0.3]	**0.1**
	max depth[None, 3, 4, 5, ..., 10]	**7**
	min child weight[1, 2, 5]	**2**
	scale pos weight[1, 5, 10]	**5**
	objective[binary:logistic, binary:hinge]	**binary:logistic**
	subsample[0.5, 1]	**1**
	colsample bytree[0.5, 1]	**1**

C FEATURE IMPORTANCES

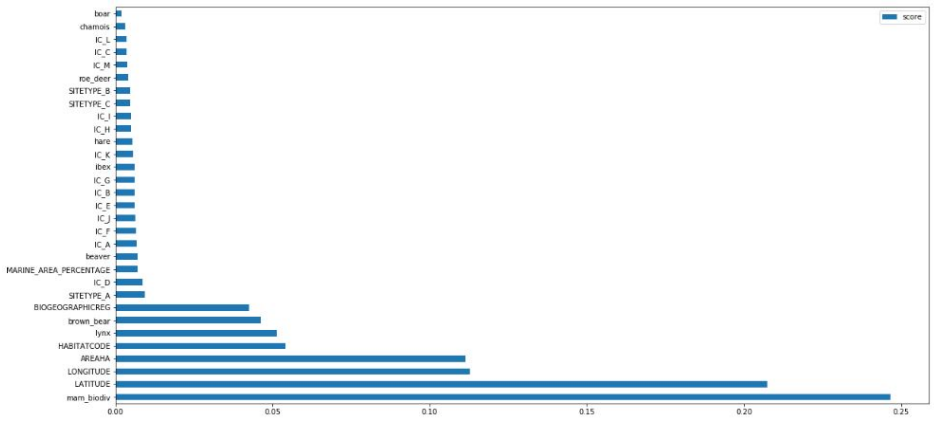

Figure 5: Feature importances RF on the original data.

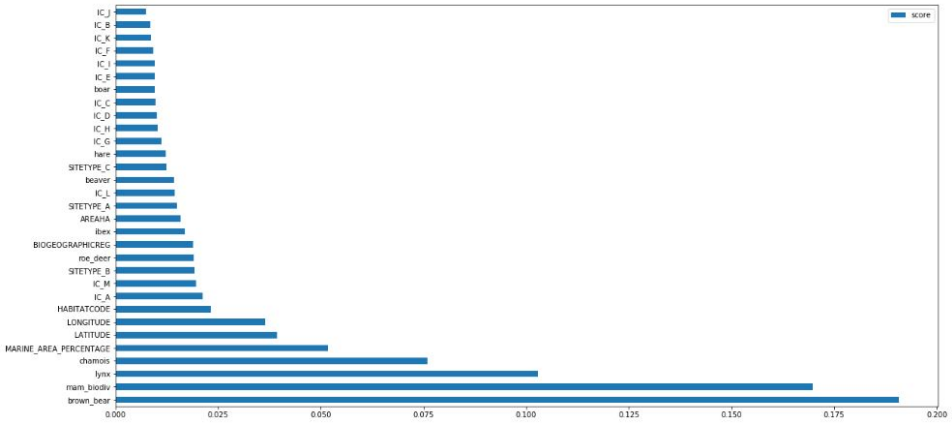

Figure 6: Feature importances XGBoost on the original data.

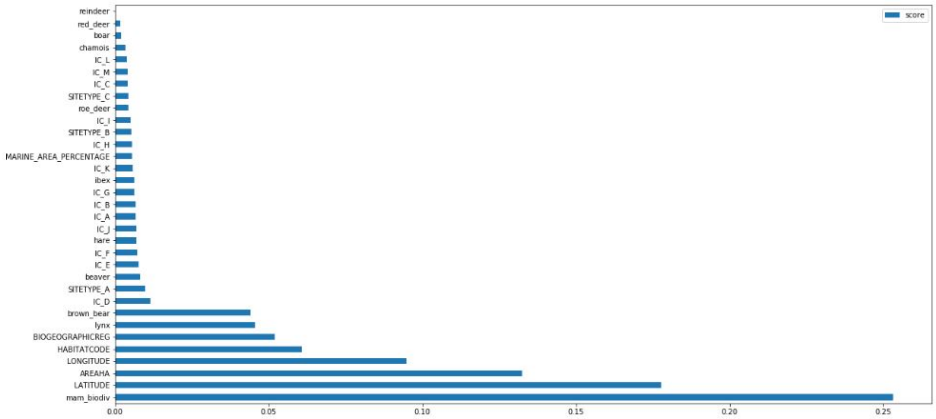

Figure 7: Feature importances RF on the balanced data.

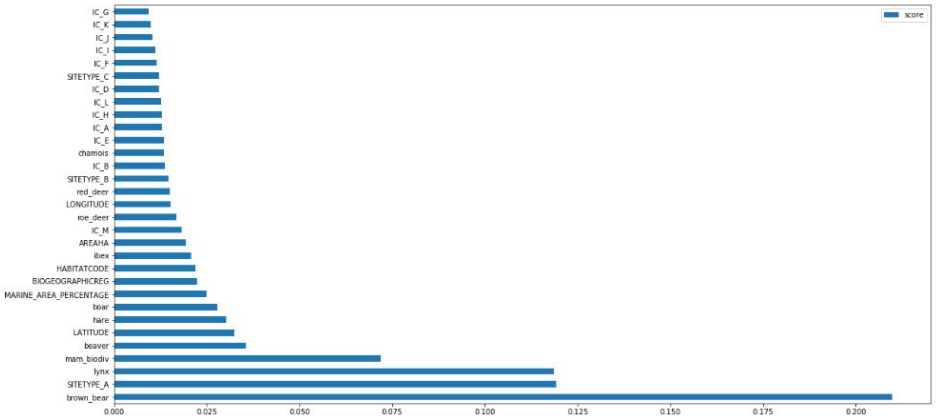

Figure 8: Feature importances XGBoost on the balanced data.