FOREST MONITOR

- Methodology to obtain the potential location of deciduous forest with high conservation values in southern Sweden

REPORT

2025 - 2

PROTECT THE FOREST

Report 2025 - 2

FOREST MONITOR
- Methodology to obtain the potential location of deciduous forest with high conservation values in southern Sweden

Author
Hainner Aparicio

Proofreading
Jon Andersson

Graphics
Hainner Aparicio & Jon Andersson

Layout
Jon Andersson

Acknowledgments
We want to express our greatest appreciation to the Postcode Lottery Foundation, to Naturkompaniet and to our private donors!

Cover photo
Jon Andersson

The report is produced by Protect the Forest 2025: www.skyddaskogen.se

INTRODUCTION

Sweden is home to a significant part of the EU's natural heritage not only in the form of valuable coniferous forests, but also older deciduous forests (DF) with high conservation values (HCV). Older deciduous forests often have unique flora, fungi, and fauna associated with them. To meet national and international environmental goals and EU laws and regulations, Sweden must protect all forests with high natural values and restore strategic areas around them.

An important step in the work to protect and restore vital forest ecosystems is to locate known and potential conservation values. This greatly streamlines inventory efforts and contributes to a unique overview of the values in forest landscapes.

The map tool Skogsmonitor.se presents data on known and potential conservation values for the entire country, linked to older forests and so-called continuity forests, forests that have not previously been clear-cut. To find potential older forests and continuity forests, Skogsmonitor has used historical aerial photographs and satellite images covering the entire country and then classified these using inventory data and algorithms that utilize findings of species of conservation interest.

There are challenges in using classic analysis of aerial and satellite images, as these do not always capture deciduous forests with potentially high conservation values (DFs with potential HCV) in a satisfactory way. New tools, like machine learning, offer a great opportunity by utilizing computational power to analyze massive loads of data from different sources in a consolidated/rapid yet detailed and robust manner.

We have used a machine learning approach to produce a map layer with predictions of where the most valuable deciduous forests can be found in southern Sweden. The map layer of deciduous forests with potential conservation values was created using a random forest classifier trained on satellite images, vegetation maps, terrain data, and data from tree and forest surveys.

The purposeful methodology attempted in this study and its results help locate these important forests within reasonable accuracy of over 90 percent (Accuracy > 0.9; Kappa > 0.8; RMSE < 0.3).

In this report we describe the methods used for the prediction of locations of DFs with potential HCV. We describe the datasets construction, pre-processing steps, modelling and the post processing steps that were carried out.

METHODOLOGY

Study area and data sources

The study area covers the southern part of Sweden except for the counties Värmland, Dalarna and Gotland, see **Figure 1**.

Figure 1. Map showing the county-wise coverage of the study area.

Reference data

The rate of success or failure in detecting various forest features with any kind of algorithm relies heavily on the construction of a solid reference dataset and well-chosen variables for modelling. We will give an exhaustive explanation of how we went about choosing the data and how we prepared the data sets for the analysis.

We used two different open data sources, where we downloaded datasets which we later mined, filtered and cleared to select specific places where the deciduous forests with high conservation value have been registered/reported. These forests correspond in ecological characteristics with our targeted forests (DFs with HCV). Therefore, they served as reference locations for creating a consistent reference dataset and further training and testing of the machine learning (ML) model.

Tabel 1. Data sources that were used to supervise the training procedure.

Data source	Forest type	Relevance
Naturvarden (Skogsstyrelsen)	Alsumpskog (ALSUMP)	High
	Aspskog (ASPSKOG)	High
	Bokskog (BOKSKOG_ONV)	High
	Hassellund (HASSLUND)	Medium
	Hedädellövskog (HEDÄDEL)	High
	Kalklövskog (KALKLÖV)	High
	Lövnaturskog (LÖVSKOG)	High
	Lövskog (LÖVSKOG_ONV)	High
	Lövskogslund (LÖVLUND)	Medium
	Lövskogslund/Hagmarksskog (LÖVLUND_ONV)	High
	Lövsumpskog (LÖVSUMP)	High
	Sekundär lövnaturskog (SEKNSKOG)	High
	Sekundär ädellövnaturskog (SEKÄDEL)	High
	Ädellövnaturskog (ÄDELLÖV)	High
	Ädellövskog (ÄDELBEST)	High
	Ädellövskog (ÄDELLÖV_ONV)	High
	Ädellövskog (ÄDELSKOG)	High
	Ädellövsumpskog (ÄDELSUMP)	High
	Örtrik allund (ALLUND)	High
	Övriga lövträd (ÖLÖVTRÄD)	High
Biotoskydd (Skogsstyrelsen)	Alkärr	High
	Hassellundar och hasselrika skogar	Medium
	Örtrika allundar	Medium
Nyckelbiotoper (Skogsstyrelsen)	Alsumpskog (ALSUMP)	High
	Aspskog (ASPSKOG)	High
	Bokskog (BOKSKOG_ONV)	High
	Hassellund (HASSLUND)	Medium
	Hedädellövskog (HEDÄDEL)	High
	Kalklövskog (KALKLÖV)	High
	Lövnaturskog (LÖVSKOG)	High
	Lövskog (LÖVSKOG_ONV)	High
	Lövskogslund (LÖVLUND)	Medium
	Lövskogslund/Hagmarksskog (LÖVLUND_ONV)	High
	Lövsumpskog (LÖVSUMP)	High
	Sekundär lövnaturskog (SEKNSKOG)	High
	Sekundär ädellövnaturskog (SEKÄDEL)	High
	Ädellövnaturskog (ÄDELLÖV)	High

Tabel 1, continued. Data sources that were used to supervise the training procedure.

Data source	Habitat type	Relevance
	Ädellövskog (ÄDELBEST)	High
	Ädellövskog (ÄDELSKOG)	High
	Ädellövsumpskog (ÄDELSUMP)	High
	Ädellövträd (ÄDELTRÄD)	High
	Örtrik allund (ALLUND)	High
	Övriga lövträd (ÖLÖVTRÄD)	High
Naturvardsavtal (Skogsstyrelsen)	Lövsumpskog	High
	Triviallövskog asp	High
	Triviallövskog björk	High
	Triviallövskog med ädellövinslag	High
	Triviallövskog övrigt	High
	Ädellövskog bok	High
	Ädellövskog ek	High
	Ädellövskog övrigt	High
Naturtypskartan – NNK (Naturvårdsverket)	9080 - Lövsumpskog	High
	9110 - Näringsfattig bokskog	High
	9130 - Näringsrik bokskog	High
	9160 - Näringsrik ekskog	High
	9161 - Näringsrik ekskog - Ek-avenbokskog	High
	9162 - Näringsrik ekskog - Ek-hassellund	High
	9170 - Torr ekskog	High
	9180 - Ädellövskog i branter	High
	9190 - Näringsfattig ekskog	High
	9750 - Svämlövskog (91E0)	High
	9760 - Svämädellövskog (91F0)	High
	9820 - Obestämd ädellövskog (9020, 9850, 9860)	High

For details on the specific data sources that were used to build the reference dataset, see **Table 1** below.

The reference datasets in **Table 1** were compiled and then prepared for the ML modelling. We used the Random Forest algorithm which is built on an assemblage of multiple decision trees. However, decision trees are sensitive to class imbalance and the classifier shows better results when it is trained with datasets where the number of data points per class is more balanced. Therefore, a balancing procedure was required.

The landscape was sampled on target areas (labelled class 1 = protected DF) against all other types of forests (labelled class 0). Finally, to avoid the influence of edge effects, sample points within a 100 meter buffer from boundary lines were subtracted. Note that forest masking was performed throughout all sampling implementations. Consequently, sampling of non-forest was avoided. Reasonable balance was attained once the difference between number of sampled points per class reached under 25%.

Once the sampling scheme was implemented and the geodata sets with the modelling variables ready (this last step will be explained in detail in the next section), variable's values were extracted for each sampled-labelled point. Then our reference dataset was ready. Depending on the size of the county, between 400,000 and 1,000,000 points were sampled per county.

Modelling data

The characteristics that make our target forest type valuable are influenced by factors other than the tree species assemblage alone, e.g. tree age, biomass, land use, terrain etc. Therefore, we explored and mined all the available open-source data that could help us understand the interactions between these factors as a predictor of DFs with HCV.

Multispectral satellite imagery with its different bands and their correlation to vegetation signatures allowed us to infer vegetation cover and the state of the vegetation. Topographic factors such as elevation and slope influence soil profiles and water fluxes, and therefore they could help us discern where our target forests thrive.

Several high-quality studies from the Swedish Forest Agency, SFA and the Swedish University of Agricultural Sciences, SLU have been carried out to gather data on the status of Swedish forests. Such data on forest status could also be of value to decipher the presence of our targeted forests.

For our methodology implementation, we also used data on general dendrometry on biomass, tree height and soil moisture from the SFA, and the tree volume of various deciduous tree species from the SLU. Data from the Swedish National Land Cover Database (NMD) was used to make a visual investigation of the ML resulting rasters. However, the NMD data was not used to train our models. In **Table 2** below, the specific data sources.

Table 2. Predictors used for modelling.

Data type		Source	Unit	Notes
Satellite data	Sentinel 2 – band 2 - 8 (8 & 8a)	Copernicus hub	DN	Sentinel 2 L2A* Resolution between 10 – 20 m Captured 2023-2024, cloud free scenes from late autumn. Accessed via: https://dataspace.copernicus.eu/browser/
Topographic data	Elevation Slope	The Swedish Geographical Survey	meters degrees	Point cloud data were accsessed via the Swedish Geographical Survey's web portal, Geotorget: https://geotorget.lantmateriet.se/ The data were processed to a DEM with 10 meter resolution.
Dendrometry	BEECHVOL_XX_P_15 BIRCHVOL_XX_P_15 DECIDUOUSVOL_XX_P_15 OAKVOL_XX_P_15	The Swedish University of Agricultural Sciences (SLU)	m3/ha	Accessed via SLU's open data service: https://gis.slu.se/data/slu_forest_map/ Resolution 10 m
	Biomass Tree height	The Swedish Forest Agency	Ton/ha dm	Accessed via the Swedish Forest Agency's open data service: https://www.skogsstyrelsen.se/e-tjanster-
Soil type	Soil moisture		wtd	

***** Level-2A provides Bottom of Atmosphere (BOA) reflectance, which are corrected for atmospheric effects. Level-2A products are useful for applications requiring accurate surface reflectance data, such as vegetation monitoring and land cover classification.

The Sentinel 2 - L2A band sets and geospatial data processing was carried out using Python 3.7, QGIS and ArcGIS.

Machine learning approach

The accurate detection of DFs with HCV using remote sensing is a complex task. As mentioned before, the particular characteristics that make this and other forest types valuable for nature conservation are influenced by multiple factors. Methods that can capture problems on relatively large areas are needed for an analysis on the spatial scale of this study. Supervised machine learning (ML) is a suitable and successfully tested approach in studies of this spatial scale. It is a method to investigate the relevance of many factors from a variety of data sources in a systematic manner. Several ML-algorithms have already been successfully used to predict forest types, but so far few have been made with the aim of detecting DFs with HCV.

We used the "scikit-learn" library in Python to implement the ML approach and investigated the relationship between variables with a cross-correlation and principal component analysis (PCA) – examples can be ob-

served in **Figure A - C**. Afterwards, we used all derived variables as predictor information for the classification with supervised machine learning.

All labelled data (class 1 = protected DF and 0 = other forest) was used to train several machine learning algorithms. With the library "lazy predict" many basic ML models were built to determine which models would work better without any parameter tuning. This included ML algorithms like Random Forest classifier, AdaBoost classifier and logistic regression. After a thorough comparison of these algorithms, the Random Forest

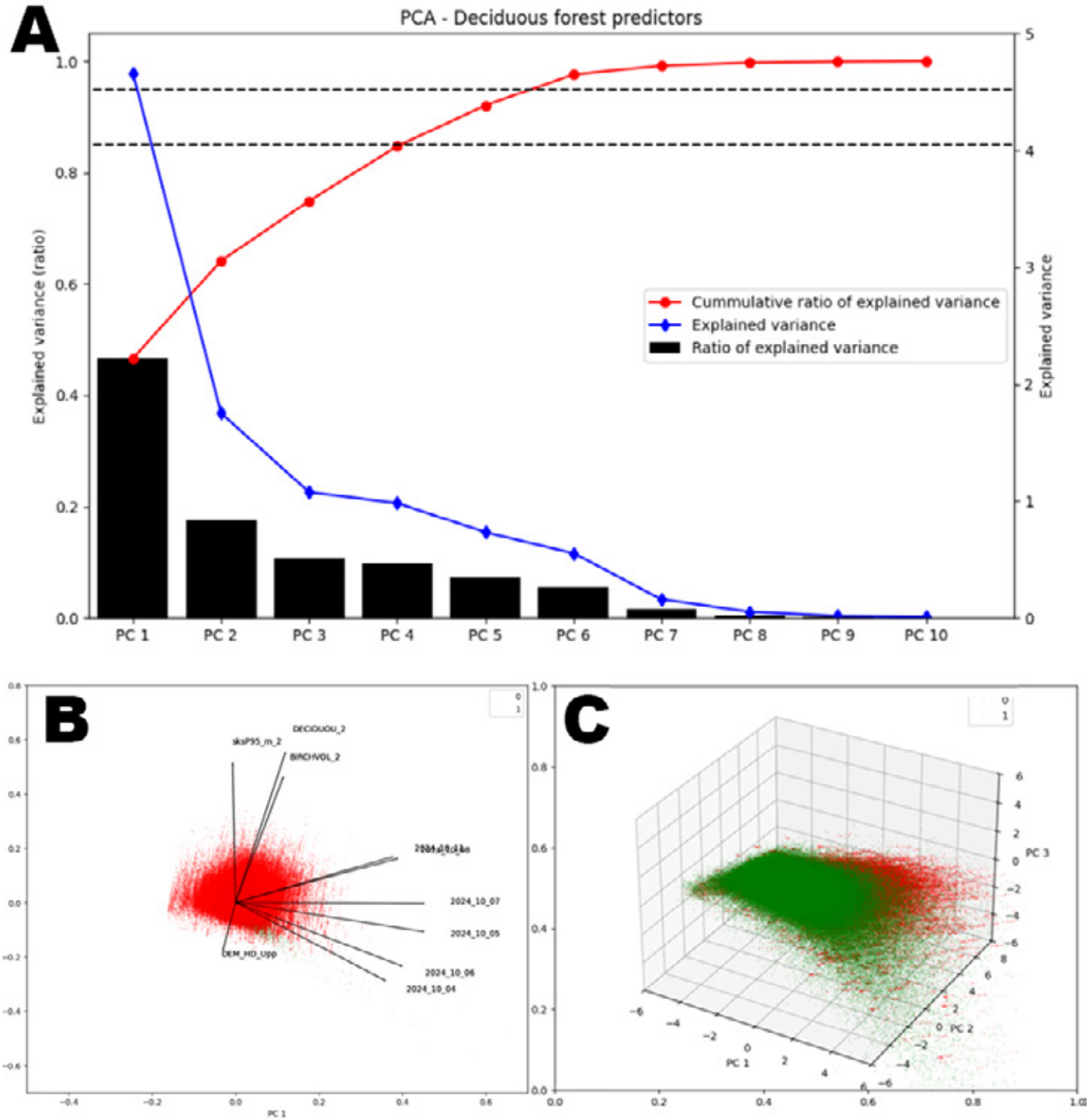


Figure 2 A - C. In A, an example of a scree plot from the PCA on data for one of the 13 county sections that were mapped in this study. In B and C, the 2D (left) and 3D (right) projection of the multidimetional vectors (predictors) and the achieved segmentation of sampled points by class (prediction).

classifier came out as the best choice. Based on this result, we chose to use this classifier.

The random forest classifier was trained (0.7 train: 0.3 test partitioned) and its performance was evaluated. To optimize the classifier, we tuned the hyper parameters and then trained the algorithm again. We used a random grid search with cross validation to define a grid of hyperparameter ranges, and randomly sampled from the grid, performing K-Fold cross validation with each combination of values. The best parameter combination was used in the final model, see **Appendix 1 - 4** for examples.

The performance of the predictions was assessed based on the testing partition of the dataset through the following metrics: overall accuracy, balanced accuracy, F1 score, recall, precision, R2, RMSE and Kappa-values. To evaluate the predictor's relevance in the results, the importance of each predictor was also calculated.

For a full description of the workflow of the study, see **Figure 3** below.

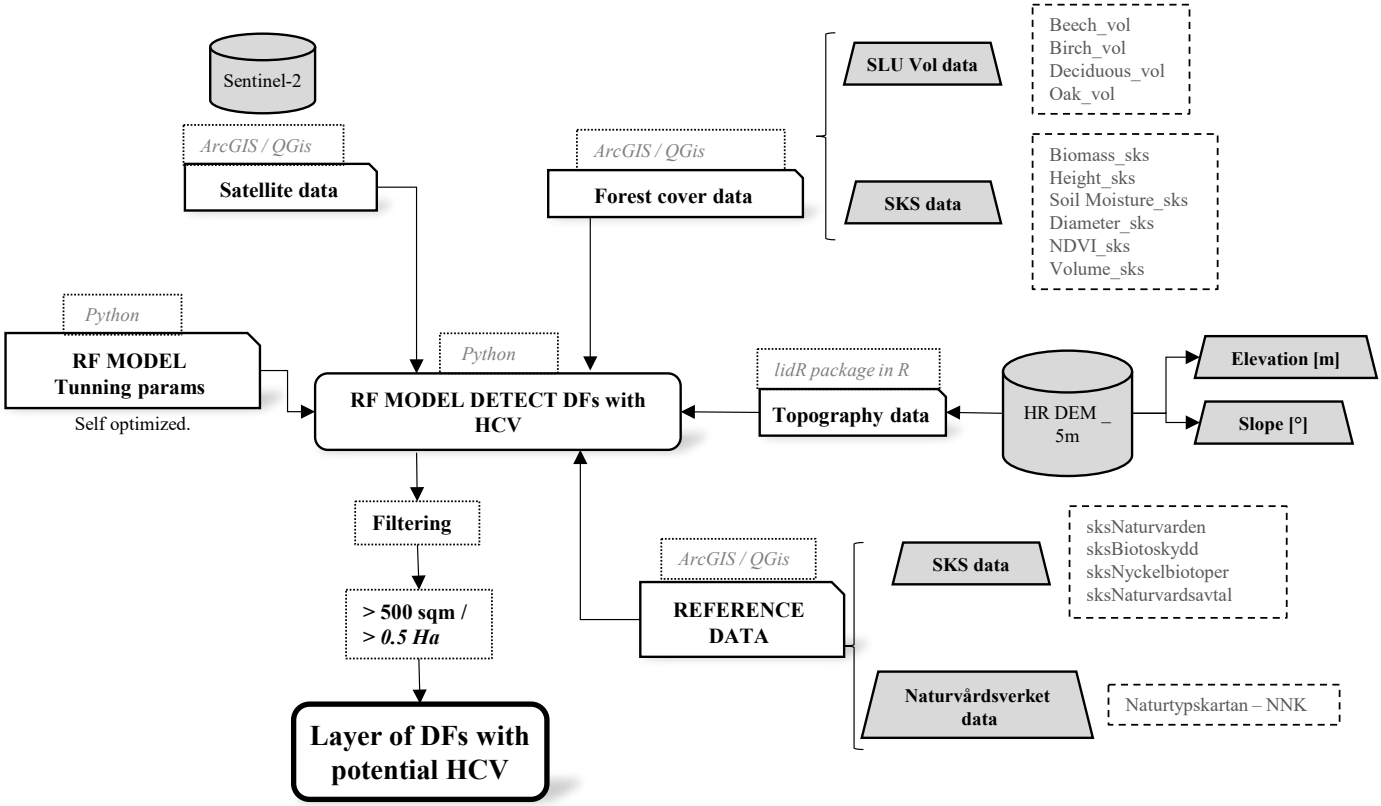


Figure 3. Workflow diagram that illustrates the general approach of our study.

RESULTS

Once the accuracy metrics were calculated, we could conclude that we achieved good accuracy values and acceptable to good error ranges (Accuracy > 0.9; Kappa > 0.8; RMSE < 0.3). According to our results, approximately a little more than 289,000 hectares of DFs with potential HCV lay in the studied area, from those, a mere 18.5% (53,519 hectares) lay inside protected areas.

Skogsmonitor would like to emphasize that it is important for everyone who uses the new deciduous forest layer to understand that this is a prediction, or forecast, of where in the landscape there could be DFs with potential HCV.

The degree of accuracy determines how valuable these types of predictions are for nature conservation work,

environmental monitoring, and in terms of streamlining the prioritization of what should be inventoried in the field.

There is no scientific consensus on what constitutes “good” accuracy in forest classification studies using machine learning. However, similar studies in this field vary from acceptable and valuable results above 70% to excellent accuracy, which usually exceeds 90–95%. High accuracy depends on factors such as the complexity of the classification task and the quality and type of data used.

Our results have a high degree of accuracy. Although they are usually correct, we have also seen examples of accumulations of standing dead conifers being detected as defoliated large deciduous trees in some cases.

For a comparison between known areas with HCV DFs and predicted presense of such forests, see **Figure 4 A-B**.

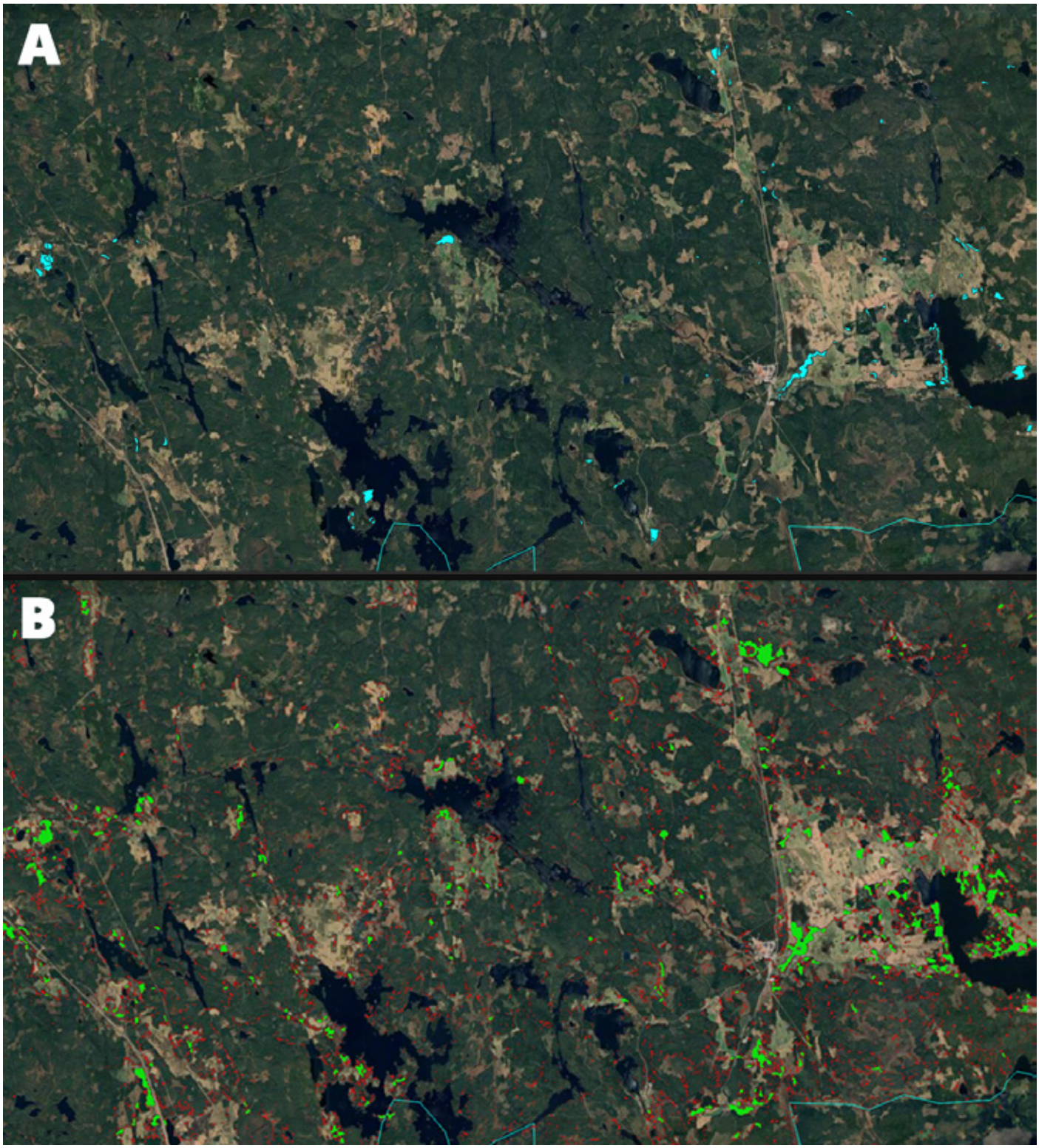


Figure 4 A - B. In A, formally protected deciduous forests in light blue and in B, the ML predictions in light green. The area covers a section of southeast Örebro county.

APPENDIX

jupyter DeciduousForests_01 Last Checkpoint: 5 hours ago

File Edit View Run Kernel Settings Help

Trusted

Code - Open in... Python 3 (ipykernel)

Parameter Tuning

[110]:
```
## create parameter grid ##

#https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74
#check: https://scikit-learn.org/stable/modules/grid_search.html

from sklearn.model_selection import RandomizedSearchCV

n_estimators = [int(x) for x in np.linspace(start = 25, stop = 200, num = 10)] # Number of trees in random forest
max_features = ['auto', 'sqrt'] # Number of features to consider at every split
max_depth = [int(x) for x in np.linspace(2, 10, num = 5)] # Maximum number of levels in tree
max_depth.append(None)
min_samples_split = [2, 5, 10] # Minimum number of samples required to split a node
min_samples_leaf = [1, 2, 4] # Minimum number of samples required at each leaf node
bootstrap = [True, False] # Method of selecting samples for training each tree

# Create the random grid
random_grid = {'n_estimators': n_estimators,
               'max_features': max_features,
               'max_depth': max_depth,
               'min_samples_split': min_samples_split,
               'min_samples_leaf': min_samples_leaf,
               'bootstrap': bootstrap}
print(random_grid)
```

```
{'n_estimators': [25, 44, 63, 83, 102, 122, 141, 161, 180, 200], 'max_features': ['auto', 'sqrt'], 'max_depth': [2, 4, 6, 8, 10, None], 'min_samples_spli
t': [2, 5, 10], 'min_samples_leaf': [1, 2, 4], 'bootstrap': [True, False]}
```

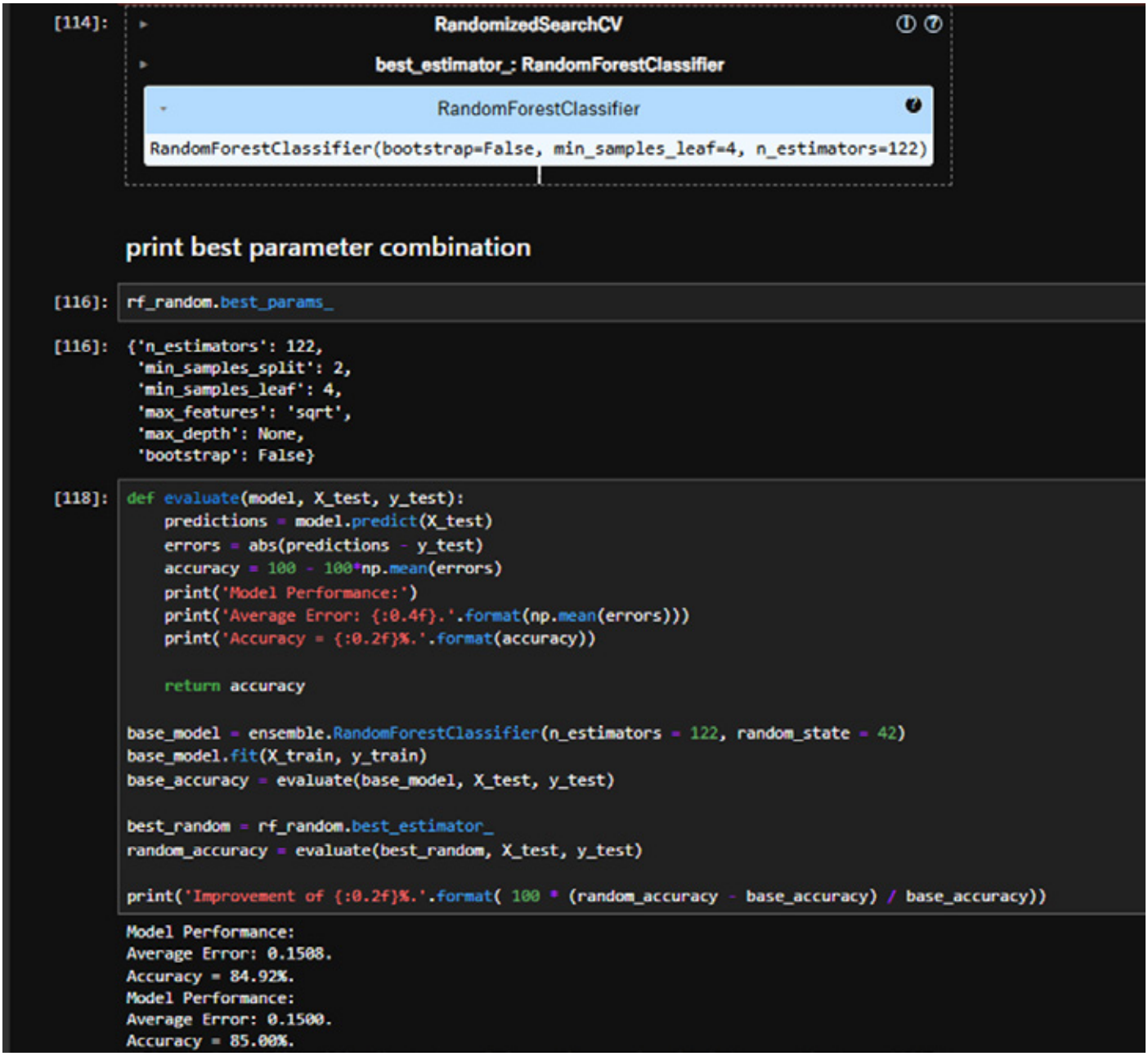
[112]:
```
## Use the random grid to search for best hyperparameters ##

## duration: 45 min ##

# First create the base model to tune
rf = ensemble.RandomForestClassifier()
print(rf.get_params()) #list all hyperparameters
```

```
{'bootstrap': True, 'ccp_alpha': 0.0, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features': 'sqrt', 'max_leaf_nodes': None, 'max_
samples': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'monotonic_cst': None, 'n_e
stimators': 100, 'n_jobs': None, 'oob_score': False, 'random_state': None, 'verbose': 0, 'warm_start': False}
```

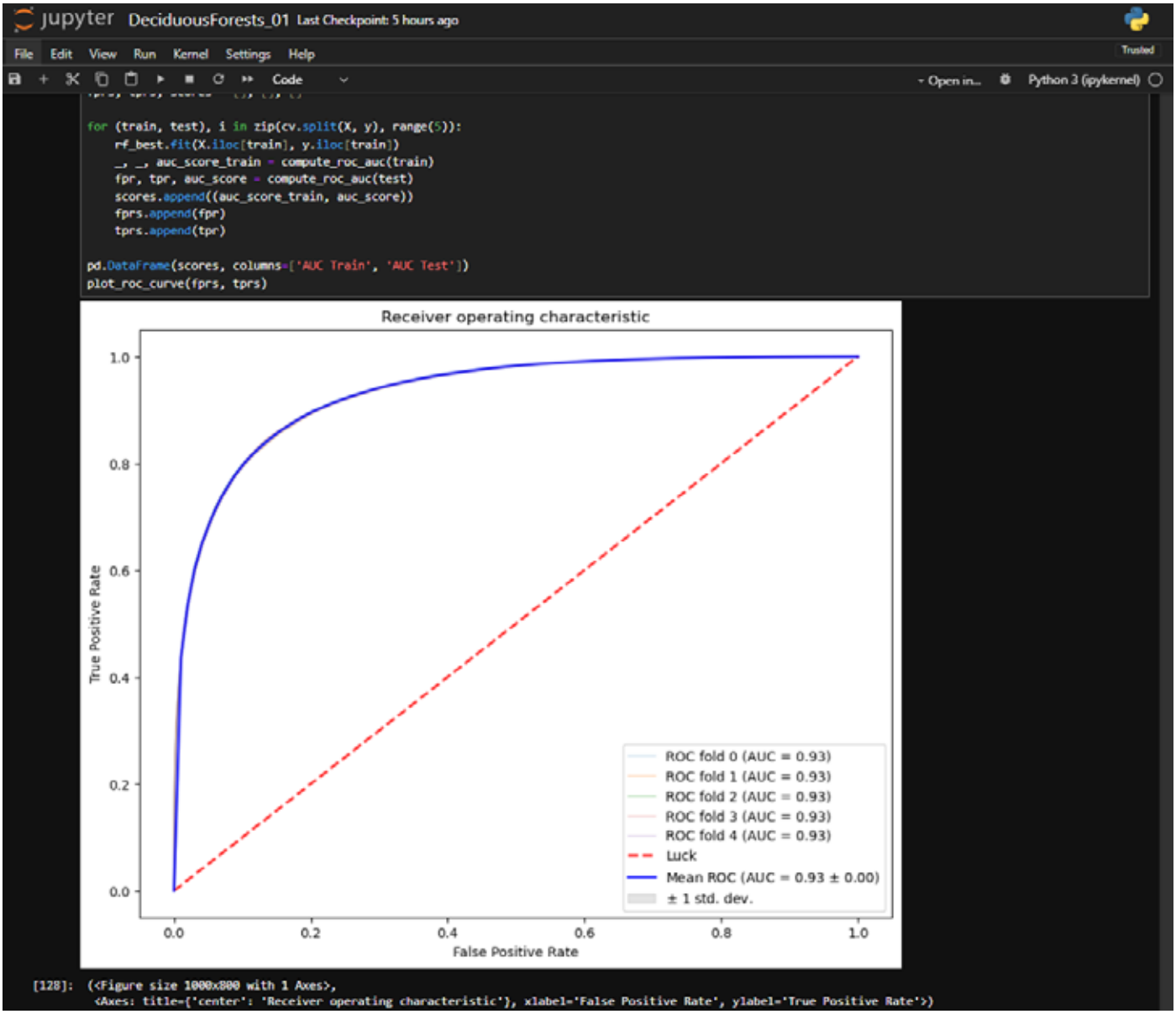
[114]:
```
# Random search of parameters, using 3 fold cross validation,
# search across 100 different combinations, and use all available cores
rf_random = RandomizedSearchCV(estimator = rf, param_distributions = random_grid, cv = 3, verbose=2, random_state=42)
# Fit the random search model
rf_random.fit(X_train, y_train)
```

```
Fitting 3 folds for each of 10 candidates, totalling 30 fits
```

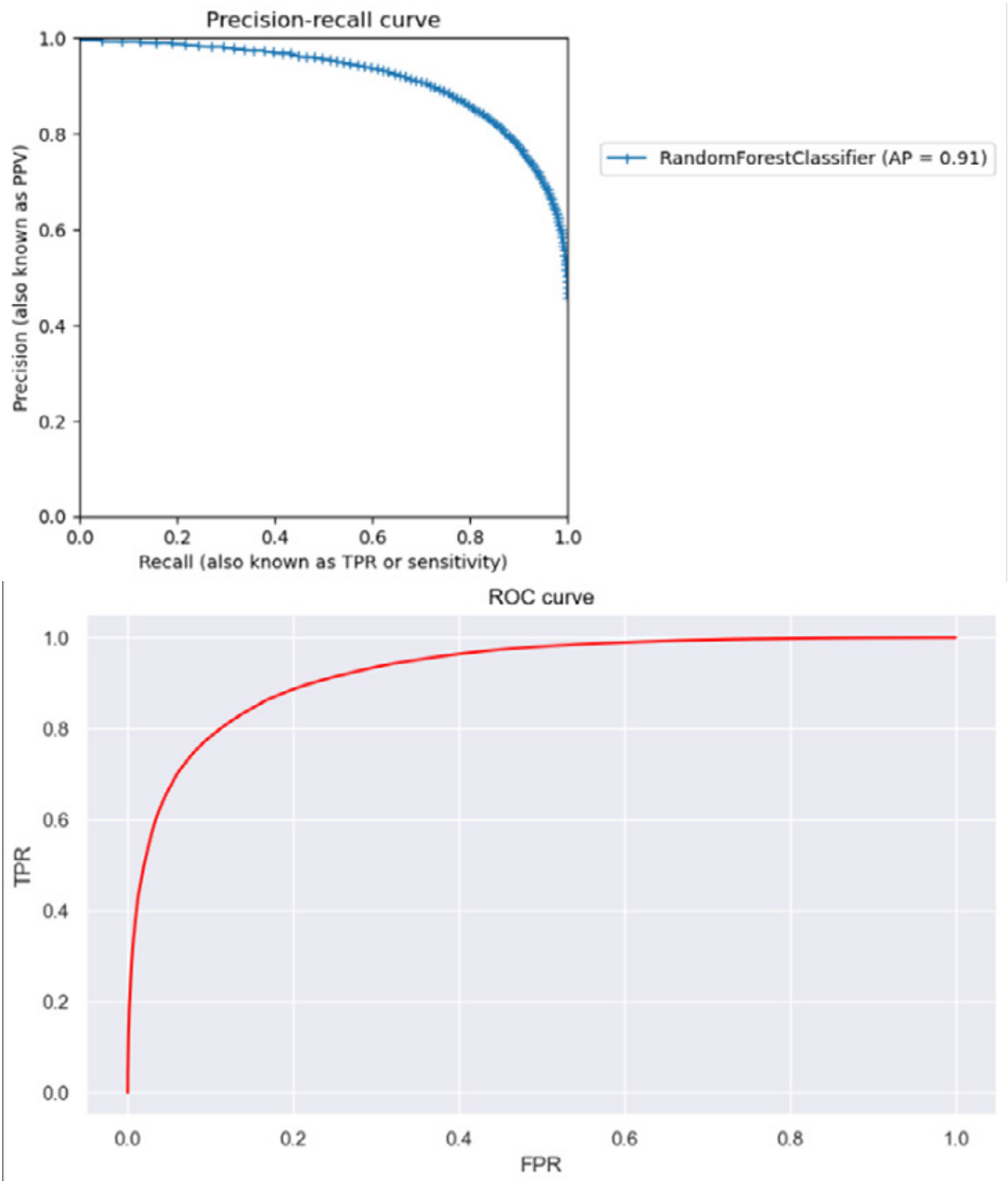
Appendix 1. Screenshot showing the script run to find out the best parameter's tuning of the RF model.



Appendix 2. Screenshot showing the ML results of best classifier, best parameter combination and its performance.



Appendix 3. Screenshot showing the ROC curve of the RF model results for one county section.



Appendix 4. Precision-recall curve and ROC curve.

www.skyddaskogen.se

CONTACT US

skogsmonitor@skyddaskogen.se
www.skogsmonitor.se