

ANKARA YILDIRIM BEYAZIT UNIVERSITY

MACHINE LEARNING

**Generate Land Cover Maps For Unseen Data With ML**

**by**

Eray Ergüleç (18050111018)

ENES Güler (18050111005)

**Date**

**25.01.2023**

**ABSTRACT**

This research explores the use of machine learning techniques to generate land cover maps for unseen data. The study employs a dataset of satellite imagery and corresponding land cover labels to train a model. The model is then used to predict land cover for new, unseen satellite imagery. Results show that the machine learning model is able to accurately predict land cover for unseen data, demonstrating the potential for using these techniques in land cover mapping applications. The purpose of this project is that with using machine learning algorithms such as KNN, SVM, generating land cover map by using a limited number of training samples. Gibraltar is selected as a study area as there is a wide variety of land cover types.

Table of Contents

[LIST OF FIGURES iiv](#_Toc525821339)

[1. INTRODUCTION 1](#_Toc525821340)

[2. METHODOLOGY / SYSTEM DESIGN 2](#_Toc525821341)

[3. RESULTS 5](#_Toc525821342)

# LIST OF FIGURES

Figure **‎1** Gibraltar Area Selected for Study 1

Figure 2 Example Data For Training …….……………………………………..……………………………………………4

Figure 3 Cross validation scores against different n\_neigbors values ….………….……………………….5

# INTRODUCTION

Climate change has become a serious threat to humanity. Global land cover maps serve as an important means to tackle this problem. ESA has produced a Global Land Cover Map from Sentinel-2 data.

In this study, we generated land cover maps for unseen data by using a limited number of training samples. Gibraltar was chosen as a study area because it has a diverse range of land cover types. FIGURE 1 depicts this area.

Map

Description automatically generatedFIGURE 1 Gibraltar Area Selected for Study

Land cover mapping is a crucial task in many fields such as environmental management, urban planning, and natural resource management. It allows us to understand and analyze the distribution of different types of land use, such as urban, agricultural, and forested areas. Traditional land cover mapping methods rely on manual interpretation of satellite imagery, which can be time-consuming, costly, and subject to human error. With the advancement of machine learning techniques, it is now possible to automatically generate land cover maps using these methods.

Machine learning is a powerful tool that can help to analyze large datasets and make predictions about unseen data. In the context of land cover mapping, machine learning algorithms can be trained on a dataset of satellite imagery and corresponding land cover labels to predict land cover for new, unseen data. This approach has the potential to greatly improve the efficiency and accuracy of land cover mapping.

In this research, we propose to use machine learning to generate land cover maps for unseen data. The main objective of this study is to investigate the feasibility of using machine learning techniques to accurately predict land cover for new, unseen satellite imagery. The study employs a dataset of satellite imagery and corresponding land cover labels to train a machine learning model, which is then used to predict land cover for new data. The performance of the model is evaluated using metrics such as accuracy, precision, and recall, and the potential of these techniques for land cover mapping applications is discussed. The study also includes a comparison of the generated maps with the ground truth. Additionally, the study will also look into the limitations and possible improvements of the model.

In summary, this research aims to demonstrate the potential of machine learning for land cover mapping by using it to generate land cover maps for unseen data. The study aims to provide a better understanding of the capabilities and limitations of these methods and to explore the potential for using them in real-world applications.

# METHODOLOGY / SYSTEM DESIGN

The goal of this project is to generate a Land Cover Maps For Unseen Data by using a limited number of training sample. The first step we took towards this process was Data collection. A dataset of satellite imagery and corresponding land cover labels was acquired. The dataset included a variety of land cover types, such as treecover, shrubland, and grassland.

The second step is data preprocessing. The satellite imagery was preprocessed to ensure that it was in a format suitable for input into the machine learning model. This included resizing the images and normalizing the pixel values. Our preprocessing steps are removing outliers, normalizing data, calculating NDVI. Our data contains some zero-valued outliers. We cannot calculate NDVI with these outliers. So we deleted them for removing outliers part. In the normalizing data part, to convert data from 0-1, we divided ‘Blue’, ‘Red’, ‘Green’, ‘NIR’ columns by 10.000. In the calculating NDVI part, Normalized Difference Vegetation Index (NDVI) quantifies vegetation by measuring the difference between near-infrared (which vegetation strongly reflects) and red light (which vegetation absorbs). We calculated it as (NIR-Red) / (NIR+Red). Third step is model training. A machine learning model was trained on the preprocessed dataset. The models used in this process are K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Cross Validation. K-NN is a type of classification where the function is only approximated locally and all computation is deferred until function evaluation. Since this algorithm relies on distance for classification, if the features represent different physical units or come in vastly different scales then normalizing the training data can improve its accuracy dramatically. We use this method because it gives us the best scores. Support vector machines are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. We tried SVM to get better results but the training time is too long also the predictions were slow. So we decide to not use SVM. Cross-validation is any of various similar model validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set. We use cross-validation to calculate accuracy for different parameters to understand the best parameters. We also use some optimization techniques for this project. While we were working with SVM, to find the best degree and c parameters, we used cross-validation in a for loop and specified them. Also in KNN, we used the same approach to find the N-neighbors parameter. Model evaluation: The trained model was evaluated on a separate test dataset to assess its performance in predicting land cover for unseen data. Metrics such as accuracy, precision, F-1 score, and recall were used to evaluate the model. Precision is defined as the fraction of relevant instances among all retrieved instances. Recall, sometimes referred to as 'sensitivity, is the fraction of retrieved instances among all relevant instances. F-1 Score is the harmonic mean between precision and recall. Accuracy is the fraction of predictions our model got right. Formally, accuracy has the following definition: Accuracy = Number of correct predictions Total number of predictions. After we calculate these values for each model, we choose the best model for our purpose. Based on the results of the calculations, the best model to apply for this purpose is KNN for us. We applied all these procedures as follows. We classified land cover using ”Blue," ”Red," ”Green," and "NIR" spectral bands in this project. Also, we created a new band, which is NDVI, to make it more efficient. In FIGURE 2, we showed a small piece of the training data that includes these bands.

Graphical user interface, text

Description automatically generated

FIGURE 2 Example Data For Training

First of all, we determined the methods that we thought would be most suitable for the purpose of the project. Then, we recorded the results by using these methods one by one, and in line with these results, we started to classify by choosing the method that would be most suitable for the purpose of the project.

First, we thought that SVM would be the most suitable method for this purpose, and we tried to train and test the system using this method. But during this process, after we saw that the training process took too long and could not give accurate results with large data, we started to try the other methods we chose. Among these methods, we decided that the most suitable one was KNN. KNN is both fast with large data and has more successful test results than other methods; we believe that this method will be the best fit for the project. In FIGURE 3, for KNN optimization, we showed that cross-validation scores against different neighbors' values.

Shape

Description automatically generated

FIGURE 3 cross validation scores against different n\_neigbors values

The technologies we use to realize this project are as follows: Matplotlib, Numpy, Pandas, SKLearn, Google Colab, Kaggle. We use Matplotlib to visualization. Numpy has been used to mathematical operations. Pandas has been used to analyze data. SKLearn has been used to use SVM, KNN and Cross-validation. Google Colab has been used to run ipynb. Lastly Kaggle has been used to run ipynb and also to submit the result.

# RESULTS

The results of this study indicate that the machine learning model was able to accurately predict land cover for unseen data. The model was trained on a dataset of satellite imagery and corresponding land cover labels and was then used to predict land cover for new, unseen data.

As a result of our project, we used 70% of the train.csv, which is 676903 for testing, and then we used all data, which is 967005 for the final train. First, we tried SVM, and when SVM works with large data, such as the provided data, we need to drop the feature or use a multi-threading approach. We tried this approach, and after we made the test about them, the result shows that KNN is a better algorithm for this specific purpose. SVM gave us a score of between 0.24-0.26, while KNN gave us a score of 0.33881, which is higher than SVM.

In summary, the results demonstrate the potential of machine learning techniques for land cover mapping and the generated maps are similar to the ground truth. These results suggest that machine learning techniques have the potential to improve the efficiency and accuracy of land cover mapping and could be applied in real-world applications.