Potential Coffee Production Hot-spots Using Machine Learning Techniques: Nagaland and Manipur, India

MSc Research Project
Data Analytics

Nitish Sharma
Student ID: x21145157

School of Computing
National College of Ireland

Supervisor: Dr Catherine Mulwa

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Nitish Sharma
Student ID:	x21145157
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Dr Catherine Mulwa
Submission Due Date:	1st February 2023
Project Title:	Potential Coffee Production Hot-spots Using Machine Learning Techniques: Nagaland and Manipur, India
Word Count:	7329
Page Count:	28

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Nitish Sharma
Date:	1st February 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Potential Coffee Production Hot-spots Using Machine Learning Techniques: Nagaland and Manipur, India

Nitish Sharma
x21145157

Abstract

Coffee is arguably one of the most consumed and traded beverages worldwide. The coffee plant is particularly climatically sensitive, requiring certain soil, altitude, and temperature conditions. The seeds of coffee fruits resembling cherries are roasted and prepared for export. As coffee's demand has steadily increased over the years, poor yields can disrupt the supply chain. Less area remains suited for coffee plantations due to global warming. Therefore, crops must be relocated to new places to accommodate the rising demand. Only the states of Karnataka, Kerala, and Andhra Pradesh grow coffee in India, making it one of the world's main coffee producers. The climate of the North East Indian States, Chhattisgarh, and sections of Maharashtra is also conducive to the growth of coffee plantations. On local PCs, classic research has employed GIS software and machine learning on raster data. This study investigates the use of geographical data and machine learning with cloud computing to identify alternative habitats. This discovery will enable the development of similar models for other crops and, as a result, the research will be scaled up using the cloud computing capability. While this approach largely eliminates the scalability difficulties of conventional geospatial analysis, cloud platforms are still undergoing development and offer fewer algorithms at present.

1 Introduction

1.1 Background and Motivation :

There is a rapid expansion of coffee culture around the world, with consumption increasing in both coffee-importing and producing countries. Coffee production in the world is led by Brazil, followed by Vietnam. India is also one of the leading producers of coffee. In India, it is grown primarily in three regions Karnataka, Kerala, and Tamil Nadu. New coffee-producing areas have been established in Andhra Pradesh and Odisha. In India, the Western and Eastern Ghats are prime coffee-growing territory because of the dense natural shade they provide. Often inter-cropped with spices such as cardamom, cinnamon, clove, and nutmeg as per coffee board of india [1]. Coffee plantations thrive on a slope, growing altitudes ranging between 1,000 m (3,300 ft) to 1,500 m (4,900 ft) above sea level for Arabica (premier coffee), and 500 m (1,600 ft) to 1,000 m (3280 ft). Ideally, both Arabica and Robusta need well-drained soil conditions rich in organic matter that

[1] https://www.indiacoffee.org/

is slightly acidic (pH 6.0–6.5). Slopes of Arabica tend to be gentle to moderate as per [2], while Robusta slopes are gentle to fairly level.

Many Indian farmers rely only on agricultural earnings to make ends meet. However, there are crops that are more profitable than the traditional ones. Coffee is one of the world's most lucrative crops, earning it the nickname "black gold." Farmers in India and other countries can benefit from coffee plantations by raising their earnings and stimulating economic growth as coffee consumption continues to climb. Coffee is a climate-sensitive plant, and as a result, its output has suffered as a result of global warming. Farmers have relocated their plantings to higher elevations on the slope in response to rising temperatures. As the demand for coffee continues to climb, the amount of land that can be used for coffee cultivation has decreased due to climate change. Even if coffee production is on the rise, less land is suited for farming, thus new locations must be identified. The study's primary objective is to identify potential new coffee-growing regions in India, focusing on the northeastern states of Nagaland and Manipur.

1.2 Research Question

This research attempts to evaluate the land and climate suitability for coffee production in various regions of India utilizing geospatial data consisting of raster images and vector data consisting of geo referenced coffee plantations in India. This project investigates the land, climatic, and other factors that influence coffee production by utilizing data from public domain websites, the cloud computing capacity of Google Earth Engine, and a machine learning-based method to identify possible coffee production hotspots. A hotspot is a location or area where the climatic conditions provide an optimum home for coffee plants, making the area suited for coffee cultivation. Using classification algorithms based on supervised machine learning, this study identifies clusters with comparable geoclimatic characteristics. These are the groupings of land that are excellent for coffee cultivation. A potential hotspot is identified as an area with a favorable environment based on the results of the algorithms.

RQ : With geospatial and remote sensing data for climatologic (temperature, rainfall), topographic (soil, terrain, elevation), and edaphological (soil type, soil organic content, soil pH) factors, it is possible to identify promising new locations for coffee production in India whilst using machine learning classification algorithms (smileRandomForest, smileGradientBoostTrees, smileCart) on cloud-based platforms ?

Sub RQ: Based on the identified potential coffee production regions can we subset the results to find hot-spots in the north eastern states of Manipur and Nagaland using classification algorithms on geospatial data.

1.3 Research Objective

To ensure that previously used methodologies and findings can be leveraged and incorporated into this research, as well as to attempt to identify substantial gaps in the reviewed literature, the research objectives incorporate and evaluate a critical evaluation of current literature on the themes. This was done in order to ensure that this research can

[2] //www.indiacoffee.org/

leverage and incorporate previously used methodologies and findings. Additional aims are outlined in Table 1, which is titled ”Research Objectives.”

Research Objective	Description
1	A critical review of literature on suitability studies on coffee production.
2	Studying of Coffee production trends across different states of India.
3	Assessment of Potential Coffee-Production Hot-spots in India using machine learning approach
3.a	Implementation and evaluation of Random Forest on geospatial data
3.b	Implementation and evaluation of Gradieng Boosting on geospatial data
3.c	Implementation and evaluation of Classiffication and Regression trees on geospatial data
4	Preparation of habitat distribution map of India
4.a	Preparation of habitat distribution map of Manipur and Nagaland
5	Preparation of a potential hotspot map of India.
5.a	Preparation of potential hotspot map of Manipur and Nagaland
6	Comparison of developed models (4)

Table (1) Research Objectives

2 Related Work

Coffee is a relatively less researched crop. It has not evolved to the extent as other crops. Thus there are only two major species of coffee being used around the world Arabica and Robusta . Coffee is a highly climate sensitive plant and global warming has impacted the production of coffee globally. This literature review evaluates various suitability studies for coffee production at various places around the world. This section comprises of various subsections(i) A review of factors affecting coffee production. All these factors have affect coffee production in some different proportions. These proportions can be determined using AHP and Machine learning based methods . Section 2 discusses (ii) A review of GIS and AHP based methods. Section 3 discusses (iii) A review of machine learning based methods (iv) A comparison of all the methods.

2.1 Factors determining coffee production:

As it is a tropical plant, coffee has highly precise requirements for the environment in which it must grow in order to be successful. According to what is said in, the environmental parameters that are optimal for coffee plantations can be broken down into three categories: climatological factors, physiographic factors, and edaphological variables (Salas López et al.; 2020). The climatological factors include things like average temperature, minimum temperature, maximum temperature, relative humidity, and dry period. Other climatological elements include these things as well. Edaphological parameters include things like soil acidity or pH, soil texture, stoniness, and organic matter.

These things are all related to the soil in some way. Elevation, the slope of the terrain, and the aspect of the terrain are the components that make up the physiographic factors. Other than these characteristics, a plantation should also be located in close proximity to a water supply in order to satisfy the water requirements, and the area that is being checked should not be in a national park or other type of designated forest. Different species of coffee are affected differently by these characteristics.

Robusta grows best at higher temperatures ranging from 20 degrees Celsius to 30 degrees Celsius with a relative humidity of 70–80 percent, while Arabica requires temperatures between 18 and 21 degrees Celsius. Along with having soil that is rich in organic matter, a seasonal rainfall of approximately 125 to 130 centimeters is required for a coffee crop to be effective. It takes a coffee plant around three years to mature enough to produce coffee fruit. According to what is stated in, the longer days and cooler temperatures that prevail during the dry season encourage the formation of flower buds (Hoffmann; 2018). The dry season causes flower buds to enter a state of dormancy, and plants do not produce flowers during this time. Following this period, the plants go through a period of pollination and flowering that lasts for about two weeks. This is followed by the maturation of coffee cherries. According to what is stated in, the full process takes approximately six to nine months for Arabica and nine to eleven months for Robusta (Wintgens et al.; 2004).

The rise in average worldwide temperature that can be attributed to human-caused climate change has also had an impact on the world's coffee plants. When temperatures rise, farmers relocate their plantings to higher elevations on the hill, where the weather is cooler. This results in a reduction in the amount of land that is suitable for plantation and farming. According to the findings of a number of studies conducted in this region, the amount of land suitable for coffee farming could decrease by as much as 50 percent by the year 2050 for both Arabica and Robusta varieties (Adhikari et al.; 2020). The majority of research that investigates the effects of climate change is conducted on a national or regional scale. According to one global impact assessment that was carried out (Bunn et al.; 2015), it is expected that the plantations will move towards higher altitudes worldwide. In addition to this, it forecasts a fifty percent decrease in the amount of suitable area. According to the findings of another study that was conducted in a same manner (Haryuni et al.; 2019) area, while Robusta could lose twenty-five percent. The hypothesis that Robusta can withstand higher temperatures for longer is called into question by this research as well. According to the findings of several studies, it may be beneficial to migrate the plants to circumstances resembling shaded tropical forests and to higher altitudes.

2.2 An analysis of existing GIS and AHP based suitability Analysis:

This research aims to find suitable area for coffee plantation based on its Climatological , physiological and Edaphological factors. These data related to these factors can be availed from open Geographical Information systems (GIS),open source satelite imagery. These data can be fed to Analytical Hierarchical Process (AHP) based models to determine the suitability of an area . AHP is a multicriteria decision method that creates a hierarchy for the decision .Each factor is assigned a weight in AHP based analysis. Many studies have been done to ascertain suitability of an area for an agricultural crop cultivation using AHP and GIS (Mendas and Delali; 2012),(Dedeoğlu and Dengiz; 2019),(Bhagat et al.; 2009),(Pilevar et al.; 2020),(Tashayo et al.; 2020),(Ostovari et al.; 2019),(Mandal et al.; 2020),(Salas López et al.; 2020) and (Pham et al.; 2021). One such study has been performed by (Rono and Mundia; 2016).This research has determined land suitability of ElgeyoMarakwet County in Kenya for coffee production. It takes into consideration various factors and finds a model to predict coffee production using GIS data, Satelite imagery using GIS modelling and Analytic Hierarchy Process (AHP) based models. This study found that Arabica coffee is more suited to be grown in the study area than Robusta coffee (Rono and Mundia; 2016). While the study has taken into factors all the necessary climatological factors , It does not account for increasing temperatures.

Similar study has been performed in Jamaica (Mighty; 2015) . This study evaluated the factors such as Elevation, Temperature, Geology, Soil Types, Slope, Precipitation, Distance to road and Distance to waterways. This study uses GIS systems , Multi Criteria Decision Analysis, and Analytical Hierarchical Process. This study found that Geology, precipitation and temperature are the most significant factors for coffee production. According to this study the most suitable areas for coffee were found to be mountains of central and eastern Jamaica (Mighty; 2015). This study is also based on GIS data and AHP based modelling techniques. Coffee plants take 5 to 7 years to blossom. The temperature can further increase in a span of 5 -7 years. This research does not takes into account the future temperature.

Another research done using AHP and GIS to find land suitability analysis for production of Potato Crop in the Jucusbamba and Tincas Microwatersheds (Iliquín Trigoso et al.; 2020). This research concluded that climatological, edaphological, topographical and socioeconomical factors effect potato production most in the area. AHP and GIS based models can be combined to determine suitability for a wide variety of purposes and not just agricultural 6 purpose. (Parry et al.; 2018) has performed land suitability analysis for urban services planning in Srinagar and Jammu urban centers in Jammu and Kashmir, India. (Doke et al.; 2021) has performed Geospatial mapping of groundwater potential zones in a hardrock basaltic terrain in India. Above mentioned researches also do not take into consideration the increase in temperature.

2.3 Review of Machine Learning approaches used in suitability studies

AHP can be integrated with Artificial Neural Networks(ANN) to predict results with superior accuracy. Numerous researches have employed AHP and ANN in conjugation . One such research done by (Taha and Rostam; 2011) combines AHP with Bayesian Network. In another research ANN has been used along with AHP by (Al-Barqawi and Zayed; 2008) to predict water sources. Some other researches like land suitability analysis done by (Oh et al.; 2011) for urban growth and prediction of ground water reserve area in a near saline water way use similar methodology. In another approach by (de Carvalho Alves et al.; 2022), Machin learning has been used to predict coffee production hotspots based on remote sensing and Mineral nutrition monitoring. This research uses samples of leaves and soil to determine the mineral and nutrient content and uses. Climatic data is collected through satellite data such as Landsat. It uses Classification and regression trees (CART), Random Forest algorithms. The research is confined to Arabica species only.

Research (Chemura et al.; 2021) discusses the impact of climate change on coffee production in Ethiopia. This research finds that , the area suitable for cofee will remain stable under all scenarios and there will be minimal impact. The research analyses both current and future climatic data and finds the impact on Ethiopian coffee production, The research however does not find any significant impact and does not do any suitability analysis on potential areas.

2.4 Summary of the researches studied

The researches discussed in this section are summarized in the table Table 2 – Summary of Research. This completes the research objective 1 specified in Table 1 - Research Objectives.

Year	**Title**	**Method Used**
2021	AHP-GIS and MaxEnt for delineation of potential distribution of Arabica coffee plantation under future climate in Yunnan, China	GIS, AHP, Maxent
2020	Land Suitability for Coffee (Coffea arabica) Growing in Amazonas, Peru: Integrated Use of AHP, GIS and RS	GIS, AHP
2022	The role of machine learning on Arabica coffee crop yield based on remote sensing and mineral nutrition monitoring	GIS, Classification and regression trees (CART), Random Forest
2015	Modeling the climate change impacts on global coffee production	GIS , Maxent, SVM, Random Forest
2022	Coffee-Yield Estimation Using High-Resolution Time-Series Satellite Images and Machine Learning	Satelite Imagery, SVM, Random Forest, Multi Linear Regression

Table (2) Summary of Research

3 Methodology

The data that was gathered for this study was stored in csv files as vectors and geographic raster pictures respectively. The geographical coordinates of the coffee plantations in India are included in the csv vector files. After that, the data is processed and altered so that it may be used for the project. In the section titled "Implementation," the processes of "data pre-processing" and "data transformation" are broken down in detail. The collected data is then input into machine learning models so that the clusters may be identified. After this step, the application of supervised machine learning classification is carried out in order to ascertain the prospective habitats that are suited for coffee production. The results of the algorithms are used to generate a map of possible hotspots. After examining the data at a more abstract level, a more in-depth analysis is carried out on the states of Nagaland and Manipur in order to further hone in on the possible hotspot spots.

3.1 Cloud Based Coffee Distribution model

For the purpose of determining the most suitable environment for coffee plantations in India, this study makes use of a modified CRISP-DM model that is based on the cloud and is suitable for species distribution models. This model is referred to as a cloud-based coffee distribution model. This research is carried out in a cloud setting with the intention of exploiting the capability of cloud computing on big raster images, which would normally be challenging to comprehend on a standard computer. In order to identify possible hotspots, this research takes a number of different aspects into consideration. The methodology is summed up in the diagram labeled "Figure 1 - Cloud Based Coffee Distribution Model."

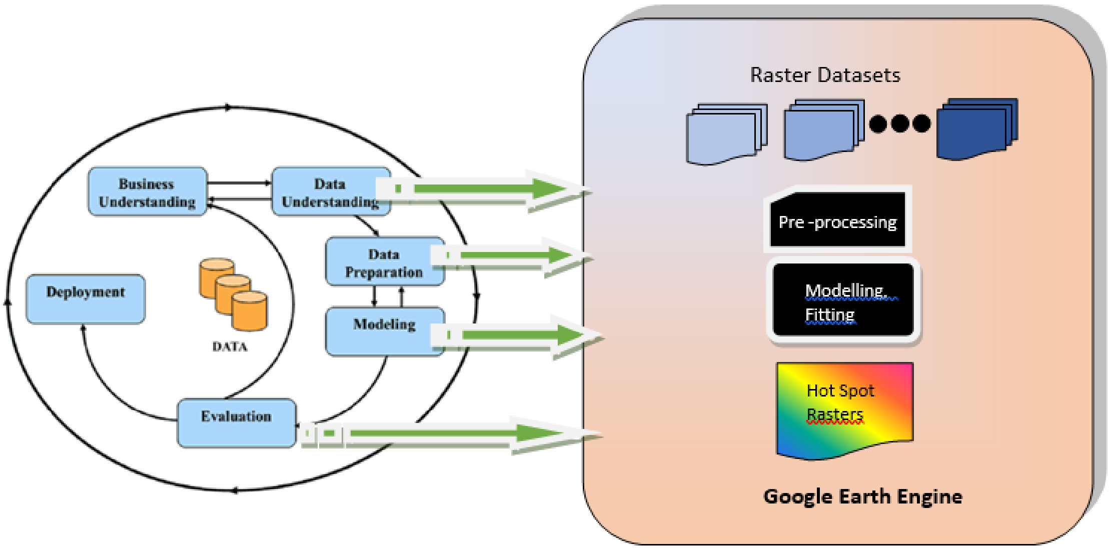


Figure (1) Cloud Based Coffee Distribution Model

3.2 Process Flow

This investigation seeks India's next coffee manufacturing hubs. The first stage is learning about coffee cultivation and production factors. Next, each aspect's data is collected. The Google Earth Engine compiles raster data for geospatial research from many publicly available sources. The GBIF website [3] is used to obtain location coordinates. For trend analysis, Tableau visualizes Coffee Board of India data. Additional dataset information is in the "Implementation" chapter. Coffee output is highly influenced by factors. To meet machine learning model requirements, these data are processed and displayed. A hotspot is a categorized region that has been researched to see if it can grow coffee. Evaluation and research are being done on classification algorithm outcomes. Figure 2 shows the project flow.

[3] https://www.gbif.org/

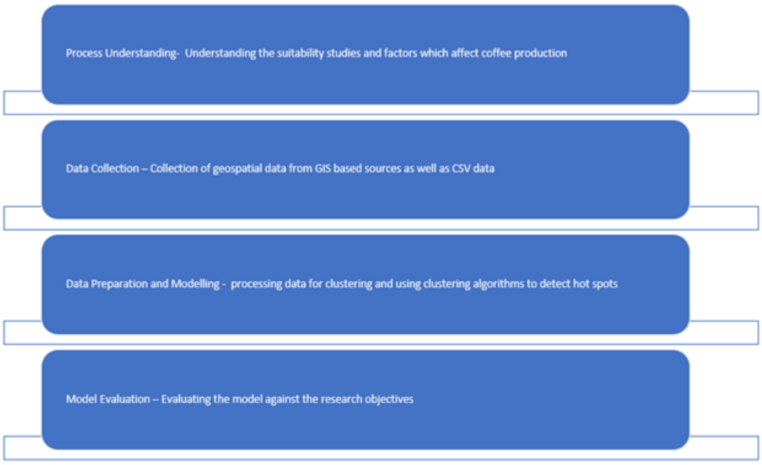

Process Understanding- Understanding the suitability studies and factors which affect coffee production
Data Collection – Collection of geospatial data from GIS based sources as well as CSV data
Data Preparation and Modelling - processing data for clustering and using clustering algorithms to detect hot spots
Model Evaluation – Evaluating the model against the research objectives

Figure (2) Project Flow

4 Design Specification

Figure 3 demonstrates the Three-Tier Architecture that was used in this assessment exercise. The Data Layer is located at the very bottom of the structure, and it serves as the foundation for the collecting of data points. This layer obtains its information for geo locations of coffee plantation places in India from GBIF[4], and it obtains its information for coffee production from the Coffee Board of India [5]. On top of the data layer is where the business logic layer is located. This layer is home to all of the data pre-processing and modelling code, as well as the business logic that drives those processes. The Random Forest Classifier, the Gradient Boosting Classifier, and the CART Classifier provided by earth engine and are used in this implementation. The Mean Area Under Curve (Mean AUC) measure is utilized in order to perform the analysis on the classification's final results. The final products of this layer consist of evaluation metrics as well as TIFF images that show the possible distribution of hotspots. The business model layer visualizations are supplemented by an additional layer called the data visualization layer, which is the topmost layer. Tableau and Python are included in this layer in order to facilitate data visualization. The Raster package in Python is used to read and interpret TIFF pictures, whilst Tableau is utilized for visualizing coffee production statistics.

[4] https://www.gbif.org/

[5] https://www.indiacoffee.org/

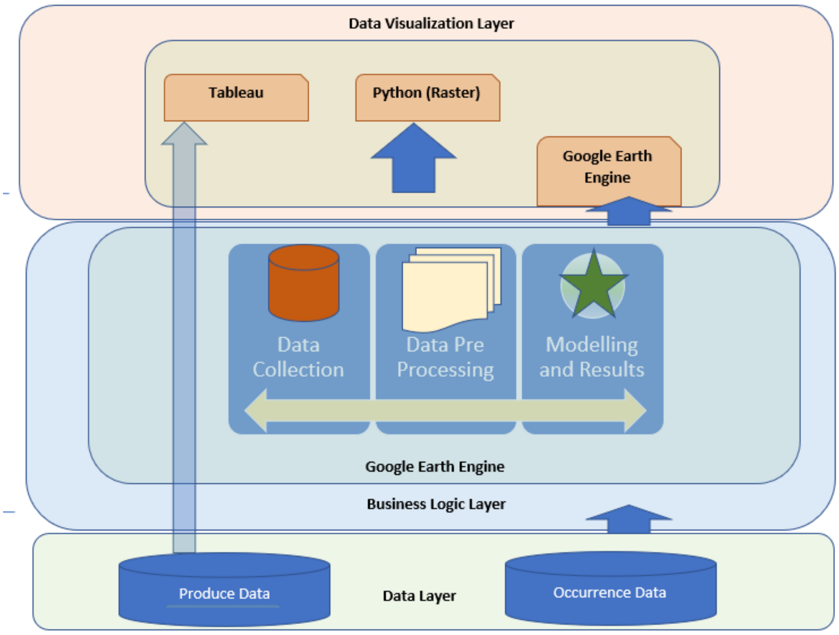


Figure (3) Design Specification

5 Implementation

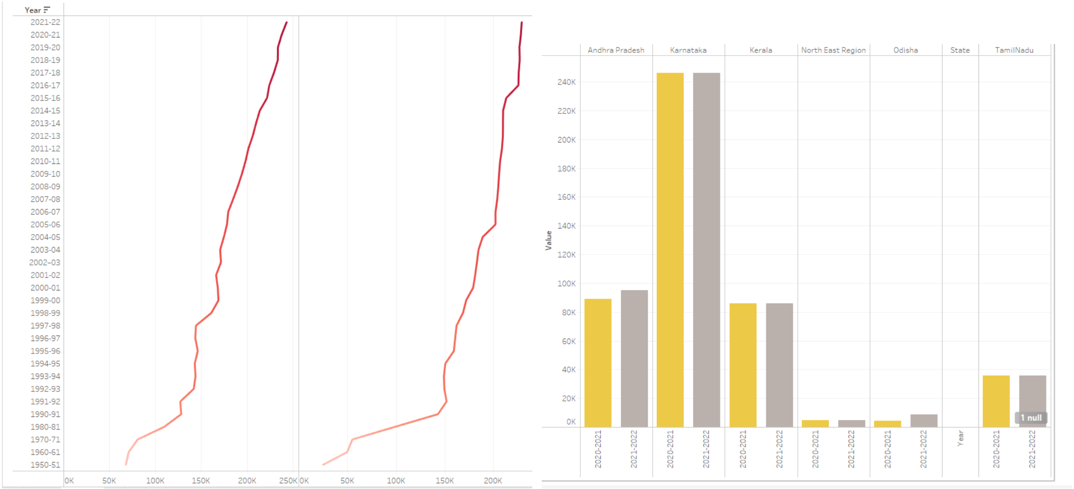
5.1 Introduction

Species distribution models are helpful in finding probable habitat for a species. They (SDMs) are numerical tools that combine observations of species occurrence or abundance with environmental estimates (1(Bellard et al.; 2012). Traditional studies have employed a variety of tools as discussed in section 2. This implantation utilizes cloud computing power of Google Earth Engine (GEE). The occurrence data and produce data is gathered in the base tier. The business logic is written in earth engine using java script which forms the business model and mapping tier. This tier also caters to visualizations created during various phases in the implementation. The top most tier is data visualization tier where visualizations are created in tableau utilizing the produce data sourced from CBF[6] and using raster library in python to read the TIF files exported after the implementation.

5.2 Gathering data from Datasets

Geo referenced coffee farms or locations forms the basis of Occurrence data in the implementation is downloaded using python from GBIF[7]. A total of 55 geo referenced locations are taken into taken. It is essential to study current trends and analyse the production of coffee in India. The produce is gathered at taluka level from coffee board of India[8]. It has data spanning from 1950 till 2022 at a state level. Analysing this data in tableau suggests that rtae of arabica production is slowing down and non traditional states of Orissa and Andhra Pradesh have shown a significant growth as shown in figure 4.

[6] https://www.indiacoffee.org/
[7] https://www.gbif.org/
[8] https://www.indiacoffee.org/



(a) Year Wise Coffee Production Of Arabica And Robusta

(b) State Wise Coffee Production for Year 2021 And 2022

Figure (4) Coffee Production Data

Figure 5 depicts the geo referenced coffee occurrence locations or presence locations. The presence points are represented in red dots in the map. As most of the coffee is produced in the southern states of India in Karnataka, Kerala and Tamil Nadu, Prescence points are concentrated around these states.

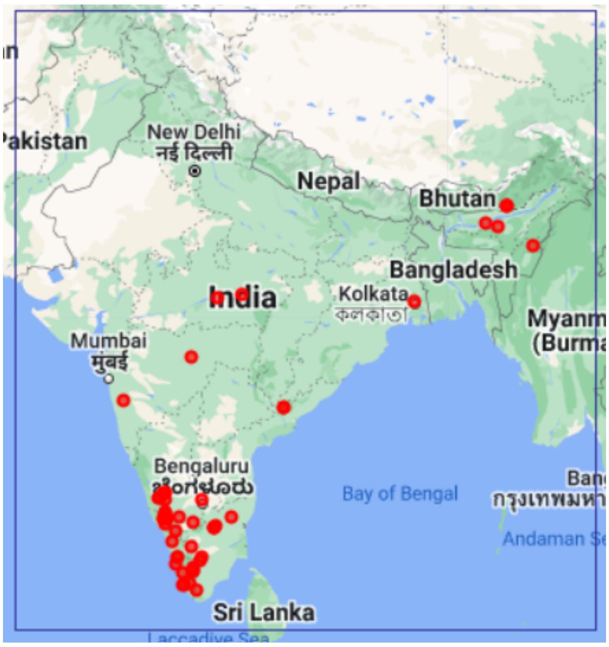


Figure (5) Prescence Points

Coffee Board of India has listed factors crucial to coffee production [9] .These factors include soil condition, slopes, elevation, aspect, Temperature ,Annual Rainfall, Blossom Showers and Backing Showers. For a good coffee habitat deep, fertile soil, rich in organic matter, well drained and slightly acidic (Ph6.0-6.5) is needed. The slopes need to be

[9] https://www.indiacoffee.org/coffee-regions-india.html

gentle to moderate with elevation of 1000 - 1500 m for Arabica and 500 – 1000 m for Robusta and aspect of North, East and North East for both breeds. It also mentions the annual rainfall should be between 1600-2500 mm for Arabica coffee and 1000-2000 mm for Robusta coffee. It also suggests that annual temperature range should be 15 -25 degree Celsius for Arabica and 20 – 30 degree Celsius for Robusta coffee. All these data have been taken into consideration in the implementation of the species distribution model. There are numerous raster and vector datasets in form of Image ,ImageCollection and FeatureCollection available on google earth engine. Images are single images where as ImageCollections are stacks or sequence of images. FeatureCollection are usually vector data containing attributes related to a geographical location which can be a polygon, point , line or any other vector data .As the data can be uncertain during covid times, The data for this Implementation takes predictor values from Jan 2018 to Dec 2020 . The predictors used in this implementation and GEE datasets used to get their values are listed in Table 3 – Datasets.

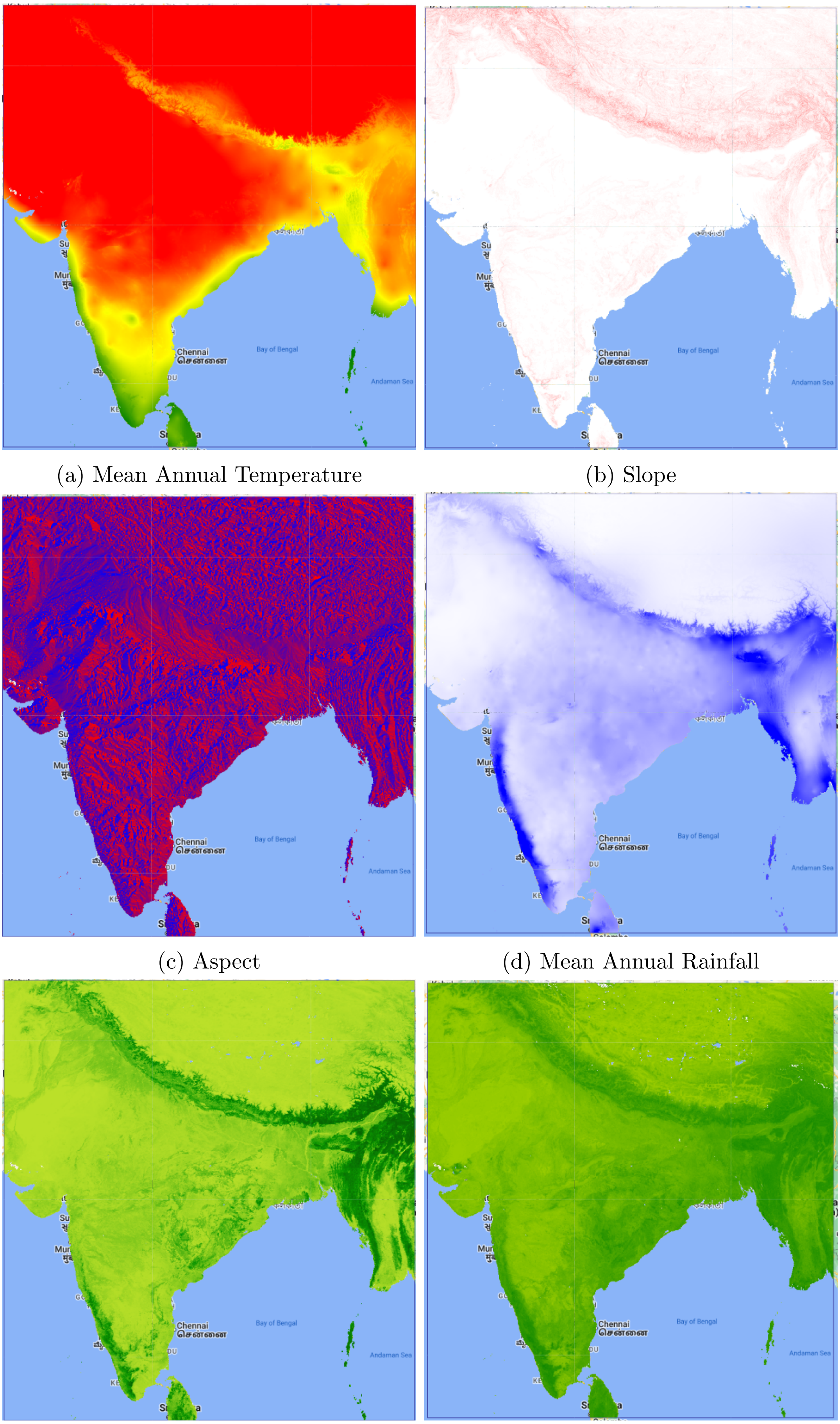
Predictor	**Dataset Type/ Dataset**	**Dataset Description**
Temperature	ImageCollection/ WORLDCLIM/V1/MONTHLY WorldClim V1	Bioclim provides bioclimatic variables that are derived from the monthly temperature and rainfall in order to generate more biologically meaningful values. It contains 4 bands tmin for minimum temperature, tmax for maximum temperature, tavg for average temperature and perc for mean annual precipitation. The data is spread across these four bands averaged on a monthly basis. For calculation of the temperatures a mean of the temperature is taken for a filtered collection spanning between Jan 2018 – Dec-2020.
Elevation	Image/CGIAR/ SRTM90$_V$4	The original intention behind the production of the Shuttle Radar Topography Mission (SRTM) digital elevation collection was to deliver consistent and high-quality elevation data with a scope that was as close to worldwide as possible. This particular iteration of the SRTM digital elevation data has been processed to remove any gaps in the data as well as make its use more straightforward. It is a single band image consisting of elevation.
Aspect	Image/USGS/SRTMGL1$_0$03	The digital elevation data produced by the Shuttle Radar Topography Mission (Farr et al.; 2007) is the result of an international research effort that attempted to obtain digital elevation models on a scale that was close to global. With a resolution of 1 arc-second, this SRTM V3 product, also known as SRTM Plus, is made available by NASA JPL. Aspect is calculated using feature engineering on elevation data by applying terrain functions on elevation data and singling out aspect band.
Slope	Image/USGS/SRTMGL1$_0$03	Slope is calculated from elevation data using feature engineering on elevation data.Terrain function is applied and slope band is selected .
pH$_{Scale}$	Image/ OpenLandMap/SOL/SOL$_PH - H2O_USDA - 4C1A2A_M/v02$	pH of the soil measured in H2O at six different depths (0, 10, 30, 60, 100, and 200 cm) with a resolution of 250 meters.

Predictor	Dataset Type/ Dataset	Dataset Description
Soil Organic Content	Image/ OpenLandMap/SOL/ SOL$_{O}RGANIC - CARBON$ $_{U}SDA - 6A1C_{M}/v02$	The amount of organic carbon in the soil, expressed as x 5 g/kg, was measured at six different depths (0, 10, 30, 60, 100, and 200 cm) with a resolution of 250 m. a conclusion reached after compiling soil data from around the world.
Rainfall	ImageCollection/WORLDCLIM/V1/BIO	The WorldClim Version 1 Bioclim offers bioclimatic variables that are obtained from the monthly temperature and rainfall in order to provide more biologically meaningful data.It contains of data in various bands. Rainfall data used in this implementation is contained in band bio12 i.e Mean Annual Rainfall and bio13 being used to find rainfall during rainy season.
Hillshade	MODIS/006/MOD44B	The Terra MODIS Vegetation Continuous Fields (VCF) product is a representation of surface vegetation cover estimations on a global scale that is done at the sub-pixel level. Developed to continually portray the terrestrial surface of the Earth as a fraction of fundamental vegetation characteristics. Hillshade is present as a band in the image collection.
Percent Tree Cover	MODIS/006/MOD44B	Percent Tree cover is calculated by selecting percent tree cover band in the above image collection. For this the collection is filtered by taking mean for a time period of Jan - 2018 to Dec to 2020.

Table (3) Datasets

5.3 Data Pre Processing and Feature Engineering

Occurrence Data is uploaded as assets in Google Earth Engine. The occurrence data consists of species information and coordinates. While uploading the data columns containing coordinates are specified. In order to limit the potential effect of geographic sampling bias on the model output due to data aggregation resulting from multiple nearby observations, data points belonging to same pixel are observed and only one random point is selected from them. For this we remove the duplicate observations in the sample. After this we are left with 50 unique Occurrences. The data points are spread across the country hence we form an area of interest around these data points by creating a 1 km buffer. It is essential to check that there is not a substantial correlation between any of the predictor variables, as this could lead to collinearity. In order to take this into account, we estimate the Spearman correlation among the predictor variable values at 5000 randomly chosen sites. It is found that minimum temperature, mean temperature and maximum temperature are correlated . Similarly rainfall in rainy season and mean annual rainfall are also correlated. So these predictors are removed and only remaining predictors are taken into consideration. All these predictors are plotted in Figure 6 below.



(a) Mean Annual Temperature

(b) Slope

(c) Aspect

(d) Mean Annual Rainfall

(e) Percentage tree cover

(f) Soil pH

Figure (6) Predictors

The regions in which pseudo-absences can be formed are identified. The model is populated with pseudo absences. To narrow the area for the formation of pseudo-absences, a two-step environmental profiling strategy is utilized. First, a k-means clustering based on Euclidean distance is conducted on the presence data, which generates random pseudo-absences within the pixels that are assessed as being more dissimilar to the presence data. In order to study the data a grid is created around the area of Interest. In this case study, a block repeated split-sample cross-validation strategy is employed to randomly divide data for model training and validation (2,3). Using random block splits, many model iterations are executed to generate training and validation data sets.

5.4 Modeling

This solution employs a repeating (10-times) spatial block cross validation technique, arbitrarily dividing 50x50km geographical blocks for model training (70 percent) and validation (30 percent) while generating pseudo-absences at random for each iteration.

In order to conduct an analysis on each of the data sets, an equal number of pseudo-absences and occurrence data were constructed for each. In order to guarantee that there will be an equal number of pseudo-absences, an arbitrarily large sample of 50,000 random points will first be generated. This will allow for pixels to be discarded that fell outside land area (within the ocean or lakes) within the spatial blocks that will be used for the training of the model. After getting rid of these masked pixels, the number of pseudo-absence points for a given data set were restricted so that it equalled the number of presence points. After that, a random forest model, a gradient boosting model, and a CART Classification model are used to determine how well the model fits the data. These models are executed for a total of five times, after which the mean habitat appropriateness of all five runs is determined.

6 Evaluation

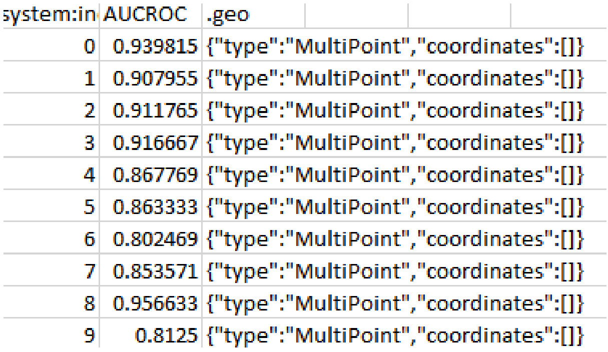
6.1 Introduction

GEE provides 4 classification techniques namely smileRandomForest, smileGradientTreeBoost, smileCart and smileNaiveBayes. smileNaiveBayes returns poor accuracy and has not been used in this implementation. This Implementation uses cases of smileRandomForest, smileGradientTreeBoost, smileCart classification implementations. Accuracy of these models is assessed using Area Under the Curve of the Receiver Operator Characteristic (AUC-ROC) (Fielding and Bell; 1997) and the Area Under the Precision-Recall Curve (AUC-PR) (Sofaer et al.; 2019) for each run using the validation data sets. Mean AUC-ROC and Mean AUC-PR for the n iterations for each models is calculated thereafter. A potential distribution map is plotted for each of these models. The result is clipped to the states of Nagaland and Mizoram to study the results.

Sensitivity (correct predictions of the occurrence) and Specificity or correct predictions of the absence are extracted . These metrics require defining a threshold. For each iteration, the threshold that maximizes the sum of sensitivity and specificity is used. Both sensitivity (the ability to make accurate forecasts of the occurrence) and specificity (the ability to make accurate predictions of the absence) are retrieved and exported to google drive or other cloud account. For this implementation the exports are rendered to google drive.

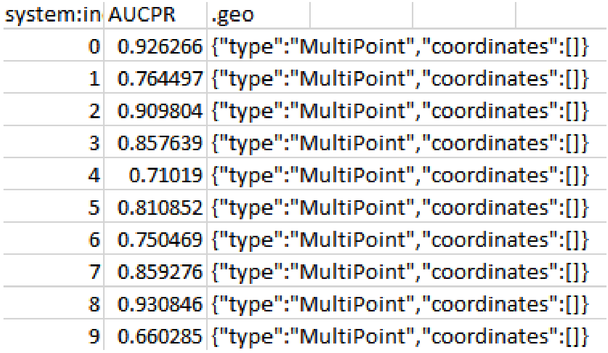
6.2 Case Study 1:Evaluation of results for smileRandomForest classifier

According to 4 A random forest is a set of tree predictors designed in such a way that each tree in the forest is dependent on the values of a random vector that is sampled randomly from each tree in the forest but follows the same distribution for each tree. When there are a lot of trees in a forest, the generalization error tends to get smaller and smaller until it reaches some sort of cap. For smileRandomForest classification in GEE numberOfTrees parameter which is the number of decision trees to create, has been set to 500. The number of decision trees to create. minLeafPopulation which tells the classifier only to create nodes whose training set contains at least this many points, has been set to 10. bagFraction ,the fraction of input to bag per tree has been kept to default value of 0.5. The tables Table 4 and Table 5 show the AUC-ROC and AUC-PR metrics respectively for the 10 iterations in the random forest classifier.



system:in	AUCROC	.geo
0	0.939815	{"type":"MultiPoint","coordinates":[]}
1	0.907955	{"type":"MultiPoint","coordinates":[]}
2	0.911765	{"type":"MultiPoint","coordinates":[]}
3	0.916667	{"type":"MultiPoint","coordinates":[]}
4	0.867769	{"type":"MultiPoint","coordinates":[]}
5	0.863333	{"type":"MultiPoint","coordinates":[]}
6	0.802469	{"type":"MultiPoint","coordinates":[]}
7	0.853571	{"type":"MultiPoint","coordinates":[]}
8	0.956633	{"type":"MultiPoint","coordinates":[]}
9	0.8125	{"type":"MultiPoint","coordinates":[]}

Figure (7) AUCROC Random Forest



system:in	AUCPR	.geo
0	0.926266	{"type":"MultiPoint","coordinates":[]}
1	0.764497	{"type":"MultiPoint","coordinates":[]}
2	0.909804	{"type":"MultiPoint","coordinates":[]}
3	0.857639	{"type":"MultiPoint","coordinates":[]}
4	0.71019	{"type":"MultiPoint","coordinates":[]}
5	0.810852	{"type":"MultiPoint","coordinates":[]}
6	0.750469	{"type":"MultiPoint","coordinates":[]}
7	0.859276	{"type":"MultiPoint","coordinates":[]}
8	0.930846	{"type":"MultiPoint","coordinates":[]}
9	0.660285	{"type":"MultiPoint","coordinates":[]}

Figure (8) AUCPR Random Forest

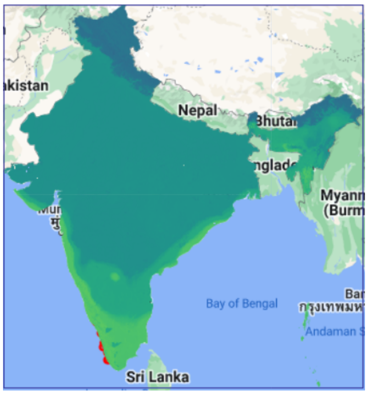


Figure (9) Habitat Suitability Random Forest

The Mean AUCROC obtained in the 10 iterations is 0.88324 and Mean AUCRPR obtained in 10 iterations is 0.81801 . The most suitable habitat and potential hotspot map predicted by Random Forest is shown in figure 12, figure 13 and figure 14. Figure 12 shows distribution map for India and Figure 13 and 14 show the maps for the states of Nagaland and Manipur. This accomplishes objective 3.a of this research outlined in table 1.

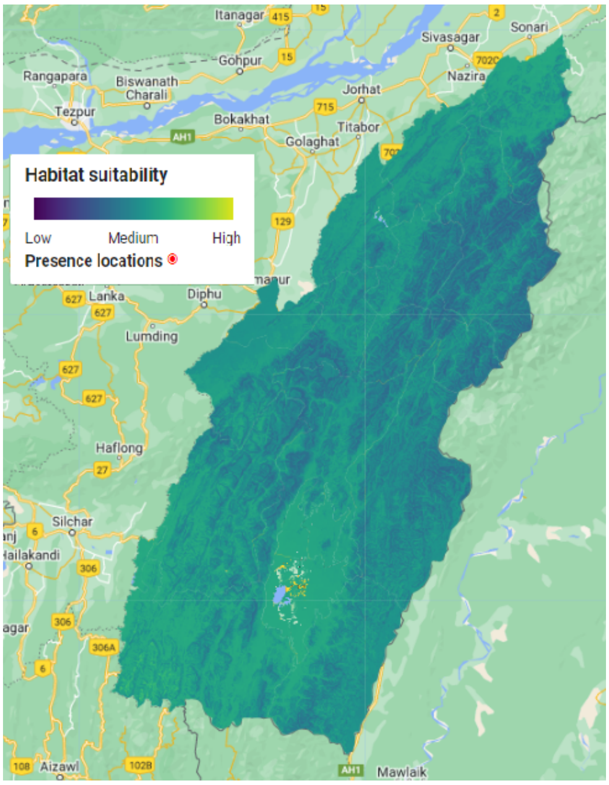


Figure (10) Habitat Suitability Random Forest in Manipur and Nagaland

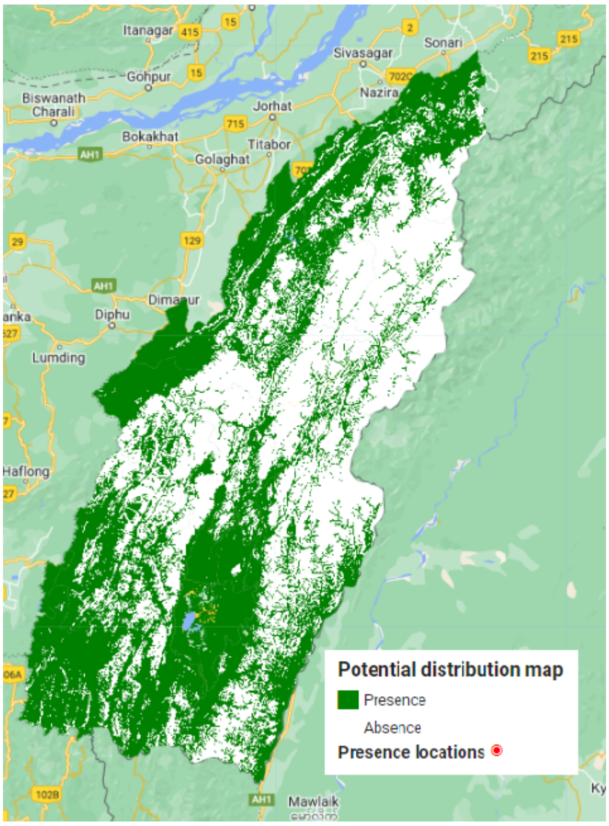
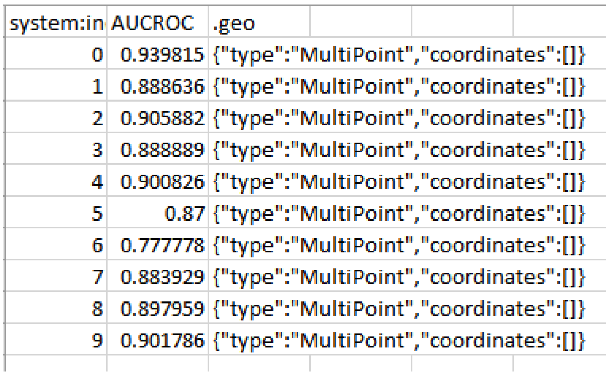


Figure (11) Potential Hotspots Random Forest in Manipur and Nagaland

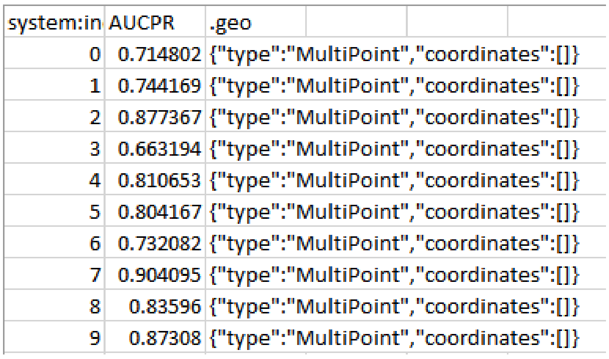
6.3 Case Study 2:Evaluation of results for smileGradientTreeBoost classifier

Gradient boosting are a class of machine learning algorithms that can be applied to a variety of different problems, including regression and classification. It provides a prediction model in the form of an ensemble of low-powered prediction models, which are decision trees. These are known to outperform the results of random forests and provide better results. For this implementation , numberOfTrees which is number of decision trees to create has been kept to 500. Other parameter values have been left to default. Table 6 and Table 7 show the AUC-ROC and AUC-PR metrics respectively for the 10 iterations using the Gradient Boosting classifier.



system:in	AUCROC	.geo
0	0.939815	{"type":"MultiPoint","coordinates":[]}
1	0.888636	{"type":"MultiPoint","coordinates":[]}
2	0.905882	{"type":"MultiPoint","coordinates":[]}
3	0.888889	{"type":"MultiPoint","coordinates":[]}
4	0.900826	{"type":"MultiPoint","coordinates":[]}
5	0.87	{"type":"MultiPoint","coordinates":[]}
6	0.777778	{"type":"MultiPoint","coordinates":[]}
7	0.883929	{"type":"MultiPoint","coordinates":[]}
8	0.897959	{"type":"MultiPoint","coordinates":[]}
9	0.901786	{"type":"MultiPoint","coordinates":[]}

Figure (12) AUCROC - Gradient Boosting



system:in	AUCPR	.geo
0	0.714802	{"type":"MultiPoint","coordinates":[]}
1	0.744169	{"type":"MultiPoint","coordinates":[]}
2	0.877367	{"type":"MultiPoint","coordinates":[]}
3	0.663194	{"type":"MultiPoint","coordinates":[]}
4	0.810653	{"type":"MultiPoint","coordinates":[]}
5	0.804167	{"type":"MultiPoint","coordinates":[]}
6	0.732082	{"type":"MultiPoint","coordinates":[]}
7	0.904095	{"type":"MultiPoint","coordinates":[]}
8	0.83596	{"type":"MultiPoint","coordinates":[]}
9	0.87308	{"type":"MultiPoint","coordinates":[]}

Figure (13) AUCPR - Gradient Boosting

The Mean AUCROC obtained in the 10 iterations is 0.88555 and Mean AUCRPR obtained in 10 iterations is 0.89595 . The most suitable habitat and potential hotspot map predicted by Gradient Boosting is depicted.

Figure 15 shows distribution map for India and Figure 16 and 17 show the maps for the states of Nagaland and Manipur. This accomplishes objective 3.b of this research outlined in table 1.

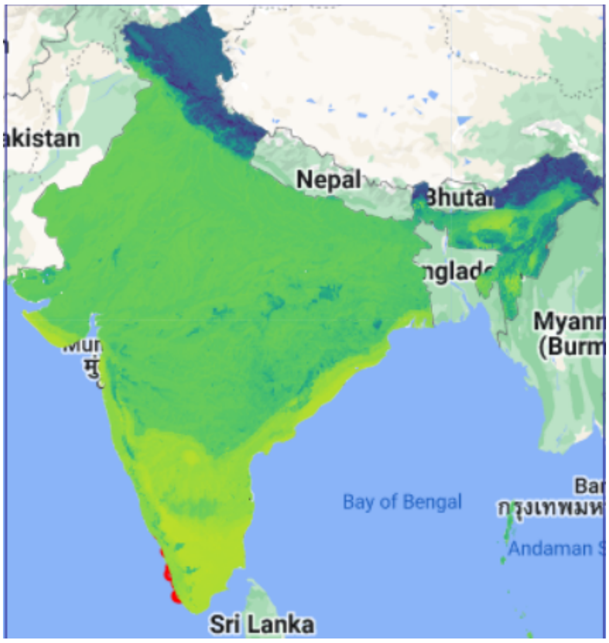


Figure (14) Habitat Suitability Gradient Boosting

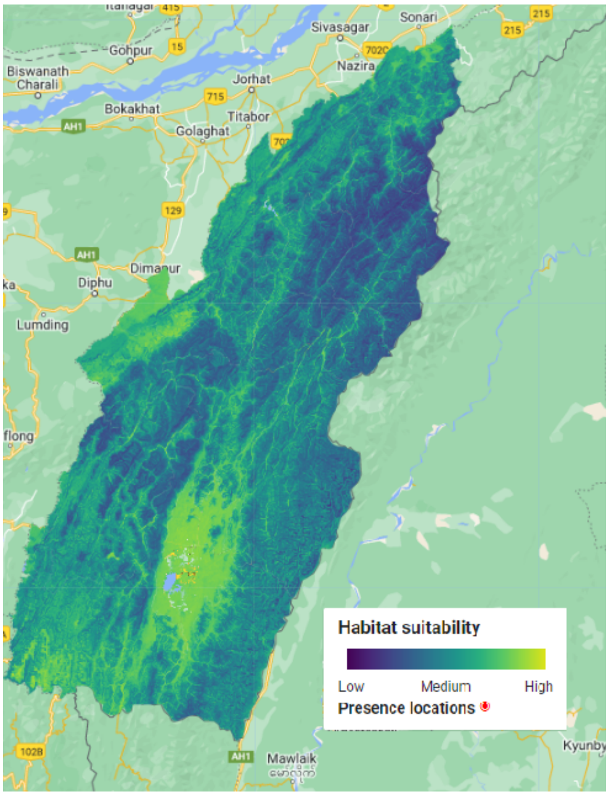


Figure (15) Habitat Suitability Gradient Boosting

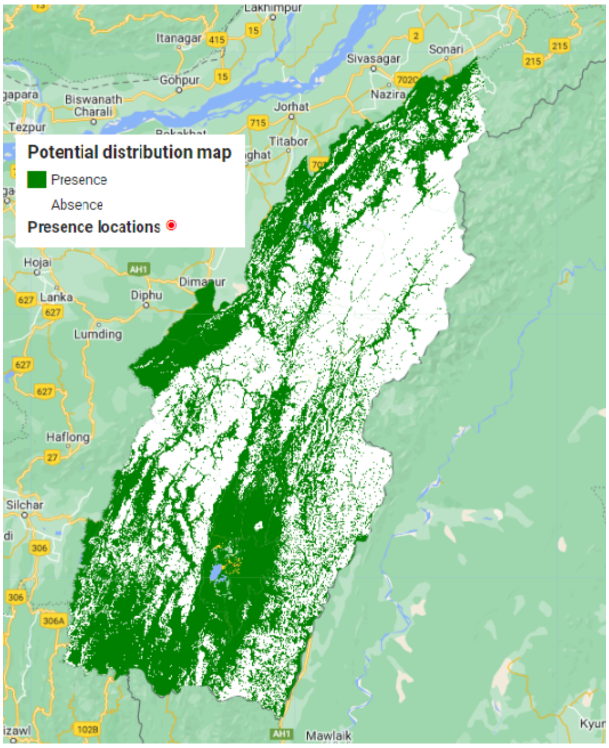
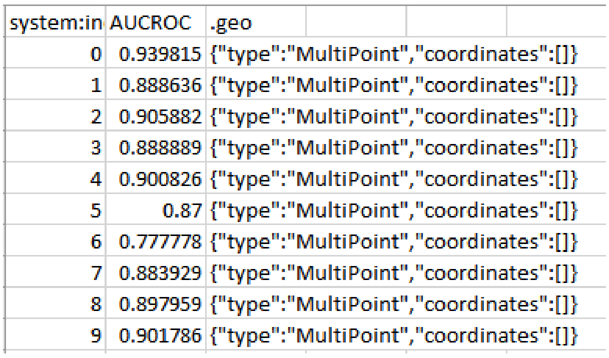


Figure (16) Potential Hotspots Gradient Boosting

6.4 Case Study 3:Evaluation of results for smileCart classifier

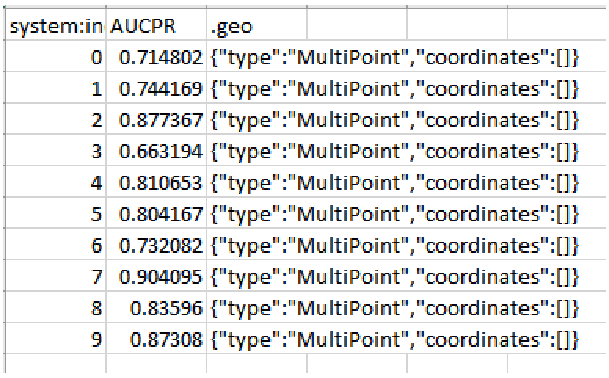
CART stands for Classification And Regression Trees and can be used both for classification and regression. CART relies on Gini impurity for finding the best tree route. For this implementation , default values have been passed to run the algorithm. Table 8 and Table 9 respectively show the AUC-ROC and AUC-PR metrics for the 10 iterations in the random forest classifier.



system:in	AUCROC	.geo
0	0.939815	{"type":"MultiPoint","coordinates":[]}
1	0.888636	{"type":"MultiPoint","coordinates":[]}
2	0.905882	{"type":"MultiPoint","coordinates":[]}
3	0.888889	{"type":"MultiPoint","coordinates":[]}
4	0.900826	{"type":"MultiPoint","coordinates":[]}
5	0.87	{"type":"MultiPoint","coordinates":[]}
6	0.777778	{"type":"MultiPoint","coordinates":[]}
7	0.883929	{"type":"MultiPoint","coordinates":[]}
8	0.897959	{"type":"MultiPoint","coordinates":[]}
9	0.901786	{"type":"MultiPoint","coordinates":[]}

Figure (17) AUCROC - CART

The Mean AUCROC obtained in the 10 iterations is 0.78365 and Mean AUCPR obtained in 10 iterations is 0.80865. The most suitable habitat and potential hotspot map predicted by CART is depicted . Figure 18 shows distribution map for India and Figure 19 and 20 show the maps for the states of Nagaland and Manipur. This accomplishes objective 3.c of this research outlined in table 1.



system:in	AUCPR	.geo
0	0.714802	{"type":"MultiPoint","coordinates":[]}
1	0.744169	{"type":"MultiPoint","coordinates":[]}
2	0.877367	{"type":"MultiPoint","coordinates":[]}
3	0.663194	{"type":"MultiPoint","coordinates":[]}
4	0.810653	{"type":"MultiPoint","coordinates":[]}
5	0.804167	{"type":"MultiPoint","coordinates":[]}
6	0.732082	{"type":"MultiPoint","coordinates":[]}
7	0.904095	{"type":"MultiPoint","coordinates":[]}
8	0.83596	{"type":"MultiPoint","coordinates":[]}
9	0.87308	{"type":"MultiPoint","coordinates":[]}

Figure (18) AUCPR - CART

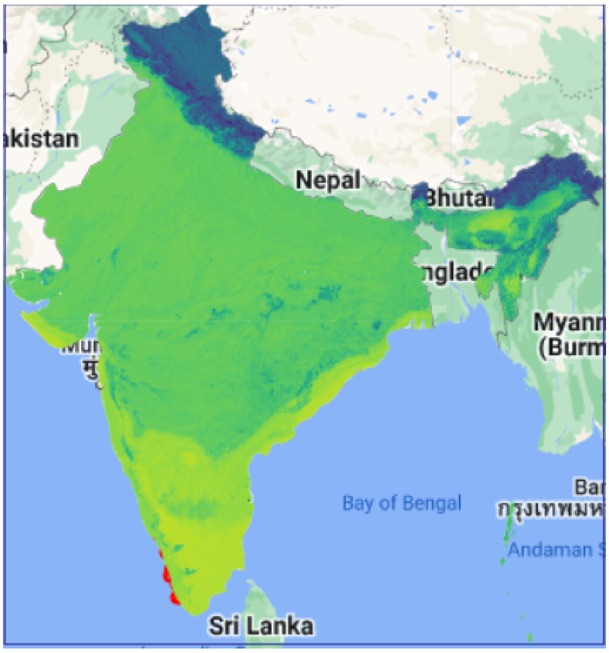


Figure (19) Habitat Suitability Cart

The objectives 3 ,3.a ,3.b ,3.c ,4 ,4.a ,5 , and 5.a that were outlined in table 1 have been successfully accomplished with the creation of a map depicting the distribution of habitat and possible hotspots.

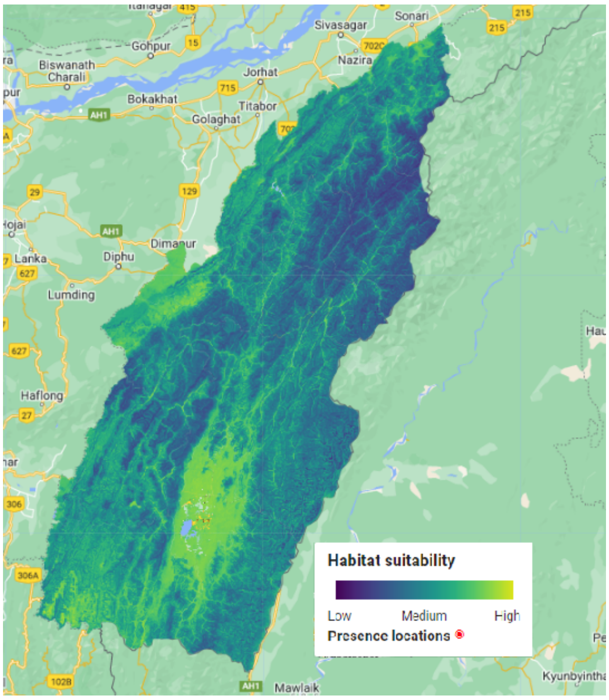


Figure (20) Habitat Suitability CART

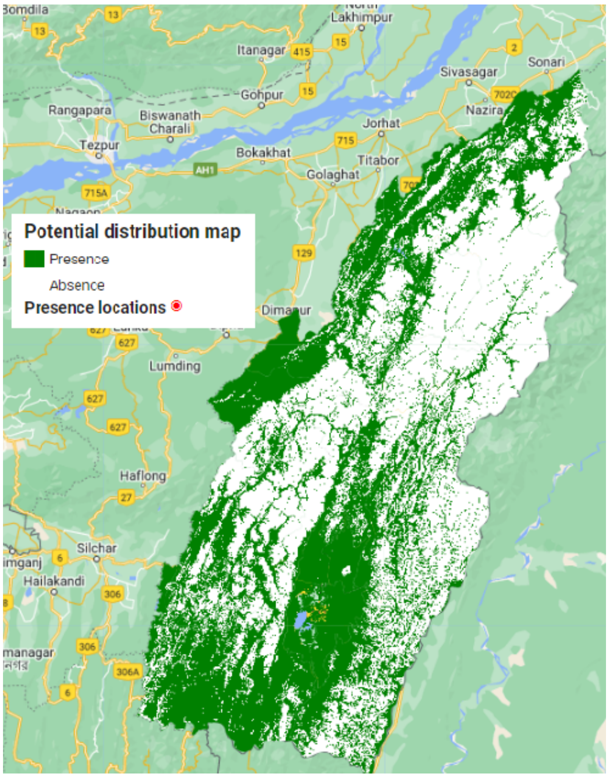
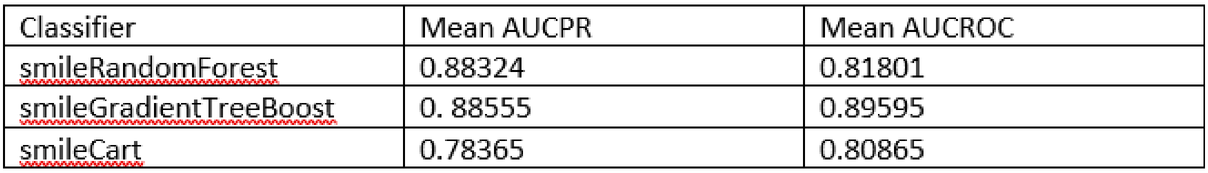


Figure (21) Potential Hotspots CART



Classifier	Mean AUCPR	Mean AUCROC
smileRandomForest	0.88324	0.81801
smileGradientTreeBoost	0. 88555	0.89595
smileCart	0.78365	0.80865

Figure (22) Results

6.5 Discussion

This implementation has successfully performed a classification and potential hotspot mapping in the state of Manipur and Nagaland. While performing the implementation all the classifiers have been studied and case studies to evaluate their performance has been done. The Mean AUCPR and Mean AUCROC for all the three classifiers used in the project are summarized in below figure. This completes the objective 6 of this research mentioned in table 1.

7 Conclusion and Future Work

The major purpose of this initiative was to investigate and identify areas in India that have the potential to become important hubs for the production of coffee.This fundamental objective has been completed successfully. As a result of this research, classifiers have been successfully deployed to locate environments in India that are appropriate for coffee cultivation. smileRandomForest, smileGradientBoostTrees, and smileCart are some of the methods that were utilized in this process.Following the discovery of possible distribution in the north-eastern region of India, suitability maps for the states of Nagaland and Manipur have been generated and rendered for driving. In table 1, all of the objectives that needed to be accomplished in order to answer the research question have been accomplished. The cloud-based geospatial analysis that was utilized in this project possesses a tremendous amount of potential in terms of its ability to work with big raster data sets. With the growing number of public data sets on google earth engine, problems can be solved more specifically.This research has taken into consideration current climatic conditions.

With raster datasets serving as the foundation, this project has implemented big data and cloud-based processing methods for machine learning in Google Earth Engine. This project also made use of the Usaga module for Python in order to collect occurrence data and view TIFF files that were generated by the earth engine. In addition to that, it takes use of Tableau for the depiction of historical trends in coffee output throughout all of India's states. Because of the introduction of this change, there have been various increases in a variety of skills, some of which are as follows: Python, Javascript, and Python for the technical elements; Remote sensing; Geographic Information System; Geospatial; Machine Learning. This application has also resulted in a significant increase in knowledge regarding coffee plantations, the elements that influence it, and the crop cycles. In addition to this, it investigates the application of machine learning on a cloud-based platform by making use of big data in order to investigate the uncountable ways in which land utilization could be improved and optimized for farming and other activities.

With the increase in global warming the coffee hot spots identified disappear.Currently earth engine does not have a future climate data set.With the use of future climate data sets, hot spots to move cultivation can be predicted and it remains to be done as a future work in this project.The research has also not taken into consideration soil nutrients.Currently earth engine is slow and the google drive only allows 15gb of storage, restricting rendering of large data sets.With more soil related data and ability to render images with high resolutions more detailed hotspots can be predicted.

Acknowledgement

I would want to take this opportunity to thank my mentor, Dr. Catherine Mulwa, for all of her assistance and direction in ensuring that I was able to finish my research project on schedule. She has consistently provided me with guidance that has been of tremendous assistance to me in overcoming the challenges I faced while working on my dissertation.

References

Adhikari, M., Isaac, E. L., Paterson, R. R. M. and Maslin, M. A. (2020). A review of potential impacts of climate change on coffee cultivation and mycotoxigenic fungi, *Microorganisms* **8**(10): 1625.

Al-Barqawi, H. and Zayed, T. (2008). Infrastructure management: Integrated ahp/ann model to evaluate municipal water mains' performance, *Journal of Infrastructure Systems* **14**(4): 305–318.

Bellard, C., Bertelsmeier, C., Leadley, P., Thuiller, W. and Courchamp, F. (2012). Impacts of climate change on the future of biodiversity, *Ecology letters* **15**(4): 365–377.

Bhagat, R., Singh, S., Sood, C., Rana, R., Kalia, V., Pradhan, S., Immerzeel, W. and Shrestha, B. (2009). Land suitability analysis for cereal production in himachal pradesh (india) using geographical information system, *Journal of the Indian Society of Remote Sensing* **37**(2): 233–240.

Bunn, C., Läderach, P., Ovalle Rivera, O. and Kirschke, D. (2015). A bitter cup: climate change profile of global production of arabica and robusta coffee, *Climatic change* **129**(1): 89–101.

Chemura, A., Mudereri, B. T., Yalew, A. W. and Gornott, C. (2021). Climate change and specialty coffee potential in ethiopia, *Scientific reports* **11**(1): 1–13.

de Carvalho Alves, M., Sanches, L., Pozza, E. A., Pozza, A. A. and da Silva, F. M. (2022). The role of machine learning on arabica coffee crop yield based on remote sensing and mineral nutrition monitoring, *Biosystems Engineering* **221**: 81–104.

Dedeoğlu, M. and Dengiz, O. (2019). Generating of land suitability index for wheat with hybrid system aproach using ahp and gis, *Computers and Electronics in Agriculture* **167**: 105062.

Doke, A. B., Zolekar, R. B., Patel, H. and Das, S. (2021). Geospatial mapping of groundwater potential zones using multi-criteria decision-making ahp approach in a hardrock basaltic terrain in india, *Ecological Indicators* **127**: 107685.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L. et al. (2007). The shuttle radar topography mission, *Reviews of geophysics* **45**(2).

Fielding, A. H. and Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models, *Environmental conservation* **24**(1): 38–49.

Haryuni, H., Dewi, T. S. K., Suprapti, E., Rahman, S. F. and Gozan, M. (2019). The effect of beauveria bassiana on the effectiveness of nicotiana tabacum extract as biopesticide against hypothenemus hampei to robusta coffee, *International Journal of Technology* **10**(1): 159–166.

Hoffmann, J. (2018). *The World Atlas of Coffee: From beans to brewing-coffees explored, explained and enjoyed*, Mitchell Beazley.

Iliquín Trigoso, D., Salas López, R., Rojas Briceño, N. B., Silva López, J. O., Gómez Fernández, D., Oliva, M., Quiñones Huatangari, L., Terrones Murga, R. E., Barboza Castillo, E. and Barrena Gurbillón, M. Á. (2020). Land suitability analysis for potato crop in the jucusbamba and tincas microwatersheds (amazonas, nw peru): Ahp and rs–gis approach, *Agronomy* **10**(12): 1898.

Mandal, V. P., Rehman, S., Ahmed, R., Masroor, M., Kumar, P. and Sajjad, H. (2020). Land suitability assessment for optimal cropping sequences in katihar district of bihar, india using gis and ahp, *Spatial Information Research* **28**(5): 589–599.

Mendas, A. and Delali, A. (2012). Integration of multicriteria decision analysis in gis to develop land suitability for agriculture: Application to durum wheat cultivation in the region of mleta in algeria, *Computers and electronics in agriculture* **83**: 117–126.

Mighty, M. A. (2015). Site suitability and the analytic hierarchy process: How gis analysis can improve the competitive advantage of the jamaican coffee industry, *Applied Geography* **58**: 84–93.

Oh, H.-J., Kim, Y.-S., Choi, J.-K., Park, E. and Lee, S. (2011). Gis mapping of regional probabilistic groundwater potential in the area of pohang city, korea, *Journal of Hydrology* **399**(3-4): 158–172.

Ostovari, Y., Honarbakhsh, A., Sangoony, H., Zolfaghari, F., Maleki, K. and Ingram, B. (2019). Gis and multi-criteria decision-making analysis assessment of land suitability for rapeseed farming in calcareous soils of semi-arid regions, *Ecological indicators* **103**: 479–487.

Parry, J. A., Ganaie, S. A. and Bhat, M. S. (2018). Gis based land suitability analysis using ahp model for urban services planning in srinagar and jammu urban centers of j&k, india, *Journal of Urban Management* **7**(2): 46–56.

Pham, M. P., Vu, D. D., Tong, T. H., Thi, M. H. N. and Sandlersky, R. (2021). Integrated use of ahp-gis-remote sensing predicting potential areas of coffee plants: a case study of buffer zone of ta dung nature park, vietnam, *E3S Web of Conferences*, Vol. 285, EDP Sciences, p. 02022.

Pilevar, A. R., Matinfar, H. R., Sohrabi, A. and Sarmadian, F. (2020). Integrated fuzzy, ahp and gis techniques for land suitability assessment in semi-arid regions for wheat and maize farming, *Ecological Indicators* **110**: 105887.

Rono, F. and Mundia, C. N. (2016). Gis based suitability analysis for coffee farming in kenya.

Salas López, R., Gómez Fernández, D., Silva López, J. O., Rojas Briceño, N. B., Oliva, M., Terrones Murga, R. E., Iliquín Trigoso, D., Barboza Castillo, E. and Barrena Gurbillón, M. Á. (2020). Land suitability for coffee (coffea arabica) growing in amazonas, peru: integrated use of ahp, gis and rs, *ISPRS International Journal of Geo-Information* **9**(11): 673.

Sofaer, H. R., Hoeting, J. A. and Jarnevich, C. S. (2019). The area under the precision-recall curve as a performance metric for rare binary events, *Methods in Ecology and Evolution* **10**(4): 565–577.

Taha, Z. and Rostam, S. (2011). A fuzzy ahp–ann-based decision support system for machine tool selection in a flexible manufacturing cell, *The International Journal of Advanced Manufacturing Technology* **57**(5): 719–733.

Tashayo, B., Honarbakhsh, A., Akbari, M. and Eftekhari, M. (2020). Land suitability assessment for maize farming using a gis-ahp method for a semi-arid region, iran, *Journal of the Saudi Society of Agricultural Sciences* **19**(5): 332–338.

Wintgens, J. N. et al. (2004). *Coffee: growing, processing, sustainable production. A guidebook for growers, processors, traders, and researchers.*, WILEY-VCH Verlag GmbH & Co. KGaA.