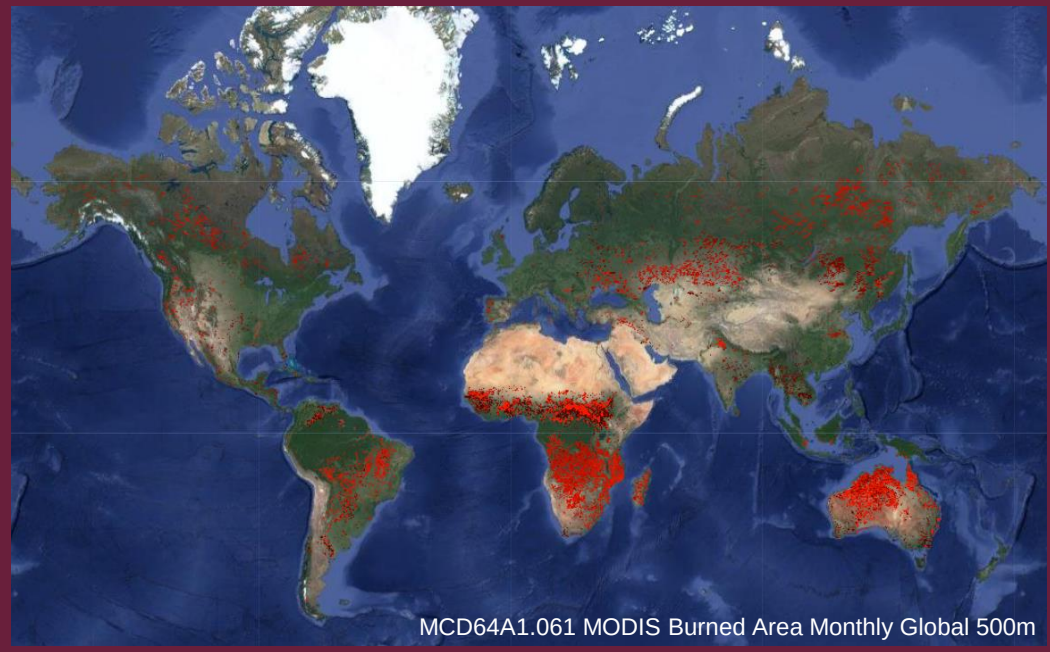
Analysis Of Factors Affecting Wildfire Distribution

Geoinformatics Project

2023-2024



POLITECNICO MILANO 1863

Input Products

Google Earth Engine

GEE Collections

Selected Variables

Variable	Description
Slope	Slope of the terrain
Aspect	Aspect (direction the slope faces) of the terrain
Wind Speed	Wind speed
Temperature (Max)	Maximum temperature
Temperature (Min)	Minimum temperature
Water Deficit	Climate water deficit
Precipitation	Precipitation accumulation
Soil Moisture	Soil moisture
NDVI	Normalized Difference Vegetation Index (NDVI)
NDMI	Normalized Difference Moisture Index (NDMI)
Land Cover	Land cover type from MCD12Q1 dataset
Human Impact Index	Human impact index

Table 2: Environmental, Climatic and Social Variables

External Products

QGIS

Hexagon Map Grid

- Derived from QGIS
- 300 km horizontal and vertical spacing
- Representing geographic areas at a broader scale
- Uploaded in GEE as asset

Shapefiles

- 7 products for each continent :
1. Europe
2. Africa
3. Asia
4. North America
5. South America
6. Australia
7. Oceania

- Each uploaded as asset on GEE

figshare

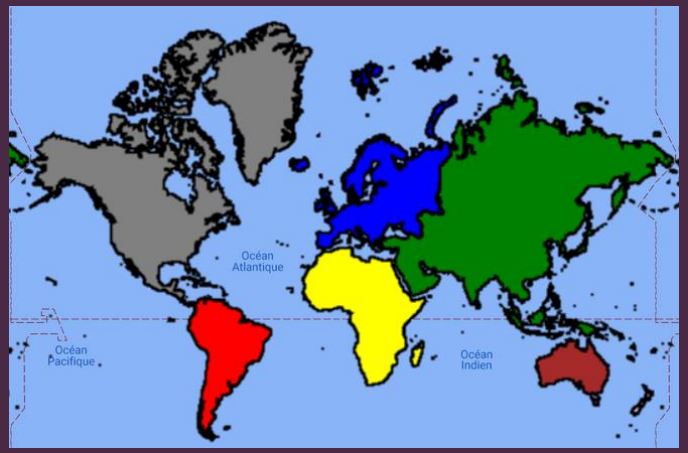
Hexagon Grid Map

Data Preprocessing

- Clipping the Area of Interest in continents

Reasons

- Avoiding the finite amount of RAM available for computation.
- Excluding the pixels representing oceans and seas from the processing step.



- Removing Cloud Cover and Shadows from Landsat images

Reasons

- Computing the vegetation indexes (NDVI and NDMI)
- Their presence can introduce noise and errors in the results.

```
function maskL8clouds(image) {
  var qa = image.select('QA_PIXEL'); // Selecting the QA band from the Landsat 8 image.
  var cloudBitMask = 1 << 4; //Defining a bitmask to isolate the cloud bit in the QA band.
  var shadowBitMask = 1 << 3;//Defining a bitmask to isolate the shadow bit in the QA band.
  //Creating a binary mask where pixels with cloud or shadow are set to 0.
  var mask = qa.bitwiseAnd(cloudBitMask).eq(0).and(
                 qa.bitwiseAnd(shadowBitMask).eq(0));
  return image.updateMask(mask)
```

Process:

- Using the **QA band** from the Landsat images.
- Defining **bitmasks** to isolate the cloud and shadow bit in the QA-band.

Data Preprocessing

- Aggregation of variables

BAND	Description	Unit
soil	Soil Moisture	mm
tmmn	Minimum temperature	degrees Celsius
tmmx	Maximum temperature	degrees Celsius
def	Climate water deficit	mm
pr	Precipitation accumulation	mm
vs	Wind-speed	m/s

Table 1: Climatic factors affecting wildfires

Collection

TerraClimate: Monthly Climate and Climatic Water Balance for Global Terrestrial Surfaces, University of Idaho

Environmental and Climatic variables

- Wind Speed
- Max Temperature
- Min Temperature
- Water Deficit
- Precipitation
- Soil Moisture

Preprocessing

- Filtering to a time-range of 13 years (2010-2023)
- Aggregation using median values
- **.median()** function: calculates the median value for each pixel

Data Preprocessing

- Aggregation of variables

Collections

1. USGS Landsat 8 Collection 2 Tier 1 TOA Reflectance
2. MCD12Q1.061 MODIS Land Cover Type Yearly Global 500m

Vegetation indices and Land Cover

- NDVI
- NDMI
- Land Cover

Preprocessing

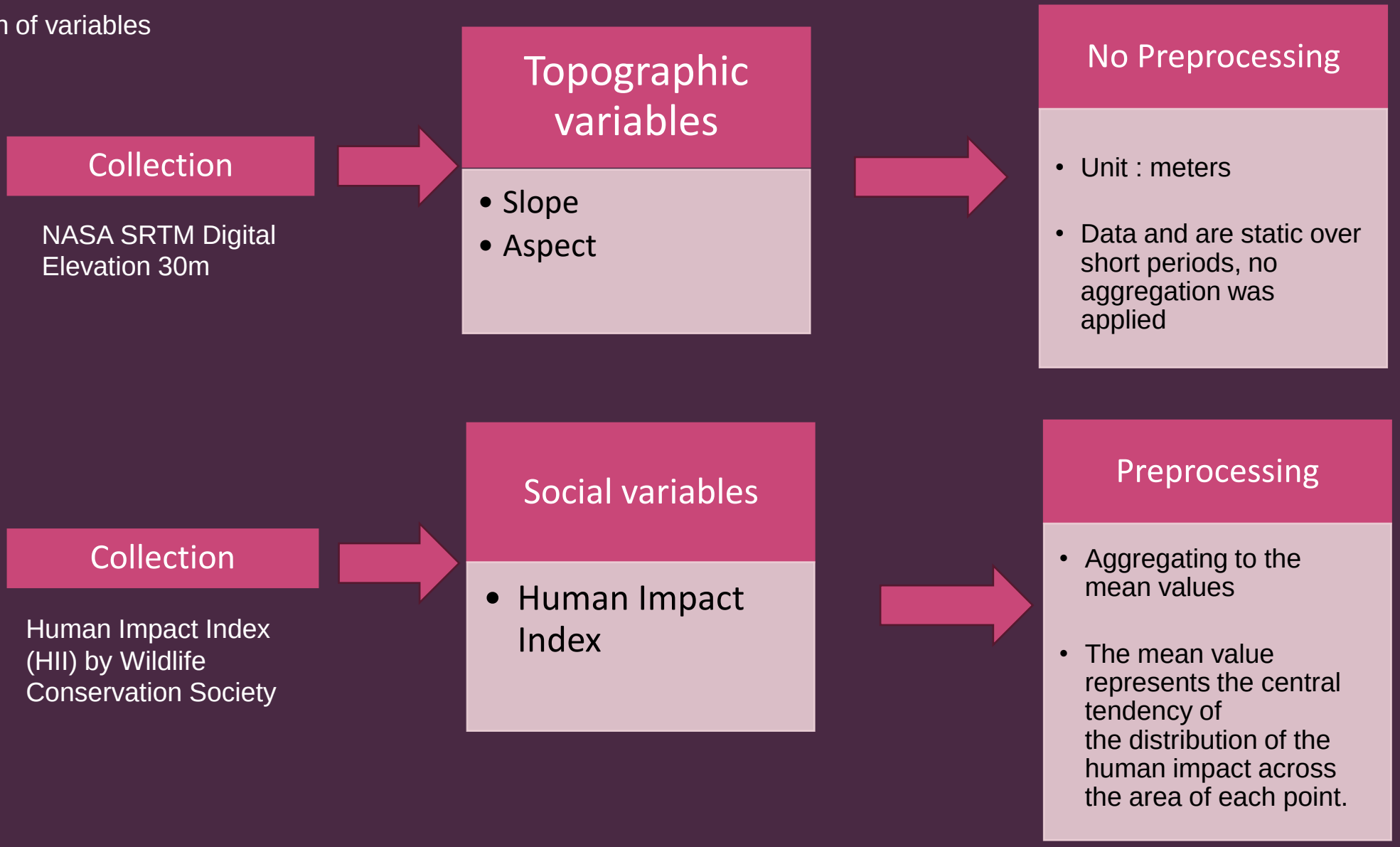
- Filtering to a time-range of 13 years (2010-2023)
- Single composite image representing the median values for each pixel
- **.median()** function is applied to the processed image collection of Landsat

No Preprocessing

- Represented by categorical values
- Taking the true values rather than aggregating them.
- Reprojected to match the CRS of other products.

Data Preprocessing

- Aggregation of variables



Final Product

Selected Variables → Predictor Image

Variable	Description
Slope	Slope of the terrain
Aspect	Aspect (direction the slope faces) of the terrain
Wind Speed	Wind speed
Temperature (Max)	Maximum temperature
Temperature (Min)	Minimum temperature
Water Deficit	Climate water deficit
Precipitation	Precipitation accumulation
Soil Moisture	Soil moisture
NDVI	Normalized Difference Vegetation Index (NDVI)
NDMI	Normalized Difference Moisture Index (NDMI)
Land Cover	Land cover type from MCD12Q1 dataset
Human Impact Index	Human impact index

Table 2: Environmental, Climatic and Social Variables

- Combines all the preprocessed variables.
- Variables are considered as **bands** of this image.
- Serves as an input for wildfire risk assessment.
- Used in the process of sampling to generate training points feeding the algorithms.

Approaches

First Approach

- Creating a binary mask of burned and non-burned pixels
- Performing stratified sampling over an area

Advantages

- Less computational time.

Second Approach

- Creating pure fires binary mask
- Creating an area image and performing zonal statistics
- Extracting sample points over hexagons

Advantages

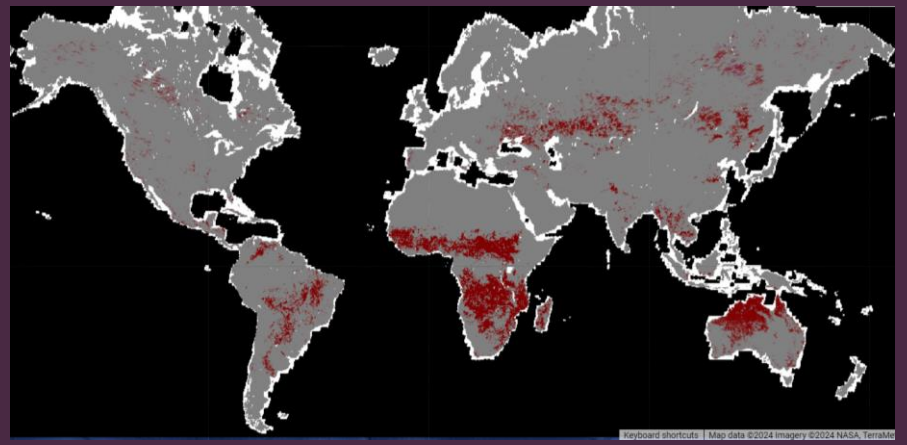
- Handling raster data easier.
- Modified for usage in smaller areas of interest.
- Simplifying the sampling process.

First Approach

1 Creation of Burn Mask from MODIS Data:

- Temporal distribution of the analysis (13 years):
  start date (January 1, 2010)
  end date (September 1, 2023)
- Derived from the 'BurnDate' band of the MODIS dataset
- Identifies areas with at least one occurrence of burning.

2 Creation of Land Data Mask:



- Extracting a land data mask from the Global Forest Change dataset .

- Filtering out irrelevant information and focusing the analysis on areas where wildfires might have occurred.

3 Unmasking burned areas and creating Binary Mask

Binary representation of burned and non-burned pixels:
White - Burned pixels
Black – Unburned pixels

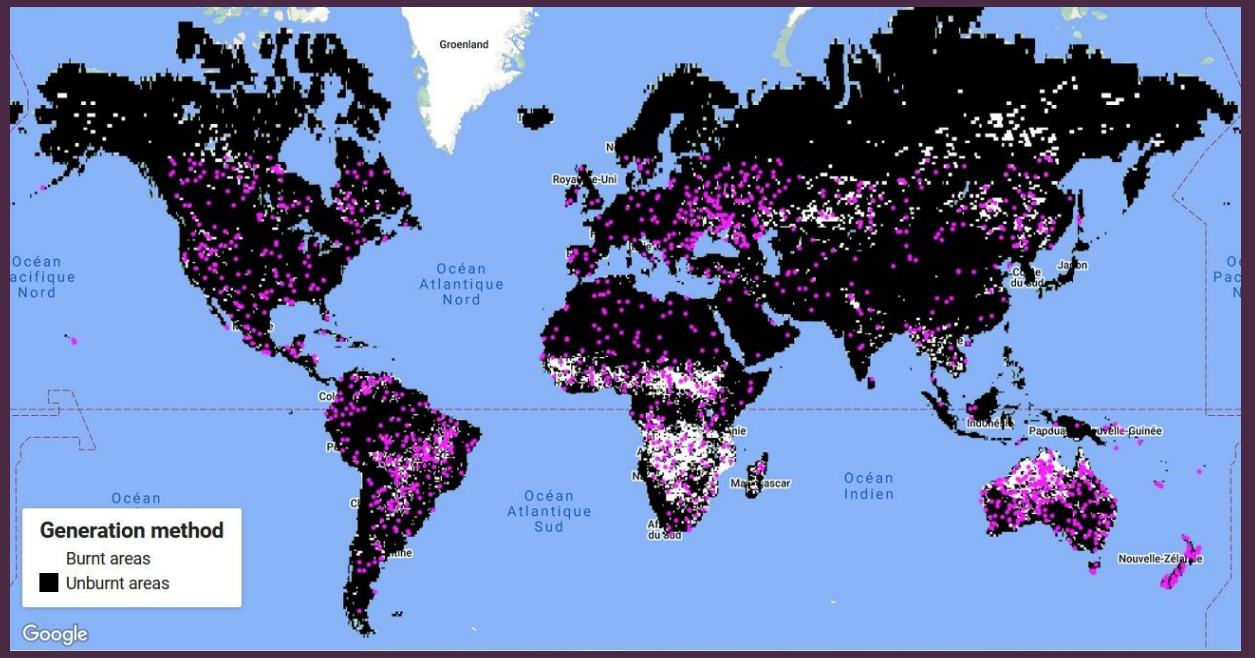
First Approach

4 Stratified Sampling

Predictor Image

Binary Mask

- To distinguish which training points are going to belong to burned pixels and which to unburned pixels.
- Adding band "Burn" as a band of the predictor image and use it for stratification.
- The points are sampled from the predictor image.



- The scale for sampling same as the one of the binary mask.

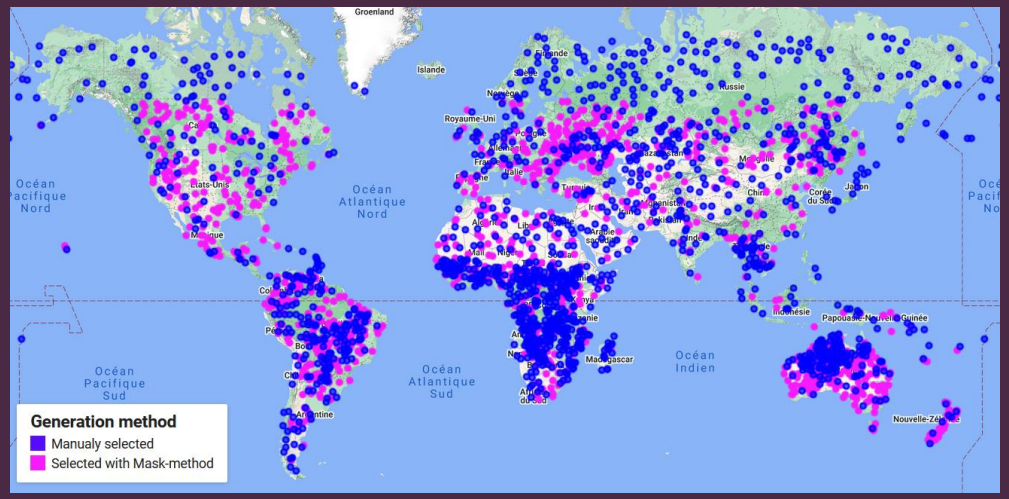
- Sampling 100 points for each class of the "Burn" band:
  - 100 points on the burned pixels, with Burn value = 1
  - 100 points on the unburned pixels, with Burn value = 0.

First Approach

5 Analysis of the dataset

Features of the Dataset

feature	type object
point id	int 8
Land cover	int 8
Human Impact Index	float 64
NDMI	float 64
NDVI	float 64
aspect	float 64
precipitation	float 64
slope	float 64
soil moisture content	float 64
wind speed	float 64
water deficit	float 64
max temperature (°C)	float 64
min temperature (°C)	float 64
burn	boolean



Datasets collected from GEE and used for the Random Forest Algorithm in Jupyter notebook

- In total 1251 points.

- Each point, represented by a row in the DataFrame, contains values for all the 14 features which are the columns of the DataFrame.

- Burn category (True) = 601 points.
- Unburned category (True) = 650 points.

First Approach

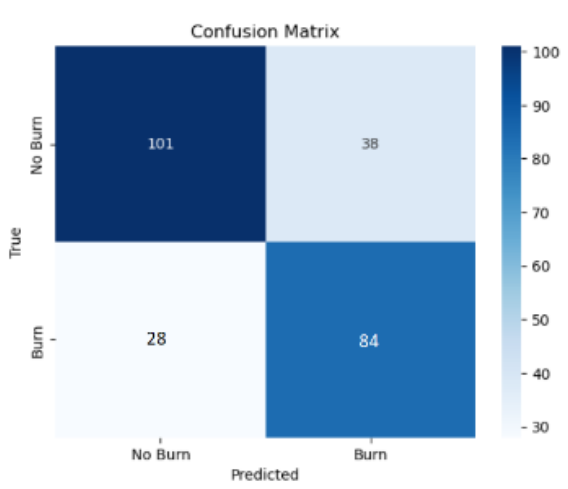
5 Analysis of the dataset (Random Forest Algorithm)

First dataset accuracy and feature importance

Model's accuracy :	0.7370
factor	importance
wind speed	0.104316
precipitation	0.095617
NDMI	0.087743
Soil Moisture Content	0.087285
Max Temperature	0.085794
NDVI	0.085042
slope	0.083885
water deficit	0.077022
Human Impact Index	0.076335
Min Temperature	0.075240
Land Cover	0.074702
aspect	0.067018

- Feeding the model with only the points generated from stratified sampling.

- 80% training set vs 20% testing set

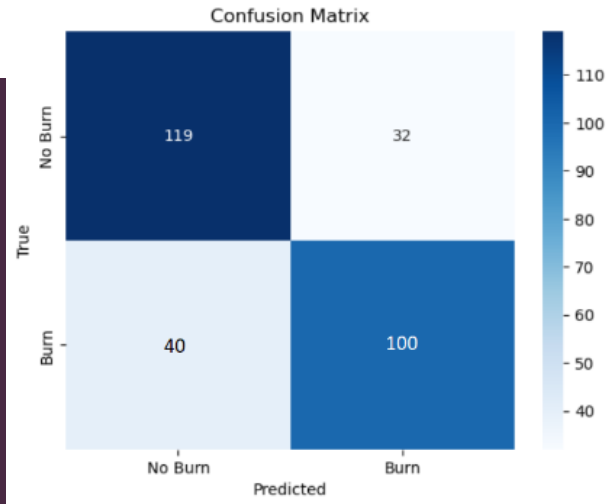


Confusion matrix of first dataset

Second dataset accuracy and feature importance

Model's accuracy :	0.7526
factor	importance
WindSpeed	0.100515
NDVI	0.095257
NDMI	0.089032
Precipitation	0.087268
HumanImpactIndexMean	0.085429
SoilMoist	0.083466
TempMax	0.083202
Slope	0.081115
LandCover	0.079303
WaterDeficit	0.078512
TempMin	0.073839
Aspect	0.063061

- Feeding the model with points taken manually from GEE.

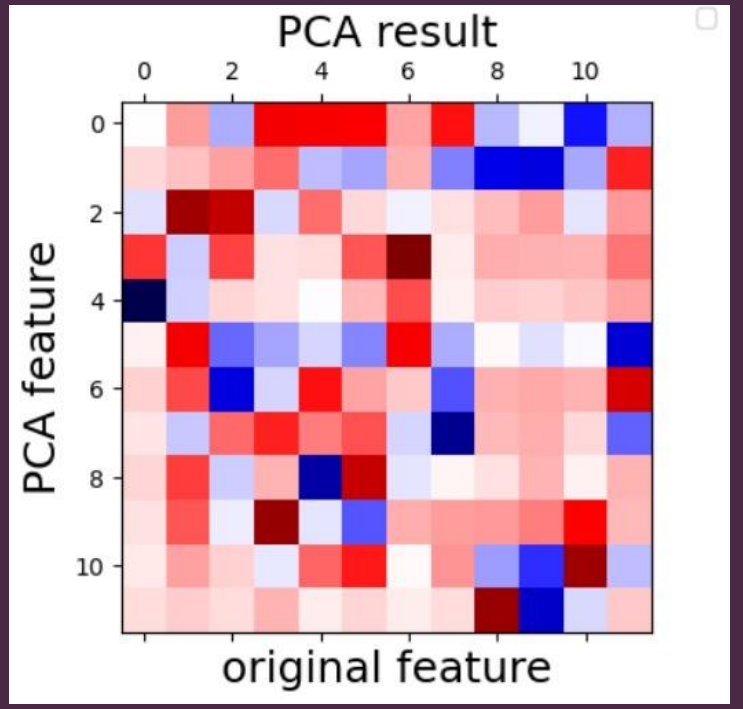


Confusion matrix of second dataset

First Approach

6 Analysis of the dataset (Principal Component Analysis (PCA))

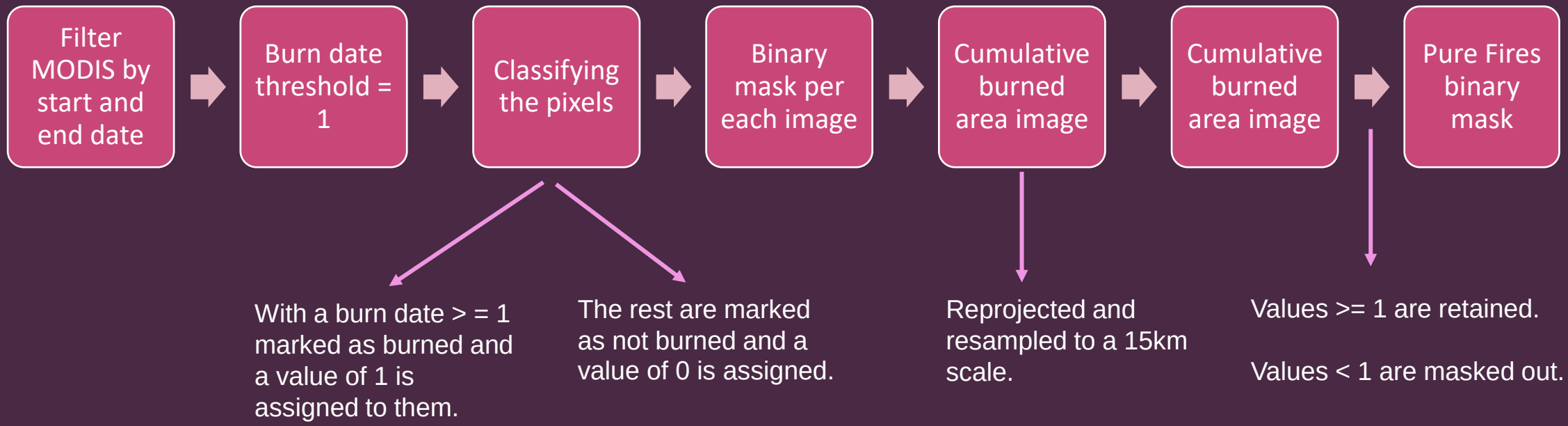
Model's accuracy:	0.7012
component	importance
pc1	0.103537
pc11	0.102431
pc2	0.098119
pc6	0.097604
pc4	0.087990
pc7	0.082340
pc10	0.075084
pc9	0.074980
pc8	0.072754
pc12	0.071393
pc3	0.070099
pc5	0.063672



- The accuracy of the model decreased and no improvement was observed.

Second Approach

1. Pure fires binary mask:



Second Approach

2 Dividing the hexagons of the grid:

Map Hexagon Grid

Africa Example

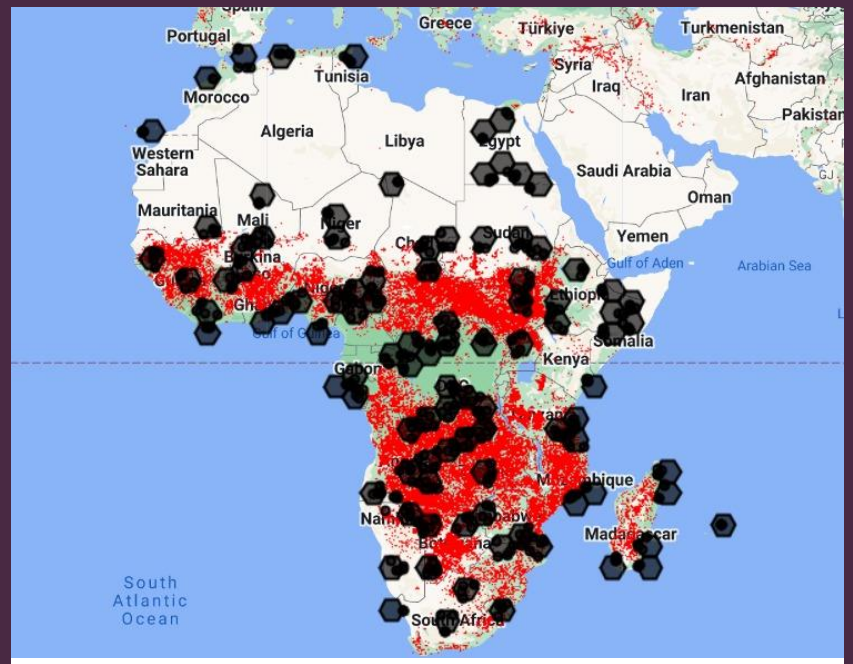
- The hexagons for each continent are divided into:
1. hexagons with burned pixel sum not 0.
2. hexagons with burned pixel sum 0.

→ Reduced to an image with a band called ‘unburned’.

Second Approach

3 Sampling Points into hexagons and creating Feature Collections:

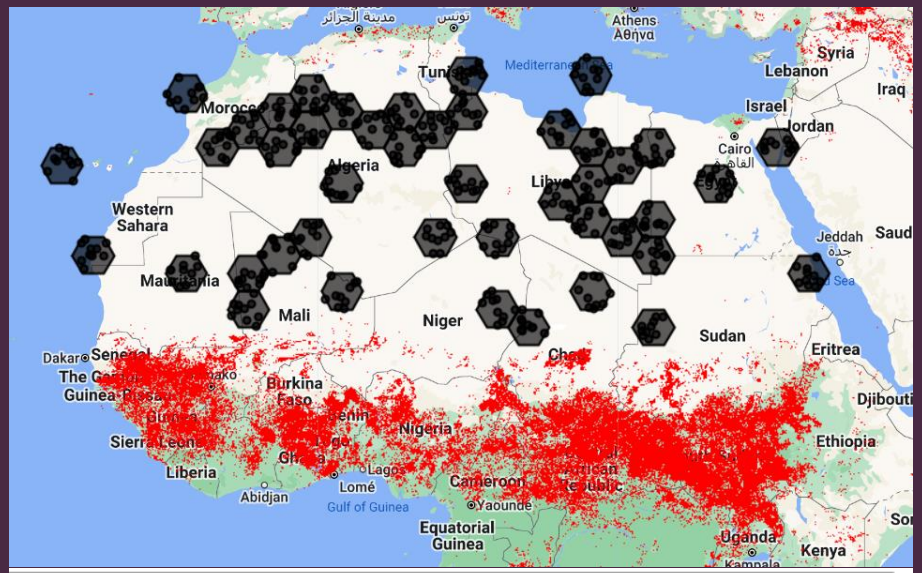
Sampling points in hexagons with burn sum not 0.



Continent	Burned	
	Hexagons	Points within each hexagon
Africa	50	5
Asia	50	10
Australia	30	5
Europe	30	5
North America	30	5
South America	50	5
Oceania	20	5

Hexagon and Burned point distribution by continent

Sampling points in hexagons with burn sum 0.



Continent	Non-burned	
	Hexagons	Points within each hexagon
Africa	50	10
Asia	50	20
Australia	3	100
Europe	50	20
North America	50	20
South America	44	20
Oceania	50	20

Hexagon and UnBurned point distribution by continent

Second Approach

4 Analysis of dataset:

Features of the Dataset

feature	type object
point id	int 8
Land cover	int 8
Human Impact Index	float 64
NDMI	float 64
NDVI	float 64
aspect	float 64
precipitation	float 64
slope	float 64
soil moisture content	float 64
wind speed	float 64
water deficit	float 64
max temperature (°C)	float 64
min temperature (°C)	float 64
hazard	int

burned points if int = 1
unburned points, if int = 0

Splitting the dataset in training and testing sets

Training set : 80%
Testing set: 20%

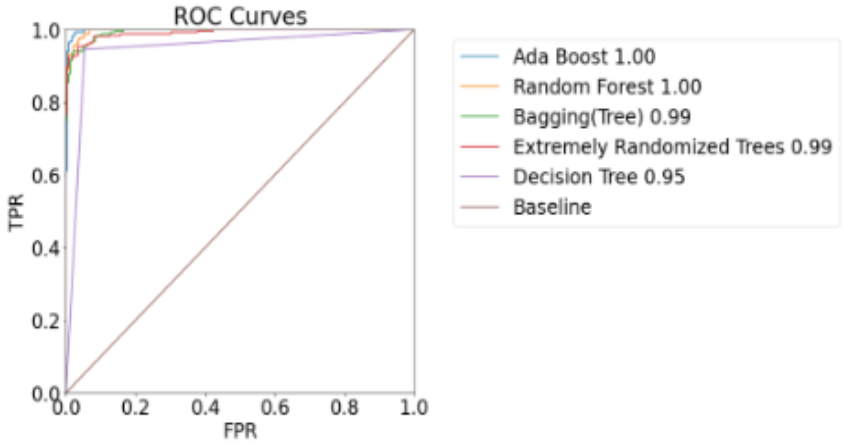
Cross- validation method : Stratified K Fold with 30 folds.

Chosen classification models for the analysis:

1. Random Forest
2. Decision Tree
3. Bagging Tree
4. Extremely Randomized Tree
5. Ada Boost

Metrics Calculation for each model:

1. Cross- validation accuracy
2. Test accuracy
3. Precision
4. Recall
5. F1 score
6. Confusion Matrix



Receiver Operating Characteristic curve and Area Under Curve

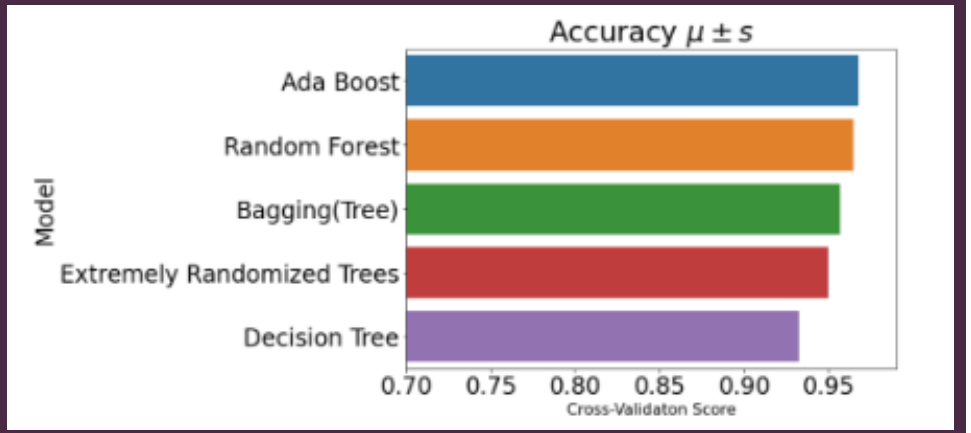
Graphical representation, that shows the trade-off between:
- the True Positive Rate (TPR) and
- the False Positive Rate (FPR).

Metric	Decision Tree	Bagging (Tree)	Random Forest
Accuracy (CV)	92.2% ± 2.4%	95.3% ± 2.8%	96.3% ± 2.6%
Accuracy (Test)	94.9%	96.8%	97.6%
Precision	94.9%	96.8%	97.6%
Recall	94.5%	95.3%	94.9%
F1 Score	94.7%	96.0%	96.2%

Classification metrics for:
1. **Decision Tree**
2. **Bagging Tree**
3. **Random Forest**

Each bar represents the mean accuracy obtained during cross-validation.

Quick assessment of the performance of each model in predicting wildfires based on the listed features.



Visual comparison of cross- validation accuracy scores for different classification methods.

Metric	Extremely Randomized Trees	Ada Boost
Accuracy (CV)	95.0% ± 2.2%	97.3% ± 1.6%
Accuracy (Test)	95.6%	97.2%
Precision	95.6%	97.2%
Recall	94.5%	96.5%
F1 Score	95.1%	96.9%

Classification metrics for:
1. **Extremely Randomized Trees**
2. **Ada Boost**

Second Approach

4 Analysis of dataset:

Method	Extremely Randomized Trees		Ada Boost	
Confusion Matrix	214	11	218	7
	14	241	9	246

Confusion matrix for:
1. **Extremely Randomized Trees**
2. **Ada Boost**

Method	Decision Tree		Bagging (Tree)		Random Forest	
Confusion Matrix	212	13	217	8	219	6
	14	241	12	243	13	242

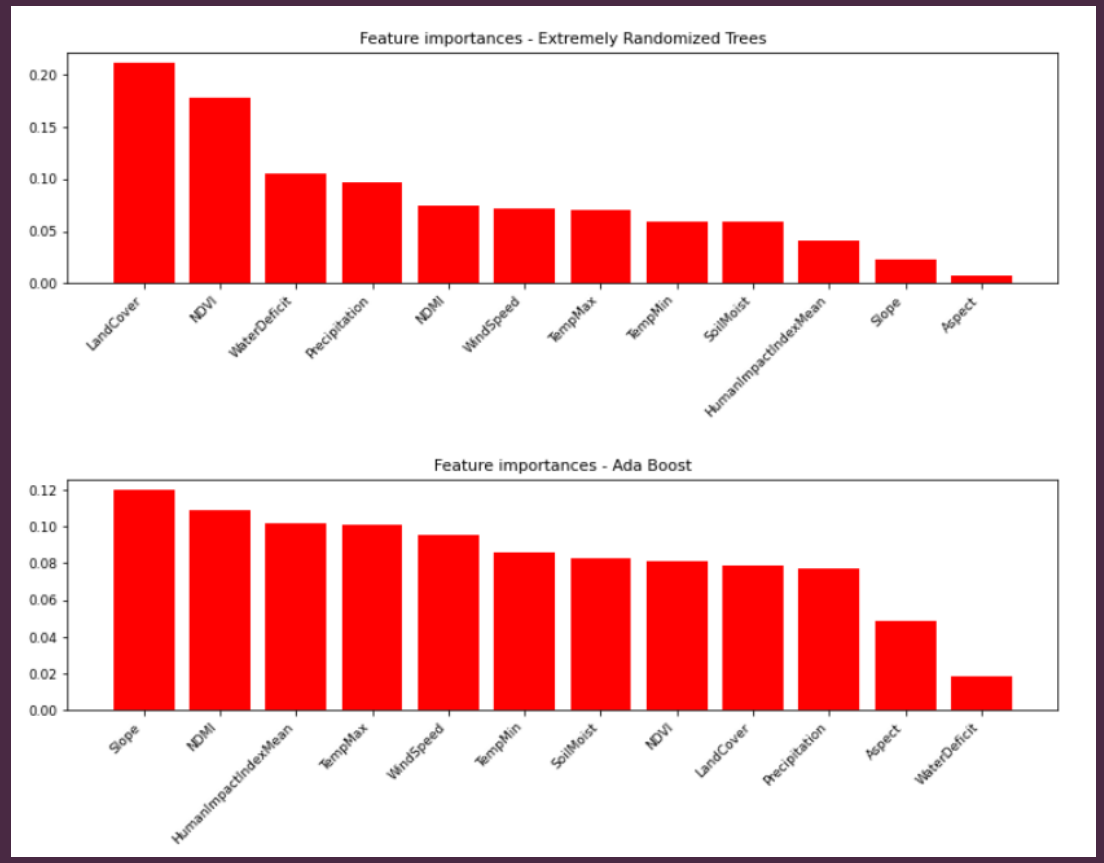
Confusion matrix for:
1. **Decision Tree**
2. **Bagging Tree**
3. **Random Forest**

T test used to compare the classifiers' performance:

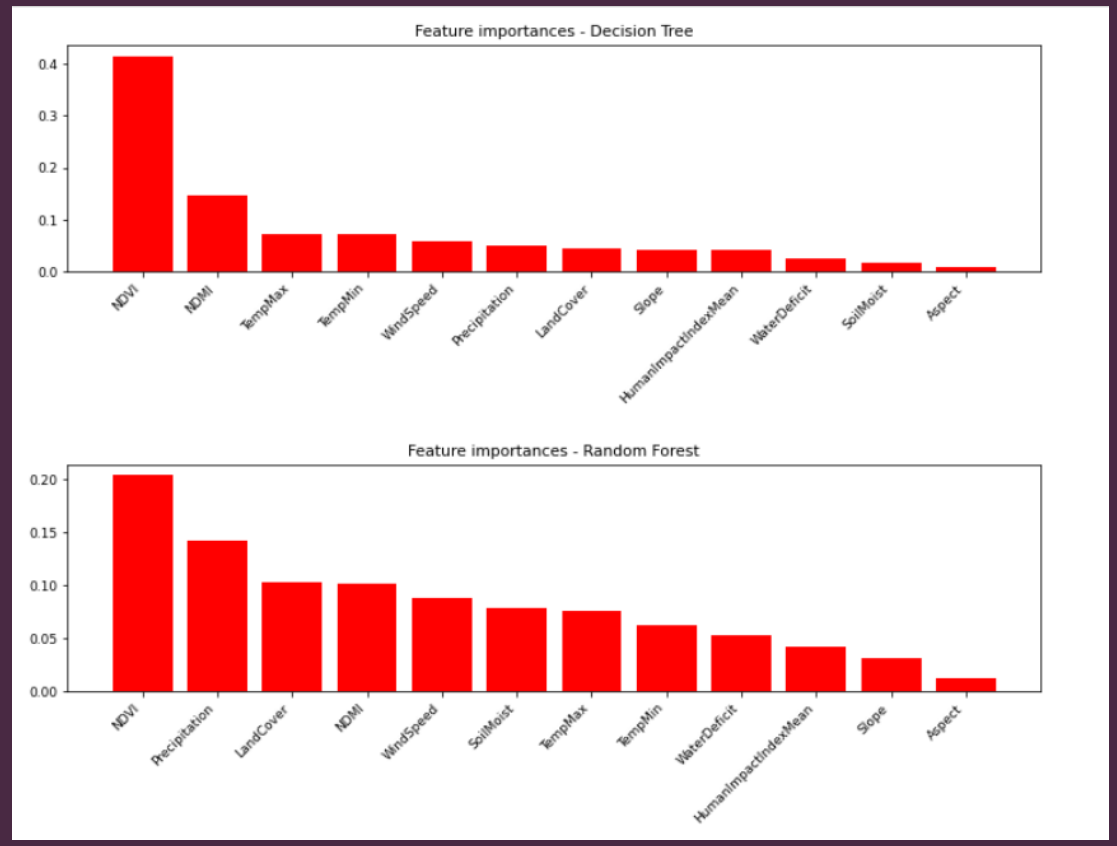
```
  Bagging(Tree) vs       Ada Boost => Difference is statistically significant (cf 95.00 p-value=0.0354)
  Random Forest vs Extremely Randomized Trees => Difference is statistically significant (cf 95.00 p-value=0.0118)
Extremely Randomized Trees vs       Ada Boost => Difference is statistically significant (cf 95.00 p-value=0.0031)
```

Second Approach

4 Analysis of dataset:



Feature importance plot of Decision Tree and Random Forest



Feature importance plot of Extremely Randomized Trees and Ada Boost

Conclusions

- Two different methods were developed to create random points on the burnt and non-burned area.
- These two methods both have their advantages

- During the process of work there where several difficulties.

- Random forest was the classifier giving the best performances among all the classifiers we trained.

- Further analysis of the result could show that all the factors that were considered in the analysis have an impact on the prediction of fires, each of them in a different scale of importance.

- Further developments may be applied to contribute in:

1. The refinement of classification models,
2. Temporal analysis
3. Spatial resolutions
4. Incorporation of additional factors
5. Integration with Real-Time data
6. Climate Change Impacts

Thank you for your attention!

Worked by:
Abbes Madeleine: Madeleinesarah.abbes@mail.polimi.it
Rajabi Forough: Forough.rajabi @mail.polimi.it
Zallemi Nikolina: nikolina.zallemi@mail.polimi.it