Preparing for wildfires avoidance by analyzing land cover maps

Bachelor of Science Thesis - Creative Technology

Julia van der Geest - s2403250
Supervisor: dr. A. Kamilaris
Critical observer: dr. J.W. Kamminga
Date: 08/07/2022

Abstract

The aim of this project was to find a method that is best suitable to predict wildfire susceptibility in Cyprus. This method needed to take factors into account on which it could base its predictions. Therefore, the selection of these factors was also of interest for this project. Based on background research that was conducted at the beginning of this project a list of 12 factors was created. This list was turned into a dataset containing the information of each factor for different data points in Cyprus. The dataset was then used to train two models, logistic regression and random forest. In the end, the logistic regression model produced an overall accuracy of 70%, the AUC of the model was 0.69 meaning that it is decently able to make the distinction between the positive (wildfire occurred) class and negative (no wildfire occurred) class. However, preconditions were set of an accuracy above 80% which resulted in the model not satisfying as a method suitable to predict wildfire susceptibility in Cyprus. Random forest on the other hand had an overall accuracy of 88% with an AUC of 0.89 and therefore did meet the precondition of accuracy above 80%. Lastly, for each model, the feature importance was plotted which resulted in precipitation being the most impactful factor in predicting the wildfire susceptibility in both models as well as NDVI, altitude, forest density, and temperature .

Acknowledgments

First of all, I would like to thank my supervisor, Andreas Kamilaris for his guidance and enthusiasm. During the past months he never told me what to do, instead he nudged me in the right direction encouraging me to make my own decisions and justify them.
I would also like to thank his associates, Chirag Padubidri and Asfa Jamil, who helped me with gathering the right information regarding the factors used in this project.

Table of contents

Chapter 1 Introduction

In the past year, Cyprus was hit by a major wildfire taking the lives of four people and forcing the inhabitants of ten villages out of their houses. It was said to be the worst wildfire to have hit Cyprus within decades and it sure won't be the last [1]. But Cyprus was not the only country to get hit. In July of 2021 California was struck by a wildfire damaging 963.309 acres of land [2]. This fire was called the Dixie fire and to get it under control, officials invested roughly 600 million dollars. It is said to be the most expensive suppression campaign in California History [3]. In addition, wildfires are not only increasing in intensity but also in numbers. According to an article by Jones M.W., this is due to human-induced climate change, which promotes the conditions on which these wildfires depend, enhancing their likelihood of happening as well as their severity [4].

Yet, it is the combination of climate change and wildfire suppression that is the most dangerous. When wildfires became public enemy number one - approximately 100 years ago - action had to be taken. However, instead of letting wildfires do their thing and only managing it when they became out of control, every wildfire had to be extinguished. No matter the size, the prime focus became to extinguish all wildfires. By doing so, the forest had time to grow and grow which has led to our current situation where a 'small' wildfire would be equal to a 'large' wildfire 100 years ago due to the increased levels of forest fuel [5][6]. Meaning that, once a wildfire does occur, its outcome will be much more catastrophic [7].

To reduce the amount, and intensity of wildfires the prevention plans should be optimized. However, to do so, not only should we look at effective prevention methods but also at which areas need it the most. As to which areas need it the most, a landcover map depicting wildfire risk should be created.

To create this map multiple methods should be explored as well as which factors have the biggest influence on wildfires. In this regard, this literature review will focus on finding methods that could predict the wildfire susceptibility of Cyprus as well as the factors that should be taken into account. Next to that, a first glance will be taken at a few prevention methods. Guiding this paper is the following research question: "What method would be best suitable to predict the wildfires susceptibility of Cyprus, and which factors should be taken into account?". To this end, the paper will start by addressing the factors that should be included when predicting the wildfire risk. After doing so, it will delve into possible methods that could be used to predict the wildfire risk.

To summarize, the following research question and sub-questions will be answered:

- What method would be best suitable to predict the wildfires susceptibility in Cyprus, and which factors should be taken into account?:
    - What methods for predicting wildfire susceptibility are out there?
    - Which factors were found to influence this susceptibility and should therefore be taken into account?

Chapter 2 Background Research

2.1 Landcover Map

2.1.1 Methods and Factors

A great number of methods exist for generating a wildfire susceptibility map, all with their own benefits and downfalls depending on the study area and factors considered. Zhao et al. [8] performed a case study in China using the Analytic Hierarchy Approach (AHP) in combination with 8 factors. The factors taken into consideration were altitude, slope, aspect, TWI (Topographic Wetness index), temperature, distance to roads, distance to populated areas, and NDVI. Based on their findings they concluded that these 8 factors were effective in generating the risk model. However, they did recommend the use of a larger number of factors to improve accuracy. Which could increase the current accuracy of 77%.

Novo et al. [9] also used the AHP method, but instead of 8 factors, they used 9. Namely, elevation, slope, aspect, NDVI, fuel model type, FWI, distance to roads, distance to settlements, and HFR (Hystorical Fire Regimes). In this study, they defined 2 factors as particularly having a strong influence on forest fire ignition, FWI, and NDVI. Whereas HFR, elevation, slope, and aspect were found to have a lower impact. For future work, they recommend the use of 'land use' as an additional factor and a more sensitive AHP in combination with multi-criteria decision analysis (MCDA) technology for more accurate results.

Busico et al. [10] made use of 12 factors in their case study in combination with AHP. These factors were precipitation, temperature, altitude, slope, aspect, soil, forest type, distance from roads, distance from settlements, distance from agricultural fields, distance from water sources, and distance from springs. Just like Novo et al. [9], Busico et al. [10] found that land use as a factor could improve the accuracy of AHP. Busico et al. also concluded that both land use and climate factors are the main factors influencing forest fire occurrence. Lastly, for future work, the use of multiple time series was advised for better accuracy.

Just like Zhao et al. [8], Nikhil et al. [11] also made use of only 8 factors (land cover types, slope angle, aspect, TWI (Topographic Wetness Index), distance from settlement, distance from road, distance from tourist spots, distance from anti-poaching camp shed) in combination with AHP. in this case study it was found that angle of slope and land cover type did not show significant influence on fire occurrence. TWI and distance to roads and settlements on the other hand demonstrated a strong correlation with forest fire occurrence. Lastly, the accuracy of the method was concluded to be 79.5%.

Sari [12] conducted a case study comparing AHP, VIKOR, and TOPSIS. In this study, they made use of 17 factors (aspect, slope, elevation, CTI (Compound Topographic Index), wetness, precipitation, temperature, wind speed, humidity, power lines, roads, settlements, buildings, land use, rivers, forest type, forest density). It was found that distance to power lines, forest type, slope, and forest density have a high impact on forest fire occurrence. This contradicts the findings of Nikhil et al. [11] who found that angle of slope did not show significant influence. Based on the results of the study, the VIKOR method was found to have the highest accuracy of 89.54% as opposed to the 86.94% and 88.99% of TOPSIS and AHP respectively. Lastly, Sari

[12] suggested that, for future work, it should be decided in advance if the distance to roads will be taken as a negative or positive factor since roads are also vital for early intervention in forest fires.

Other methods that are used for wildfire susceptibility mapping are FR, SE, EBF, BLR, KLR, and CNN. Starting with FR and SE, Pourtaghi et al. [13] conducted a study in Iran comparing the two. For their study, they used 14 different factors (slope degree, slope aspect, TWI (Topographic Wetness Index), TPI (Topographic Position Index), plan curvature, wind effect, annual temperature, annual rainfall, Soil texture, distance to roads, distance to rivers, distance to villages, NDVI, Land use map). They found that NDVI, Land use, and annual temperature have the biggest influence on forest fire occurrence, supported by the findings of Novo et al. [9] and Busico et al. [10]. When comparing the accuracy of both methods, SE turned out to have higher accuracy as opposed to FR with 83.16% accuracy instead of 79.85%. However, the study does advise using a method that is simple and has high accuracy. Coming back to the comparison between FR and SE, FR is more suitable for large amounts of data since it can quickly and easily process that in GIS. SE on the other hand is a good option when only small sample sizes are available as it enables the determination of the least-biased probability distribution with limited knowledge and data.
As for EBF, 2 case studies were conducted by Pourghasemi [14] and Nami et al. [8], both in Iran. Starting with Pourghasemi [14] who compared EBF to BLR using 15 factors (Distance to rivers, distance to roads, distance to villages, slope degree, slope aspect, altitude, plan curvature, TPI (Topographic Position Index), TWI (Topographic Wetness Index), land use, NDVI, soil texture, wind effect, annual precipitation, annual temperature). In his study, he found that an important constraint of EBF is that if there is no value for disbelief in a certain class, then it indicates that there is no forest fire occurrence in the same class. Leading to the conclusion that BLR performs slightly better than EBF with an accuracy of 81.93% as opposed to 74.3%. Nami et al. [15] on the other hand conducted a case study solely into EBF using 1 factor less as opposed to Pourghasemi [14] (Slope, aspect, altitude, plan curvature, TWI (Topographic Wetness Index), TRI (Topographic Roughness Index), temperature, rainfall, evapotranspiration, LULC (land use/cover), soil type, proximity to rovers, proximity to roads, proximity to settlements). It was found that the overall accuracy of the map was 81.03%, which is higher than Pourghasemi's [14] accuracy for EBF. They did find another weak point of EBF which refers to the neglect between predictor variables relationships which can lead to the assumption that all variables carry equal weight, which can decrease the accuracy of the map. Next to that, they found that climate variables and temperature have the highest effect on the likelihood of wildfires occurring. However, for future research, they suggest making use of more factors to increase the accuracy.
Another method that could possibly be used is Kernel Logistic Regression (KLR), which was studied by Bui et al. [16] in Vietnam using 10 factors (slope, aspect, TWI, distance to roads, distance to populated areas, land cover, NDVI, surface temperature, wind speed, and rainfall). This model turned out to have an overall accuracy of 89.29% on the training dataset and 81.25% on the validation dataset. And performs slightly better than the SVM (Support Vector Machine) model. Furthermore, it was found that NDVI, TWI, land use, and surface temperature have the highest predictive powers for forest fires, which is supported by the findings of [9]-[11]

[13][15]. However, to increase the accuracy more factors need to be used. Bui et al. [9] suggested humidity and drought.

Lastly, Zhang et al. [17] conducted a case study in China using CNN (Convolutional Neural Network), which is a feed-forward neural network whose parameters are trained using classic stochastic gradient descent based on the backpropagation algorithm. In this study, 14 factors were used: elevation, slope, aspect, average temperature, average precipitation, surface roughness, average wind speed, maximum temperature, specific humidity, precipitation rate, forest coverage ratio, NDVI, distance to roads, and distance to rivers. From the study, it was found that precipitation rate, specific humidity, and maximum temperature did not satisfy the critical values and could therefore be excluded from this study. Temperature, wind speed, surface roughness, precipitation, and elevation on the other hand were considered the most important factors that influence forest fire occurrence. The overall accuracy of this model was found to be 91% in training and 82% in validation. CNN was also tested against other models (random forests (RF), support vector machine (SVM), multilayer perceptron neural network (MLP), and kernel logistic regression (KLR)), but still came out as the best performing method with the highest overall accuracy of (87.92%) for the validation dataset, followed by RF (84.36%), KLR (81.23%), SVM (80.04%), and MLP (78.47%). The advantages of CNN were described as follows:

1. Because the CNN can consider the correlation of adjacent spatial information, it has advantages in the study of problems with spatial and geographical correlation characteristics
2. CNN preserves the spatial relationships between pixels by learning the internal feature representations from factor vectors
3. CNN reduces the number of weights that need to be trained and the computational complexity of the network through weight sharing

However, CNN does require many training samples, which might be a problem for areas with fewer data. Furthermore, it relies heavily on high-end machines compared to traditional ML algorithms that can run on low-end machines. Next to that, the training time of the CNN is longer than those of traditional ML models. For further studies, the deficiency of its interoperability needs to be studied.

2.1.2 Case Studies Literature Matrix

To summarize the information presented in the previous section a literature matrix is presented below.

	Study Area	**Climate**	**Factors**	**Methods**	**Accuracy**
[8] Zhao et al.	Laoshan National Forest Park, Nanjing, China	Subtropical monsoon	altitude, slope, aspect, TWI, temperature, distance to roads, distance to populated areas, NDVI	AHP (Analytic Hierarchy Approach)	77%
[9] Novo et al.	Galicia, Spain	Sub-Mediterranean	elevation, slope,	AHP (Analytic	~

			aspect, NDVI, fuel model type, FWI, distance to roads, distance to settlements, HFR	Hierarchy Approach)	
[10] Busico et al.	Campania region, Southern Italian Peninsula	Mediterranean	precipitation, temperature, altitude, slope, aspect, soil, forest type, distance from roads, distance from settlements, distance from agricultural fields, distance from water courses, and distance from springs	AHP (Analytic Hierarchy Approach)	~
[11] Nikhil et al.	Parambikulam Tiger Reserve, Kerala, India	~	land cover types, slope angle, aspect, TWI (Topographic Wetness Index), distance from settlement, distance from road, distance from tourist spots, distance from anti-poaching camp shed	AHP (Analytic Hierarchy Approach)	79.5%
[12] Sari	Mugla province, Turkey	Warm and temperate	aspect, slope, elevation, CTI (Compound Topographic Index), wetness, precipitation, temperature, wind speed, humidity, power lines, roads, settlements, buildings, land use, rivers, forest type, forest density	AHP (Analytic Hierarchy Approach) TOPSIS VIKOR	AHP → 88.99% TOPSIS → 86.94% VIKOR → 89.54%
[13] Pourtaghi et al.	Minudasht forests, Golestan province, Iran	Temperate/ semi-humid	slope degree, slope aspect, TWI, TPI, plan curvature, wind effect, annual temperature, annual rainfall, Soil texture, distance to roads, distance to rivers, distance to villages, NDVI, Land	FR (Frequency Ratio) SE (Shannon's Entropy)	FR → 83.16% SE → 79.83%

			use map		
[14] Pourghasemi	Golestan province, northern Iran	Temperate/ semi-humid	Distance to rivers, distance to roads, distance to villages, slope degree, slope aspect, altitude, plan curvature, TPI (Topographic Position Index), TWI (Topographic Wetness Index), land use, NDVI, soil texture, wind effect, annual precipitation, annual temperature	EBF (Evidential Belief Function) BLR (Binary Logistic Regression)	EBF → 74.30% BLR → 81.93%
[15] Nami et al.	Hyrcanian ecoregion, northern Iran	Semi-dessert	Slope, aspect, altitude, plan curvature, TWI (Topographic Wetness Index), TRI (Topographic Roughness Index), temperature, rainfall, evapotranspiration, LULC (land use/cover), soil type, proximity to rovers, proximity to roads, proximity to settlements	EBF (Evidential Belief Function)	81.03%
[16] Bui et al.	Cat Ba National Park Area, Hai Phong City, Vietnam	Tropical monsoon	slope, aspect, TWI, distance to roads, distance to populated areas, land cover, NVDI, surface temperature, wind speed, and rainfall	KLR (Kernel Logistic Regression)	81.25%
[17] Zhang et al.	Yunnan province, southwestern China	Plateau-type tropical monsoon	elevation, slope, aspect, average temperature, average precipitation, surface roughness, average wind speed, maximum temperature, specific humidity, precipitation rate, forest coverage ratio, NDVI, distance to roads, and distance to rivers	CNN (Convolutional Neural Network)	87.92%

2.1.3 Conclusion

Nine methods of susceptibility mapping were discussed. All methods used a variety of factors to assure their accuracy. All factors used at least once in a case study are altitude, slope, elevation, slope angle, aspect, plan curvature, TWI (Topographic Wetness Index), CTI (Compound Topographic Index), TPI (Topographic Position Index), TRI (Topographic Roughness Index), temperature, precipitation, wetness, evapotranspiration, humidity, wind speed, wind effect, distance to roads, distance to populated areas, distance to settlements, distance from agricultural fields, distance from watercourses, distance from springs, distance from tourist spots, distance from anti-poaching camp shed, distance from power lines, distance from buildings, distance from rivers, NDVI, fuel model type, soil, soil texture, forest type, forest density, forest coverage ratio, land cover types, land use, FWI (Fire Weather Index), HFR (hystorical Fire Regimes), which are 39 factors in total.

Based on all the case studies observed, FWI, NDVI, land use, climate factors, TWI, distance to roads, distance to settlements, distance to power lines, forest type, slope, forest density, soil, temperature, wind speed, surface roughness, precipitation, and elevation are deemed highly influential on wildfire occurrence. HFR, topographic factors (slope, elevation, aspect), slope angle, land cover type, precipitation rate, humidity, and max temperature on the other hand were found to not be influential on wildfire occurrence. When comparing the 2 lists some factors are on both lists. Therefore, the final list of factors that were all deemed influential on wildfire occurrence is FWI, NDVI, land use, climate factors, TWI, distance to roads, distance to settlements, distance to power lines, forest type, forest density, soil, temperature, wind speed, surface roughness, precipitation, and elevation. However, it should be kept in mind that these results are based on specific regions in different countries. When generating a set for Cyprus an open mind should be kept.

Next to the factors, the case studies discussed made recommendations for the methods and discussed some weak points. For AHP, Novo et al. [9] recommended the use of sensitive AHP in combination with multi-criteria decision analysis (MCDA) to improve accuracy. Busico et al. [10] recommended the use of multiple time series to improve accuracy. Sari [12] found that VIKOR is more accurate than AHP, but that AHP is in turn more accurate than TOPSIS. Pourtaghi et al. [13] compared FR and SE and found SE to be more accurate. However, FR can process large amounts of data, whereas SE is more useful when only a small sample size is available. Furthermore, Pourtaghi et al. [13] recommend the use of an easy method with high accuracy. Pourghasemi [14] compared EBF and BLR and found that BLR has higher accuracy. Nami et al. [15] dove deeper into the functioning of EBF and found that the model neglects the relationship between predictor variables decreasing its accuracy. Lastly, Zhang et al. [17] used CNN and found that it requires many training samples, relies heavily on high-end machines, and its training time is longer as opposed to traditional ML models.

2.2 Methods of Suppression and Prevention

When it comes to suppression and prevention of wildfires many methods are already out there, amongst which are forest road networking, prescribed burning, and targeted grazing. What follows, this paper provides advantages and disadvantages concerning these three methods.

2.2.1 Forest Road Networking

Forest road networking is the first example of both a prevention and suppression method. According to Alcubierre et al. [18], road networking in difficult-to-access areas could be a useful strategy for pre-suppression, also called prevention. Stefanović et al. stated that forest roads represent the basic infrastructure necessary to effectively prevent and suppress wildfires [19]. This claim is supported by a review by Laschi et al. [20] which also dives deeper into its benefits and downfalls. In their review, they refer to a french paper by Croisé and Crouzet which mentioned that forest road networking could allow for continuous surveillance of the areas at risk of wildfires. This statement was supported by Larjavaara [21] who stated that dense road networks enable fast detection of wildfires. Furthermore, Laschi et al. [20] refer to the findings of an Italian paper by Calvani et al. from 1999 which mentions that good forest road planning allows for a very quick response time in the case of a wildfire. Demir et al. [22] build on that stating that a lack of forest roads would lead to a decrease in quick and early suppression. Besides methods for successful suppression, they could serve as fast evacuation routes for citizens as well as easy emergency access in case of casualties [20]. Next to that, Chuvieco and Congalton [23] found that they can serve as fire breaks limiting the fire rate of spread.

However, extra roads bring with them extra maintenance as well as increased human activity in previously difficult-to-access areas. This increased human activity, according to Cardille et al. [24] leads to a higher probability of ignition. Which is supported by the findings of Laschi et al. [20] who reviewed multiple papers regarding the statement of Cardille et al. [24], which stated that places closer to roads are at higher risk due to human accessibility.

All in all, even though forest road networking could increase the number of wildfires, they serve as a good method to suppress wildfires faster due to better access and continuous surveillance of the area.

2.2.2 Prescribed Burning

Another useful method for wildfire prevention is prescribed burning which could reduce surface fuels. When talking about wildfire prevention and suppression in their paper, Éric et al. [25] state that to prevent wildfires, management should reduce surface fuels, prune trees, and thin the stand. Agee and Carl [26] build upon these basic principles of forest fuel reduction. In their paper, they refer to reduction treatments such as reduction of surface fuels and decreasing crown density amongst others. They mention that a possible tool to accomplish these forms of reduction could be prescribed burning. According to Wade and Lunsford [27], prescribed burning is, to deliberately set fire to forest fuels to reduce the fuel load, which in turn results in

smaller, and better to control wildfires.

However, prescribed burning is still not a widely used method to prevent wildfires as a result of its following weaknesses. First, the results of a SWOT analysis performed by Marino et al. [28] showed that foresters and policymakers are skeptical about its benefits. This skepticism is supported by the many downfalls of prescribed burning presented in the literature review of Fernandez and Botelho [29]. According to their findings, prescribed burning has the best results in areas where no extreme weather conditions occur like strong winds. Furthermore, they fear that the longevity of prescribed burning is no longer than five years since nature will sooner or later regain its fuel load. Both these worries are supported in a study conducted by Weston et al. [30], which found that prescribed burning can effectively reduce fuel loads for only several years after the treatment, in mild weather conditions.

Despite this skepticism, prescribed burning still shows promising results. Fernandez and Botelho [29] refer to a study conducted by van Wagtendonk in Sierra Nevada California, which found that the average fireline intensity was reduced by 76% using prescribed burning, and its burned area by 37%. In addition, Duane et al. [31] found that prescribed burning plans have the ability to decrease the number of high-intensity wildfires and extreme fire events. Furthermore, a study conducted with two scenarios, no-treatment, and a combination of prescribed burning and tree pruning and thinning, showed that the treatment scenario did slow the fire growth and allowed for quicker containment. The no-treatment scenario on the other hand was estimated to have costs seven times higher. Pacheco and João [32] supports this statement, suggesting that countries employing prescribed burning could potentially save millions.

Based on this information you could argue that based on the cost alone it would be beneficial to use prescribed burning. However, if the effects would indeed only last for a short amount of time and the method is only effective in mild weather conditions, it could be of best interest to implement a more effective method first.

2.2.3 Targeted Grazing

Lastly, targeted grazing is supposed to be an effective method to suppress and prevent wildfires decreasing the rate of spread. According to Launchbaugh and Walker [33] targeted grazing, also called prescribed grazing, is the use of livestock grazing in a specific area, at a specific time (season and duration) to accomplish vegetation and landscape goals. This is a possible method to reduce the risk and size of a wildfire [33]. Rouet-Leduc [34] supports this claim, calling it a promising management strategy to avoid fuel build-up and mitigate wildfires. This mitigation strategy was further explored by Bruegger et al. [35] using the BehavePlus fire model, which found that targeted grazing reduced flame lengths and thereby the potential cost of suppressing a wildfire. Next to that, the results of their study showed a 60% decrease in the fire rate of spread in grass communities and over 50% decrease in grass/shrub communities.
Furthermore, since livestock tends to graze more in some areas than others, they create firebreaks [36], which can slow a wildfire down or even stop it. Next to that, it can serve as a

safe passage for firefighters [36]. When attempting to create a firebreak, livestock should be confined to a specific strip of land surrounded by fences [36].

However, the effects of targeted grazing are slow. Generally, it takes about three years to see a difference [33]. Besides the slow effect, it also requires a lot of maintenance, since grasses, bushes, and trees do grow back [36]. The biggest side effect though would be that, according to Bachelet et al. [37], the lack of grasses growth due to grazing allows trees to thrive, since there is more water available. Lastly, targeted grazing was found to be most effective in grass communities and under moderate weather conditions [35].

Overall, while targeted grazing is a useful method to reduce the risk of wildfires by creating firebreaks and reducing the forest fuel, its effects are slow and could encourage the growth of trees. Like prescribed burning, the use of a more effective method should be explored prior to targeted grazing.

2.2.4 Conclusion

The methods discussed were forest road networking, prescribed burning, and targeted grazing. All have their pros and cons. Forest road networking serves as a good method to suppress wildfires faster due to better access and continuous surveillance. However, roads lead to increased human activity, which was correlated with a higher probability of ignition. Prescribed burning serves as a method to effectively reduce surface fuels. However, its effectiveness is highest in areas where no extreme weather conditions occur. Next to that, its effects would only last for a short amount of time since nature will sooner or later regain its fuel load. Having said that, it could still be beneficial to use this method since it is found to be more cost-effective as opposed to no treatment at all. Lastly, targeted grazing is useful for creating firebreaks and reducing forest fuel. However, its effects are slow and could encourage the growth of trees.
In conclusion, all three methods discussed could prove effective when added to wildfire management policies. Forest road networking would be a preventive method as it would serve its purpose in the long term. However, it can also be used for suppression as it provides access to wildfires. Prescribed burning on the other hand would be a suppression method since its effects will only last for a short amount of time. Lastly, targeted grazing would be a bit of both (prevention and suppression) as its effects will not be visible fast. However, due to the regrowth of vegetation, its effects will not be long-term and asks for continuous upkeep.

Chapter 3 Methods and Techniques

In this section, the general structure of this project will be discussed. To do this in a structured manner the Creative Technology Design Process, by Mader and Eggink [38], will be used. In figure 1, this process is depicted. The phases of which will be discussed in further detail later on. This design process contains key elements of two pre-existing classical approaches, divergence and convergence, and spiral. Starting with the first approach. It consists of two different phases, divergence, and convergence. In the divergence phase, the design space is defined and in the convergence phase, the design space is reduced until a certain solution is reached. This approach is integrated into three of the four phases that the Creative Technology Design Process contains namely the ideation, specification, and realization phase. The second approach, spiral, refers to the sequence of design steps taken to reduce this design space. This approach is integrated all throughout the ideation and specification phases.

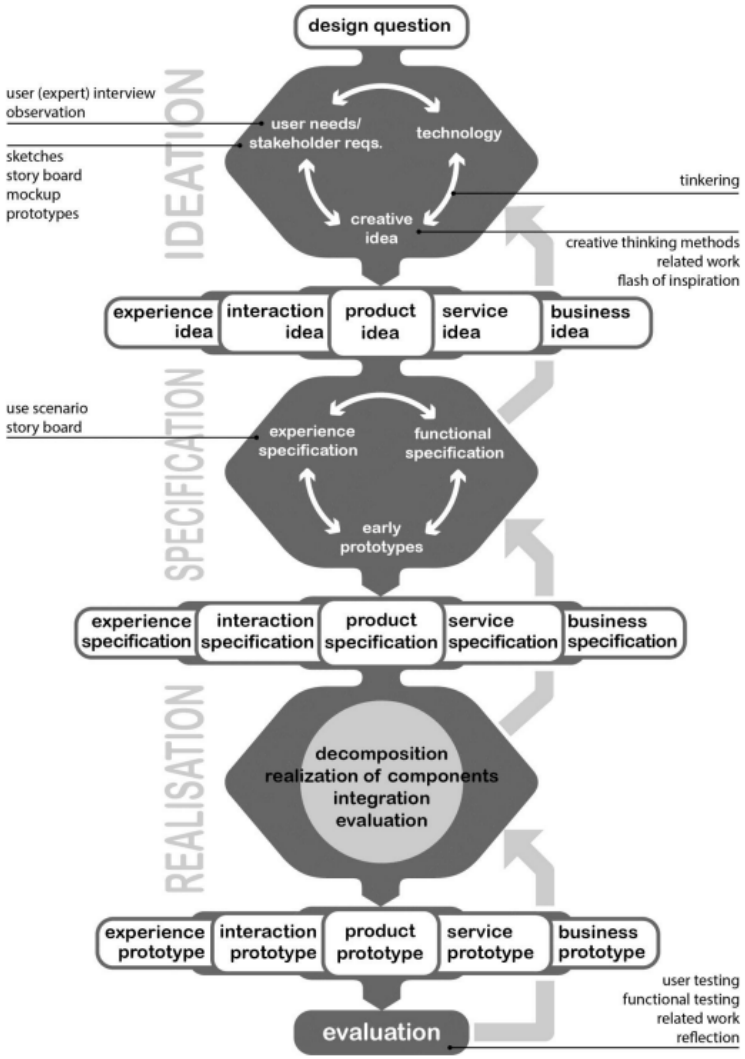

Figure 1: Creative Technology Design Process [38].

3.1 Ideation

During the ideation phase, the problem is defined, relevant information is gathered and ideas are formed [38]. At the end of this phase, a more narrowed-down version of the project idea is established. The problem this project is trying to address is the increasing number and intensity of wildfires by finding a method best to predict the probability of a wildfire occurring in Cyprus which can then be used to create a wildfire susceptibility map. Furthermore, the susceptibility of an area is dependent on multiple factors. Therefore, a better understanding is needed of which factors have the highest influence. These methods and factors must be explored through different case studies broadening our knowledge and understanding. The goal is then to narrow the number of methods down to one or two. However, to narrow the methods down another aspect should be considered, the climate of Cyprus. Each case study takes place in a different place in the world, meaning different climates. When choosing a method, it would be ideal if it was tested in an area with the same climate as Cyprus. Next to that, a list of factors must be created. This list will consist of all factors used at least once in a case study.

3.2 Specification

In the second phase, the specification phase, the final list of factors will be defined. This will be done based on the results of the case studies discussed in the background research. If multiple case studies deem a factor unimportant it will be excluded from the final list of factors that will be used in this project. Next to that, data on factors that cannot be obtained for Cyprus will also be excluded. Furthermore, the methods chosen in the ideation phase will be further explored. After the specification phase, the function of the method should be clearly defined. Lastly, preconditions will be determined which can, later on, be used to validate the models used in this project. These preconditions will be based on both the case studies as well as the desires of the supervisor.

3.3 Realization

The realization phase is used to create the dataset to be used using ArcMap 10.8. This dataset will contain the information of each factor for each datapoint in Cyrus. Next to that, the two models selected/explained in the specification phase will be implemented in python and trained using the dataset. The trained models can later on be used to create the wildfire susceptibility map.

3.4 Evaluation

During the last phase, the evaluation phase, the functioning of the models is tested. To this end, the accuracy of the methods is discussed and justified by a confusion matrix and AUC-ROC curve. Furthermore, the importance of each factor is plotted and discussed.

Chapter 4 Ideation

As discussed in chapter three, in the ideation phase a method needs to be chosen and a list of factors has to be created. The list of factors will be created based on the background research. Each factor used at least once will be included in the list. However, let us start with the method to be selected. In order to make this decision, the climate of Cyprus needs to be taken into consideration.

Cyprus is known to have an intense Mediterranean climate [39] with long dry summers and mild winters. In the summer months, the maximum temperatures range from 27 to 38 degrees. In the winter months, the minimum temperatures range from 0 to 5 degrees. The humidity in Cyprus ranges from average to slightly low at night during the winter days. During the summer it is very low.

So which method should be selected? In chapter two, the background research, nine methods were discussed in form of case studies. When going through the case studies, AHP showed promising results. It received recommendations on how to improve the accuracy, but no large disadvantages were mentioned. Next to that, most case studies using this method were conducted in areas with similar climates to Cyprus which could increase its chances of also working properly for Cyprus. It was therefore thought to be a good method to start with. However, it should be kept in mind that the functioning of the method is very data-dependent. Meaning, the promising results in the case studies do not imply much when it comes to this project.

Besides the methods, the factors also had to be discussed. After cross-referencing all ten case studies the list of factors used at least once is as follows:

- altitude
- slope
- elevation
- slope angle
- aspect
- plan curvature
- TWI (Topographic Wetness Index)
- CTI (Compound Topographic Index)
- TPI (Topographic Position Index)
- TRI (Topographic Roughness Index)
- temperature
- precipitation
- wetness
- evapotranspiration
- humidity
- wind speed
- wind effect
- distance to roads

- distance to populated areas
- distance to settlements
- distance from agricultural fields
- distance from water courses
- distance from springs
- distance from tourist spots
- distance from anti-poaching camp shed
- distance from power lines
- distance from buildings
- distance from rivers
- NDVI
- fuel model type
- soil
- soil texture
- forest type
- forest density
- forest coverage ratio
- land cover types
- land use
- FWI (Fire Weather Index)
- HFR (Hystorical Fire Regimes)

In the specification phase this list of parameters will be narrowed down based on the available data of Cyprus and further findings of the case studies.

Chapter 5 Specification

5.1 Factors

Based on the literature reviewed, the following factors were found to have the biggest influence on wildfire susceptibility:

- FWI (Fire Weather Index)
- NDVI
- Land use
- Climate factors
- TWI (Topographic Wetness Index)
- Distance to roads
- Distance to settlements
- Distance to powerlines
- Forest type
- Forest density
- Soil
- Temperature
- Wind speed
- Surface roughness
- Precipitation
- Elevation

When going through the list of 39 factors again, slope and distance to tourist spots were added. Distance to tourist spots was added since the increased human activity was correlated with an increased risk of wildfires. Slope was added based on the idea that the angle of elevation could influence the chances of a wildfire jumping crowns.

To come to the final list a better understanding was needed of what was actually meant by some of the factors. FWI is a method used to indicate fire danger. However, since the point of this project is to create this myself using an algorithm it was excluded from the final list. NDVI represents the type of vegetation in an area [40]. Where high values refer to rainforests and low values to rocks. Land use is used to describe the use of that land. For example, agriculture or recreation. Climate factors are factors that influence the weather and weather conditions [41]. However, since temperature, precipitation, and wind speed are already included in the list as individual factors, the general term climate factors can be excluded. Forest type is a group of forest ecosystems of generally similar composition. However, since the forest type is relatively similar in the entire country of Cyprus (pines, cedrus, and platanion) its influence on the susceptibility map would be very small. For this reason, it was excluded from the final list. Forest density indicates which percentage of an area is covered by trees and is included in the final list. Surface roughness or TRI (Topographic Roughness Index) expresses the amount of elevation difference between adjacent cells of a DEM [42]. However, since we already included slope and elevation as factors there was no need to add TRI to the list. Lastly, soil was interpreted as the

moisture content of the soil. When going through the list of 39 factors again CTI (Compound Topographic Index) was found, which is a measure of soil moisture potential. Since the definition is the same, soil was renamed CTI. However, after further investigation, it was found that CTI and TWI are interchangeable [43] leading to the exclusion of CTI from the final list.

Based on this reasoning the final list of factors is as folows:

- Elevation
- TWI (Topographic Wetness Index)
- Temperature
- Precipitation
- Wind speed
- Distance to roads
- Distance from powerlines
- NDVI
- Land use
- Slope
- Distance to populated areas/settlements
- Distance from tourist spots
- Forest density

5.2 Method(s)

Based on the literature reviewed, AHP would be a suitable method to use for this project due to its overall lack of cons and its suggestions for improvement. However, after diving into the functioning of the method it was found to be less reliable as opposed to other options. The way AHP works is that it uses a pairwise comparison matrix to calculate the weight each factor carries in the prediction of the occurrence of a wildfire. Each factor is given a value from 1 to 5 representing its priority over another value. However, the classification of this priority is to be decided by the client or in this case myself making this process biased. Therefore, I opted to go for a different method. After consulting with my supervisor the method used is Logistic Regression, and specifically binary logistic regression since the ground truth of this project is in binary form. This ground truth refers to whether or not a wildfire has occurred in a specific location and is presented in either 0, meaning no fire occurred, or 1, when a fire has occurred. Furthermore, we are dealing with multiple binary logistic regression since more than 1 factor is considered.

If Logistic Regression does not provide desired results Random Forest will be used to hopefully get better accuracy since this model is stronger as it uses multiple decision trees to base its prediction on. Note, Random Forest can be used for regression and classification problems [44]. Classification refers to a situation where the outcome is discrete in nature or categorical. The outcome fits into a category like ‘true’ or ‘false’ [45]. Regression on the other hand does not. The outcome is continuous in nature [45]. Since this project aims at classifying each area of Cyprus ranging from low susceptibility to high susceptibility, Random Forest Classification will be used.

5.2.1 Logistic Regression

Let us take a closer look at the functioning of logistic regression. Logistic regression is a "Supervised machine learning" algorithm that can be used to model the probability of a certain event happening [46]. There are multiple types of logistic regression, amongst which multiple logistic regression. When multiple independent variables are considered when predicting the output, which is the case for this project.

In order to understand logistic regression one must understand linear regression. Linear regression attempts to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation/best fit line to the observed data [47]. The sum of distances of all the data points to the line has to be minimal. Logistic regression attempts to do the same. However, in this case, the data is fitted to a logit function.

In other words, logistic regression attempts to predict the probability of an event happening by fitting data to a logit function. The smaller the sum of distances to the line, the higher the accuracy. It plots this line using the following formula: $y(x) = \beta_0 + \beta_1 x + \ldots + \beta_n x$, where β_n represents the coefficient of each factor x.

5.2.2 Random Forest

Random forest on the other hand is an ensemble learning method composed of multiple decision trees [48]. The output is based on the majority of voting. It can outperform any of the individual models due to the ensemble of decision trees.

The multiple trees run in parallel without interaction amongst them. There should be as little correlation amongst the trees as possible so they can protect each other from their individual errors. In order to achieve this bagging and feature randomness are used.
Bagging, or bootstrap aggregation, allows each tree to randomly sample from the dataset. Feature randomness is applied when splitting a node of the tree. In this case, the tree can only choose from a random subset of features to base the split on [48].

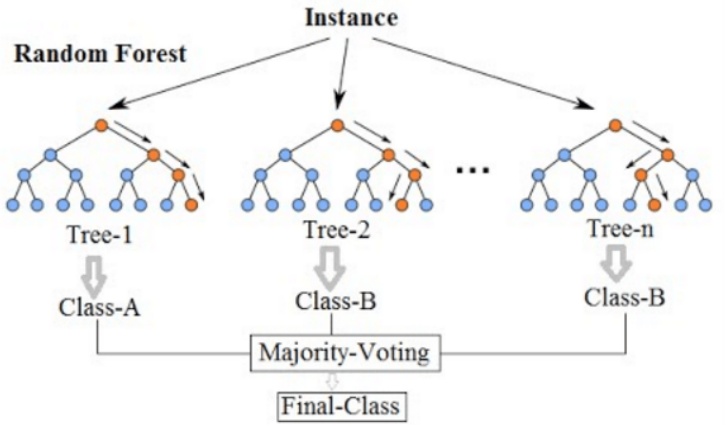

Figure 2: Random Forest model [49].

5.3 Preconditions

As mentioned in the introduction, prevention plans should be optimized. These renewed prevention measures should then be put in place starting with the areas that are at the highest risk of wildfires. This project attempts to find methods that best predict this risk and if one would suffice it can be used to create a wildfire susceptibility map. This map could then be used to locate these areas facing the biggest risk. However, if this map is used to base the distribution of resources on, it better be as accurate as possible. Since it could otherwise lead to areas of high risk being classified as low risk. This could in turn lead to an out-of-control wildfire, due to the lack of prevention methods put in place, taking the lives of both inhabitants and wildlife.
However, since 100% accuracy is close to impossible a minimum accuracy must be put in place to determine the usefulness of a method. Ideally, the method would have an accuracy of above 80% since this would mean an improvement of the already existing map generated by my supervisor. However, due to the possible inaccuracies in the functioning of the model, stakeholders, like fire departments, should be informed. They could take this information into consideration when deciding how to distribute their resources.
As for the number of factors, this is not dependent on the method used and therefore should not reflect the validity of the methods. However, it is an improvement mentioned in a lot of the case studies. Therefore one should strive to use as many as possible.

Chapter 6 Realization

6.1 Dataset Generation

Based on the final list of factors the dataset to be used in training the models should be generated. In order to do this data of each individual factor for Cyprus has to be gathered as well as the information used as the ground truth. The majority of this data was presented to me by two associates, Asfa Jamil and Chirag Padubidri, of my supervisor. However, the data for windspeed and precipitation I had to gather myself. Asfa provided me with two links [50][51] containing the information. A geodata downloader [52] was used to access the information and a geodata converter [53] to convert the information into a shapefile readable by ArcMap. However, the data for both factors was not complete. It only covered certain points of Cyprus. When converting the shapefile to a raster it would mean that the majority of Cyprus would have null values for the uncovered points as can be seen in figure 3.

Figure 3: Raster image of windspeed.

In order to solve this problem interpolation was used. Specifically natural neighbor interpolation, which predicts the value of a point based on the closest input samples. A Voronoi diagram is constructed for each input point. Next, a new Voronoi polygon is created for the interpolation point, and the proportion of overlap is used as the predicted value [54].
After performing the interpolation the raster of windspeed presents as follows (see figure 4). As you can see, each point is now covered, meaning each point contains a value for windspeed.

Figure 4: Raster image of interpolated windspeed.

The same steps were taken for precipitation leaving us with the raster presented in figure 5.

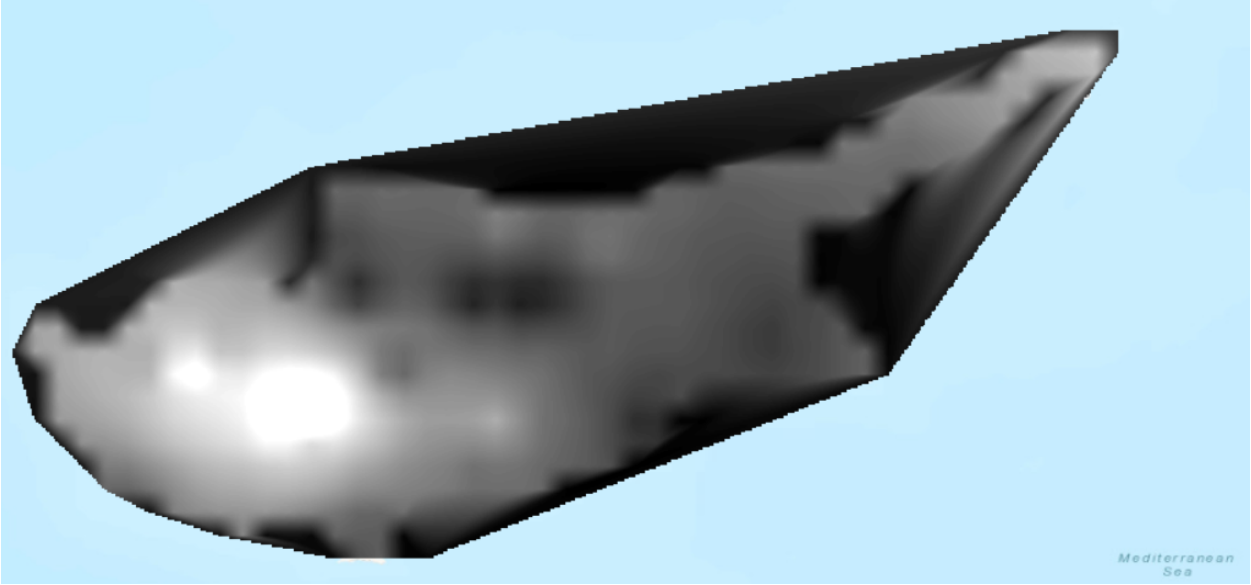

Figure 5: Raster image of interpolated precipitation.

The next step was to generate the attribute table, which would later be used as the input dataset for each method. Since the ground truth was delivered in two separate files, one containing historical fire points and the other 'no fire' points, this step had to be executed twice.
The tool used in ArcMap to generate these tables was the Extract Multi Values to Points tool. It takes a shapefile as input point feature, in this case either the historical fire or no fire file. Furthermore, it takes all the factors as input rasters. It then combines all the information of the factors in the attributes table of the ground truth.

After both attributes tables were generated they needed to be merged to create the final dataset. This was done using the merge tool in ArcMap, which takes two shapefiles as input and merges them.

Next, the dataset was exported and fine-tuned in excel before turning it into a CSV file. This finetuning needed to take place since the dataset contained columns that were not needed to calculate the weights such as ID number or the date on which a wildfire started. In the end, the CSV file contained the following columns:

- Fire, containing the ground truth, so either 0 or 1
- Windspeed
- Precipitation
- Altitude
- Temperature (the average yearly temperature from 1981 to 2010)
- Canopy height, representing the forest density
- TWI, Topographic Wetness Index
- Slope
- NDVI, Normalized Difference Vegetation Index
- DTPL, Distance to Power Lines
- DTR, Distance to Roads
- DTTS, Distance to Tourist Spots
- DTPA, Distance to Populated Areas

Note, land use is not present in this dataset due to the size of the file being to large and the fact that most of Cyprus was classified with the same value. Therefore, not adding significant impact to the prediction.

6.1.1 Import Dataset into Python

Since this step is the same for each method it will be discussed once within this section. First, the dataset is imported and separated into our feature and target column, x and y.

```
data = pd.read_csv('Data.csv')
x = data.iloc[:, data.columns != 'FIRE']
y = data.FIRE
```

Next, the dataset is divided into training and test sets with an 80 to 20 ratio, so 80% of the dataset will be used for training and 20% for testing.

```
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.20,
random_state=5, stratify=y)
```

After which the ‘X’ sets are scaled using a minmax scaler, which scales the data to a value between zero and one.

```
scaler = MinMaxScaler(feature_range=(0,1))
```

The next step involves the implementation of the model which will be discussed in the next section.

6.2 Logistic Regression

As discussed in chapter 5, logistic regression is a supervised machine learning algorithm that can be used to predict the probability of an event happening. It does so by fitting a logit function to the data, where the distance from all the data points to the line has to be minimal.

6.2.1 Logistic Regression Model

The model is built in python using the logistic regression model of sklearn and optimized using its parameters. In this case, the maximum number of iterations was altered since without doing so the model would crash due to it reaching this number before finishing its predictions. Next, the model was fitted with the scaled 'X' training set and the 'y' training set.

```
model = LogisticRegression(max_iter=10000)
model.fit(X_train_scaled, y_train)
```

After the model is fitted, the coefficients can be printed which represent the weights that will be used to generate the wildfire susceptibility map.

```
print(pd.DataFrame({'feature':list(x.columns),'feature_importance':[i for i
in model.coef_[0]]}))
```

For the full code, see Appendix A.

6.2.2 Map Generation

Based on the weights generated by the python code, the wildfire susceptibility map is generated. Figure 6 shows the feature importance of each factor of the logistic regression model in python. The feature importance represents the weights, which represent the impact a factor has on the prediction of the probability of a wildfire occuring.

Feature	Feature Importance
Altitude	3.672317
Windspeed	0.369040
Precipitat	4.121611
DTPA	1.589780
DTPL	-0.589951
DTR	-0.457049
DTTS	-0.341228
NDVI	3.190523
Slope	1.465866
TWI	-0.082883
Forest density	-3.784703
Temperature	2.773461

Figure 6: Output weights of Logistic Regression.

In ArcMap these weights were used in the raster calculator as followed, implementing the function mentioned in 5.2.1:
(- 10.06190526) + ("Altitude.tif" * 3.672317) + ("Windspeed" * 0.369040) + ("Precipitation" * 4.121611) + ("DTPA.tif" * 1.589780) + ("DTPL.tif" * - 0.589951) + ("DTR.tif" * - 0.457049) + ("DTTS.tif" * - 0.341228) + ("NDVI.tif" * 3.190523) + ("Slope.tif" * 1.465866) + ("TWI.tif" * - 0.082883) + ("CY_CANOPY.tif" * - 3.784703) + ("avg_yearly_temperature_1981_2010.tif" * 2.773461)

The output of this calculation contains the wildfire susceptibility of Cyprus and is presented in a raster image as can be seen in figure 7, where green represents low susceptibility and red high susceptibility.

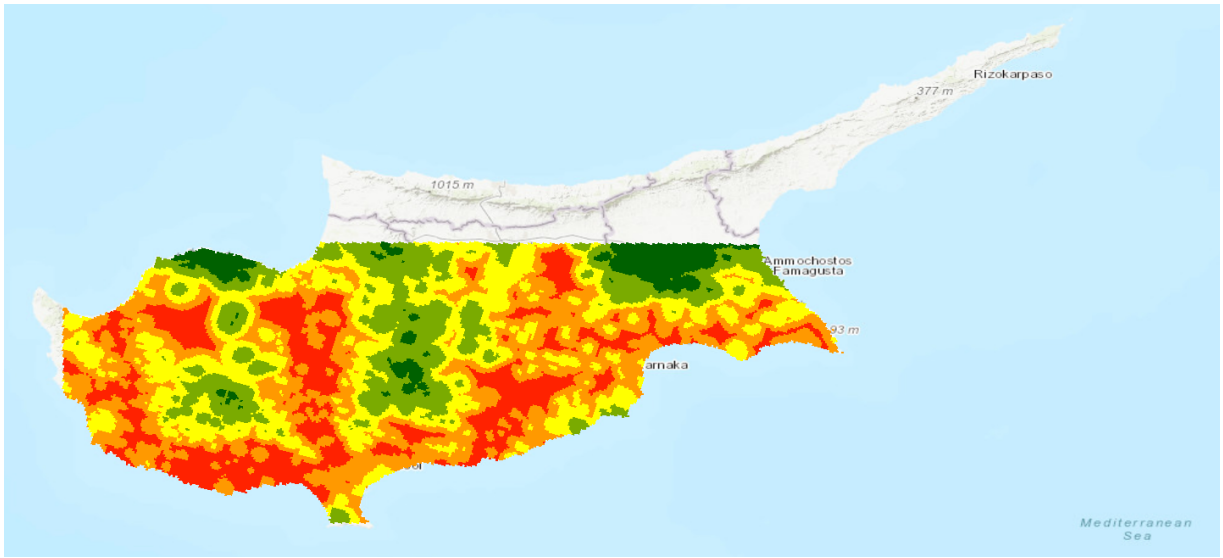

Figure 7: Wildfire susceptibility of Cyprus using logistic regression.

6.3 Random Forest Classification

As discussed in chapter 5, random forest is an ensemble learning method composed of multiple decision trees. The output is based on the majority of voting amongst these trees. These trees run in parallel and due to bagging and feature randomness there is as little correlation amongst them as possible.

6.3.1 Random Forest Classification Model

The Implementation of the random forest classification model is similar to that of logistic regression. The only difference takes place in building the model, since a different model is used.

```
model = RandomForestClassifier()
```

For the full code, see Appendix B.

6.3.2 Map Generation

Since the mapping of this method is not as simple as using the feature importance/weights in combination with a logit function, it was not accomplished within this project. However, a closer look was taken into the prediction process behind random forest. As mentioned in chapter 5.2.2, random forest is an ensemble learning method. It uses multiple decision trees to make a prediction of either 0 or 1. Zero representing the event, a wildfire, not happening. And one representing the event, a wildfire, happening. Since we used the classification model, the output was determined based on majority of voting. However, since a value of zero and one did not suffice for the risk map a closer look was taken at the predict_proba function in python. This function returns two values between 0 and 1 which represents the risk of no wildfire occurring and a wildfire occurring. However, in order to map this method, random forest, in ArcMap a better understanding is needed of its implementation in generating susceptibility maps.

Chapter 7 Evaluation

For each method three validation methods are used. Starting with the accuracy of the training and test sets. As mentioned in chapter 6, the training and test sets were created using the train_test_split() function of sklearn.model_selection, and was split using an 80 to 20 ratio respectively.

```
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.20,
random_state=5, stratify=y)
```

The accuracy of the training set is calculated using the score method present in python.

```
train_acc = model.score(X_train_scaled, y_train)
```

In order to calculate the accuracy of the test set, first a prediction set has to be created. This is done using the predict function present in python.

```
y_pred = model.predict(X_test)
```

Using this prediction set the accuracy of the test set is calculated using the accuracy_score method from sklearn.metrics.

```
test_acc = accuracy_score(y_test, y_pred)
```

The test accuracy is most important since it represents the accuracy of our model when used on data that was not used for training the model.

Next to the accuracy, a confusion matrix is also generated using the confusion_matrix() method of sklearn.metrics.

```
cm = confusion_matrix(y_test, y_pred)
```

The confusion matrix gives an overview of which percentage of the data was correctly classified and which percentage was wrongly classified. Further diving into what percentage was classified as a true or false negative (0) and a true or false positive (1).

Lastly, an AUC_ROC Curve is plotted using the roc_curve method and roc_auc_score of sklearn.metrics.

```
fpr, tpr, _ = roc_curve(y_test, y_pred)
```

```
auc = metrics.roc_auc_score(y_test, y_pred)
```

The ROC (Receiver Operator Characteristic) curve is a evaluation metric for a binary classification problem [55]. It plots the true positive rate (TPR) against the false positive rate (FPR). The closer the curve is to the top left corner of the graph, the better its performance [56]. The AUC (Area Under Curve) summarizes the performance of a classifier to a single measure. It represents the ability of a classifier to distinguish between classes [55]. The higher the AUC, the better the performance of the model. When the AUC is equal to one, the model can perfectly distinguish between the positive (1) and negative (0) class. When 0.5<AUC<1, the classifier has a high chance of being able to distinguish between the positive and negative classes. When the AUC is equal to 0.5 the classifier is predicting randomly or the same class for each data point.

For the complete code of each evaluation method, see Appendix C.

7.1 Logistic Regression

7.1.1 Accuracy

In figure 10 the accuracies of both the training and test set for logistic regression can be seen.

```
The accuracy for training set is 70.01967592592592
The accuracy for test set is 70.34413580246913
```

Figure 10: Training and test accuracies of logistic regression.

Since the accuracy of the test set is only 70% a closer look was taken as to how fast the model reaches this accuracy. In order to do so, the accuracy of the model was plotted against the number of data points which resulted in the following graph present in figure 11.

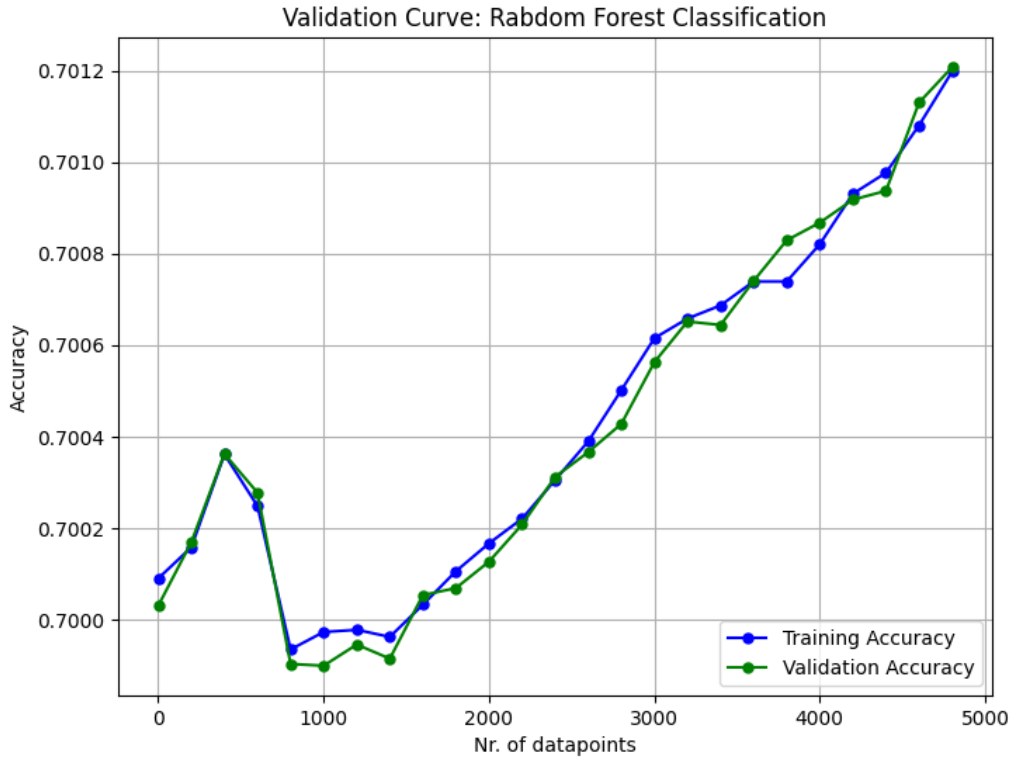

Figure 11: Accuracy of logistic regression plotted against the number of data points.

As can be seen, after roughly 5000 data points the model has already reached its optimal accuracy. The use of another method would be advised to try and improve this accuracy.

7.1.2 Confusion Matrix

The confusion matrix of the logistic regression model can be seen in figure 12.

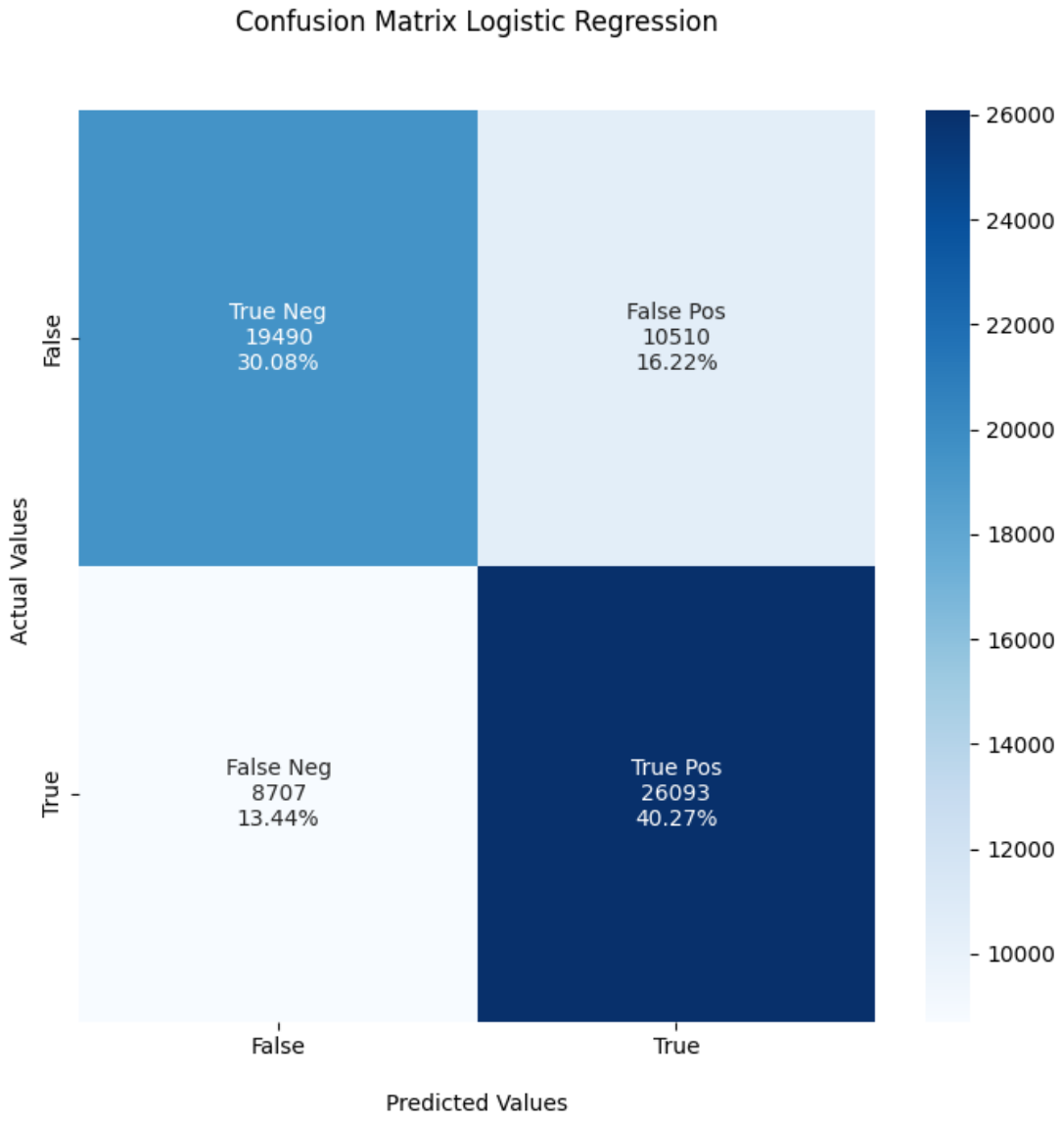

Figure 12: Confusion Matrix of logistic regression.

As can be seen, 16.10% of the dataset is falsely classified as negative. Meaning that a wildfire has occurred in that spot but was classified as no wildfire occurred. Furthermore, 13.72% were falsy classified as positive. Meaning that no wildfire occurred in that location, but was classified as if a wildfire did occur there. The percentage of data that was correctly classified is roughly 70% which supports the findings of the test sets accuracy of 70.34%.

7.1.3 AUC-ROC Curve

In figure 13 the ROC curve of the logistic regression model is depicted including the AUC score.

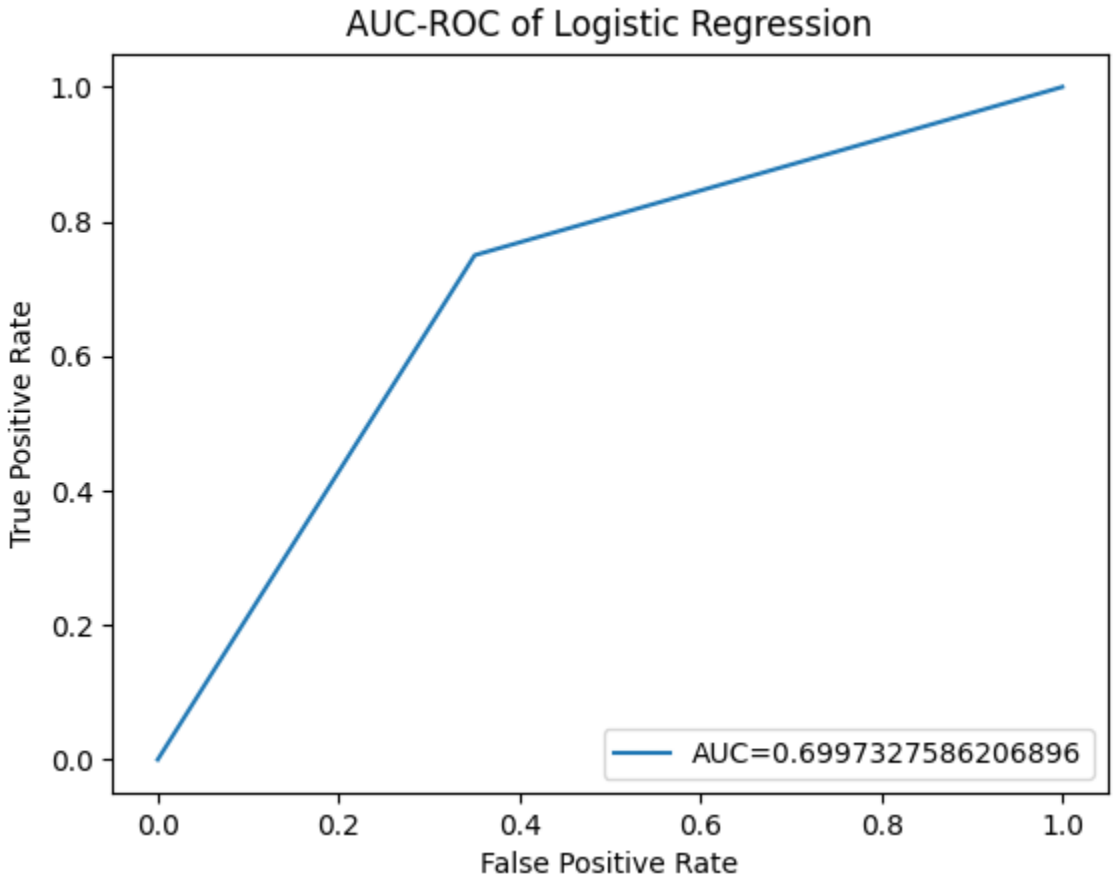

Figure 13: AUC-ROC curve of logistic regression.

As can be seen in figure 13, the AUC value is roughly 0.6997. As discussed earlier, an AUC between 0.5 and 1 indicates that the model has a high chance of being able to distinguish between the positive and negative classes. The AUC is not equal to 1 so there is the opportunity for improvement, but overall a good result. Next to that, it supports the test accuracy found earlier.

7.1.4 Feature Importance

Figure 14 shows the feature, or factor, importance of each factor in determining the susceptibility of an area to wildfires.

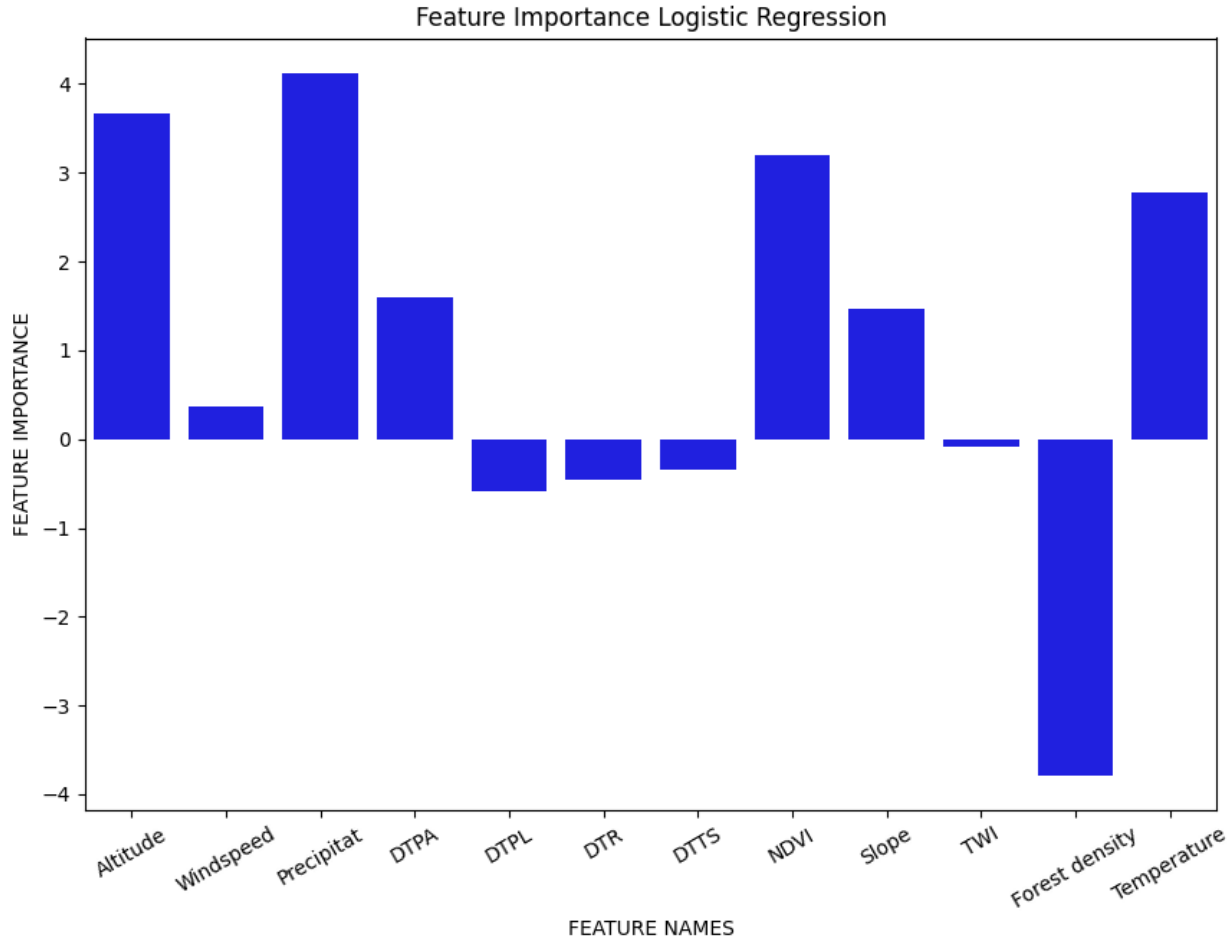

Figure 14: Feature importance of logistic regression.

As can be seen in figure 14, there are 4 factors that have a negative impact on predicting the risk of wildfires of which forest density has the largest negative impact. Precipitation on the other hand has the largest positive impact on predicting the wildfire occurrence probability.

7.2 Random Forest Classification

7.2.1 Accuracy

In figure 15 the accuracies of both the training and test set of random forest classification can be seen.

```
The accuracy for training set is 99.96180555555556
The accuracy for test set is 88.64197530864197
```

Figure 15: Training and test accuracies of random forest classification.

Since the accuracy of the test set is above 80% the results are acceptable when compared with the precondition. However, to answer the question if optimization of this model should be attempted a validation curve was plotted which plots the accuracy of the model against the number of trees. The results can be seen in figure 16.

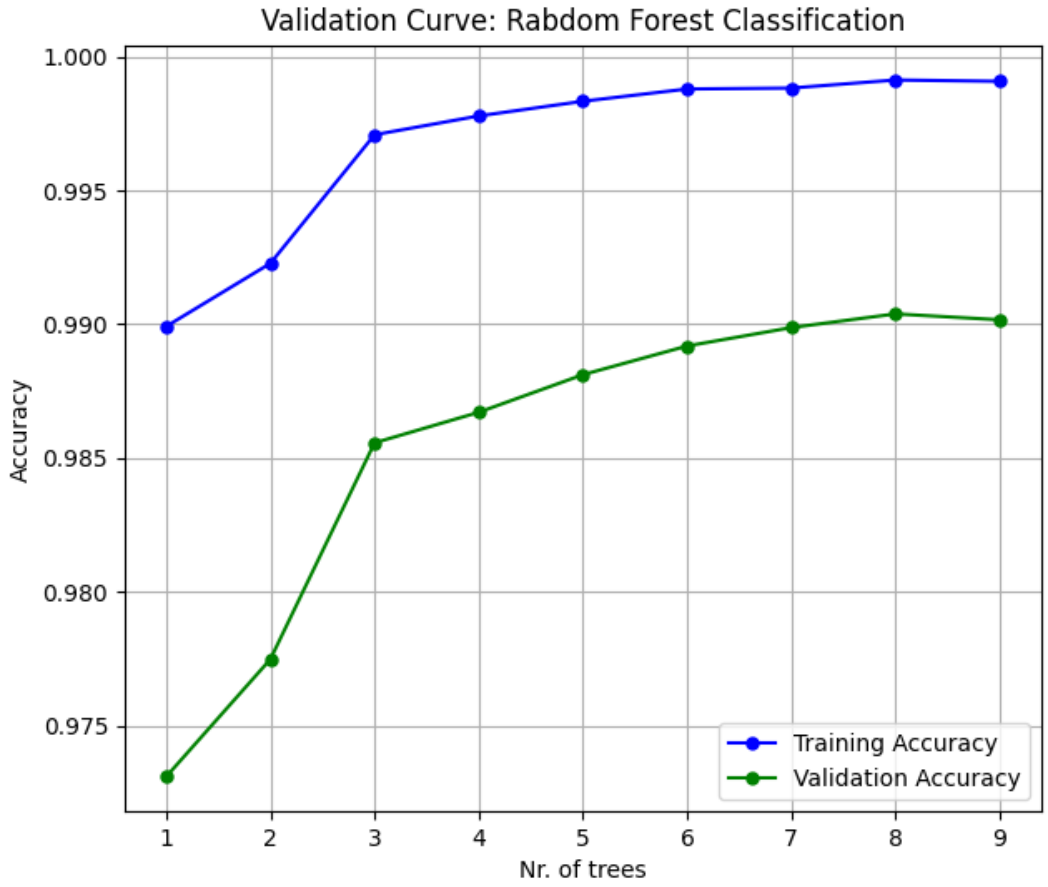

Figure 16: Validation curve of Random Forest.

As can be seen in figure 16, after roughly seven trees, the model has reached its optimal accuracy. Optimization could take place however significant improvement is not expected.

7.2.2 Confusion Matrix

The confusion matrix of the random forest classification model can be seen in figure 17.

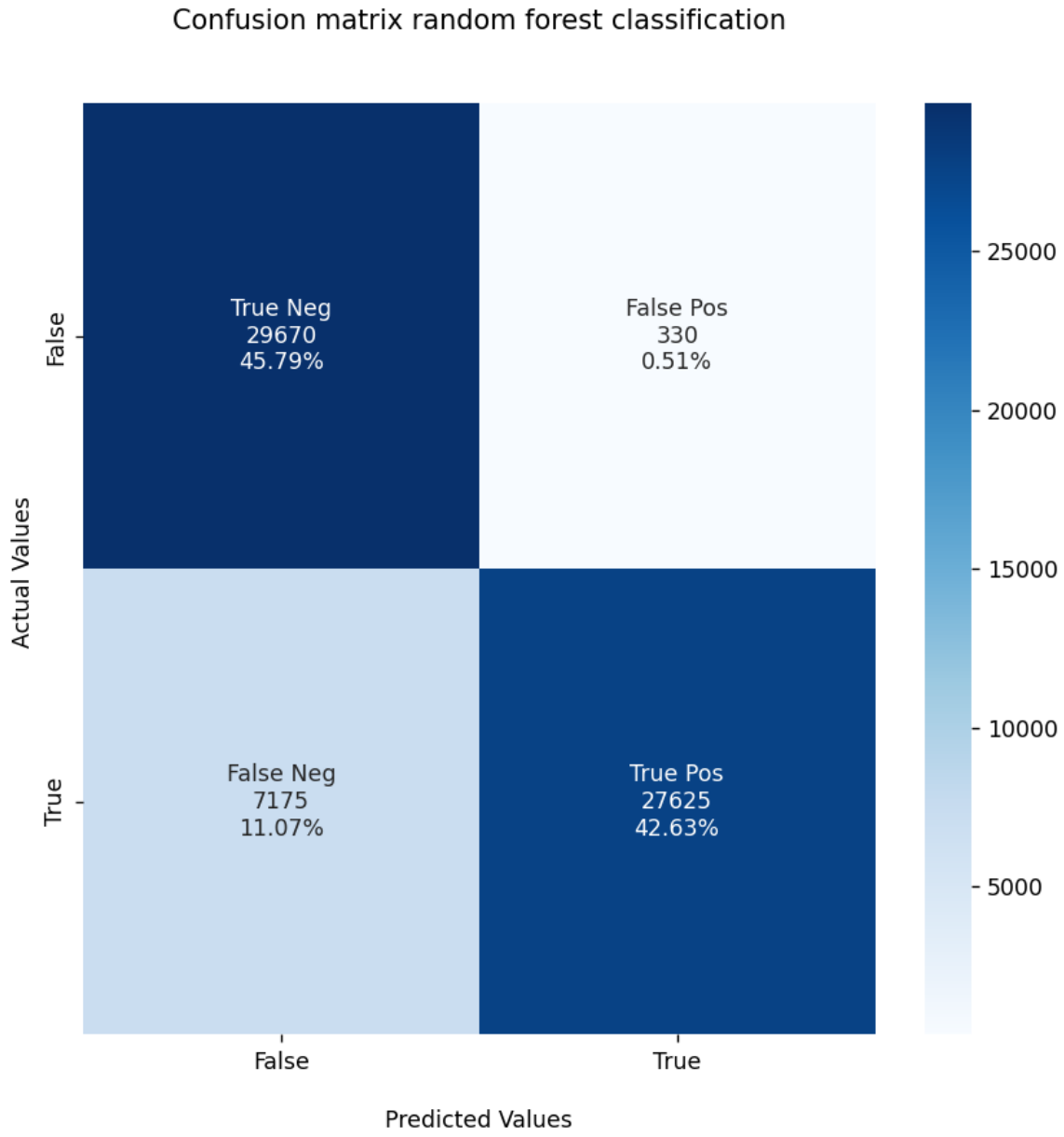

Figure 17: Classification matrix of random forest.

As can be seen, 11.07% of the dataset is falsely classified as negative. Meaning that 11.07% of the data points as classified as if no wildfire occurred in that location. However, based on the actual value of that location it should have been classified as a wildfire has occurred there. Next to that, 0.51% is falsely classified as positive. Meaning that it was classified as if a wildfire had occurred in that location even though no wildfire has occurred. Furthermore, its correctly classified percentage supports the test accuracy of 88% as seen in figure 15.

7.2.3 AUC-ROC Curve

In figure 18 the ROC curve of the random forest classification model is depicted including the AUC score.

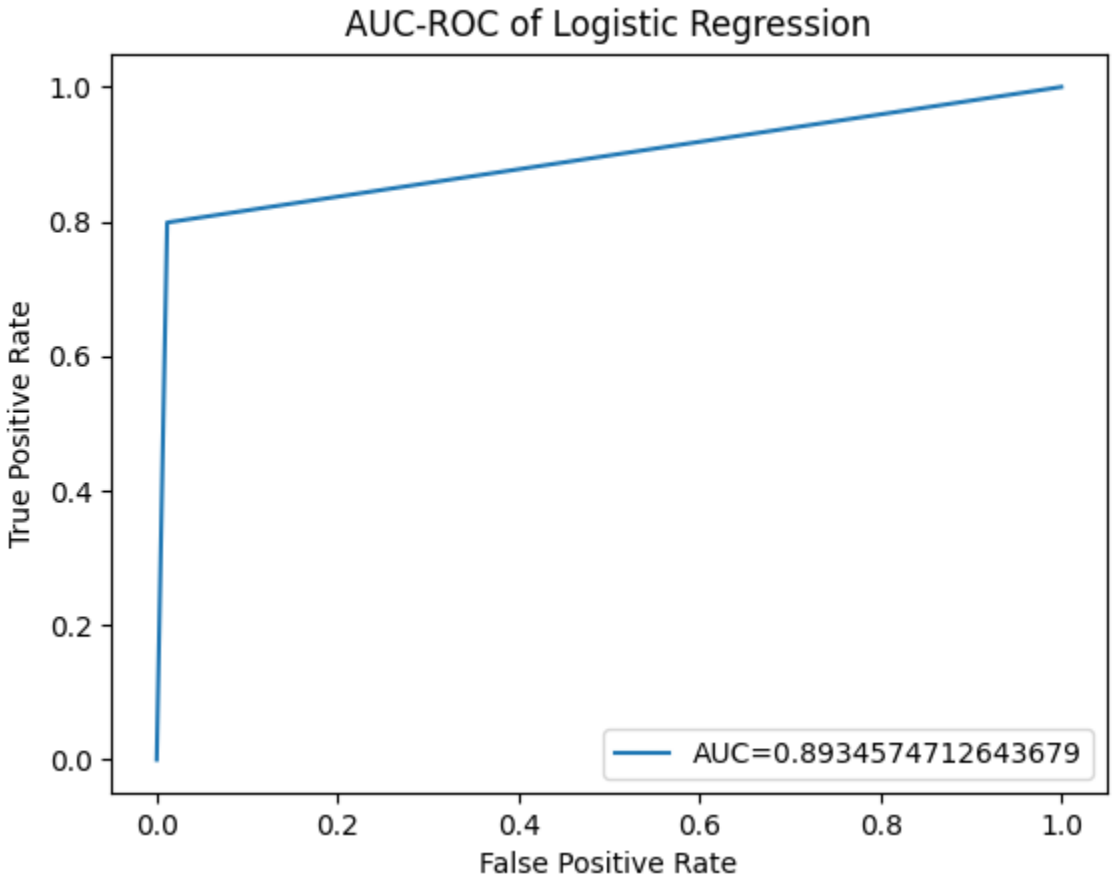

Figure 18: AUC-ROC curve of random forest classification.

As can be seen in figure 18, the AUC value is roughly 0.8934. As discussed earlier, an AUC between 0.5 and 1 indicates that the model has a high chance of being able to distinguish between the positive and negative classes. The AUC of random forest is higher than that of logistic regression and therefore a better model to use in creating a wildfire susceptibility map.

7.2.4 Feature Importance

Figure 19 shows the feature, or factor, importance of each factor in determining the susceptibility of an area to wildfires.

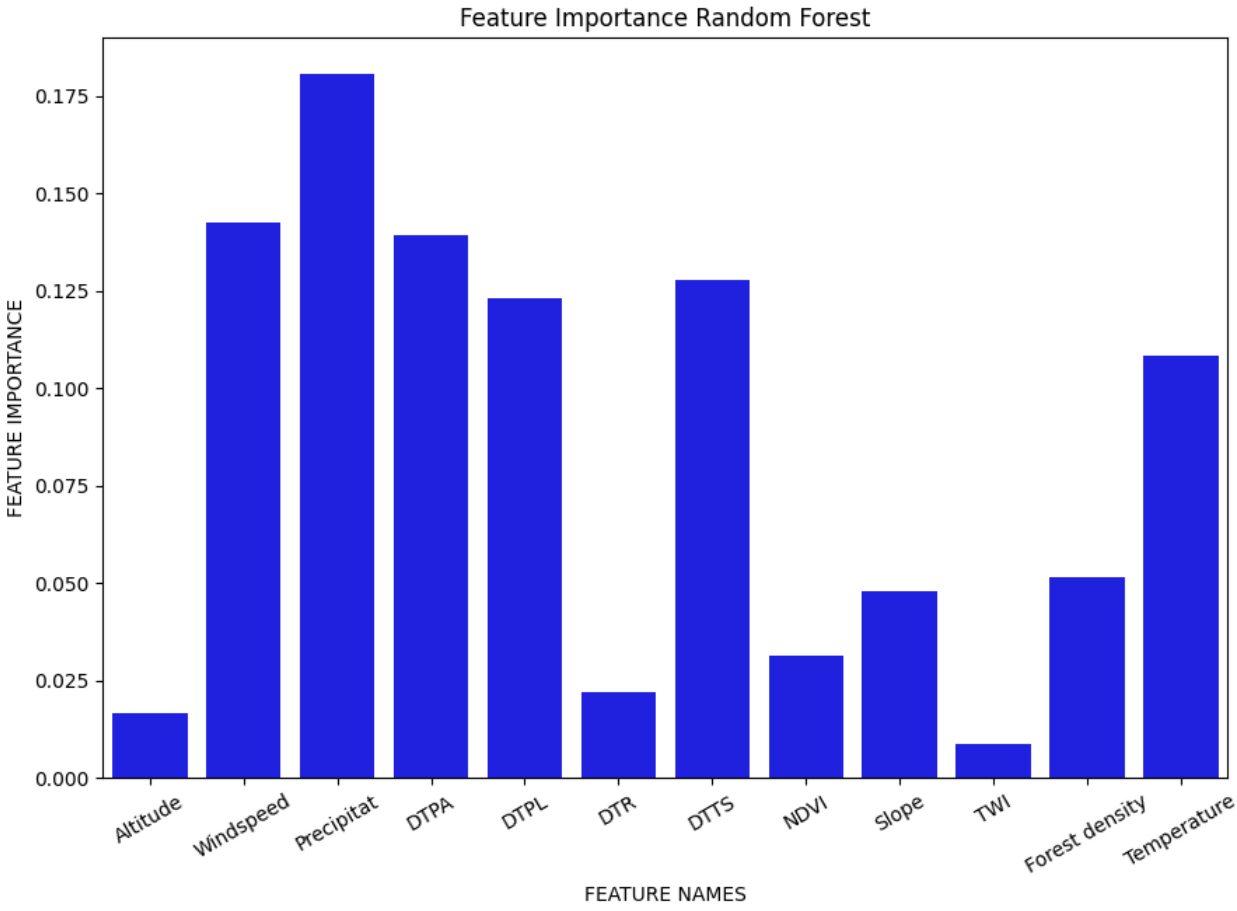

Figure 19: Feature importance of random forest classification.

As shown in figure 19, no factors with a negative feature importance exist. Meaning, each factor has a positive impact on determining the risk of a wildfire occurring. Just like with logistic regression, precipitation has the highest impact. This is expectable since rain is the enemy of fire.

Chapter 8 Conclusion & Discussion

The goal of this paper was to answer the main research question: What method would be best suitable to predict the wildfires susceptibility of Cyprus, and which factors should be taken into account? To answer this question, a better understanding was needed of which methods are out there and which factors influence wildfires. To do so multiple case studies were reviewed. In those studies, nine methods were discussed including the factors they used to predict the risk of a wildfire occurring. Based on the results the methods showed and the location in which they were tested, AHP was selected. It had good recommendations as to how to improve its accuracy and no major disadvantages were mentioned. This method however presented a problem in that the classification of the priority of a factor was based on human judgment. This could lead to unreliable results and therefore the method was discarded. After going back over the background research and discussing it with my supervisor the following two methods were deemed useful: logistic regression and random forest.

As for the factors, a selection was made based on the results of the studies discussed in the background research. A list of 16 factors was comprised of which each was deemed influential on wildfire susceptibility. This list was then refined based on further research, excluding some factors due to their irrelevance in Cyprus or based on the fact that it was already contained in the list under different names. This left a list of 13 factors to be considered.

After the selection of the factors and methods, the dataset was composed, in ArcMap 10.8. During this process, the factor land use was excluded due to its file size and the fact that most of Cyprus was used to the same extent. The final dataset was then used during the implementation of the models. Starting with the logistic regression model, the overall accuracy of the test set was 70.344%. It falsely classified 29.66% of the dataset where 13.44% was wrongly classified as negative and 16.22% as positive. Even though the AUC value of the model was satisfactory, between 0.5 and 1, it was not enough to meet our preconditions of accuracy above 80%. When taking a closer look at the progression of the accuracy against the number of data points it was found that the overall accuracy was already accomplished after roughy 5000 datapoint and therefore an attempt to optimize this model is not recommended.

Random forest on the other hand had an overall test set accuracy of 88.64%. This percentage satisfies the precondition of >80% and therefore this model is satisfactory. These results were supported by the AUC score of 0.89 and the confusion matrix which found that only 11.58% of the dataset was falsely classified.

Another interesting result is the factor importance which was plotted for both models (see figure 20). This graph shows which factors had a positive or negative impact on determining the risk of a wildfire occurring. The feature importance of logistic regression showed that precipitation, altitude, NDVI, temperature, DTPA, and slope had the biggest impact ranging from biggest impact to lowest impact respectively. Random forest on the other hand classified their importance as follows: precipitation, windspeed, DTPA, DTTS, DTPL, temperature, forest density, slope, NDVI, DTR, altitude, and TWI, again ranging from high to low impact

respectively. Similar to logistic regression random forest found that precipitation had the highest impact. Next to that, random forest also ranked temperature, windspeed, DTPL, DTTS, and DTPA high on the impact list.

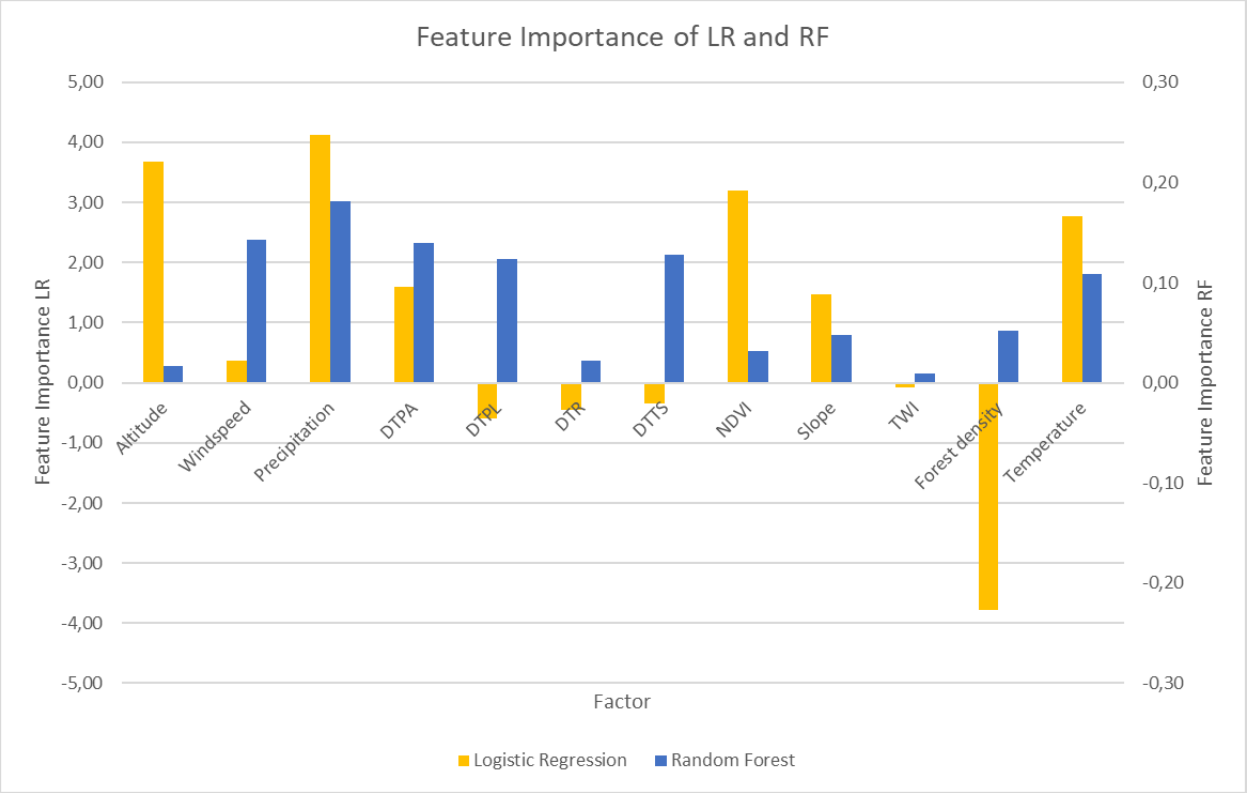

Figure 20: Feature importance of logistic regression and random forest.

Having said that, random forest found each factor to have a positive impact whereas logistic regression classified five factors as having a negative impact meaning they negatively impact the prediction of wildfire susceptibility.

To answer the research question, the random forest model is best suitable to predict the wildfire susceptibility of Cyprus. It meets the preconditions set in chapter 5.3 which took into account the literature reviewed and the preferred outcome of the supervisor. As for which factors should be taken into account. Based solely on the results of the feature importance of both models, precipitation, altitude, NDVI, temperature, DTPA, and slope should definitely be taken into account. The remaining factors can also be taken into account since the random forest model, which presented the highest accuracy, found them to be of importance. However, to verify this another method should be implemented and used to compare results.

Chapter 9 Future work

Given the current state of the research conducted the following need to be addressed in future work. First of all, the accuracy of the models should be attempted to be improved. Especially that of random forest since it has a higher potential to be improved. To do that one could include more factors since an increase in factors is related to an increase in data on which the model can base its predictions. Next to that, a look can be taken into tuning hyperparameters as it could also potentially improve accuracy. Additionally, since the creation of the wildfires susceptibility map of random forest is not as arbitrary as that of logistic regression it was not covered in this paper. Therefore, this map should be created, as part of future work, within the ArcMap environment. As for the logistic regression map, this can be shared with stakeholders such as firefighters to be used to educate them on the risks they are facing. After which it can be used to reevaluate the distribution of their resources. However, as mentioned in 5.3, the stakeholders should be informed of the possible inaccuracies due to its accuracy not being 100%.

Furthermore, as already mentioned in chapter 8, to validate the importance of each factor a different method like SVM (Support Vector Machine) should be implemented. Comparing the current importances which those of SVM would provide a better-informed evaluation.

Lastly, as discussed in this paper there are multiple methods to prevent or suppress wildfires. However, time did not allow for this project to also consider which methods should be put in place and more specifically in what locations. This is something that can be done as future work. Besides the methods that are already discussed, other methods should also be investigated. Next to that, a closer look should be taken into what resources are available in Cyprus. This could help decide which methods are optional and which are not.

Bibliography

[1]A. France-Presse, "'Pure hell': Cyprus hit by worst forest fire in decades," *The Guardian*, Jul. 04, 2021. https://www.theguardian.com/world/2021/jul/04/cyprus-says-deadly-forest-fire-close-to-being-under-control

[2]"Dixie Fire | Welcome to CAL FIRE," *www.fire.ca.gov*, Oct. 25, 2021. https://www.fire.ca.gov/incidents/2021/7/13/dixie-fire/

[3]B. McDonald, S. Burrous, E. Weingart, and M. Felling, "Inside the Massive and Costly Fight Against the Dixie Fire," *The New York Times*, Oct. 11, 2021. [Online]. Available: https://www.nytimes.com/interactive/2021/10/11/us/california-wildfires-dixie.html

[4]M. Jones, A. Smith, R. Betts, J. Canadell, I. Prentice, and C. Le Quéré, "Climate Change Increases the Risk of Wildfires," Jan. 2020. [Online]. Available: https://www.preventionweb.net/files/73797_wildfiresbriefingnote.pdf

[5]P. Hessburg, "Why wildfires have gotten worse -- and what we can do about it," *Ted.com*, May 2017. https://www.ted.com/talks/paul_hessburg_why_wildfires_have_gotten_worse_and_what_we_can_do_about_it

[6]M. Jurvélius, "HEALTH AND PROTECTION | Forest Fires (Prediction, Prevention, Preparedness and Suppression)," *Encyclopedia of Forest Sciences*, pp. 334–339, 2004, doi: 10.1016/b0-12-145160-7/00277-5.

[7]K. Hagmann, P. Hessburg, and S. J. Prichard, "How years of fighting every wildfire helped fuel the Western megafires of today," *The Conversation*, Aug. 02, 2021. https://theconversation.com/how-years-of-fighting-every-wildfire-helped-fuel-the-western-megafires-of-today-163165

[8]
P. Zhao, F. Zhang, H. Lin, and S. Xu, "GIS-Based Forest Fire Risk Model: A Case Study in Laoshan National Forest Park, Nanjing," *Remote Sensing*, vol. 13, no. 18, p. 3704, Sep. 2021, doi: 10.3390/rs13183704.

[9]
A. Novo, N. Fariñas-Álvarez, J. Martínez-Sánchez, H. González-Jorge, J. M. Fernández-Alonso, and H. Lorenzo, "Mapping Forest Fire Risk—A Case Study in Galicia (Spain)," *Remote Sensing*, vol. 12, no. 22, p. 3705, Jan. 2020, doi: 10.3390/rs12223705.

[10]
G. Busico, E. Giuditta, N. Kazakis, and N. Colombani, “A Hybrid GIS and AHP Approach for Modelling Actual and Future Forest Fire Risk Under Climate Change Accounting Water Resources Attenuation Role,” *Sustainability*, vol. 11, no. 24, p. 7166, Dec. 2019, doi: 10.3390/su11247166.

[11]
S. Nikhil *et al.*, “Application of GIS and AHP Method in Forest Fire Risk Zone Mapping: a Study of the Parambikulam Tiger Reserve, Kerala, India,” *Journal of Geovisualization and Spatial Analysis*, vol. 5, no. 1, May 2021, doi: 10.1007/s41651-021-00082-x.

[12]
F. Sari, “Forest fire susceptibility mapping via multi-criteria decision analysis techniques for Mugla, Turkey: A comparative analysis of VIKOR and TOPSIS,” *Forest Ecology and Management*, vol. 480, p. 118644, Jan. 2021, doi: 10.1016/j.foreco.2020.118644.

[13]
Z. S. Pourtaghi, H. R. Pourghasemi, and M. Rossi, “Forest fire susceptibility mapping in the Minudasht forests, Golestan province, Iran,” *Environmental Earth Sciences*, vol. 73, no. 4, pp. 1515–1533, Jul. 2014, doi: 10.1007/s12665-014-3502-4.

[14]
H. R. Pourghasemi, “GIS-based forest fire susceptibility mapping in Iran: a comparison between evidential belief function and binary logistic regression models,” *Scandinavian Journal of Forest Research*, vol. 31, no. 1, pp. 80–98, Aug. 2015, doi: 10.1080/02827581.2015.1052750.

[15]
M. H. Nami, A. Jaafari, M. Fallah, and S. Nabiuni, “Spatial prediction of wildfire probability in the Hyrcanian ecoregion using evidential belief function model and GIS,” *International Journal of Environmental Science and Technology*, vol. 15, no. 2, pp. 373–384, Jun. 2017, doi: 10.1007/s13762-017-1371-6.

[16]
D. Tien Bui, K.-T. Le, V. Nguyen, H. Le, and I. Revhaug, “Tropical Forest Fire Susceptibility Mapping at the Cat Ba National Park Area, Hai Phong City, Vietnam, Using GIS-Based Kernel Logistic Regression,” *Remote Sensing*, vol. 8, no. 4, p. 347, Apr. 2016, doi: 10.3390/rs8040347.

[17]
G. Zhang, M. Wang, and K. Liu, “Forest Fire Susceptibility Modeling Using a Convolutional Neural Network for Yunnan Province of China,” *International Journal of Disaster Risk Science*, vol. 10, no. 3, pp. 386–403, Sep. 2019, doi: 10.1007/s13753-019-00233-1.

[18]P. Alcubierre *et al.*, “Prevention of Large Wildfires using the Fire Types Concept,” Mar. 2011. [Online]. Available: https://archief.nipv.nl/wp-content/uploads/sites/2/2022/03/201103-efi-prevention-of-large-wildfires-using-the-fire-types-concept.pdf

[19]B. Stefanović, D. Stojnić, and M. Danilović, “Multi-criteria forest road network planning in fire-prone environment: a case study in Serbia,” *Journal of Environmental Planning and Management*, vol. 59, no. 5, pp. 911–926, Jun. 2015, doi: 10.1080/09640568.2015.1045971.

[20]A. Laschi *et al.*, “Forest Road Planning, Construction and Maintenance to Improve Forest Fire Fighting: a Review,” *Croatian Journal of Forest Engineering*, vol. 40, pp. 207–219, Feb. 2019.

[21]M. Larjavaara, “Climate and forest fires in Finland – influence of lightning-caused ignitions and fuel moisture,” *Dissertationes Forestales*, vol. 2005, no. 5, Jan. 2005, doi: 10.14214/df.5.

[22]M. Demir, A. Küçükosmanoğlu, M. Hasdemir, T. Ozturk, and H. H. Acar, “Assessment of forest roads and firebreaks in Turkey,” *African Journal of Biotechnology*, vol. 8, no. 18, pp. 4553–4561, Sep. 2009.

[23]E. Chuvieco and R. G. Congalton, “Application of remote sensing and geographic information systems to forest fire hazard mapping,” *Remote Sensing of Environment*, vol. 29, no. 2, pp. 147–159, Aug. 1989, doi: 10.1016/0034-4257(89)90023-0.

[24]J. Cardille, S. Ventura, and M. Turner, “Environmental and Social Factors Influencing Wildfires in the Upper Midwest, United States,” *Ecological Application*, vol. 11, no. 1, pp. 111–127, Feb. 2001, doi: 10.2307/3061060.

[25]E. Rigolot, P. Fernandes, and F. Rego, “Managing wildfire risk: prevention, suppression,” *EFI Discussion Paper*, vol. 5, pp. 50–52, Dec. 2015.

[26]J. K. Agee and C. N. Skinner, “Basic principles of forest fuel reduction treatments,” *Forest Ecology and Management*, vol. 211, no. 1–2, pp. 83–96, Jun. 2005, doi: 10.1016/j.foreco.2005.01.034.

[27]D. D. Wade, *A Guide for Prescribed Fire in Southern Forests*. U.S. Department of Agriculture, Forest Service, Southern Region, 1989. Accessed: Apr. 13, 2022. [Online]. Available: https://books.google.nl/books?hl=en&lr=&id=x2YTAAAAYAAJ&oi=fnd&pg=PA1&dq=a+guide+for+prescribed+fire+in+southern+forests&ots=uhnWt5LOU5&sig=D9Qcv8Ikz1F25Re6G029HDP7RUk&redir_esc=y#v=onepage&q=a%20guide%20for%20prescribed%20fire%20in%20southern%20forests&f=false

[28]E. Marino, C. Hernando, R. Planelles, J. Madrigal, M. Guijarro, and A. Sebastián, “Forest fuel management for wildfire prevention in Spain: a quantitative SWOT analysis,” *International Journal of Wildland Fire*, vol. 23, no. 3, p. 373, 2014, doi: 10.1071/wf12203.

[29]P. Fernandes and H. Botelho, “A review of prescribed burning effectiveness in fire hazard reduction,” *International Journal of Wildland Fire*, vol. 12, pp. 117–128, Jul. 2003, doi: 10.1071/WF02042.

[30]C. J. Weston, J. Di Stefano, S. Hislop, and L. Volkova, “Effect of recent fuel reduction treatments on wildfire severity in southeast Australian Eucalyptus sieberi forests,” *Forest Ecology and Management*, vol. 505, Dec. 2021, doi: https://doi.org/10.1016/j.foreco.2021.119924.

[31]A. Duane, N. Aquilué, Q. Canelles, A. Morán-Ordoñez, M. De Cáceres, and L. Brotons, “Adapting prescribed burns to future climate change in Mediterranean landscapes,” *Science of The Total Environment*, vol. 677, pp. 68–83, Aug. 2019, doi: 10.1016/j.scitotenv.2019.04.348.

[32]R. M. Pacheco and J. Claro, “Prescribed burning as a cost-effective way to address climate change and forest management in Mediterranean countries,” *Annals of Forest Science*, vol. 78, no. 4, Dec. 2021, doi: 10.1007/s13595-021-01115-7.

[33]K. Launchbaugh and J. Walker, “CHAPTER 1: Targeted Grazing - A New Paradigm for Livestock Management.” Accessed: Nov. 04, 2021. [Online]. Available: https://rangelands.org/wp-content/uploads/2014/03/chapter_1_targeted_grazing.pdf

[34]J. Rouet-Leduc *et al.*, “Effects of large herbivores on fire regimes and wildfire mitigation,” *Journal of Applied Ecology*, vol. 58, no. 12, pp. 2690–2702, Sep. 2021, doi: https://doi.org/10.1111/1365-2664.13972.

[35]R. A. Bruegger, L. A. Varelas, L. D. Howery, L. A. Torell, M. B. Stephenson, and D. W. Bailey, “Targeted Grazing in Southern Arizona: Using Cattle to Reduce Fine Fuel Loads,” *Rangeland Ecology & Management*, vol. 69, no. 1, pp. 43–51, Jan. 2016, doi: 10.1016/j.rama.2015.10.011.

[36]C. Taylor, “CHAPTER 12: Targeted Grazing to Manage Fire Risk 10 KEY POINTS.” Accessed: Jan. 23, 2021. [Online]. Available: https://www.webpages.uidaho.edu/rx-grazing/Handbook/Chapter_12_Targeted_Grazing.pdf

[37]D. Bachelet, J. M. Lenihan, C. Daly, and R. P. Neilson, “Interactions between fire, grazing and climate change at Wind Cave National Park, SD,” *Ecological Modelling*, vol. 134, no. 2–3, pp. 229–244, Oct. 2000, doi: 10.1016/s0304-3800(00)00343-4.

[38]
A. H. Mader and Wouter Eggink, “A Design Process for Creative Technology,” *University of Twente Research Information*, pp. 568–573, 2014, [Online]. Available: https://research.utwente.nl/en/publications/a-design-process-for-creative-technology

[39]
“Climate & Weather,” *www.visitcyprus.com*. https://www.visitcyprus.com/index.php/en/practical-info/climate-weather#:~:text=Cyprus%20enjoys%20an%20intense%20Mediterranean

[40]
Remote Sensing Phenology, “NDVI, the Foundation for Remote Sensing Phenology | U.S. Geological Survey,” *www.usgs.gov*, Nov. 27, 2018. https://www.usgs.gov/special-topics/remote-sensing-phenology/science/ndvi-foundation-remote-sensing-phenology

[41]
“Climate factors,” *www.geo.fu-berlin.de*, Jun. 23, 2015. https://www.geo.fu-berlin.de/en/v/iwm-network/learning_content/environmental-background/basics_climategeography/Climate-factors/index.html#:~:text=Climate%20factors%20are%20terrestrial%20factors

[42]
“Topographic Ruggedness Index,” *www.usna.edu*, Apr. 22, 2022. https://www.usna.edu/Users/oceano/pguth/md_help/html/topo_rugged_index.htm#:~:text=General%20discussion%20of%20Roughness

[43]
“Topographic wetness index,” *Wikipedia*, Feb. 06, 2021. https://en.wikipedia.org/wiki/Topographic_wetness_index#:~:text=The%20topographic%20wetness%20index%20(TWI

[44]
A. Raj, “A Quick and Dirty Guide to Random Forest Regression,” *Medium*, Jun. 11, 2021. https://towardsdatascience.com/a-quick-and-dirty-guide-to-random-forest-regression-52ca0af157f8#:~:text=Random%20forest%20is%20a%20type

[45]
G. L. Team, “Random forest Algorithm in Machine learning: An Overview,” *GreatLearning Blog: Free Resources what Matters to shape your Career!*, Feb. 19, 2020. https://www.mygreatlearning.com/blog/random-forest-algorithm/#ClassifierVs.Regressor

[46]
H. Bonthu, “An Introduction to Logistic Regression,” *Analytics Vidhya*, Jul. 11, 2021. https://www.analyticsvidhya.com/blog/2021/07/an-introduction-to-logistic-regression/#:~:text=Logistic%20Regression%20is%20a%20%E2%80%9CSupervised

[47]
H. Bonthu, “Simple Linear Regression | Learn Simple Linear Regression (SLR),” *Analytics Vidhya*, May 23, 2021. https://www.analyticsvidhya.com/blog/2021/05/learn-simple-linear-regression-slr/

[48]
J. Huneycutt, “Implementing a Random Forest Classification Model in Python,” *Medium*, May 21, 2018. https://medium.com/@hjhuney/implementing-a-random-forest-classification-model-in-python-583891c99652#:~:text=Random%20forests%20algorithms%20are%20used

[49]
A. Chakure, “Random Forest Classification and Its Implementation,” *The Startup*, Nov. 06, 2020. https://medium.com/swlh/random-forest-classification-and-its-implementation-d5d840dbead0

[50]
“Layer: RasterT_Idw_win3_SpatialJoin (ID: 0),” *superworld-gis.cyens.org.cy*. https://superworld-gis.cyens.org.cy/server/rest/services/Hosted/winds_complete_data/FeatureServer/0

[51]
“Layer: RasterT_afr1_TableToExcel_XYTableToPoint (ID: 0),” *superworld-gis.cyens.org.cy*. https://superworld-gis.cyens.org.cy/server/rest/services/Hosted/spi_2019/FeatureServer/0

[52]
“geodatadownloader,” *geodatadownloader.com*. https://geodatadownloader.com/

[53]
“Online GIS/CAD Data Converter | SHP, KML, KMZ, TAB, CSV, ...,” *mygeodata.cloud*. https://mygeodata.cloud/converter/

[54]
“How Natural Neighbor works—ArcGIS Pro | Documentation,” *pro.arcgis.com*. https://pro.arcgis.com/en/pro-app/2.8/tool-reference/spatial-analyst/how-natural-neighbor-works.htm#:~:text=The%20algorithm%20used%20by%20the

[55]
A. BHANDARI, “AUC-ROC Curve in Machine Learning Clearly Explained,” *Analytics Vidhya*, Jun. 15, 2020. https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/

[56]
C. Chan, “What is a ROC Curve and How to Interpret It,” *Displayr*, Jul. 05, 2018. https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/

Appendix A

Logistic Regression Code

```
import pandas as pd
import numpy as np
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, accuracy_score
from sklearn.model_selection import train_test_split
import seaborn as sns
from sklearn.preprocessing import MinMaxScaler
import matplotlib.pyplot as plt
from sklearn import metrics
from sklearn.model_selection import validation_curve
from sklearn.pipeline import make_pipeline

##import data
data = pd.read_csv('Data.csv')

##seperate data into feature and target column
x = data.iloc[:, data.columns != 'FIRE']
y = data.FIRE

##divide into training and test sets
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.20,
random_state=5, stratify=y)

##scale dataset
scaler = MinMaxScaler(feature_range=(0, 1))
scaler.fit(X_train)
X_train_scaled = scaler.fit_transform(X_train)
X_test = scaler.fit_transform(X_test)

##build model
#model = LogisticRegression(C=1e12, class_weight='balanced',
solver='liblinear', max_iter=10000)
model = LogisticRegression(max_iter=10000)
model.fit(X_train_scaled, y_train)
y_pred = model.predict(X_test)

##weights
feature_importance=pd.DataFrame({'feature':list(x.columns),'feature_importa
```

```
nce':[i for i in model.coef_[0]]})
print(feature_importance)
#Plot Searborn bar chart
fi = sns.barplot(y=feature_importance['feature_importance'],
x=feature_importance['feature'], color='blue')
fi.set_xticklabels(fi.get_xticklabels(),rotation = 30)
#Add chart labels
plt.title('Feature Importance Logistic Regression')
plt.xlabel('FEATURE NAMES')
plt.ylabel('FEATURE IMPORTANCE')
plt.show()
```

Appendix B

Random Forest Classification Code

```
import pandas as pd
import numpy as np
from sklearn.metrics import confusion_matrix, accuracy_score
from sklearn.model_selection import train_test_split
import seaborn as sns
from sklearn.preprocessing import MinMaxScaler
import matplotlib.pyplot as plt
from sklearn import metrics
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import validation_curve
from sklearn.pipeline import make_pipeline

##import data
data = pd.read_csv('Data.csv')

##seperate data into feature and target column
x = data.iloc[:, data.columns != 'FIRE']
y = data.FIRE

##divide into training and test sets
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.20,
random_state=5, stratify=y)

## scale dataset
scaler = MinMaxScaler(feature_range=(0, 1))
scaler.fit(X_train)
X_train_scaled = scaler.fit_transform(X_train)
X_test = scaler.fit_transform(X_test)

##build model
model = RandomForestClassifier()
model.fit(X_train_scaled, y_train)
y_pred = model.predict(X_test)

##weights
feature_importance=pd.DataFrame({'feature':x.columns,'feature_importance':m
odel.feature_importances_})
print(feature_importance)
```

```
#Plot Searborn bar chart
fi = sns.barplot(y=feature_importance['feature_importance'],
x=feature_importance['feature'], color='blue')
fi.set_xticklabels(fi.get_xticklabels(),rotation = 30)
#Add chart labels
plt.title('Feature Importance Random Forest')
plt.xlabel('FEATURE Names')
plt.ylabel('FEATURE Importance')
plt.show()
```

Appendix C

Accuracy

```
##evaluate on train set
train_acc = model.score(X_train_scaled, y_train)
print("The accuracy for training set is {}".format(train_acc1 * 100))

##evaluating on test set
test_acc = accuracy_score(y_test, y_pred)
print("The accuracy for test set is {}".format(test_acc2 * 100))
```

Confusion Matrix

```
##confusion matrix
cm = confusion_matrix(y_test, y_pred)

group_names = ['True Neg','False Pos','False Neg','True Pos']
group_counts = ["{0:0.0f}".format(value) for value in
                cm.flatten()]
group_percentages = ["{0:.2%}".format(value) for value in
                     cm.flatten()/np.sum(cm)]

labels = [f"{v1}\n{v2}\n{v3}" for v1, v2, v3 in
          zip(group_names,group_counts,group_percentages)]
labels = np.asarray(labels).reshape(2,2)

ax = sns.heatmap(cm, annot=labels, fmt='', cmap='Blues')
ax.set_title('Confusion matrix logistic regression\n\n');
ax.set_xlabel('\nPredicted Values')
ax.set_ylabel('Actual Values ');
ax.xaxis.set_ticklabels(['False','True'])
ax.yaxis.set_ticklabels(['False','True'])

plt.show()
```

AUC-ROC Curve

```
##AUC-ROC
fpr, tpr, _ = metrics.roc_curve(y_test, y_pred)
auc = metrics.roc_auc_score(y_test, y_pred)
plt.plot(fpr,tpr,label="AUC="+str(auc))
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.legend(loc=4)
plt.show()
```