Prediction of Avalanche risk in the alps by the use of Machine Learning Algorithms. A case study in Kaprun/Mooserboden

2010695005

FH Salzburg

02. January 2022

Contents

1 Introduction

Avalanches are among the most dangerous natural disasters [1]. They threaten people and environment in seasonally snow-covered mountain regions such as the alps. In 2021, 265 recorded avalanches (recoreded on Lawis) in Austria killed 17 people and injured 5 others. A total of 41 people were involved in the events [2]. The occurrence of snow avalanches is steadily increasing due to climate change [3] [4] [5]. In order to enable preventive measures to combat avalanches, a good assessment of the impending danger is necessary. The complexity of this task has already been described in several studies. It comes from the fact that there are many potentially influancial parameters, of which in most scenarios only a few are available and and they change significantly even for small geographical differences. For example, the estimation of avalanche danger levels is mostly based on weather data, which are also predicted, results of snowpack models and data on snow instability. In addition, information on the terrain is often used for these features [5]. Avalanches have been artificially triggered by explosions for a long time. For this preventive fight against snow avalanche accidents, automatic and semi-automatic snow avalanche detection systems have been developed.Of these, infrasound-based systems are the most promising, in the early detection of slow avalanches, for example ice avalanches [6] [1]. In addition, these systems are used to detect as many avalanches as possible. Many avalanches have not been recorded, which means that the accuracy of predictions made by automatic systems is compromised. Geographic Information Systems (GIS), Hierarchy Process Methods (AHP), and Remote Sensing (RS) are used together for this purpose to assess and assess avalanche hazards. with the help of these systems avalanche hazards can be detected, but they have a high error rate. Therefore, predicting the risk of avalanches on selected slopes would be a great improvement in the precision of preventing avalanches by blasting them. It has been revealed that some machine learning can achieve good results in predicting natural disasters and explicit avalanches [3] [4] [5]. Avalanche warning services could receive promising forecasts for the following days in advance and make them available to the public. This leads to a significant increase in safety for backcountry skiers and other winter mountain sports enthusiasts. Backcountry skiing is a sport in which the athletes are skiing unmarked or unguarded areas within or outside the boundaries of a ski resort. For this, they either climb the mountain with touring skis or are brought up by helicopter or snowmobile. Having good predictions for the avalanche hazard, makes it easier and safer to plan these tours. This has led to the author's personal interest for this topic and the resulting possibilities to get high precision predictions of avalanches hazard for explizit Slopes by the use of machine learning models.

The aim of the master thesis is to predict the probability of avalanches from explicit slopes. These predictions should be made for a certain number of days in the future, for example for one week in the future. This is done by using topographical data (e.g. slope, slope orientation by cardinal directions or proximity to rivers and streams) in conjunction with meteorological data for the day to be predicted (e.g. weather, temperature, snow depth, wind direction), wich will be also used retrospective. Thus, the study is intended to provide an answer to the questions, "What data features are needed to predict a day-accurate avalanche probability for explicit slopes by the use of machine learning methods?" and "How many days in advance can these predictions be determined?".

It also is important for the trafic and safetiy of people living in avalanche dangeres parts of the alps. Also important for the workers who work in the avalanche lines on the water energy of the glockner group.

2 Related Work

In the context of avalanche risk prediction, several studies have already been conducted. Machine learning methods have been applied on this purpose. Data sets of avalanche events from smaller mountain areas were used as case studies. In the alps a dataset of avalanche events from the area around Davos, Swizerland over the last 13 years is one Example for these case studies. The study about the data from Davos aimed to predict an entire winter season of avalanche days. To get the Meteorological Data for their study the Team form the SLF, Swizerland combined the data of snow avalanche events with those of an automatic weather station and the simulated snowpack properties, like new snow depth, liquid water content, Stability indices, critical crack length, and the hand hardness, of the SNOWPACK model. A random forest model was then trained on these merged data. One finding of the resulting study is that the predictions without the snowpack factors are just as good as those without this additional data. Furthermore, they concluded that their prediction attempts were severely limited by the use of inaccurate, biased avalanche-related data and by the fact that the spatial scale was too large for their models. They came to the result that forecasting on a small spatial scale using only one avalanche warning system could work well and could be a good aid for avalanche hazard forecasting services. [7]. In another study, avalanche hazard maps were attempted using predictions from machine learning methods. The space around the watershed of Karaj, Iran was used in this paper. The machine learning methods used in this study were SVM and MDA.Also for this purpose, various meteorological data were brought in and used to train the algorithms. From the study, it was found that avalanches seem to slide out mainly in the vicinity of streams. In addition, MDA performed better in predicting avalanche danger compared to SVM. Both methods produced results with an accuracy of 0.83 for the SVM and 0.85 for the MDA. [5].

3 Methodology

3.1 Feature selection

It is a challenging and significant task in the field of data science to create machine learning models from high dimensional data sets. Machine learning research has long assumed that too many columns of data lead to a reduction in prediction quality. This phenomenon is caused by misinterpretation of the features by the algorithms used and is called overfitting. A huge amount of dimensions also increases the computation costs and can reduce the performance in total. The more dimensions a data set has, the more prevalent it will contain redundant, noisy, and unimportant features, wich lead to overfitting and increase the error rate of the learning algorithm. Therefore, it can help to focus on a small subset of really important features. [8] [9] [10] Feature selection is divided into two steps. The first step is to filter the data and reducing the feature space by removing the previously mentioned irrelevant features. In the second step, an optimal subset of features of the remaining data is created using a wrapper. [8] This can be achieved by removing redundant and unimportant data columns to get enhance performance of prediction, scalability and generalization capability in learning efficiency and avoid overfitting. If a feature does not affect the prediction quality of the learning model, it is not important for the prediction [10]. This does not mean that this feature does not contain useful data. It only indicates that it is not statistically related to other features [10]. A good feature Selection can help to get much better predictions from the machine learning models and decrease the error rate [8]. In most feature selection methods, optimization algorithms are used to build a subset of the most relevant features. This leads to better performance and better classification results. [9] The popular approaches to do this, are models, features quality measures, feature evaluation, search strategies and combinations of these [10]. Depending on how the training set is labeled, supervised (fully labeled), unsupervised (unlabeled) and semi-supervised (partially labeled) feature selection methods are used [8]. Feature selection methods can also be divided into the three groups Filter, Wrapper and Embedded Method, based on how they interact on the learning models. The Filter method selects features based on statistical factors. This method does not depend on the learning algorithms and therefore consumes much less time. For an Example there are correlation coefficient or the chi-square test. [10] [8] The Wrapper method totally depends on the classifier used, wich means it does need more computation time than the Filter method. On the other hand the best subset of features comes directly based on the results of the classifier and they are more accurate than the filter methods.[10] [8] Wrapper models mostly use the accuracy rate and the classification error as default evaluation scores. The feature selection results of these models are often created at the same time as the results of the machine learning model. This is due to the fact that the learning model is embedded in the feature selection.[8] An examples for this method are genetic algorithms. The third variant is the wrapper method. This performs better than the other two because it requires less computation time and makes collective decisions based on hybrid learning or ensemble learning. An example of such a method is the random forrest. [10] [8] Feature selection methods should have a small time and space complexity and do not generate a lot of overhead, but must also have a high learning accuracy [8].

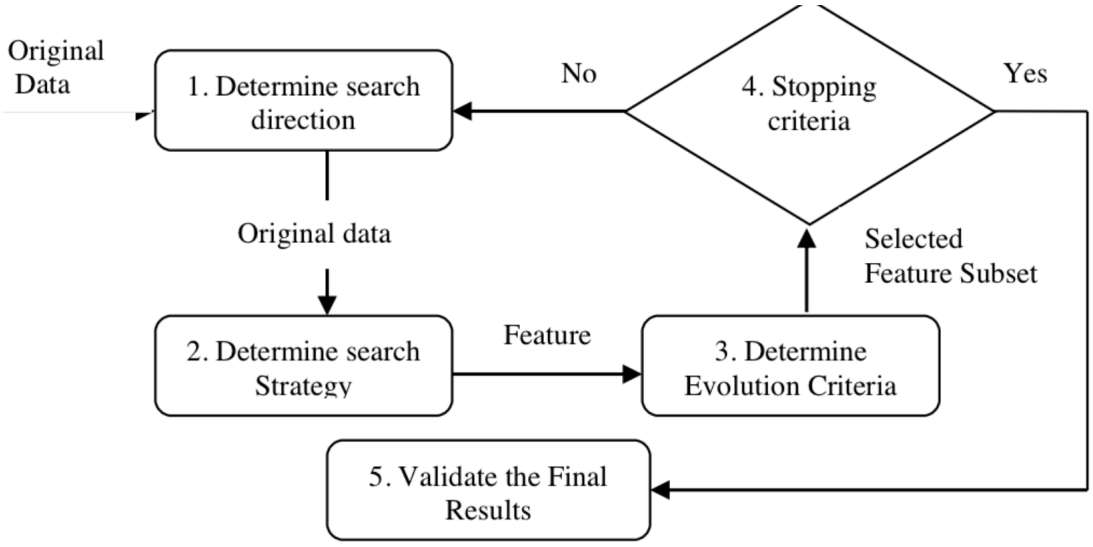

Figure 1: Stages in FS-process. [10]

Figure 1 shows the 5 steps of a feature selection process. The process starts by the search direction, wich can be forward, backward or random. The second step is to define wich of the three search strategies, randomized, exponential or sequential, should be used. After that the feature selection method selects features by the use of the evaluation criteria. To reduce cost, computation time and complexity, it is important to specify a stopping criteria. It defines the point on wich the method should break.[10] For example the depth of a decision tree defines the maximum number of branches, nodes and leafs of the tree, wich also defines its maximum complexity. After the feature selection algorithm finished it search process, the results must be validated. For this step there are a lot of methods. For example cross validation or confusion matrix. [10] How important parameters were considered to be for the avalanches in each study seems to be strongly related to what parameters were available and wich machine learning models have been used for the study. As an example, in the study in Iran, wich is described in the article ”Snow avalanche hazard prediction using machine learning methods” [5], elevation was not ranked as particularly important for prediction, whereas in a study in India reported in the paper ”Parameter importance assessment improves efficacy of machine learning methods for predicting snow avalanche sites in Leh-Manali Highway, India” [4], it has been ranked as the second most important feature. In the First Study, more additional meteorological and geographic parameters were available, which appear to be more important than the elevation [5] [4]. Because of its low computation time for highly dimensional datasets and good results, in context of this thesis genetic algorithm is used as search strategy. In case of this thesis decision trees, logistic regression and SVM are used as classifiers for a Genetic Algorithm. The next three chapters describe these as well as the genetic algorithm in detail and give an understanding about how they are used to find the important features of an dataset. The Decision tree and Logistic regression models are also used to give an general understanding about statistical importance of all features in the dataset.

3.1.1 Decision Tree

”Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.” [11] Decision tree are a representation method based on knowledge about the features of a dataset to represent classification rules [12]. Decision Trees use a set of if-than-else rules to decide wich value to predict. These models are good to understand, interpret and visualisable because they use white box models in wich every step is a boolean logic and easy explainable [11]. A standard decision tree starts with a root node, does have some branches as well as child nodes and leaves [12]. The root node splits the set by a rule on the features wich provides the best classification of the instance. This goes recursive till the max depth is reached or the classification is completed. So a branch is the path from the root node to the leaf. The leaf at the end of an branch, wich represent the class labels of the feature to predict. [12] They can also handle categorical data. It’s also possible to use them to predict multiple values at the same time, wich is a typical problem in supervised machine learning called the Multi-output problem [11]. Because Decision Trees are likely to overfit if used on high dimensional datasets, they are no option to be used as an alternative prediction method for this work. But if used with a low tree depth, they can give a use a good understanding about the importance of some individual features for the prediction of multiple or specific parameters [11]. The deeper the nodes are in the tree, the less important the features they represent are for the classification. In addition, the decision tree contains only parameters that contribute to the classification. Therefore, not only the importance of the features for classification can be determined, but also whether they can be used for classification at all. [12] This advantages of decision trees make them also useful for feature selection. In the case of the study, wich is represented in the article: ”Feature selection using Decision Tree and classification through Proximal Support Vector Machine for fault diagnostics of roller bearing” [12] Decision trees are used for the feature selection. Similar to this work, the study was concerned with a classification problem. It was not about the prediction of natural disasters but about the fault diagnosis of bearings.

3.1.2 Logistic Regession

The Logistic Regression is a classification training algorithm, wich is often used in the field of predictive analytics. It is also a supervised and discriminative machine learning model. [13] Because of the fact that it appreciates the probability that an event will happen, the resulting value must be between 1 and 0. [13] The logistic regression and linear regression models are two of the most popular models in the field of data science, as they are very easy to execute and require little computation time. linear regression is used to find the correlation between two features. This is done by drawing the line that best fits through a number of data points. This line is calculated using the least square method. Similar to this behavior the logistic regression is used to calculate the correlations between multiple features and the variable to be determined. [13]

3.1.3 Genetic algormithm

Genetic algorithm is an evolutionary based adaptive optimization search methodology. They contain to the category of wrapper models. As a Wrapper it is used for the second step of feature selection to find an optimal subset of features for the learning algorithm [8]. Like a lot of other technical inventions, the functionality of genetic algorithms is inspired by nature. For example Neuronal Networks are inspired by the functionality of the human brain. Genetic algorithms resemble the Darwinian natural selection and evolution of species. They use this mechanisms to optimize modeling problems and get a good subset of features. Genetic algorithms simulate the natural selection of species [14] . This means only the species who survive environmental changes can become another generation. Each generation represents a population of individuals. Each of this individuals represents a single solution for the problem and is defined by a genetic string wich is build out of chromosomes wich represent encoded features. [14] [15] Genteic algorithms are also able to handle huge dimensional datasets efficiently because of their exploitational and explorational characteristics [14]. The algorithm starts by creating a random generated population, wich happens by generating a number of chromosomes. After that step a classification model is constructed based on the combination of variables of each chromosome. This classification model is validated, on each chromosome, with an k-fold cross validation by the use of statistical scores like the accuracy score. The fitter chromosomes have a higher chance to get passed on to the next generation. After that the genetic algorithm selects and recombinates the chromosomes by the validation of the scores from parent and offspring to get a new population. It depends on the stop criteria whether the algorithm stops or runs the same cycle again with the new population. The algorithm needs an stopping criteria on wich it will stop processing new generations. [15] [16] [14]

Jianjiang Lu and Tianzhong Zhao and Yafei Zhang [14] describe the three main operations for the process of a Genetic search methodology, wich are selection, crossover and mutation operation, as follows. The selection operation searches for the strongest N individuals from the current population. These are used as parents for the next generation of individual solutions. The crossover operation is spliced into three steps. At first it generates C_N^2 pairs of combinations between all parent individuals. Secondly it generates the two numbers $a(0 < a < m)$ and $b(0 < b < ma)$, in wich m represents the length of each chromosome, a indicates the start position of the crossover operation and b is the length of the crossover operation. For the last step, it is assumed for each parent pair $C_1^t = \{w_k\}$ and $C_2^t = \{w'_k\}$ with $k = a + 1, \ldots,$ where $a + b$ are two gen groups. To generate two new individuals for the pool of individuals, wich is used in the mutation operation to generate a new population, the gens in the range of $[(a + 1), (a + b)]$are exchanged. The exchange is carried out on the basis of the crossover rate P_c as follows [14]:
$C_1^t + 1 = w_1, k, C_2^t + 1 = w_2, k$, where $w_1, k = \gamma * w'_k + (1-\gamma) * w_k, w_2 k = \gamma * w_k + (1-\gamma) * w'_k$, in this context γ is a predefined constant. The Mutation operation takes the, in the separation operation, created individuals into a pool with the parent individuals so that the variation in the new population is guaranteed. The K worst individuals out of this pool get a small mutation rate P_m. After that a number of genes are pickt, from every individual, by the mutation operation and a new offspring is generated as following: when a gene $w_k(w_k \in [0, 1]$ is mutated and its next generation isw'_k, the mutation operation is [14]:

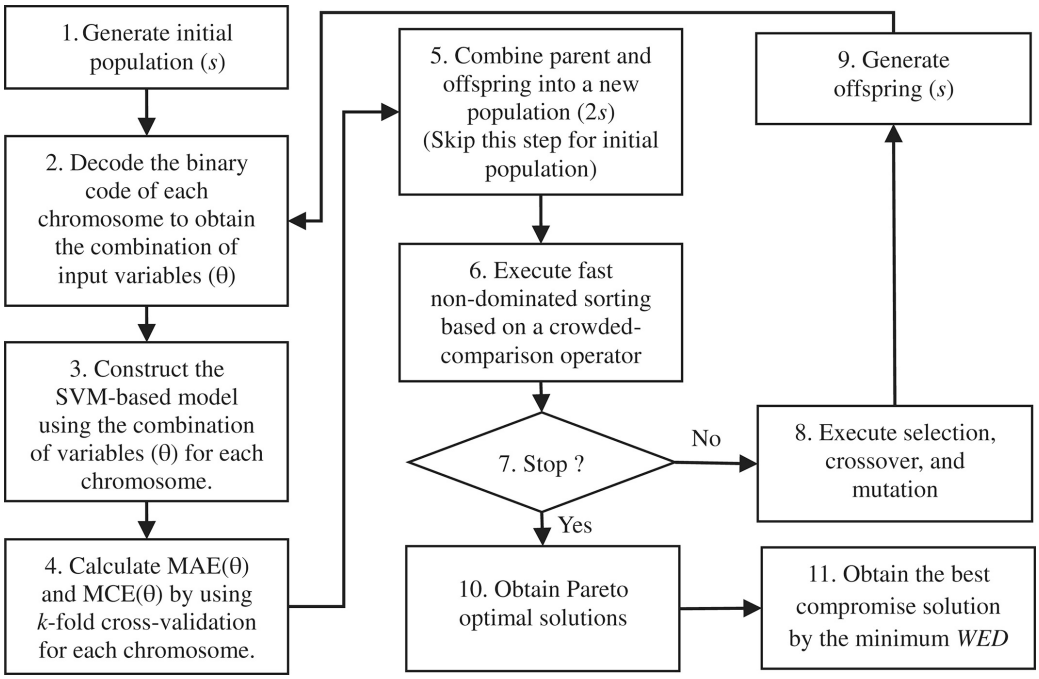

Figure 2: Flowchart of an genetic algorithm based on an SVM classifier [16]

$$w'_k = \begin{cases} w_k + \Delta(t, 1 - w_k), & \eta = 0 \\ w_k + \Delta(t, w_k), & \eta = 1 \end{cases} \quad [14]$$

The variable η is a random number wich can be ether '1' or '0' and the return value of the function $\Delta(t, \gamma)$ is in range $[0, \gamma]$ [14].

$$\Delta(t, \gamma) = \gamma(1 - r^{(1 - \frac{t}{M})}) \quad [14]$$

r is a number wich is randomly chosen in the range of $[0, 1]$. Furthermore, t shows the value of the iterations. M represents the maximum of iterations and p indicates the predefined mutation parameter. Caused of this functions, the genetic algorithm mutates in earlier generations more than in later ones.[14]
In the article "Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS", Petro Liashchynskyi and Pavlo Liashchynskyi tested Grid Search, Random Search and Genetic Algorithm on the CIFAR-10 Dataset. They concluded that the Genetic algorithm took more time, but also produced better results. With larger numbers of features, it was even faster than the other two. [15]

The authors of the article "Predictor selection method for the construction of support vector machine (SVM)-based typhoon rainfall forecasting models using a non-dominated sorting genetic algorithm" [16] used the genetic algorithm in combination with the SVM classifier for the prediction of typhoons. This natural disasters are dependent, as well as snow avalanches, on meteorological and topographical data.

3.2 Validation

curve (ROC) analysis to scrutinize the sensitivity, specificity, and accuracy
Cross-validation

3.3 Machine Learning Models

In order to achieve adequate results, a series of machine learning models will be trained in the context of the thesis. In the past, some models have already proven their worth in predicting natural disasters. For example, the SVM (support vector machine) and the MDA (multivariate discriminant analysis) models. They are useful for detecting subtle patterns in complex data sets and Flexible in handling data of different dimensions. SVM models are desgined to deal with high dimensional data. Thats one aspect why they have already been used to predict natural disasters, such as earthquakes, floods, typhoons, drought, landslides and avalanches [5] [4] [17]. MDA forms efficient linear combinations of independent variables. MDAs have not been used that often to predict natural disasters, but shows superior performance compared to SVM in the case study in the Karaj water conservation area in predicting avalanche risk levels [4].

3.3.1 SVM (Support Vector Machine)

Support Vector Machines are supervised machine learning algorithms based on the statical learning theory. Supervised learning algorithms are given a number of input features and the parameters to be predicted with labels. Each feature can also be considered as a dimension in a hyper-plane. The SVM creates a hyper-plane to split the hyper-space into two or more parts. This depends on how many classes are to be predicted. So the SVM can be applied to cases of the multi-class problem just like decision trees. To minimise the generalisation error the SVM tries to seperate the classes with the maximum margin.[12]

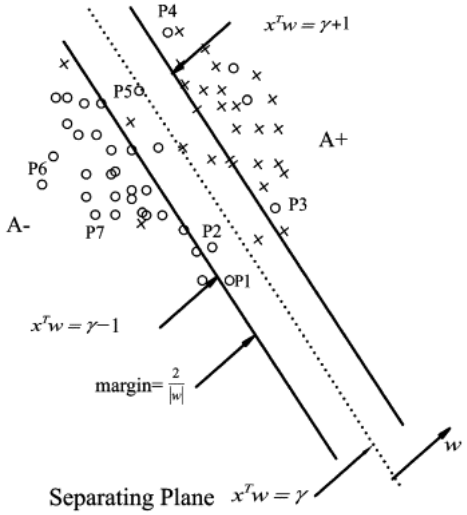

Figure 3: Standard SVM Classifier. [12]

3.3.2 MDA (multivariate discriminant analysis)

3.3.3 Neuronal Network

3.3.4 Model evalutaion

Sensitivity analysis, AUC statistic [5]

4 Data

4.1 Origin of the data

The avalanche warning service of the energy company VERBUND AG, headquartered in Kaprun, Salzburg, provides the data required to carry out this work. VERBUND represents Austria's leading energy company and is Europe's largest producer of electricity from hydropower. The company states that it obtains almost 100% of its electricity from renewable sources. The electricity is mainly generated from hydro, wind and photovoltaic power plants. In addition, this is supported by gas-fired thermal power plants [18].

The data were recorded in the vicinity of the Mooserboden storage power plant for 39 avalanche strokes. The background for the exceptionally accurate recordings of avalanches is the need for the most accurate possible prediction of avalanches in the 39 avalanche lines around the power plant. Accurate forecasting is of great importance to ensure the safety of the employees working in the areas of the avalanche lines. The power plant is part of the Kaprun power plant group, which includes both pumped and storage power plants. The power plant group is operated by Verbund Hydro Power and is located in Salzburg on the edge of the Hohe Tauern at 2040 meters above sea level and is surrounded by the over 3000 meter high mountains of the Glockner group.[19]

4.2 Meteorological Data

The Data Table Mooser_Wetter_Daten includes meteorological data for each day in the months of november to may in the period from 1953 to 2022. The table contains data about the air temperatures at the times 7:00, 14:00 and 19:00 as well as the snow temperature, the wind direction, the wind force, the snow sinking depth, the day weather as well as the weather from the day before, the snow depth, the precipitation, as well as the avalanche degree.

4.3 Avalanche related data

The Table Allg_Lawinen_Katalog represents general data on all recorded avalanches in the 39 avalanche lines of the area for the same period as the meteorological data from the table Mooser_Wetter_Data. More precisely, it contains data such as the type of avalanche, the old ID of the avalanche line where the avalanche went down, the time and date when the avalanche was recorded, the volume of the avalanche, the general weather conditions at the time of the avalanche, as well as general data on the snowpack and the danger level on the day of the avalanche. The Meteorological Data shown in this Table are not used in this study, because the data from the table Mooser_Wetter_Daten are homogeneous and available for each day of the winter season. Another table of the database named kaplawstr contains additional information about the avalanche lines. For this work, the old and the new code of the avalanche lines are used from this table

4.4 Topographical Data

The Topographical Data is recorded in the Database Table TOPP. This table contains several rows for each avalanche line, wich can be identified by the new avalanche line code. The rows record minimum, maximum and mean slope, as well as the orientation and height of the slope. The table also contains various other data columns. These are not used in the further course of the work, since they cannot be assigned to the individual avalanche lines in general, but are connected with individual avalanches, which are not allocated to them in the context of this work.

4.5 Data Composition

The database tables Allg_Lawinen_Katalog (Avalanche related data), kaplawstr (contains the old as well as the new avalanche line IDs), TOPP (Topographical Data) and Mooser_Wetter_Daten (Meteorological data) wich are already described in the previous chapters were merged into a homogeneous data set in the context of this master thesis. This process is explained in detail in this chapter. The tables include data from 1944 to 2022. the quality of data increased with the years. This can be shown above all by the fact that in the seasons before 1989/1990 in average, there are 22.0769 per season recorded wich is a lot less than from this season to 2022. In that inteval the average of recorded avalanches is 81.636. This can also be seen in Table 1 wich shows the sum of avalanches recorded per season. With the exception of a few outliers, hardly any avalanches were recorded in these years, and in some cases none at all. In order to increase the homogeneity of the data, all data outside the period from 1989 to 2022 were removed from the database tables. Subsequently to this measure the kaplawstr table has been merged to the Allg_Lawinen_Catalog table using the old avalanche line ID. This adds to each avalanche the associated new avalanche code and avalanche name, which are used as additional ID. The connection is necessary because the TOPP table, which represents the topographic data for the avalanche routes, does not contain the old avalanche route ID. In the course of this step, all lines that were labeled with the avalanche line name ”all avalanches” also have been removed. These are not included in the kaplawstr table, since this does not represent an exact departure of an avalanche in one of the avalanche lines, but only states that in many of the avalanche lines small avalanches have departed.

```
TOPP = kaplawstr['Code_neu'].apply(lambda x: TOPP.loc[TOPP['Lawinencode']
    == x].mean())
```

Listing 1: calculation of TOPP data for every avalanche line

In the second step, the average values for all columns from the associated avalanches were calculated from the TOPP table for each avalanche line. The python code shown in Listing 1 demonstrates this process. By this measure, one row is created for each avalanche line. The table contained without this procedure a total of 1905 rows. In the default state, the table could not have been connected to the other data tables. Another way to get only one row per avalanche stroke would be to select a random value for the respective stroke. The reason for taking the average value is that there are not the same number of lines for all avalanche lines and the values of the individual lines per avalanche line do not differ greatly from each other. Thus, the average

Intervall	Avalanche
1956/ 1957	11
1957/ 1958	4
1958/ 1959	7
1959/ 1960	62
1964/ 1965	8
1965/ 1966	8
1966/ 1967	15
1967/ 1968	9
1968/ 1969	3
1969/ 1970	20
1970/ 1971	22
1971/ 1972	0
1972/ 1973	67
1973/ 1974	31
1974/ 1975	78
1976/ 1977	0
1979/ 1980	27
1980/ 1981	40
1981/ 1982	29
1982/ 1983	27
1983/ 1984	0
1984/ 1985	10
1985/ 1986	24
1986/ 1987	37
1987/ 1988	17
1988/ 1989	18
1989/ 1990	66
1990/ 1991	55
1991/ 1992	133

Intervall	Avalanche
1992/ 1993	80
1993/ 1994	47
1994/ 1995	93
1995/ 1996	3
1996/ 1997	19
1997/ 1998	18
1998/ 1999	90
1999/ 2000	128
2000/ 2001	89
2001/ 2002	124
2002/ 2003	84
2003/ 2004	92
2004/ 2005	97
2005/ 2006	100
2006/ 2007	40
2007/ 2008	86
2008/ 2009	79
2009/ 2010	52
2010/ 2011	52
2011/ 2012	150
2012/ 2013	121
2013/ 2014	55
2014/ 2015	75
2015/ 2016	66
2016/ 2017	67
2017/ 2018	133
2018/ 2019	177
2019/ 2020	67
2020/ 2021	130
2021/ 2022	26

Table 1: recorded Avalanches per Season

value represents the entirety of the lines per stroke consistently. The topographic data from the newly assembled TOPP table was then merged to the entire dataset using the new avalanche ID. Subsequently, these avalanches were assigned to the daily recorded meteorological data of the Mooser_Wetter_Data table by an outer join, so as result there is at least one row per day in the dataset. In cases where several large avalanches have occurred at the same day, the dataset contains one row per avalanche and each Includes the meteorological data for this day.

In order to train a machine learning algorithm for the prediction of avalanches for topographically defined slopes in conjunction with the meteorological data available for this work, the topographical data must also be mapped onto the days without avalanches. The algorithm needs this information, as the data set would otherwise only contain topographic data directly related to avalanches. This would mean, for example, that the slope inclination could not become a feature for the prediction.

```
for i in gesamt_df.index:
    if(pd.isnull(gesamt_df['meanExpo'][i])):
        sample = TOPP.sample(1)
        gesamt_df['meanExpo'][i] = sample['meanExpo'].values[0]
        gesamt_df['meanSlope'][i] = sample['meanSlope'].values[0]
        gesamt_df['stdDevSlope'][i] = sample['stdDevSlope'].values[0]
        gesamt_df['MinSlope'][i] = sample['MinSlope'].values[0]
        gesamt_df['MaxSlope'][i] = sample['MaxSlope'].values[0]
        gesamt_df['Altitude'][i] = sample['Altitude'].values[0]
```

Listing 2: mapping random sample lines of topographical data onto the rows of non avalanche days

The consequence of this is that the topographic data must also be mapped to the days without avalanches. Because these days are not connected to an avalanche line ID and an even distribution on the slopes on these days is required, random lines from the calculated mean values of the TOPP table were mapped to them. Listing 2 shows this process in form of the corresponding Python code. The resulting dataset maps 7055 rows and 139 columns. 2728 of these rows are recorded avalanches departures. The dataset contains columns that are redundant, empty, sparsely filled or contain information wich can not be used to train a machine learning algorithm. Figure 2 shows a heatmap of the Nan values in the dataset. As mentioned in chapter 3.1 about feature selection, these columns have a huge cost in computation time, decrease the prediction performance and increase the error rate. This requires the measure to remove all columns with these characteristics.

After dropping this features the Dataset includes 46 columns. In addition, a new column was added to the dataset, which contains either a 1 in case of an avalanche or a 0 in case no avalanche has occurred. This column is added to make it possible to predict whether an avalanche will occur or not. To train a machine learning algorithm on predicting whether an avalanche will go down from a particular slope, all features that can be used to directly and without any other features determine whether an avalanche will descend must be removed from the dataset.

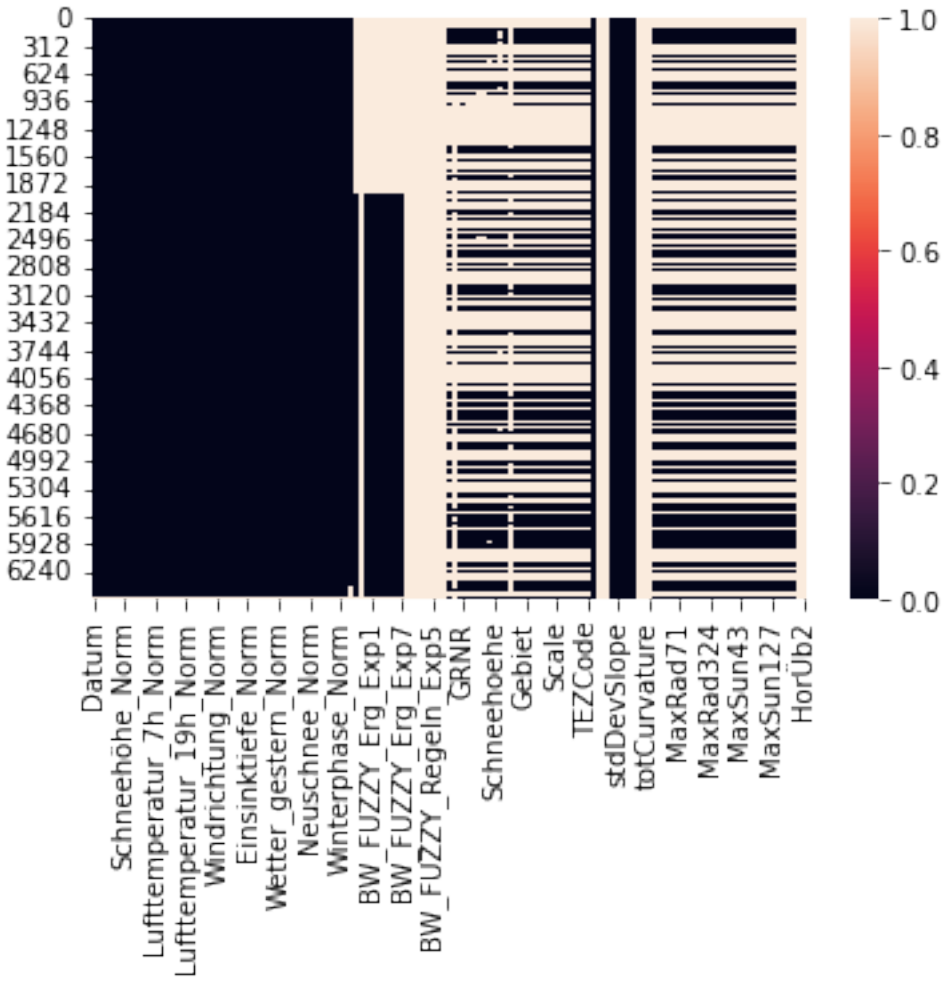

Figure 4: Heatmap to show the distribution of Nan values in the dataset

5 Feature Selection

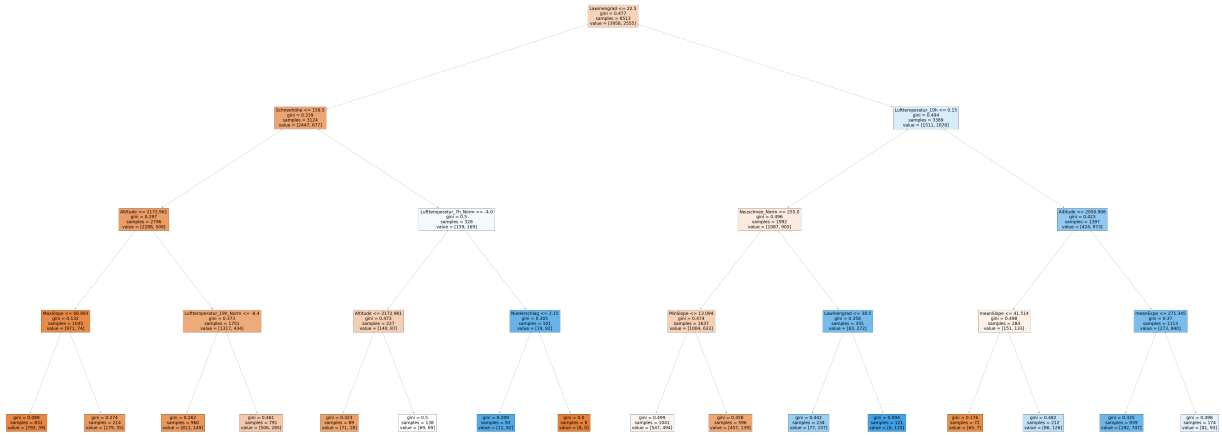

Figure 5: Decision Tree for Avalanche classification with one value to predict

Listings

List of Figures

References

[1] V. by Hermann Maurer, version 3, 2019. [Online]. Available: https://austria-forum.org/af/AustriaWiki/Lawine?version=3 (visited on 01/15/2022).

[2] L. S.L.O.-L.V.L.K.U. W. Lawinenwarndienst Tirol Lawinenwarndienst Steiermark. (2022). Lawinen Ereignisse, [Online]. Available: https://lawis.at/incident/ (visited on 01/10/2022).

[3] Y. L.G. B. Eric Martin Gérald Giraud, „Impact of a climate change on avalanche hazard“, Annals of Glaciology, vol. 32, pp. 163–167, 2001. [Online]. Available: https://doi.org/10.3189/172756401781819292.

[4] B. D. V. Anuj Tiwari Arun G., „Parameter importance assessment improves efficacy of machine learning methods for predicting snow avalanche sites in Leh-Manali Highway, India“, Science of The Total Environment, vol. 794, 2021. [Online]. Available: https://doi.org/10.1016/j.scitotenv.2021.148738.

[5] A. M.F.S.-H.V.P.S. S. Bahram Choubin Moslem Borji, „Snow avalanche hazard prediction using machine learning methods“, Journal of Hydrology, vol. 577, 2019. [Online]. Available: https://doi.org/10.1016/j.jhydrol.2019.123929.

[6] T. Thüring, M. Schoch, A. van Herwijnen, and J. Schweizer, „Robust snow avalanche detection using supervised machine learning with infrasonic sensor arrays“, Cold Regions Science and Technology, vol. 111, pp. 60–66, 2015, ISSN: 0165-232X. DOI: https://doi.org/10.1016/j.coldregions.2014.12.014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165232X14002419.

[7] B. R. Stephan Harvey Alec van Herwijnen, „Statistical Nowcast of Avalanche Activity at the Regional Scale“, 2016. [Online]. Available: https://arc.lib.montana.edu/snow-science/item/2437.

[8] J. Cai, J. Luo, S. Wang, and S. Yang, „Feature selection in machine learning: A new perspective“, Neurocomputing, vol. 300, pp. 70–79, 2018, ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2017.11.077. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231218302911.

[9] M. Allam and M. Nandhini, „Optimal feature selection using binary teaching learning based optimization algorithm“, Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 2, pp. 329–341, 2022, ISSN: 1319-1578. DOI: https://doi.org/10.1016/j.jksuci.2018.12.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1319157818306463.

[10] B. Venkatesh and J. Anuradha, „A Review of Feature Selection and Its Methods“, Cybernetics and Information Technologies, vol. 19, no. 1, pp. 3–26, 2019. DOI: doi:10.2478/cait-2019-0001. [Online]. Available: https://doi.org/10.2478/cait-2019-0001.

[11] (2022), [Online]. Available: https://scikit-learn.org/stable/modules/tree.html (visited on 04/13/2022).

[12] V. Sugumaran, V. Muralidharan, and K. Ramachandran, „Feature selection using Decision Tree and classification through Proximal Support Vector Machine for fault diagnostics of roller bearing“, Mechanical Systems and Signal Processing, vol. 21, no. 2, pp. 930–942, 2007, ISSN: 0888-3270. DOI: https://doi.org/10.1016/j.ymssp.2006.05.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0888327006001142.

[13] (2022), [Online]. Available: https://www.ibm.com/topics/logistic-regression (visited on 05/21/2022).

[14] J. Lu, T. Zhao, and Y. Zhang, „Feature selection based-on genetic algorithm for image annotation“, Knowledge-Based Systems, vol. 21, no. 8, pp. 887–891, 2008, ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2008.03.051. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095070510800097X.

[15] P. Liashchynskyi and P. Liashchynskyi, „Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS“, CoRR, vol. abs/1912.06059, 2019. arXiv: 1912.06059. [Online]. Available: http://arxiv.org/abs/1912.06059.

[16] .

[17] A. Pozdnoukhov, R. Purves, and M. Kanevski, „Applying machine learning methods to avalanche forecasting“, Annals of Glaciology, vol. 49, 107–113, 2008. DOI: 10.3189/172756408787814870.

[18] (2022), [Online]. Available: https://www.verbund.com/de-at/ueber-verbund (visited on 04/01/2022).

[19] (2022), [Online]. Available: https://www.verbund.com/de-at/ueber-verbund/besucherzentren/kaprun (visited on 04/05/2022).