Materials and methods

2.1: Characteristics of the study area and data set.

The area of study is in the Province of Chañaral (26.4 °S), in the Copiapo region, Chile, between longitudes 70.7°E and 69.5°E and latitudes 26.2°S and 27°S (Fig. 1). In topographic and physical terms, the average altitudes range between 1000 and 1500 m above sea level. The geology of the Salado river basin is distinguished by an ancient basement composed of Palaeozoic metamorphosed sedimentary rocks (Chañaral epimetamorphic complex). This complex is located near the coast and has been eroded over time by several minor streams, especially by the Salado river. In addition, there are strips of Palaeozoic intrusions, such as the batholiths of the Quebrada del Castillo. The region has experienced intense volcanic and tectonic activity, reflected by the presence of several intrusive bodies and two major north-south regional fault systems (the Atacama Fault System to the west and the Domeyko Fault System to the east). In the study area, volcanosedimentary rocks of the La Negra Formation and the Punta del Cobre Group outcrop. These geological units are found in the centre and south of the intermediate depression and to the west of the basin. On the other hand, the Cenozoic era is characterised by sedimentary rocks that represent the ancient relief at the foot of the mountain range. These rocks denote greater compaction due to their age and are associated with the Miocene-Pliocene. The more recent natural and anthropogenic deposits, which show a lower degree of compaction, outcrop on the slopes and riverbeds that cross the basin. These deposits are variable in composition and are generally thin in narrower areas and thicker in more open areas. Landslides have mainly occurred on these geological formations.

The type of landslides found in this region correspond to alluviums and debris flows, and are Rainfall-Induced Landslides. According to the climatic classification, the area of study is in the transitional area between the hyper-arid and semi-arid zones (González, 2018). Rainfall averages 1.7 mm per year in the lower zone of the basin and can reach up to 52 mm per year in the upper zone, which 135 accumulates between June and August (Juliá et al., 2008) and the average temperature is between 10 and 20° C for the coastal and medium arid zones and -1.7° C in the high-altitude zone (Antonioletti et al., 1972).

To examine the correlation between the spatial prediction of landslides and the relevant influencing factors, it is imperative to consider the most recent landslide records available. Consequently, to establish a comprehensive and reliable inventory for the study area, information from previous studies will be used and subsequently verified by laboratory analysis. Location of landslides was recorded through the analysis of the Sernageomin database, bibliographic search and interpretation of aerial photographs, satellite images extracted from Sentinel Hub and the Copernicus repository, as well as the use of the Google Earth program. The information is managed together through the free software Qgis version 3.22 and scripts created in R that contain the necessary instructions for the geospatial processing of the data and the extraction of characteristics. For the execution of the work, 86 locations of landslide events in the area of study were used, and another 86 locations were used as points where no landslides occur, to balance the data collection. Also, in (Dou, 2019) it is suggested that

by using samples from the landslide scarp polygon, it is possible to increase the accuracy of the model. In this work, the center of the landslide body was being used to characterize the phenomenon. Therefore, to increase the dataset, it was decided to take 10 samples (pixels) within the scarp polygon of each landslide, and in the case of the non-landslide points, a polygon surrounding the initial point was created, and from this, 10 samples are taken within this polygon, which allows to multiply by 10 the studied dataset, and thus provide the statistical robustness needed in this work.

2.2: Conditioning factors and selection of factors

In this study, 22 factors associated with the landslides were used, which were obtained mainly through the DEM of the AW3D30 project (Tadono et al., 2016) and their respective analysis through R and multiple geoprocessing packages, in addition to satellite images received from the LANDSAT 9 campaign (Masek et al., 2020), which include eleven spectral bands that can be combined to identify characteristics in the ground. On the basis of the factors previously mentioned, the selected factors are the slope angle and orientation, the curvature, the elevation above sea level, the profile and plane curvature, the valley depth (VD), the stream power index (SPI), the topographic wetness index, and the slope length. Additionally, the normalized indices NDVI, GNDVI, EVI, NDMI, BSI, NDWI, NDGI were obtained from images from February 2022. Then, for the machine learning analysis and the respective algorithms, the data relating to each category associated with each factor are included and then analysed using R. Table I summarizes the factors and the types of classes into which they can be classified, as a reference. Also, figure 3 and 4 show thematic maps of all the factors presented in this study along with landslide inventory.

2.2.1: IGR technique

Susceptibility assessment depends on the contributing elements. There are multiple techniques to determine the capacity of predictive elements that play a role in the occurrence of landslides, such as the gain ratio(Nithya and Duraiswamy, 2014), the significance of relief (Ahmad and Dey, 2005), and the information gain ratio (IGR) (Chapi et al., 2017). In this research, the latter has been chosen as the metric to quantify the predictive power of the contributing elements. The IGR methodology is used to identify the most relevant elements among the 23 contributors previously discussed in the field of research. Let consider F as the data set used for training, containing an initial sample n. Let the set n(Mi,F) represent the number of samples in the training set F that belong to the class Mi (which can be landslide or non-landslide). Consequently, the following equation can be established (Abedini et al., 2019b):

“Mathematical framework”

2.2.2: Correlation calculation

In addition to the IGR, the Pearson correlation factor will be used to eliminate features that are correlated with each other, since this can cause noise and not contribute to the model's performance. This test is important to evaluate the dependence between conditioning factors. Generally, Pearson's correlation establishes the ratio between the covariance of a pair of factors and the product of their standard deviations (Dou et al., 2019).

2.3: Modeling using Machine Learning

Machine learning corresponds to an empirical approach for both classification and regression in non-linear systems. Such systems can be multivariable, involving literally thousands of variables. In machine learning, if there is sufficient data, a training data set is built covering as much of the system's parameter space as possible. Typically, a random subset of the data is set aside for completely independent validation. Machine learning is ideal for handling those problems where theoretical knowledge is still incomplete, but for which a certain number of meaningful observations and other data are available (Lary et al., 2016).

2.3.1: SVM

It corresponds to an algorithm based on statistical learning theory, used in regression and classification problems (Vapnik, 1999). In the work in progress, the classification mode will be used. The main characteristic of this method is that during the learning process the algorithm transforms the initial space to a higher dimensional one, which allows the establishment of hyperplanes that are able to separate easily and thus classify new examples (Kavzoglu et al., 2015). In addition, it can work with nonlinear problems thanks to the incorporation of a Kernel, whose performance is controlled by the value (Tien Bui et al., 2016). The model´s precision is also controlled by the C regularization parameter. Both parameters can be fine-tuned using the grid search technique (Kavzoglu et al., 2015) or using the random search.

2.3.2: LR

This method has been mainly used in the last decade in susceptibility assessment (Budimir et al., 2015), since it has proven to be very useful as base model when a new one is being tested (Chang et al., 2019). Logistic regression is the equivalent of linear regression, which uses a non-linear transformation to estimate class. It can calculate the weights of each conditioning factor as independent variables based on the binary dependent variable at a certain level of statistical confidence (Shirzadi et al., 2012). Advantages of the method include: (1) It does not require the data set to have a normal distribution, (2) The independent and dependent variables can be either continuous or discrete, and (3) It does not assume that the variables have the same statistics in their variances (Dou et al., 2019).

2.3.3: RF

The Random Forest (RF) is one of the most used methods in machine learning (Breiman, 2001). The model generates multiple classification trees to then obtain a final weighted score (Breiman et al., 2017). The algorithm adds diversity among classification trees by alternating data and further modifying the set of explanatory factors arbitrarily over the various processes of tree induction (Arabameri et al., 2020). The hyperparameters that are necessary for the growth of the tree are the number of trees k and the number of predictive factors used to split the nodes (m). The OOB error (out of bag) is characterized as the percentage of the total number of objects that are misclassified, therefore it is a rational estimate of generalization error. The OOB error is estimated at the moment of building the model. In (Breiman, 2001) it is mentioned that the random forest creates a limiting value for the generalization error. Such error

often declines as the number of trees grows. In turn, k must be large enough to allow such convergence. The method calculates the value of the predictive variable by examining how much the error declines as the data are permuted for that variable while holding constant for the others. The growth in error corresponds to the value of the explanatory variable (Breiman, 2001). One of the main advantages of the random forest is its resistance to overtraining and the development of many trees where there is no risk of overfitting. Therefore, there is no need to rescale, transform or change the algorithm. For the predictors, the random forest is not too affected by outliers and deals missing values automatically (Crippen, 1990).

2.3.4: Xgboost

This method originated from the boosting tree gradient algorithm(Friedman, 2001). It uses regularized boosting techniques to reduce overfitting, thus improving the accuracy of the model. XGBoost it can scale in diverse scenarios, handle sparse data, use scarce computational resources with high performance, have extensive and detailed documentation, and be simple to implement (Chen and Guestrin, 2016). This algorithm has won multiple contests (Chen and Guestrin (2016), Nielsen (2016)), it has extensive hyperparameters that when synchronized substantially improve the model. XGBoost is an extension of the gradient boosting algorithm. The main idea of a boosting algorithm is to combine several weak learners sequentially to achieve better performance (Hastie et al., 2009). The method uses several classification and regression trees (CART) and integrates them using the boosting gradient method. XGBoost is made up of three aspects that differentiate it from the other algorithms, being: (i) an objective function regularized for better generalization, (ii) a boosting gradient tree for additive training, and (iii) a columnar subsampling to prevent overfitting (Chen and Guestrin, 2016).

2.3.5: Hyperparameter Optimization

Hyperparameters correspond to the values that are setup before data training and generally affect the performance of predictions generated. These actions improve the performance by fine-tuning these hyperparameters (Todorov and Billah, 2022). For the search of optimal hyperparameters, it is common in literature to use the grid search, which was used in this work.

2.3.6: Cross Validation

In order to increase , it is necessary to carry out a cross-validation process for the model validation. It refers to the process of repeatedly dividing the data set into a training set and a test set, where the former is used to fit a model, which is applied to the test set. When comparing the predicted values with the known values of the test set, it is possible to obtain a statement with reduced bias on the model's ability to generalize the model to unknown data. In this case, a 100-fold repeated 5-fold cross-validation is used, which means randomly dividing the data into five partitions to be used once as a test set. This ensures that each observation is used in the test set, which requires the fitting of five models. Subsequently, the process is repeated 100 times. In each iteration the cut of data shall be different. In summary, this leads to 500 models, where the average measure performance (in this case the AUC value) measures the overall model´s predictive power (Lovelace et al., 2019). When applying the 5-fold cross-

validation method, it is equivalent to dividing the data set by considering 80% for training and 20% for validation/testing. Unlike traditional methods, each of the folds is used at some point for training and also for validation. Therefore, the ROC curves presented correspond to the averages obtained, considering that this process was repeated 100 times to achieve greater statistical robustness.

2.3.7: Model validation

Validation performance is a critical step within a modeling procedure; thus, several statistical indices have been suggested and used. In this work, the ROC curve will be used which is a basic measure in this type of evaluations (Pham et al., 2018). The plot is constructed with specificity and sensitivity on the x and y axis respectively (Pham et al., 2018), (Shirzadi et al., 2012). Currently, the predictability of landslides in the respective area is examined by using a curve under the ROC curve (AUC) (Abedini et al., 2019b). The statistic to be used for comparing the models corresponds to the average of AUC values obtained in the 500 iterations carried out through cross-validation.
Besides AUC (Area Under the Curve), the efficiency of the landslide models will be evaluated through statistical analysis by comparing the classification errors between the models. In general, a parametric test should be used for these cases. However, the values obtained from the error classification have a distribution that does not meet the normality assumption necessary for this type of test, which was tested by the test of Kolgomorov-Smirnov Berger and Zhou (2014), with a result lower than 0.05 for all models, which implies that the null hypothesis that values have a normal distribution is rejected. Furthermore, when performing the Box-Cox transformation it is also not feasible to normalize the models. Therefore, the Friedman non-parametric test will be used to determine statistical differences between the models. This test corresponds to the non-parametric equivalent of the ANOVA test and is used when the samples come from the same distribution and are paired and is used to determine whether the average of the populations is equal (Ostertagova et al., 2014). This test only shows the significant differences between the models without judging pairs between two or more models. To discriminate between the models, the Nemenyi test will be used, which corresponds to a post-hoc test whose objective is to find groups of data that are different after a global statistical test (in this case the Friedman test) has rejected the null hypothesis that the performance of the models is the same. This test carries out pairs test to measure performance (Nemenyi, 1963). Fig. 2 summarizes the design methodology respect to the procedures used this work.