Probabilistic analysis of the Martian hydrological distribution in relation to climatic variables

Problem

The planet Mars is one of the primary objects of interest in recent studies of the solar system and continues to generate profound interest due to its potential habitability and the search for traces of past or present life. Among the various scientific challenges related to the red planet, one of the crucial issues concerns the presence of water. Water is a fundamental element for life as we know it and is also essential for future exploration and colonization of Mars. An element key to our approach is the use of perennial ice, representing a solid form of water, as a direct indicator of water presence, serving as the foundation for our analysis. Our statistical study focuses on this central problem: assigning a probability to the presence of water in each quadrant of Mars based on various climatic variables to stimulate and guide hydrological research on the planet.

Data

The data bolstering our study can be obtained through the Mars Climate Database (MCD)[1], a database supported by space agencies such as ESA, CNES and NASA, which collects simulated and modeled climatic data for Mars and is already used by various research organizations for space mission planning. Through the MCD, it is possible to access the 14 climatic variables of interest (out of over 70 available) selected based on their close relevance and suitability for the objectives of our research[2]. These variables encompass "GCM perennial surface water ice", which will be used as our dependent variable, assuming values of 0 (absence of perennial surface water ice) and 1 (presence of perennial surface water ice). The data include georeferenced observations, comprising a total of 3072 measurements for each variable, that span a wide range of geographical positions. The north latitude ranges from -90º to 90º with an interval of approximately 3.8298º, while the longitude ranges from -180º to 180º with an interval of about 5.9833º. It is important to note that each of these measurements was taken at the 141.3rd degree of the Martian revolution motion; therefore, caution is recommended regarding the sensitivity of future studies to orbital position.

Method

The choice between two distinct approaches is guided by the nature of the distribution highlighted in the graphical representation. If the distribution of the variables appears approximately normal and simultaneously exhibits homoscedasticity, discriminant analysis is suggested as a classifying method. Alternatively, considering the binary nature of the object of the study (1 to indicate the presence of water, 0 to indicate its absence), the use of a logistic regression model is suggested. This model will allow predicting the probability of the presence of water on Mars, offering a robust approach to understanding the determining factors. This methodical approach will be preceded by an inevitable part of pre-processing, that includes the EDA, the data cleaning, and the initial selection of the variables of interest; this screening, will take shape thanks to the Forward Stepwise Selection (FSS) that has been preferred to the Best Subsect Selection (BSS), and that will be conducted with the Bayesan Information Criterion (BIC) rather than others[3]. Furthermore, as the last step of the scientific research there will be the evaluation of the model. Indeed, in order to assess the accuracy of the model, examine the sensitivity and specificity and avoid erroneous classifications (identified as false negatives, indicating no water when there is, and false positives, indicating water when there is not), the application of the following evaluation method is proposed: the data will undergo k-fold cross-validation[4]. This approach ensures model evaluation on independent data and helps prevent overfitting by repeatedly splitting the dataset into training and test sets during the training process. Subsequently, the combination of the confusion matrix and the ROC curve will be applied enabling a complete and reliable assessment of the validity of the model.

[1] For further details refer to appendix, Mars Climate Database.

[2] For further details refer to appendix, The 14 relevant variables.

[3] For further details refer to appendix, Subset selection method.

[4] By default, it implies a split percentage of 90%-10% between training and testing when k=10.

Implementation

The first step of our research is represented by the Exploratory Data Analysis (EDA) phase. At this stage we were able to see that the variables examined did not have missing values. However, a limited number of outliers have been highlighted; It is crucial not to consider such outliers as candidates for deletion as they may contain significant information and represent legitimate phenomena or local conditions of interest. The variable "Altitude above local surface (m)" does not provide any useful information to explain or predict the results as it is uniformly distributed. To examine the characteristics of the distributions of the variables[5], we propose the verification of normality and homoskedasticity, respectively by means of the Shapiro test, preferred for its sensitivity in detecting small deviations, and the Fligner test, preferred for its robustness in the presence of non-normally distributed data. The results of these tests showed that none of the variables we examined is distributed as a normal one and that only some of these[6] have homoskedasticity; therefore, the use of discriminant analysis as a classification method is to be excluded.

Subsequently, the second step of our study is represented by the selection of the most relevant variables for the construction of the predictive model. In this context, we choose to conduct a stepwise forward selection procedure using the Bayesian Informative Criterion (BIC) in order to identify a subset of variables that are particularly informative and significant for the purposes of our investigation. The procedure led to the selection of a group of eight variables, which were identified as particularly relevant to our study[7].

In the third step of the analysis, we proceed with the drafting of the logistic regression model where:

- GCM perennial surface water ice is our dependent variable.
- The dataset is represented by the subset created on the initial dataset by selecting only the variables identified by the forward stepwise selection.

```
> # Logistic Regression model
> glm.fit.fwd <- glm(`GCM.perennial.surface.water.ice..0.or.1.` ~ Temperature..K.+
Pression..Pa.+ Monthly.mean.surface.H2O.layer..kg.m2.+ Water.ice.column..kg.m2.+
Water.ice.mixing.ratio..mol.mol.+ Water.ice.effective.radius..m.+
H.column..kg.m2.+ H2.column..kg.m2., data = Variables.fwd, family = binomial())
```

However, the following code leads to an important warning message:

```
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occured
```

This problem, in simple terms, indicates that our model shows overconfidence in its predictions, producing extreme probabilities, i.e. values very close to 0 or 1, instead of providing intermediate values as one would expect. This behavior could be explained by analyzing the following factors:

- Class Imbalance
- Well separated classes
- Outliers
- Scalability
- Multicollinearity

We have diligently addressed the issue associated with this warning, implementing a series of operational strategies that include the application of oversampling techniques, data standardization, and fusion of affine variables[8]. Then, to complete the logistic regression model of our study, based on the coefficients[9] estimated by the model itself, we proceed to perform calculations on our data set in order to determine the estimated probabilities of belonging to class 1 of each observation. Subsequently, in order to make a final classification, we assign a specific class to each observation, based on the estimated probability value, using a threshold of 0.5 as a discriminating criterion.

[5] For graphical representations refer to appendix, Variables distribution.

[6] Which are: monthly mean surface H2O layer, monthly mean surface CO2 layer, water ice column, water ice mixing ratio, water ice effective radius. For these variables the p-value is higher than 2.2e-16.

[7] GCM perennial surface water ice, Temperature, Pressure, Monthly mean surface H2O layer, Water ice column, Water ice mixing ratio, Water ice effective radius, H column e H2 column. For further details refer to appendix, FSS subset.

[8] For further details refer to appendix, Warning management.

[9] For further details refer to appendix, Logistic regression coefficients.

Evaluation

In accordance with the goal of accurately and reliably evaluating the performance of the logistic regression model developed during the study, a 10-fold cross-validation strategy has been adopted. This method has been employed after the initial logistic regression analysis and subsequently repeated for a total of eight iterations. The main objective of this procedure is to guarantee the reliability of the model's performance measurements, providing a robust and general evaluation of its predictive capabilities on independent data. Analyzing the results of the 10-fold cross-validation and its 8 repetitions, it is observed that the error range of the model remains between 0.0086 and 0.0096[10]. This relative constancy and narrowness of the error range indicate a strong consistency in model performance across different subsets of data which represents a significant indicator of a good fit of the model, suggesting that it not only adapts well to the training data, but also maintains high performance on unseen datasets. Once we got the predictions from the model in each iteration of the cross-validation, we computed the confusion matrix to evaluate the classification performance from previous classification predictions, where each observation was assigned to a class based on the associated probability and a threshold of 0.5[11]. This threshold provided satisfactory results which, however, were optimized by a further filing by changing the threshold (now 0.074), thus obtaining the most balanced mix to ensure high sensitivity and high specificity at the same time[12]. The reduction of the threshold generates an increase in sensitivity, which goes from 87.2928% to 97.7901% and, at the same time, causes a reduction, albeit minimal, in specificity, from 99.8271% to 97.3020%. Finally, our ROC curve[13] approaches the upper left corner of the graph, suggesting that the model is able to achieve high sensitivity while maintaining a low false positive rate.

Conclusion

This project has been proposed as a monographic and methodical essay on a probabilistic analysis of the Martian hydrobiological distribution, based on various climatic variables. Thanks to the logistic regression model that has been applied in this project, we are given the estimation of the presence of perennial surface water ice across different Martian quadrants, as we can see in the heatmap (FIG 1).

In this latter scenario, the findings reveal a substantial dependency of the water presence on a set of eight climatic variables. The utilization of a logistic regression model has given the possibility to well navigate through the complexity on Martian climatic conditions with a very good accuracy based on a 10-fold cross-validation. By the study, one of the critical insights that emerged has been the heterogeneous distribution of water ice across Mars. This heterogeneity is crucial for future exploration and colonization efforts, as it is points to specific areas that may have higher probabilities of water presence, key factor for life. Indeed, the study is proposed as a cornerstone in today's scientific discussion aimed at the progress, and it is based on a data-driven approach. Science is running faster, technology's too, and SpaceX has been able to host the first space flight with civilians on board. Life will become multi-planetary, and Mars is the only planet in the solar system that makes possible the human life: there can be water in some quadrants.

[10] For further details refer to appendix, 10-fold crossvalidation

[11] For further details refer to appendix, Confusion matrix

[12] For further details refer to appendix, Confusion matrix

[13] For a graphical representation refer to appendix

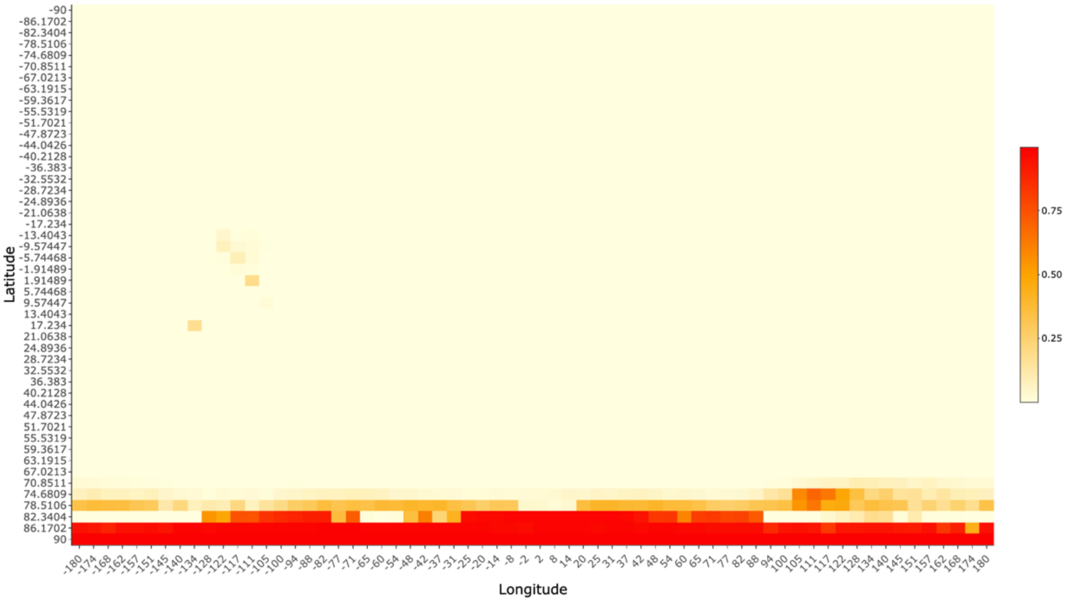


FIG 1 – Heatmap based on the logistic regression and the Martian quadrants.

Appendix

Mars Climate Database: it is a highly significant project that provides a wide range of simulated and modeled climatic data for the planet Mars. This database offers a valuable resource for researchers and scientists dedicated to studying the Martian climate and atmosphere, as well as for those involved in planning space missions to the red planet, such as rovers and space probes.

The information contained in the MCD includes details about temperature, atmospheric pressure, humidity, wind speed, and various other specific meteorological parameters of Mars. These data result from advanced simulations based on General Circulation Models (GCMs) of the Martian surface. This means that the MCD provides an accurate representation of the weather conditions on Mars, thanks to the validation of the data obtained through direct observations.

The scope of the MCD is substantial, as it covers a wide range of temporal scales and atmospheric phenomena. This encompasses yearly variations, changes in dust concentration in the atmosphere, and simulations of various extreme solar radiation conditions in the ultraviolet (EUV) range. Furthermore, the database offers a seasonal representation, with data related to 12 "typical days" spanning the entire year.

The MCD was developed jointly by the Laboratoire de Météorologie Dynamique in Paris, in collaboration with the Open University (UK), the Oxford University (UK) and the Instituto de Astrofisica de Andalucia (Spain) and received support from the European Space Agency (ESA), the Centre National d'Etudes Spatiales (CNES) of France and the National Aeronautics and Space Administration (NASA) of United States of America. Currently, the database is in version 6.1 and is freely accessible through an interactive online server.

In addition to data directly obtained from GCM models, the MCD also provides post-processing tools, such as high-resolution spatial interpolation of environmental data and methods for reconstructing their variability. These additional features make the MCD an even more valuable resource for the research and analysis of the Martian atmosphere.

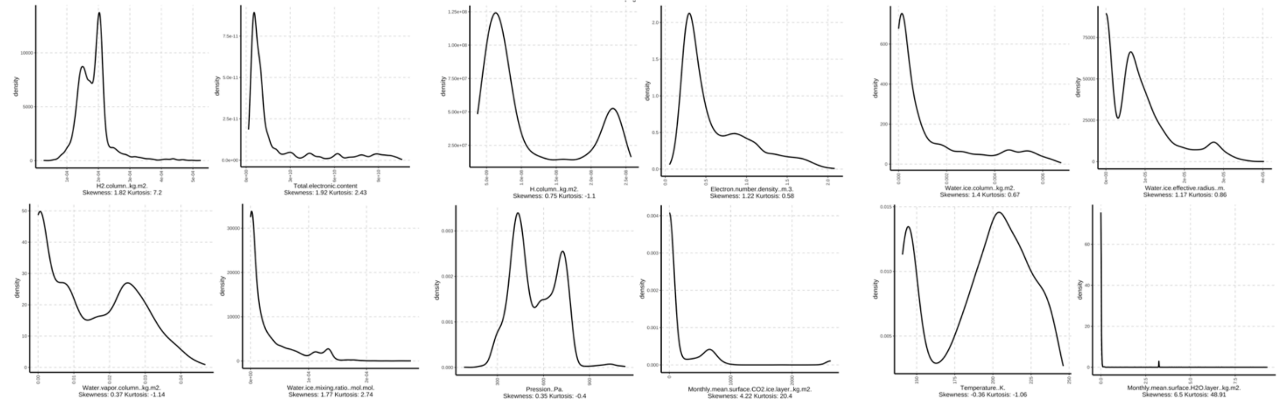
The 14 relevant variables: among the 72 variables provided by the MCD, we have chosen to analyze the 14 most relevant for our study:

- *Temperature (K)*: it is a critical variable as it determines the physical state of water on Mars, affecting the transition of water between solid, liquid, and gaseous states. Due to its greater distance from the Sun and a thinner atmosphere that retains fewer greenhouse gases, Mars is generally a colder planet than Earth; hence, it's essential to consider that water may exist in solid forms.
- *Pressure (Pa):* atmospheric pressure can influence the atmosphere's capacity to sustain water in the form of vapor. Under high-pressure conditions, water is more likely to remain in vapor form.
- *Altitude above local surface (m):* altitude relative to the Martian surface can influence atmospheric temperature and pressure.
- *Monthly mean surface CO2 ice layer (kg/m2):* the amount of carbon dioxide (CO2) ice on the surface can affect the thermal balance, which, in turn, impacts temperature and the presence of water.
- *Monthly mean surface H2O layer (kg/m2):* the quantity of water ice on the surface is directly linked to the presence of water on Mars. The greater the amount of water ice, the higher the likelihood of finding water.
- *Water vapor column (kg/m2):* the amount of water vapor in the atmosphere is significant since water in vapor form is a direct indicator of water presence in the atmosphere.
- *Water ice column (kg/m2):* the quantity of water ice in the atmosphere can influence cloud formation and precipitation possibilities.
- *Water ice mixing ratio (mol/mol):* the proportion of water ice in the atmosphere can reflect the atmospheric conditions related to water.
- *Water ice effective radius (m):* similar to the Water Ice Column, the size of water ice particles in the atmosphere can influence cloud formation and precipitation possibilities, which are, in turn, linked to the presence of water.

- *GCM perennial surface water ice (0 or 1):* the presence of permanent water ice on the surface is a direct indicator of water presence, as permanent water ice suggests the likelihood of water.
- *H column (kg/m2) and H2 column (kg/m2):* as water is composed of hydrogen and oxygen, the quantities of hydrogen and molecular hydrogen in the atmosphere can affect the availability of these elements for water formation.
- *Electron number density and Total electronic content*: these variables may be related to the Martian ionosphere and the interaction of water with solar radiation. They can influence the distribution and stability of water in Mars' atmosphere.

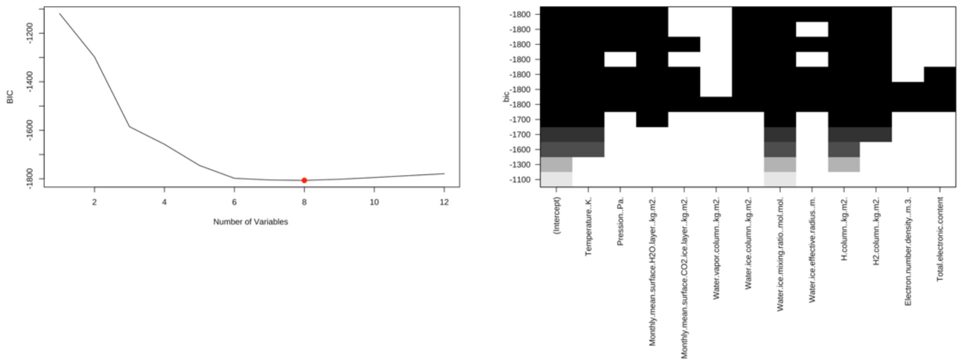
Subset selection method: The adoption of Forward Stepwise Selection (FSS) instead of Best Subset Selection (BSS) is motivated by multiple important reasons.
FSS has proven to be an efficient and scalable choice, particularly useful when dealing with datasets with a large number of potential predictors. Compared to BSS, which looks at all possible combinations of independent variables, FSS proceeds incrementally, evaluating one variable at a time and including it only if it contributes significantly to improving the performance of the model. This approach reduces the risk of overfitting training data, allowing for more robust and generalizable models.
In addition, FSS is based on objective criteria such as p-value or other model goodness indicators, making it particularly suitable when trying to obtain thrifty, yet accurate models. This method avoids the need to manually examine all possible combinations of variables, saving time and resources.
In the context of the application of the FSS, we have adopted the Bayesian Information Criterion (BIC) as the main measure to guide decision-making. The BIC applies a penalty based on the number of variables included in the model, helping to prevent overfitting and promote more parsimonious models. In addition, the BIC is based on Bayesian principles, considering the a priori probability of the different configurations of the model. This approach offers a more accurate estimate of the complexity of the model and its ability to generalize than other criteria such as adjusted R^2 and Cp.
Finally, it should be noted that the BIC demonstrates considerable robustness, especially in the presence of small sample sizes. This feature is of particular relevance when the number of observations in the dataset is small, as criteria such as adjusted R^2 and Cp may be more sensitive to random fluctuations.

Variables distribution:



Distribution of climate variables

FSS subset:

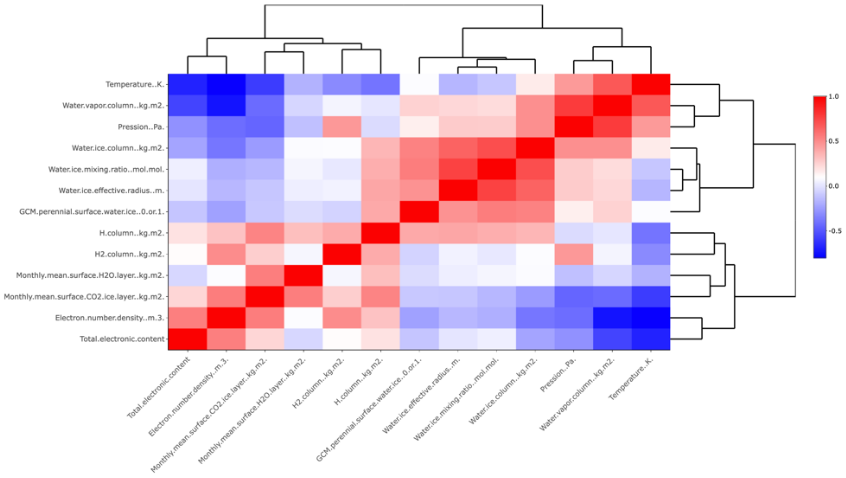


On the left, the graph representing the Bayesian Information Criterion curve as the number of variables included in the model varies, identifying the point at which the BIC reaches its minimum (subset consisting of eight variables). On the right, the representation of the model obtained with eight selected variables and its effect on it

Warning management:

- Class imbalance: as per EDA, it is possible to see that about 94% of the observations of the dependent variable "GCM perennial surface water ice" take a value of 0, leaving room for only 6% in which it records the presence of perennial surface water ice. In these cases, the logistic regression model tends to predict the majority class more accurately at the expense of the minority class. This can lead to incorrect predictions, biased coefficients, and poor generalization to minority class observations, which could be crucial in some applications.
  To solve this problem, we oversampled class 1, bringing the sample to a balance of about 50% for class 0 and 50% for class 1; however, the warning comes back.
- Well separated classes: in some cases, the predictor variables perfectly separate the two classes in your dataset. When this happens, the model can't find a proper logistic regression line that separates the classes, and it may result in fitted probabilities of 0 or 1 for some observations. For this reason, an approach based on discriminant analysis is recommended; However, in the absence of normality and homoskedasticity of the variables, as previously ascertained, such a predictive model is to be excluded.
- Outliers: These can have a significant impact on the estimators of the variables in the model. If an outlier has too much influence on a predictive variable, it could cause an estimation of the model's coefficients that tries to fit too well into the data, leading to calculated probabilities close to 0 or 1; However, as previously stated, the elimination of these anomalous points is to be excluded as they are closely linked to georeferenced (the elimination of one of these would result in the absence of information relating to a quadrant).
- Scalability: If the predictive variables take very large or very small values, it may be difficult for the logistic regression optimization algorithm to converge correctly. At the same time, if the variables have very different scales, the logistic regression model may have difficulty correctly weighing the importance of the variables during convergence. This can lead to predicted odds close to 0 or 1. This problem could be solved by standardizing data; Proceeding in this way and repeating the execution of the logistic regression code, however, the warning reappears, suggesting that scalability was not the main element that caused such difficulty in predicting the model
- Multicollinearity: If two or more independent variables have a strong correlation with each other, the model may show instability in coefficient estimates and forecasts. This phenomenon occurs because the model has difficulty in discriminating the effects of related variables.
  In our study, we evaluated the presence of multicollinearity in the variables involved in the analysis by establishing a threshold value to identify non-multicollinearity, setting it at 10. We found that some of the output coefficients exceeded this threshold, indicating the need to address this issue. To address multicollinearity, we opted for a more inclusive approach rather than eliminating multicollinear variables directly. Our strategy consisted of combining the variables based on their

correlation through the creation of a new column that summarized the indices of the combined variables.



CORRELATION MATRIX WITH DENDOGRAMS

To do this, we standardized the data to avoid scale issues and calculated the indices using an arithmetic mean with equal weights.

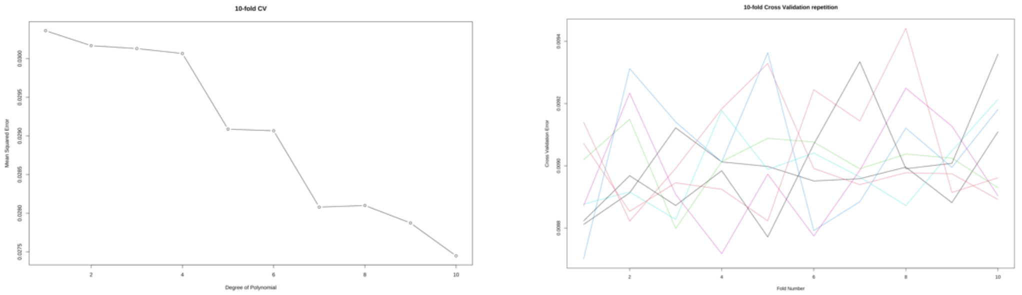
Citing an example, we combined the variables "Pressure" and "Water vapor column" and then evaluated the multicollinearity between this new combination and the other variables. We observed a significant reduction in multicollinearity, but the values obtained were not yet below the predetermined threshold. We then repeated the process for the variables "H2 column" and "Electron number density", further reducing the multicollinearity between the variables. However, we made the decision not to continue with the multicollinearity analysis and to continue our study without combining its variables. Instead of taking the variable combination approach described above, we preferred to keep the variables separate and not take any further steps to reduce multicollinearity. It is important to note that the in-depth analysis of multicollinearity and, therefore, finding an optimal configuration that allows our model to perform at its maximum possibilities, requires a higher level of knowledge; Therefore, the above approach is presented only as a possible multicollinearity treatment strategy.

Logistic regression coefficients: the coefficients revealed by the model expressed in the following table:

VARIABLES	COEFFICIENT
Intercept	-6.119607e+01
Temperature	2.495580e-01
Pressure	2.279010e-02
Monthly mean surface H2O layer	-1.324436e+00
Water ice column	-3.611653e+02
Water ice mixing ratio	3.118562e+04
Water ice effective radius	-2.806653e+04
H column	1.626360e+09
H2 column	-2.281437e+05

On the left, the variables considered by the regression model and the intercept of the regression model. On the right, the coefficients for the variables; It should be noted that the discrepancy in scale in the coefficients is due to the different order of magnitude of the data corresponding to the same variables.

10-fold crossvalidation:



On the left, the results of a single iteration of a 10-fold cross-validation procedure; The dataset is divided into 10 parts (or "folds"), with 9 parts used for model training and 1 for testing. On the right, the representation of a total of 8 repetitions of the 10-fold cross-validation procedure.

Confusion matrix:

		True status		
		0	1	Total
Predicted status	0	2886	23	2909
	1	5	158	163
	TOTAL	2891	181	3072

Confusion matrix with threshold 0.5

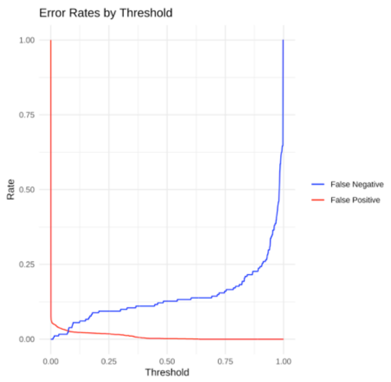
From this we can derive:

$$Err_{tr} = \frac{1}{n}\sum 1_{y_i \neq \hat{C}(x_i)} = \frac{FP + FN}{n} = \frac{5 + 23}{3072} = 0{,}9115\%$$

$$sensitivity_{tr} = \frac{TP}{P} = \frac{158}{181} = 87{,}2928\%$$

$$specificity_{tr} = \frac{TN}{N} = \frac{2886}{2891} = 99{,}8271\%$$

Through an analysis of the distribution of classification errors, it is observable that the reduction of the value of the classification threshold leads to an acceptable increase in the rate of false positives, but at the same time leads to a significant decrease in the rate of false negatives; This phenomenon highlights a substantial trade-off within our model, suggesting the convenience of lowering the classification threshold.



The selection of the new threshold was determined in view of the importance of not overly penalizing the identification of true negatives in the context of our research. The main objective of our study is to guide space exploration activities by identifying the geographical areas of Mars where water is most likely to be found, in order to optimize future missions and minimize the associated costs. However, indiscriminately identifying too many quadrants as potential areas containing water sources would result in the allocation of resources to areas where such a resource is not present, undermining the effectiveness of missions. On the other hand, identifying too few potential areas would defeat the very purpose of our study. Therefore, in order to avoid an excessive imbalance between the sensitivity and specificity of our model, we have opted for a threshold that allows us to maintain a false negative rate equivalent to that of false positives. This threshold corresponds to the point of intersection between the two distributions in the chart previously analyzed, resulting in a specific value of 0.074. This choice balances the ability of our model to identify areas with potential water presence without excessively compromising specificity or sensitivity and leads us to a new confusion matrix:

		True status		
		0	1	Total
Predicted status	0	2813	4	2817
	1	78	177	255
	TOTAL	2891	181	3072

Confusion matrix with threshold 0.074

$$Err_{tr} = \frac{1}{n}\sum 1_{y_i \neq \hat{C}(x_i)} = \frac{FP + FN}{n} = \frac{4 + 78}{3072} = 2{,}6693\%$$

$$sensitivity_{tr} = \frac{TP}{P} = \frac{177}{181} = 97{,}7901\%$$

$$specificity_{tr} = \frac{TN}{N} = \frac{2813}{2891} = 97{,}3020\%$$

ROC curve:

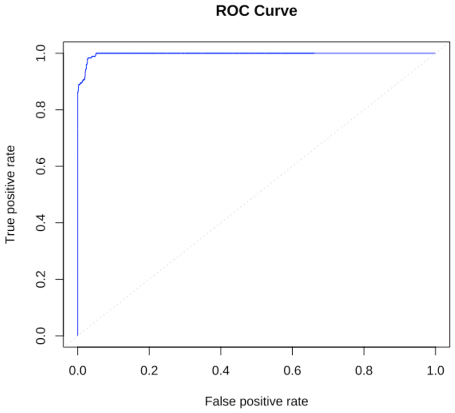


FIG. 6 – ROC CURVE