# Impact of imbalance data on Logistic Regression models to predict risk plant disease

*Fiore Juan Manuel*[1], *Suarez Francor*[2], *Balzarini Monica*,[3] Bruno Cecilia[4]

[1]juanmfiore@mi.unc.edu.ar, Unidad de Fitopatología y Modelización Agrícola (UFyMA). INTA-CONICET

[2]suarezfranco@agro.unc.edu.ar, UFyMA. INTA-CONICET

[3]monica.balzarini@unc.edu.ar, UFyMA. INTA-CONICET. Estadística y Biometría. Facultas de Ciencias Agropecuarias (FCA). Universidad Nacional de Córdoba (UNC). Argentina.

[4]cebruno@agro.unc.edu.ar, UFyMA. INTA-CONICET. Estadística y Biometría. FCA. UNC. Argentina.

The imbalance problem occurs when there is a skewed class distribution, large samples of one class, and few samples of the other class. The degree of imbalance can vary largely and when this happens the ability of the statistical model to predict the occurrence of the minority class is heavily affected. This situation is frequent in plant diseases, with the number of unhealthy plants being lower than that of healthy plants. Logistic regression (LR) models are linear models that allow us to predict a binary event from multiple predictive variables. This work aims to evaluate the impact of unbalance data on two different models: a classical LR and a more state-of-the-art version of a Logistic regression in this case boosted (BLR), on a binomial classification problem from plant disease data using climatic variables as predictors. The current study implements two under-sampling algorithms, a well-known Tomek algorithm and Condensed K-neighbours (CNN), and two oversampling algorithms: Synthetic minority oversample Technique (SMOTE) and Adaptative Synthetic Sapling (ADASYN) with three different proportions of imbalance based on two pathosystems collected on cucurbits crops. The original imbalance dataset was 5:1 in the PRSV pathosystem and 9:1 in the CMV, absent to presence, respectively. Then, we generated different levels of imbalance to change the relationship between the absence/presence of pathosystem. Thus, for each original dataset we modify the dataset with 10:3 ratio, other with 10:6 ratio, and finally 1:1 ratio (completely balanced). The original dataset was partitioned into 80% for training and 20% for validation. Then, the two models, LR and BLR, were trained and cross-validated. The experiment was reproduced six times with different seeds to ensure the reproducibility and randomness on the dataset partition for validations. The criteria to compare the model's performance to predict the risk of plant disease were the Area Under the Curve (AUC) from the receiver operator characteristic (ROC) curve, sensitivity and specificity were used as metrics. Finally, the models were tested with the remaining 20% dataset who did not participate in the training. LR with the original (imbalance) dataset was the worst performance with an AUC of 0.62, and a sensitivity of 0.28 although its specificity achieved about 0.95. LBR had an AUC of 0.75, a sensibility of 0.53, and a Specificity of 0.98. The model that most responded to the different imbalance levels was the LR. Both algorithms got the best results with SMOTE 6:10. LR AUC improved to 0.76 and LBR achieved and AUC of 0.82. leaving the 1:1 ratio in second place. This pattern of SMOTE 6:10 performing better, repeats itself in combination with the under-sampling techniques. Indicating there is an optimum threshold in which from that point the excess data produced by the algorithm generates noise. Undermining the algorithms performance.

**Keywords**: oversampling algorithms, under-sampling algorithms, cucurbits crops.