# Extended Abstract
## Please do not add your name or affiliation

| Paper/Poster Title | **The determining factors of farmers' participation in insurance schemes: a comparative analysis of machine learning tools.** |
|---|---|

**Abstract prepared for presentation at the 97th Annual Conference of the Agricultural Economics Society, The University of Warwick, United Kingdom**

**27th – 29th March 2023**


| **Abstract** | ***200 words max*** |
|---|---|
| Identifying the factors influencing insurance participation plays a vital role in ensuring its success. Many scholars have tried to unravel the question with the most varied approaches. Using Machine Learning, we address a gap in the literature by analysing 10,926 individual fieldcrop farm observations from the Italian Farm Accountancy Data Network from 2016 to 2019.We use three Machine Learning (ML) tools: Least Absolute Shrinkage and Selection Operator (LASSO), Boosting and Random Forest, and compare them to the Generalised Linear Model (GLM) usually used in insurance modelling and adopted in this study as the baseline. In addition, we have adopted a cross-validation process to determine the reliability of the findings. The results revealed the outstanding performance of ML: these tools efficiently perform variable selection with an accurate prediction of insurance compliance. In particular, Boosting performs better than other tools, especially in comparison with GLM, using a smaller set of regressors. At the same time, we have explored the reliability of variable frequency selection and the importance of each variable concerning participation. Focusing on the subset of information that best explains insurance participation could reduce the cost of designing insurance schemes. | |



| **Keywords** | Machine Learning, Insurance, Risks |
|---|---|
| **JEL Code** | G22, C55, Q18, Q12, G22<br>see: www.aeaweb.org/jel/guide/jel.php?class=Q) |


| **Introduction** | ***100 – 250 words*** |
|---|---|

External shocks, such as extreme events in weather conditions, markets or policy, significantly impact agriculture. Farmers use a variety of risk management tools to face these risks, where insurance takes the lion's share. The agricultural insurance literature has analysed several aspects concerning the relationship between farmers and insurance. Although consolidated literature has studied the effects of multiple farm-specific features in risk management strategies, it is necessary to emphasise which variables affect the adhesion in insurance schemes. This assessment is required to build or modify the structure of the insurance contract or to meet special insurance requirements based on the unique features of individual farms. Hence, this can support insurance companies and policymakers in creating contracts that satisfy farmers' needs. This study analyses the (many) characteristics that potentially influence farmers' behaviour if we consider the involvement in an insurance scheme using different Machine Learning tools. Participation choice is usually affected by a large number of variables, making the task challenging. Furthermore, these factors

are interrelated and can easily mask the influence on any prediction of adhesions. Performing an accurate prediction and recognising the factors that affect farmers' participation are the main objectives of this analysis. Unfortunately, traditional methodologies (GLM) cannot satisfactorily use this large set of variables because of problems such as multicollinearity and overfitting. Problems that ML tools could overcome.

| **Methodology** | ***100 – 250 words*** |
|---|---|

This study investigates which are the characteristics that push the farmer to adhere to insurance instruments subsidised by the Rural Development Program (RDP). Relyed on the literature, we investigate 66 characteristics, including economic, technical, financial, topographic and climatic features. The analysis utilises individual data from the Italian FADN from 2016 to 2019. We restrict the analysis to fieldcrop farms (type 1 of farming) in order to obtain subjects with comparable risk exposure. This selection yields 10,926 observations. We employ three Machine Learning (ML) tools to find the factors that affect insurance participation: LASSO, Boosting and Random Forest. These ML tools have been chosen because they select the variables with high goodness-of-fit. The performance of ML is compared to that of GLM, a model usually used in the appraisal of insurance policies. Furthermore, to evaluate the robustness of the results, we implemented a cross-validation (CV) procedure.This technique divides the entire sample into two sets: the training-set to find the model setting and the test-set used to predict the outcome by adopting the setting found previously. This procedure is performed several times.We take into account two aspects when assessing the variable selection capabilities of models: first, the frequency of selection, and second, the relative significance of each variable in the adoption of insurance. In addition, the model performance also takes goodness-of-fit into account by analysing the confusion matrices, MAE (mean absolute error), MSE (mean squared error), and RMSE (mean squared error).

| **Results** | ***100 – 250 words*** |
|---|---|

Boosting prevails in selecting variables: this model can reach high performance using only 41 variables on 66, at converse, the other models present low capacity in selection variables. The confusion matrix reports that ML overcome GLM considering sensitivity, negative prediction value, detection rate and balanced accuracy. In particular, Boosting obtain the best performance, follows by Random Forest, and Lasso. MAE, MSE and RMSE confirm the best Boosting performance, the poor outcomes reach by GLM and LASSO, and finally, Random Forest shows mixed results. Two additional aspects are under investigation, and extensive results will be provided in the full version of the paper: i) Which variables are selected the most? ii) Which variables are the most important? Preliminary results show that the crucial factors that affect insurance participation are in order of importance: farm economic size, presence of other gainful activities, amount of utilised agricultural area, kW of available machinery, production diversification (Herfindahl index), degree of intensification (as total revenue per unit of utilised agricultural area), fixed capital on total capital, and mechanical expenses.

| Discussion and Conclusion | *100 – 250 words* |
|---|---|

Although participation in an insurance scheme is a complex decision, ML ensures relatively good prediction for sure better than GLM models. Furthermore, Boosting offers better performances within the considered ML approaches by utilising a small set of regressors. Conversely, the settings of Boosting can be challenging, and the evaluation of trade-offs with performance must be necessarily considered. Nevertheless, the proposed ML tools allow us to identify the variables that affect participation in insurance, their robustness of selection and their importance.

The general conclusion show as ML is a helpful tool for exploring the factors that explain farmers' participation in insurance schemes. Furthermore, the results obtained can help design better insurance schemes and, hopefully, boost farmers' involvement. Therefore, the ML approach is a crucial step that should be done carefully considering the characteristics of the empirical case study.