



N-INFINITE

by  livepanel

Survey Processing incorporating Synthetic Responses

June 2024

Index

Survey Processing incorporating Synthetic Responses	1
Index	1
Abstract	2
Introduction	2
Methodology	4
Case study: Latin American Mobility	9
Chart 1: Sample	9
Model preparation	10
Chart 2: Average Deviation	11
Chart 3: Average Deviation Drill Down	12
N-Infinite comparison: 2024 vs 2020	13
Chart 4: Average Deviation 2020 vs 2024	13
Last conclusions	15
Bibliography	16
Appendix	17
Appendix 1: Survey structure	17
Appendix 2: Key Training Variables	20



Abstract

This paper presents the significant advancements in performance, quality, speed, and cost reduction achieved through the integration of Machine Learning (ML) methodologies in survey processing via the generation of **synthetic responses based on real users**. The study highlights LivePanel's innovative N-Infinite algorithm, which has been in development since late 2017, as a pivotal tool in this evolution. By leveraging artificial intelligence, LivePanel addresses traditional survey challenges by filling incomplete responses and reaching target quotas in difficult data collection environments. The methodology, illustrated through a case study on mobility in Latin America, demonstrates the effective use of ML techniques to enhance data quality and precision. The key findings include a marked reduction in the average deviation, highlighting the algorithm's efficiency in generating reliable data with only a third of the effort compared to traditional fieldwork. The paper underscores the potential for further advancements in predictive models and quality control measures, solidifying AI's role in revolutionizing survey methodologies.

Introduction

Reviewing the evolution in the survey industry, it has undergone significant evolution over the last few decades. Initially, surveys were conducted primarily through face-to-face interviews and paper questionnaires, which required considerable logistics and human resources effort. Thanks to technological advances, telephone surveys and later online surveys emerged, allowing for faster and more accessible data collection.

Today, machine learning has emerged as a transformative force in a variety of industries. Its application provides value in a variety of sectors from agriculture (Liakos, 2018) to medicine (Rajkomar, 2019). In recent years, its adoption and evolution have been phenomenal, ushering in a new era. Livepanel has proposed to take the next step in the evolution of surveys, taking advantage of the benefits of Artificial Intelligence, in order to solve the difficulties that this sector presents, in this



sense it has been charting a great path since **the end of 2017** with its **N- algorithm. Infinite.**

This innovation allows the following functions to be performed:

- ➔ **Complete the incomplete:** Allows to load a group of surveys abandoned by the panelists for various reasons (distraction, boredom, lack of time, etc.) and complete the missing fields based on information from other completed surveys, very quickly and accurately.
- ➔ **Complete quotas:** Promotes obtaining of valuable information in environments hostile to the survey, that allows completing the study quotas. Resolving situations such as lack of interest in the topic, availability of qualified users, representativeness of the sample with respect to the population pyramid, among others.

Through this work, it is proposed to show the improvements in performance, speed and costs that the adoption of Artificial Intelligence methods can generate, specifically referring to Machine Learning methods. For this, a case study will be developed based on a survey of individuals, and a comparison will be made of the original results, against the results that can be generated by surveys processed through Machine Learning (ML) techniques, at different response levels incorporating synthetic responses and comparing these results against the first work that LivePanel presented in 2020, and thus analyze the improvements in recent years.

In order to provide a framework for evaluating the results looked for, we understand that the following premises are valid for the Public Opinion and Markets industry at the time of writing this document:

1. **The details:** The sample size and composition requirements of the commissioning studies are evident when wanting to perform drill-downs that offer value while maintaining the significance of the information offered. It is



relevant to have the

lowest sample deviation at the aggregate level, but it is also important to be able to find richness in the required sections.

2. **Speed and cost do matter:** A survey system that does not present significant quality differences but that manages to deliver partial or total results in a fraction of the time and at a fraction of the cost is positively valued by users of field information.

Methodology

The work process consists of a five-step process that begins with the preparation of the hearings and ends with the development of comparative metrics. Step **1** involves extracting the audience obtained during fieldwork on our panelist base. In this first instance, a fundamental task is developed and that is to **check the quality** of the users' responses, since this is the **fundamental input** to obtain the desired results, AI is not immune to the principle of garbage in, garbage out. This process involves, for example, removing empty responses , normalizing free text responses with the help of Open AI, debugging inconsistent responses, or end-to-end logic checks.

After this filtering, and with the aim of making a comparison between real responses and synthetic responses, the selected cases are ordered according to the date on which they responded, and the respondents are separated into two groups, the first ones will act as input for Machine Learning, which will be the training users, and the second will be the users to predict. For example, we take the first two days of going out into the field, which corresponds to 30% of the first responses, and we use it as a training file. The remaining 70% will be real users that will be used for prediction, from which we will obtain the synthetic responses. This will be repeated for different combinations from 30% to 90% of field cases, in 10% increments.

In **step 2** , in an automatic process developed by the Livepanel tool, a comparative procedure is carried out, which allows establishing which questions within the user database, excluding those from the questionnaire on which we are going to work, were



answered by both the field cases and the audience selected for the prediction. In all operations, the age range, gender, country and socioeconomic level are included.

The precision with which the predictions will be made will depend on the quality of this data, since these questions allow us to relate field users to what we will use to predict. Here variables come into play that allow us to find out the **personality or behavior** of the person, such as their eating habits, their most used means of transportation, use of the mobile phone, among many other variables that serve this purpose. The tool can work with the data in encoded form, since the tool's own characteristics do not require text fields. This first step culminates with the selection of a large number of features to work on (on average 250 variables are usually generated in this first instance).

The **model training** that is carried out in **step 3** through an automated procedure in a LivePanel API, consists of the **evaluation of different** Machine Learning methods accompanied by a respective tuning of hyperparameters. To ensure the predictive effectiveness of the model, the k-fold-Cross Validation method is used. The models considered are: Tree and Ensemble of decision trees (Random Forest and Gradient Boosting) and Logistic Regression. The reason for choosing this variety of ML algorithms lies in the fact that they sought to use models that are linked in the greatest possible way with supervised methods, since they allow generating explanations at the level of the characteristics used. Now, when determining what each of these methods is used for, they can be grouped into two large groups: classification problems or regression problems. The first are those in which the variable to be predicted (or estimated) is categorical; Regression problems, on the other hand, aim to study continuous phenomena. In each training, between 10 and 255 machine learning models are evaluated, and the one that demonstrates the best prediction capabilities is selected after being tested with a partial population of field cases.

Each of the methods used will be briefly described below:

- ➔ The **Decision Tree** is made up of a sequential series of logical constructions, which lead to a series of possible results (forming something similar to the branches of a tree). Although it is usually used mainly for classification problems, they can also be used within regression problems (in these cases they are called regression trees) (Charbuty, 2021).
- ➔ Ensembles **consist** of a set of machine learning models. Although this class can be comprised of any machine learning model, for these cases ensembles made up purely of decision trees will be used, because it is much faster to build them. Both the *Random Forest* type is used, which consists of the formation of a large number of trees, from which a training subset is selected based on random samples, and *Gradient Boosting*, which has a similar essence with one exception: each tree tries to learn from the previous one (Dietterich, 2000).
- ➔ **Logistic Regressions** allow us to estimate the probability that a variable belongs to a categorical variable based on a quantitative one. Its main application consists of binary classifications (Nick, 2017).
- ➔ **Neural Networks** are computational models that emulate the functioning of the human brain to process information and learn from it. They use interconnected node structures to recognize complex patterns and continually improve through machine learning. They are used in multiple fields, such as speech and image recognition, analysis of large volumes of data and system automation. Their ability to adapt and solve complex problems makes them essential in the development of advanced technologies (Kumar, 2012).

However, this definition is not exhaustive. The different methods corresponding to the use of sets of decision trees can also perform regressions with good results.

That said, for each field an **optimal model** trained sequentially will be generated, which is chosen through a combination of evaluation statistics among which the **F**

Statistic was chosen, which combines both the Precision analysis -positive predictions that actually are- and *Recall* -positive predictions over the total of positive predictions- as well as this second independently. The aim of the choice of these statistics is **to consolidate efficiency in training** fields with unbalanced samples.

Step 4 performs the **sequential prediction** of the data sets also in an automated manner. In this context, the **SmartMatch** tool is incorporated . This technology arises in response to the need to mitigate the natural tendency of Machine Learning models to underestimate the participation of smaller classes in the predictions offered.

Finally, for **step 5**, two comparative activities were carried out:

- ➔ Firstly, the calculation of the average of the absolute deviations for each of the possible responses with respect to the field result. For example: for question number 1, if the results were 75% for option 1, 15% for option 2 and 10% for option 3 in the **total** number of cases in the field, if the prediction was 73% for the option 1, 16% for option 2 and 11% for option 3, the result of the general deviation will be $(2+1+1)/3 = \mathbf{0.75\%}$.
- ➔ Secondly, and pursuing the aforementioned objective of *Allowing the disaggregation of information (Drill-down) while maintaining significant sample sizes in each section and even unpublished combinations in the field work*, an analysis similar to the previous point was carried out but with **three levels of disaggregation** , namely: Country, Age and Gender for a total of **72 additional combinations** for each possible answer, for a total of 1867 cuts/response (combination of Country, Age, Gender and Possible Answer).

Imputation of missings: for field cases, there will be cuts that will not see observations in the response. There is literature that provides information on existing methods for imputing missing values from surveys, such as the work of Rosati (2017), which tries to impute through ensembles of Lasso regressions. However, the following criterion will be used: given the existence of missing

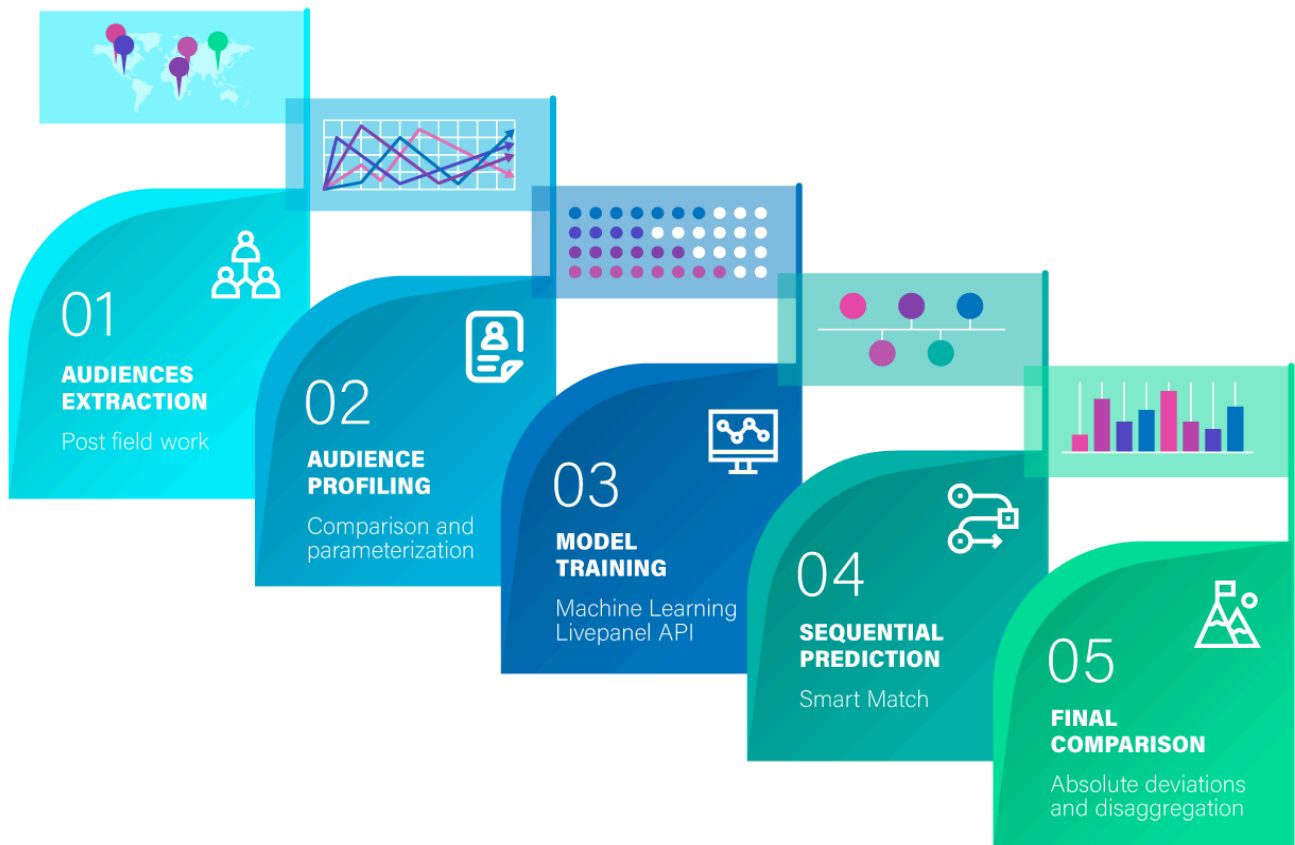


N-INFINITE

by  livepanel

values (as long as the original field has answers in it), a deviation of 50% will be imputed. The reason for this is that, if there is a value other than 0 in 100%, and the field could not predict those responses, technically we would be talking about an infinite deviation (infinite 0s enter any non-zero number). In order not to penalize the field so much in these cases, it is reduced to only 50%.

The procedure can be summarized in the following image.



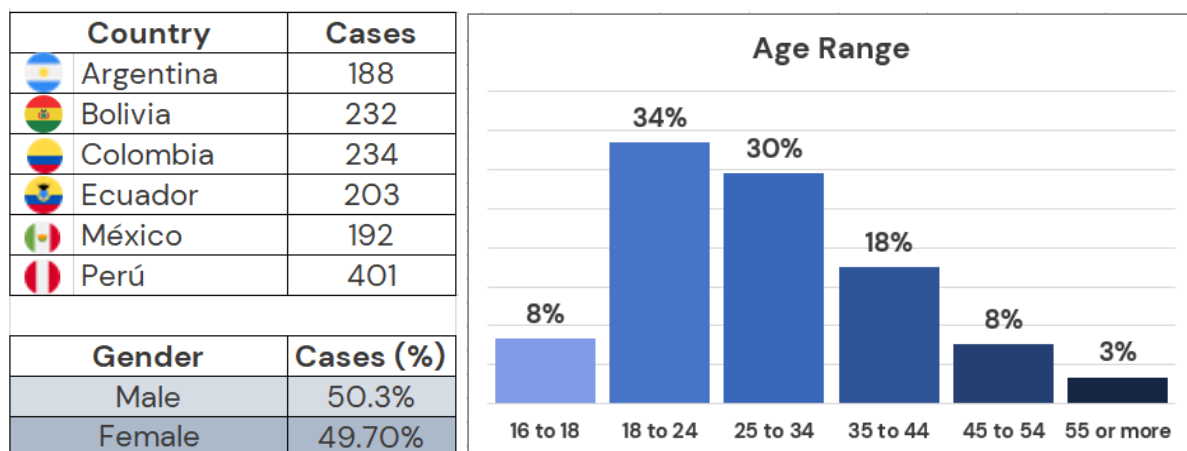
Case study: Latin American Mobility

This is a study published in the newspaper La Nación ¹that analyzes mobility in Latin America, specifically focused on sustainable mobility. Its main objective is to provide relevant information that serves business decisions in this emerging sector. Through

this study, we seek to understand current trends and the growing interest in this new form of transportation in the region.

The study was carried out for a total of 2,150 individuals distributed in various Latin American countries, in the “free fall” modality. Starting the process, a set of 6 countries was selected, which have the largest sample size and representativeness of age groups (the sample was limited to those over 16 years of age). Once **step 1** has begun , **filtering is carried out on the quality** of the responses, and 1216 cases are reached in the final sample, on which our studies will focus. Thus obtaining good quality raw material from this first instance.

Chart 1: Sample



Source: self created

¹<https://www.lanacion.com.ar/autos/amor-o-rechazo-quienes-se-comprarian-un-auto-electrico-y-que-pasa-en-la-argentina-nid03112023/>

The questionnaire included 13 questions, but for the purposes of the study, only the ten with the fewest possible response items were selected, in order to simplify the analysis of variance. This is possible since due to the format consulted there are no interrelationships between questions - a circumstance that, on the other hand, strongly conditions the following answer(s) and significantly improves the measured performance - in order to preserve the objectivity of the study. . The structure of the survey is attached in Appendix 1.

Model preparation

In the development of **step 1**, the 1216 cases were divided into various groups that will be used on the one hand as training and on the other as prediction. For example, a first study that is carried out with 30% field cases and 70% with Machine Learning predictions, here the first 365 users who responded to the survey are taken as training cases for the models, and the remaining 851 will be users of prediction. Which simulates a situation of strong demand for the tool. The same is done for different combinations until reaching 90% field users and 10% for prediction, a less demanding situation. Finally, a dataset is generated with the original responses, that is, 100% field, which will serve as a basis for comparison.

Then, starting **step 2**, by matching the common questions that users have, a total of 250 possible questions were found to use, among which are active type data sources, on the one hand data on the variables users' social profiles (gender, age, SES), other questions that serve as profiling, as well as passive data (that provided by the user when accepting the service). Within the last one can be named:

- ➔ Type of device
- ➔ Apps installed on a subset of 100 most installed apps

As previously mentioned, this step produces a large number of **variables** applicable to model training and prediction. After this, this number of questions will be reduced



N-INFINITE

by livepanel

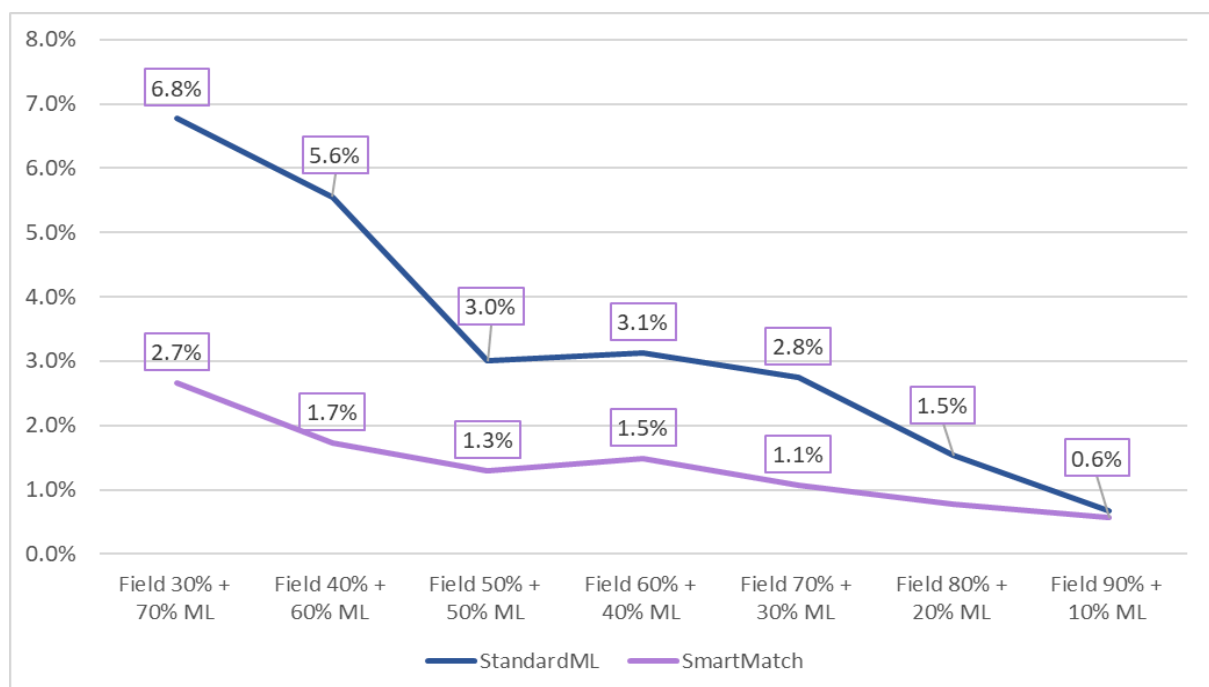
using Natural Language

Processing (NLP) or Regularization (Lasso) techniques. Which will be called **Key Training Variables (VCE)**. Some of these are attached in Appendix 2.

In **step 3**, a total of 250 models were trained, of which the model with the highest F and Recall was chosen for each of the fields to be trained, followed by the prediction corresponding to **step 4**, both with the Smartmatch function and without it. (StandardML). For **step 5**, comparisons are made both at an absolute and relative level (Drill Down):

Chart 2 measures the average deviation for different combinations of field cases and machine learning with respect to the 100% field data. As expected, as the participation of field data increases, the deviation decreases. The interesting thing here is to note that even in the most complex situations, the tool achieves low deviation values, with a value of 6.8%.

Chart 2: Average Deviation



Source: self created



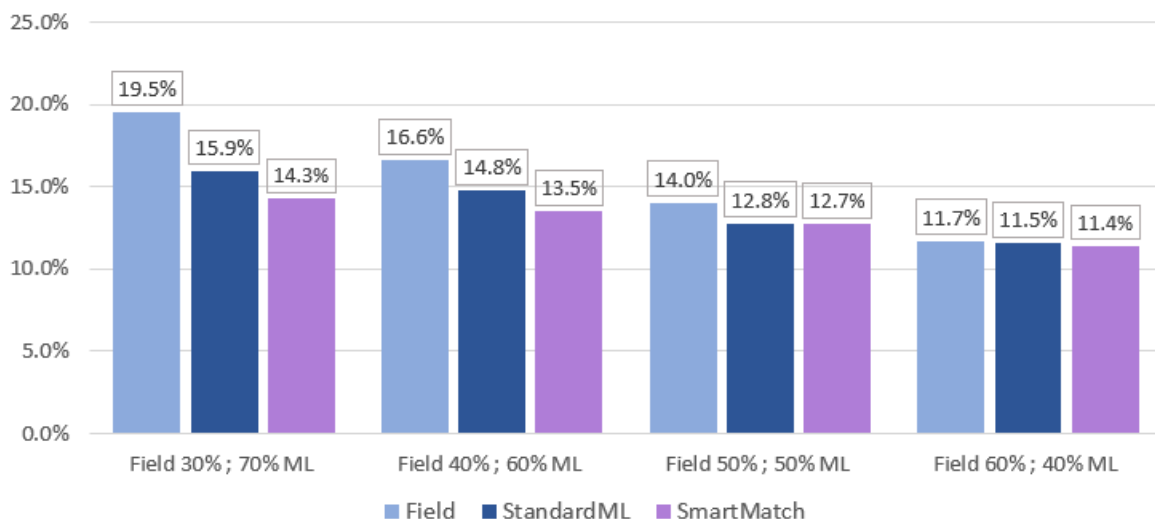
N-INFINITE

by  livepanel

Now, something very important about this graph is to see what happens with the new Livepanel development, SmartMatch. It is observed that this tool achieves a substantial improvement, and allows more accurate results to be achieved even in the most adverse situations. For example, it reaches a 2.7% deviation with only 30% of the field; to reach this same level, 70% of the field is necessary to obtain the same deviation values with Standard Machine Learning.

Next, in chart 3, we analyze what happens in relative terms, allowing drill down.

Chart 3: Average Deviation Drill Down



Source: self created

This graph is the most revealing regarding the impact of the methodology: unlike other procedures that show severe limitations when interpreting or weighting partial results to be comprehensive of a population, the use of a solid methodology that considers both responses and pre-existing information of the panelist allows not only to reproduce responses with greater precision in the scarcest sections of the field, but which gives life to a very important number of cuts that do not show responses in the partial field. Indeed, the methodology is capable of **creating value** and this value creation, by definition, can be improved over time, thanks to the accumulation of

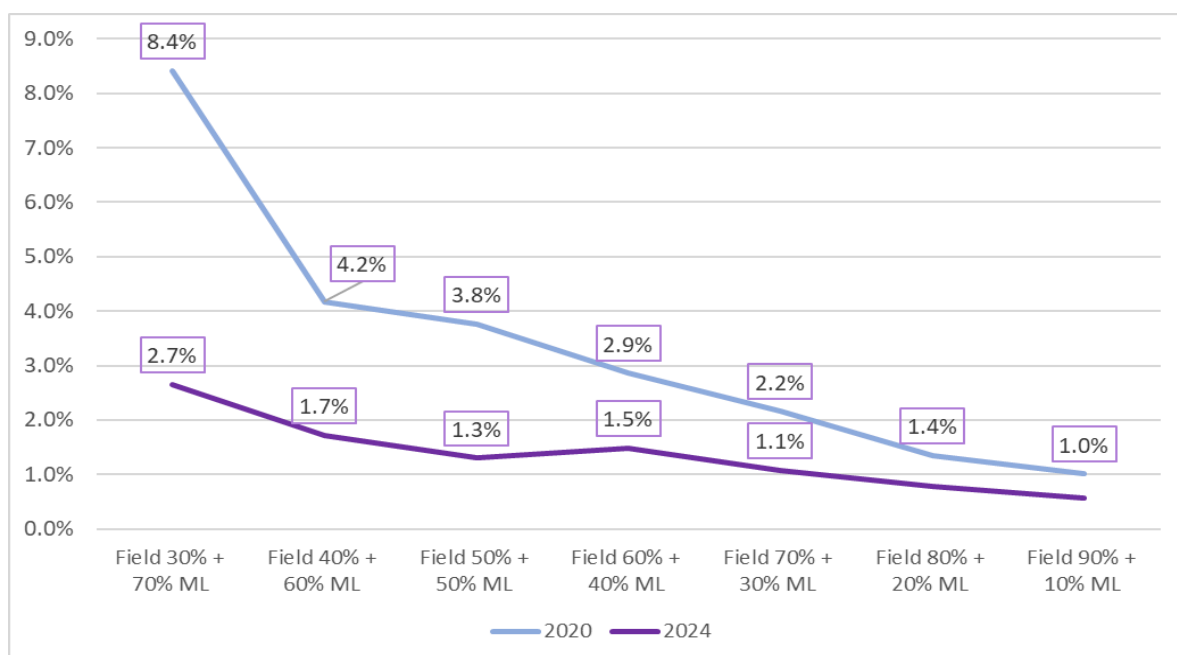
questions of a similar nature, a greater amount of passive information, improvements in the recurrent training of models. , etc.

Thanks to the **creation of new cuts** , Machine Learning tools allow obtaining a smaller deviation even by analyzing in a disaggregated way. In this comparison, an improvement is also noted with the application of SmartMatch, which allows the deviation in the study to be further reduced.

N-Infinite comparison: 2024 vs 2020

Livepanel has years of development and projects carried out with AI applied to surveys, thanks to this it has achieved enormous advances in the optimization of the tool, which allows it to reach high levels of quality, with lower costs and processing times. These advances are can be evaluated in the following graph.

Chart 4: Average Deviation 2020 vs 2024



Source: Own elaboration based on the work of Livepanel "Processing Surveys from Lookalike Audience" (2020) and data from the present study.



Even when the studies correspond to different projects, the data allow us graph the differences that exist between a tool that was in full development in 2020, and the one today that is much more robust. We observe that for the results based on 30% field and 70% synthetic responses, the average deviation decreases from 8.4% to 2.7%, and this advantage is replicated along the curve, for all combinations of field and synthetics. These results were thanks to progress in various aspects, which can be summarized as follows:

- **Better data quality:** We know the importance of raw materials for each investigation. For this reason, we have learned to make the most of both the profiling data and the answers to the questionnaire to be analyzed. This took us from discarding cases with empty responses to eliminating users who showed that they were not responding conscientiously.
- **Logic Controls:** We introduced a solution that analyzes the logic of the questionnaire from start to finish, allowing not only excellent results for each isolated variable, but also for the project as a whole.
- **Choice of models:** Experience allows us to have developed a better system, which manages to choose from a large number of models, the one that best adapts to the characteristics of each question.
- **SmartMatch:** One of the latest innovations, and one that has generated a revolution in our projects. This tool captures the anomalies typical of Machine Learning, which tends to overweight some responses, and corrects them.



Last conclusions

Throughout this work, we have demonstrated that the integration of Artificial Intelligence methodologies, specifically Machine Learning (ML) techniques, in survey processing, can offer significant improvements in terms of performance, accuracy, speed and costs.

Implementation of the N-Infinite algorithm has enabled incomplete surveys to be completed and quotas met in challenging data collection environments with high accuracy and efficiency. The results show that it is possible to obtain useful and relevant data using only a third of the effort of the traditional field, which significantly impacts the times and costs of market research.

In this study, the use of SmartMatch has been crucial, as it allows an effective synergy between field cases and predictions, minimizing anomalies and improving the quality of responses. Advances in quality controls, choice of models and questionnaire logic have resulted in a notable reduction in the average deviation, from 8.4% in 2020 to 2.7% today.

Looking ahead, it is essential to continue improving the details. Even more complex situations where the training and prediction population pyramid are significantly different can be explored to further develop the predictive models and control SmartMatch functions. Livepanel continues to study and develop improvements on various fronts, with a clear long-term vision.

Bibliography

Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, 2(01), 20-28.

Deltell, L., Claes, F., & Osteso, J. M. (2013). Predicción de tendencia política por Twitter: Elecciones Andaluzas 2012. *Ámbitos. Revista internacional de comunicación*, (22).

Dietterich, T. G. (2000, June). Ensemble methods in machine learning. In *International workshop on multiple classifier systems* (pp. 1-15). Berlin, Heidelberg: Springer Berlin Heidelberg.

Kumar, K., & Thakur, G. S. M. (2012). Advanced applications of neural networks and artificial intelligence: A review. *International journal of information technology and computer science*, 4(6), 57.

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, 18(8), 2674.

LivePanel. (2020). Processing surveys from lookalike audiences. *Esomar*, 21. Retrieved from <https://esomar.org/uploads/attachments/clopxahp100njtj3vu69hqyyy-processing-surveys-from-lookalike-audiences.pdf>

Nick, T. G., & Campbell, K. M. (2007). Logistic regression. *Topics in biostatistics*, 273-301.

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347-1358.

Rosati, G. (2017). Construcción de un modelo de imputación para variables de ingreso con valores perdidos a partir de ensamble learning: Aplicación en la Encuesta permanente de hogares (EPH). *SaberEs*, 9(1), 91-111.



N-INFINITE

by  livepanel

Appendix

Appendix 1: Survey structure

Mobility Survey

1. Are any of the modes of transportation you use hybrid, electric, or powered by renewable or sustainable fuels?
 - Yeah
 - No
 - Don't know
2. Have you considered changing your traditional vehicle for an electric or hybrid one?
 - Yeah
 - No
 - I have one
3. Are you willing to pay more for an ecological means of transportation?
 - Yeah
 - No
 - It depends on the price
4. How often do you use mobile applications related to sustainable mobility?
 - Daily
 - Weekly
 - Monthly
 - Rarely
 - Never



N-INFINITE

by  livepanel

5. Do you think your city is prepared for sustainable mobility?

- Totally agree
- OK
- Neutral
- In disagreement
- Totally disagree

Info: In what we call “Developed Countries” there is a growing use of electric vehicles, hydrogen, and sustainable fuels. Considering a certain delay in Latin America.

6. What do you think are the main factors that give rise to this delay? Are you satisfied with the service of your mobile company? - Weak economies

- Little relevant
- Very relevant

7. What do you think are the main factors leading to this delay? – Role of governments

- Little relevant
- Very relevant

8. What do you think are the main factors leading to this delay? – Disinterest of manufacturers

- Little relevant
- Very relevant

9. What do you think are the main factors leading to this delay? - Geographic characteristics

- Little relevant



N-INFINITE

by  livepanel

- Very relevant

10. What do you think are the main factors leading to this delay? - Local businesses

- Little relevant
- Very relevant



N-INFINITE

by  livepanel

Appendix 2: Key Training Variables

Social variables

Gender

Age range

SES

City

Country

Questions that serve profiling

How often do you usually renew your cell phone?

What is the main use you give to your cell phone?

Do you take the time to rate the apps you download on Google Play or the App Store?

How important are social networks in your life?

What is your favorite social network?

To capture a moment, do you prefer to take a photo or record a video?

What do you prefer more, an in-person chat every once in a while or virtual conversations (chat, comments) more often?



N-INFINITE

by  livepanel

How concerned are you
about the current situation?

How would you describe your current financial situation?

What means of transportation do you use most frequently for your daily trips?

Etc.

Passive Data

Applications

Facebook
instagram
Telegram
PayPal
Spotify
Netflix
Lemon
Twitter
Snapchat
Pinterest
Free market

Device

Type of device
cell phone company