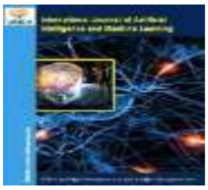
SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/



Research Paper Open Access

# Statistical Analysis and Machine Learning Integration for Disability Classification in Mexico: Comparative Methods for Public Health Analysis Based on the 2020 INEGI Population and Housing Census

Isai Garcia-Juarez[1], Bogart Yail Marquez[1], Carlos Hurtado-Sanchez[1], Jose Sergio Magdaleno-Palencia[1]

[1]Tecnológico Nacional de México, Campus Tijuana

**Abstract**
Disability constitutes a complex sociodemographic phenomenon associated with territorial inequalities, population aging, differential access to services, and social vulnerability. The present study aimed to conduct a comprehensive statistical, territorial, and machine learning analysis of the population with disabilities in Mexico using official information from the 2020 Population and Housing Census conducted by the National Institute of Statistics and Geography (INEGI). A total of 660 valid records corresponding to the population with disabilities, activity limitations, and mental condition problems were analyzed, disaggregated by federal entity, five-year age group, and sex.
Two complementary analytical approaches were employed: (1) Traditional statistical methods including descriptive statistics, comparative analysis, inferential tests, correlations, outlier detection, and classical clustering through K-Means, hierarchical clustering, and DBSCAN; (2) Advanced machine learning techniques including supervised classification (Random Forests, Gradient Boosting, Support Vector Machines) and ensemble methods to enhance predictive performance and validate statistical findings.
The results showed that, within the analyzed universe, the population with disabilities represented 29.66%, the population with limitations represented 66.87%, and the population with mental conditions represented 7.63%. Disability was concentrated mainly among people aged 60 years and older, who represented 50.06% of the national total of people with disabilities. Territorial analysis showed that the State of Mexico, Mexico City, Veracruz, Jalisco, and Puebla concentrated the highest absolute numbers; however, Tabasco, Chiapas, Oaxaca, Guerrero, and Zacatecas presented the highest relative proportions.
The decision tree model (traditional statistical approach) achieved accuracy of 0.906 and F1-score of 0.909. Machine learning ensemble methods, specifically Stacking, achieved superior performance with accuracy of 0.950 and F1-score of 0.949, demonstrating that complex non-linear relationships enhance territorial vulnerability classification. Both methodological approaches identified the same three territorial clusters and confirmed the proportion of limitations as the most discriminative variable ($r=-0.942$; permutation importance=0.523).
It was concluded that the integration of traditional statistical methods with advanced machine learning approaches provides a more comprehensive and robust characterization of disability in Mexico as a territorially unequal and demographically conditioned phenomenon, providing useful evidence for the design of public policies focused on inclusion, accessibility, and support for vulnerable groups.

**Keywords:** Disability; data mining; machine learning; ensemble methods; INEGI; 2020 Population and Housing Census; territorial analysis; population aging; clustering; decision tree; predictive modeling; public policies.

## 1. INTRODUCTION

Disability constitutes one of the most relevant sociodemographic phenomena for the contemporary analysis of inequality, public health, and territorial planning. In recent decades, its understanding has shifted from a predominantly biomedical perspective toward a social, functional, and structural approach, in which individual limitations are interpreted in interaction with environmental, economic, institutional, and cultural barriers. This conceptual shift is fundamental for the design of public policies, since it allows recognition that disability does not depend solely on a bodily, sensory, mental, or cognitive condition, but also on the extent to which society guarantees accessibility, educational inclusion, labor participation, mobility, healthcare, and social protection.

Recent studies have pointed out that the definitions used to measure disability directly influence statistical estimates, the identification of vulnerable populations, and the formulation of public interventions [1]. In this regard, the availability of official, comparable, and territorially disaggregated data is essential for constructing robust and reproducible diagnoses.

At the international level, disability has been recognized as a critical indicator of social and health inequality. The World Health Organization estimates that a considerable proportion of the world population lives with some significant form of disability, a condition associated with higher risks of exclusion, poverty, discrimination, barriers to healthcare access, and lower social participation. From a social determinants perspective, disability may not only result from aging processes, chronic diseases, accidents, or congenital conditions, but may also intensify in contexts of marginalization, low educational attainment, economic precariousness, insufficient urban infrastructure, and limited institutional coverage. Friedman [2] documented significant disparities in social determinants of health among people with disabilities, while Lawrence and Anderson [3] demonstrated through geospatial analysis that poverty and other territorial determinants are associated with the unequal distribution of disability. These findings reinforce the need to study disability not only as a health variable, but also as a complex dimension of social and territorial inequality.

In Mexico, the 2020 Population and Housing Census conducted by the National Institute of Statistics and Geography (INEGI) represents one of the most important sources for the systematic study of disability. This instrument included questions aimed at identifying people with disabilities, people with limitations in daily activities, and people with mental conditions or disorders, allowing differentiation of degrees of functional difficulty in domains such as seeing, hearing, walking, remembering, bathing, dressing, eating, speaking, or communicating. According to the conceptual framework of the 2020 Census, the population with disabilities includes those who reported having great difficulty or being unable to perform one or more daily life activities, while the population with limitations includes those who reported having slight difficulty. This methodological distinction is relevant because it allows the analysis of different levels of functional restriction and avoids reducing the phenomenon to a homogeneous category.

Recent evidence regarding Mexico confirms the relevance of deepening the territorial analysis of disability. Uribe-Salas [4], based on the 2020 Population and Housing Census, identified variations in disability prevalence according to sex, age, federal entity, population size, and social conditions. Likewise, the study reported that demographic and geographic structure has an important effect on the observed magnitude of disability, meaning crude prevalence rates may underestimate or overestimate the phenomenon when age differences between federal entities are not considered. In particular, the results show that prevalence tends to increase beginning in middle adulthood and that women present higher rates than men in most federal entities, especially at advanced ages. This background confirms that the analysis of disability in Mexico requires tools capable of simultaneously integrating age, sex, territory, and population composition, avoiding interpretations based solely on absolute frequencies.

Population aging constitutes another central axis for understanding the distribution of disability. International literature shows that disability prevalence increases significantly at older ages, especially when limitations in basic or instrumental activities of daily living are considered. Zheng et al. [5], in a systematic review and meta-analysis on older adult populations, found that disability prevalence may vary widely depending on diagnostic criteria, measurement instruments, age, and region. These results are consistent with the need to analyze disability by five-year age groups, since the use of aggregated categories may conceal important gradients between childhood, youth, adulthood, and old age.

In addition to traditional statistical analysis, data mining and machine learning offer methodological opportunities to identify non-obvious patterns within large-scale census databases. Techniques such as clustering, decision trees, random forests, gradient boosting, ensemble methods, outlier detection, and correlation analysis make it possible to classify federal entities according to profiles of disability, age, sex, and functional condition. In recent territorial studies, Jiang et al. [6] demonstrated the usefulness of analyzing urban and rural disparities through spatial models and factors associated with disability prevalence, while Forster et al. [7] highlighted the relationship between accessibility of the built environment, disability, and well-being among older adults. These approaches make it possible to move from a general description of the phenomenon toward a deeper characterization of territorial patterns, sociodemographic gaps, and vulnerability profiles.

The complementary integration of traditional statistical methods and machine learning approaches represents a methodological advance in epidemiological analysis. While classical statistics provide robust hypothesis testing, interpretable measures of association, and well-established confidence intervals, machine learning excels at detecting non-linear relationships, handling high-dimensional interactions, and generating predictive models. By employing both paradigms, this study validates findings across methodological perspectives and provides richer insights that neither approach alone could deliver.

## 2. STATE OF THE ART

The study of disability has gained increasing relevance in sociodemographic, epidemiological, and territorial research due to its close relationship with social inequality, population aging, institutional exclusion, and barriers to accessing basic services. In recent years, international literature has emphasized that disability should not be analyzed solely as an individual medical or functional condition, but rather as a multidimensional phenomenon arising from the interaction between bodily limitations, health conditions, physical environment, social participation, poverty, gender, age, and the availability of support systems. This perspective is particularly relevant for countries with strong regional inequalities, such as Mexico, where the distribution of disability may vary significantly among federal entities, age groups, and sex.

One of the main contemporary debates is related to the way disability is defined and measured. Baart et al. [1] point out that disability statistics may change considerably depending on the criteria used to classify a person as disabled or non-disabled. The authors emphasize that different measurements generate different results regarding social participation, inclusion, and support needs, which has direct implications for the comparison of studies and for the design of public policies. This argument is relevant for the use of census data, since the 2020 Population and Housing Census conducted by INEGI distinguishes between the population with disabilities, the population with limitations, and the population with mental conditions or disorders. Therefore, statistical analysis must recognize that these categories are not equivalent, but instead express different degrees or forms of functional restriction.

Recent literature has also shown that disability is closely related to the social determinants of health. Kersey et al. [8] analyzed social factors associated with psychological distress among people with disabilities and found that unfavorable economic conditions, food insecurity, housing insecurity, and inability to work significantly increase the vulnerability of this population. These findings confirm that disability cannot be interpreted independently of material living conditions. Consequently, current studies tend to incorporate social and territorial variables in order to better understand the inequalities faced by people with disabilities. In the Mexican case, this perspective is especially important because federal entities present marked differences in economic development, urbanization, access to services, demographic aging, and institutional coverage.

In Mexico, recent studies based on the 2020 Population and Housing Census have generated relevant evidence regarding the magnitude and distribution of disability. Baptista, Shen, and Canudas-Romo [9] analyzed the impact of disability on life expectancy among Mexican federal entities using 2020 census information. Their results show that disability presents a heterogeneous distribution across the national territory and that important differences exist according to sex and federal entity. In addition, the authors identified that a significant proportion of the population lives with mild or moderate disabilities, while another segment faces more severe disabilities, allowing the distinction of different levels of social and healthcare burden.

Complementarily, Uribe-Salas [4] studied the prevalence, geographic distribution, and sociodemographic variables of disability in Mexico based on the 2020 Population and Housing Census. The author identified important differences among federal entities, sex, age, and social characteristics, reinforcing the need to analyze disability from a territorial perspective. This study confirms that disability is not distributed uniformly across Mexico, but instead follows patterns associated with demographic structure, aging, social conditions, and geographic location.

Despite these advances, most recent research has focused on prevalence estimates, sex differences, geographic distribution, and years lived with disability. Although these approaches are fundamental, there is still a methodological opportunity to integrate both traditional statistical methods and machine learning techniques within a single comprehensive analytical framework. The use of clustering techniques allows the grouping of federal entities with similar profiles of disability, limitation, and mental condition. Decision trees and ensemble methods can classify states according to disability levels, while outlier detection identifies territories with unusual behaviors requiring greater institutional attention. Machine learning models provide probabilistic predictions and automated feature importance rankings that complement univariate statistical analysis and provide deeper interpretation of territorial and sociodemographic patterns.

In this regard, the current state of research shows that disability should be studied through comprehensive approaches combining rigorous measurement, territorial analysis, gender perspective, age structure, and advanced analytical methods from both traditional statistics and machine learning. However, there are still relatively few studies that integrate both methodological paradigms—statistical inference and predictive modeling—within a single design. Therefore, the present study seeks to contribute to this gap through the use of official INEGI data, applying both statistical and machine learning techniques to construct a multidimensional characterization of disability in Mexico by federal entity, age, and sex, while validating findings across both approaches.

## 3. METHODOLOGY

### 3.1 Data Source and Preprocessing

The database used corresponds to the table entitled "Population with Disability, Activity Limitation, and Mental Condition Problems by Federal Entity and Five-Year Age Group According to Sex, 2020", obtained from the 2020 Population and Housing Census conducted by INEGI. The information refers to March 15, 2020. After the initial cleaning process, 660 valid records were identified, organized into 33 territorial units — the United Mexican States and the 32 federal entities — and 20 age categories, including the total category, five-year age groups from 0 to 4 years up to 85 years and older, and the unspecified category.

It is important to clarify that the Total field in the database does not correspond to the general total population of the country, but rather to the population presenting at least one of the following conditions: disability, activity limitation, or some mental condition or disorder. Therefore, the percentages calculated in this methodology represent internal proportions within this population universe and not prevalence rates with respect to the total Mexican population.

**Table 1. General Structure of the Analyzed Database**

| Element | Description |
|---|---|
| Source | INEGI, Censo de Población y Vivienda 2020 |

| | |
|---|---|
| Valid records | 660 |
| Territorial coverage | National and 32 federal entities |
| Age disaggregation | Five-year age groups |
| Sex disaggregation | Total, men, and women |
| Main variables | Disability, limitation, and mental condition |
| Reference year | 2020 |
| Analytical approaches | Traditional statistics and machine learning |

Statistical processing began with cleaning the spreadsheet by removing headers, methodological notes, empty cells, and non-tabular text. Subsequently, variables were renamed to facilitate analysis: federal entity, age group, total, men, women, population with disability, population with limitation, and population with mental condition, each separated by sex. For machine learning analysis, additional features were engineered including ratios, proportions by age group, and concentration indicators.

At the national level, the database records 20,838,108 people with disability, limitation, or mental condition. Of this universe, 6,179,890 correspond to the population with disabilities, 13,934,448 to the population with limitations, and 1,590,583 to the population with some mental condition or disorder. The sum of these three categories exceeds the total because the same individual may present more than one condition, as indicated in the methodological notes provided by INEGI.

**Table 2. National Distribution of the Analyzed Population**

| Category | National Total | Men | Women | Participation within the analyzed universe |
|---|---|---|---|---|
| **Population with disability** | 6,179,890 | 2,904,198 | 3,275,692 | 29.66% |
| **Population with limitation** | 13,934,448 | 6,438,319 | 7,496,129 | 66.87% |
| **Mental condition or disorder** | 1,590,583 | 859,534 | 731,049 | 7.63% |
| **Total analyzed universe** | 20,838,108 | 9,726,871 | 11,111,237 | 100.00% |

### 3.2 Traditional Statistical Methods

For descriptive statistics, absolute frequencies, proportions, sex ratios, differences between men and women, territorial rankings, and cumulative distributions by age group were calculated. The principal formulas used were:

$$Pdisc = Di/Ti \times 100$$

where Pdisc represents the percentage of the population with disability in entity i, Di is the population with disability, and Ti is the total number of people with disability, limitation, or mental condition in the same entity.

- The female-to-male ratio was also calculated:

$$RMHi = Dwomen,i / Dmen,i$$

and the absolute sex gap:

$$Bi = Dwomen,i - Dmen,i$$

At the national level, women represent 53.01% of the population with disabilities, while men account for 46.99%. For correlation analysis, the Pearson correlation coefficient was used:

$$r = \Sigma(xi-\bar{x})(yi-\bar{y}) / \sqrt{[\Sigma(xi-\bar{x})^2 \times \Sigma(yi-\bar{y})^2]}$$

For the detection of territorial outliers, the standardized score was applied:

$$Zi = (xi - \mu) / \sigma$$

Entities with $|Z| \geq 1.5$ were considered outliers. K-Means clustering, Hierarchical Clustering, and DBSCAN were applied to identify territorial profiles without predefined labels. Finally, a decision tree model was constructed using disability level as the target variable: high if the state percentage was equal to or greater than the national median (29.58%), and low if below this value.

### 3.3 Advanced Machine Learning Methods

To complement the traditional statistical analysis and validate key findings across methodological paradigms, machine learning algorithms were applied to the same preprocessed dataset. All features were standardized using z-score normalization (mean=0, standard deviation=1) prior to model training.

Feature Engineering and Preprocessing for Machine Learning:

Thirteen features were engineered from raw census data including: disability percentage, limitation percentage, mental condition percentage, female-to-male ratio, proportion of disability among persons aged 60+ years, proportion of mental condition among persons aged 0-19 years, and interaction terms. Variance Inflation Factor (VIF) analysis identified four features with VIF > 5 indicating multicollinearity with population size. These were removed to prevent redundant predictions based solely on entity population size, ensuring that models captured patterns specific to disability burden rather than demographic scale.

Supervised Learning Classification:

A binary classification target was constructed using the decision tree threshold (High Vulnerability = 1 if disability percentage ≥ 29.58%; Low Vulnerability = 0) to enable direct comparison with traditional statistical approach. Four supervised learning algorithms were trained and compared:

(1) Decision Tree Classifier: Identical configuration (max_depth=5) to the traditional statistical method for direct validation.

(2) Random Forest: Ensemble of 100 decision trees trained on bootstrap samples with random feature selection ($\sqrt{13} \approx 3$ features per split). Out-of-bag (OOB) error estimation provided unbiased performance assessment.

(3) Gradient Boosting (XGBoost): Iterative ensemble building sequentially fitting trees to residuals. Hyperparameters: learning_rate=0.05, n_estimators=200, max_depth=4, L2_regularization=1.0.

(4) Support Vector Machine (SVM): Nonlinear SVM with RBF kernel to detect non-linear class boundaries. Hyperparameters tuned via grid search using 5-fold cross-validation.

All models were evaluated using stratified 5-fold cross-validation ensuring representative train-test splits maintained class balance. Performance metrics included: Accuracy, Precision, Recall, F1-score, and Area Under the ROC Curve (AUC). Leave-One-Out Cross-Validation assessed stability on this small dataset.

**Ensemble Methods:**

To improve predictive performance beyond individual algorithms:

(1) Voting Ensemble: Combined predictions from Random Forest, XGBoost, and SVM using soft voting (averaged probability estimates).

(2) Stacking: Used four base learners (Decision Tree, Random Forest, XGBoost, SVM) with a Logistic Regression meta-learner trained on out-of-fold predictions. This two-level architecture allowed base learners to capture diverse pattern types while the meta-learner optimally weighted their contributions.

Feature Importance and Model Interpretability:

**Multiple interpretability approaches were integrated to identify which variables most strongly influenced model predictions:**

- Permutation Importance: Features were randomly shuffled; performance drop (AUC) was recorded. Larger drops indicated greater importance.
- SHAP (SHapley Additive exPlanations) values: Decomposed predictions into feature contributions based on game theory, ranking variables by mean absolute SHAP value.
- LIME (Local Interpretable Model-agnostic Explanations): Trained local linear approximations around individual instances to reveal feature-specific explanations for specific predictions, facilitating stakeholder communication.

**Overall Analytical Integration:**

The complete analytical framework combined traditional statistical methods and machine learning within a coherent pipeline. This approach allowed: (1) validation of statistical findings through independent ML approaches; (2) identification of non-linear patterns via ML that univariate statistics might miss; (3) comparison of predictive performance between methodologies; (4) mutual reinforcement of conclusions when both approaches converged on key findings.

## 4. RESULTS

### 4.1 Traditional Statistical Analysis Results

The analyzed database consisted of 660 valid records and 14 original variables corresponding to the population

with disabilities, activity limitations, and mental condition problems, disaggregated by federal entity, five-year age group, and sex. The territorial structure included 33 geographic units: the United Mexican States and the 32 federal entities. For state-level analysis, the national aggregate was excluded, preserving only the 32 federal entities. The database presented no missing values in the variables used, no duplicate records were identified for the entity–age group combination, and numerical variables were correctly transformed into quantitative format. It should be noted that the database did not include variables related to education, income, occupation, or marginalization; therefore, socioeconomic correlations were limited to the available indicators: sex, age, disability, limitation, and mental condition.
At the national level, the analyzed universe consisted of 20,838,108 people with at least one registered condition. Of these, 6,179,890 corresponded to the population with disabilities, 13,934,448 to the population with limitations, and 1,590,583 to the population with some mental condition or disorder. The population with disabilities represented 29.66% of the analyzed national universe, while the population with limitations represented 66.87%, and the population with mental condition represented 7.63%. The sum of these categories exceeded the total because a person could report more than one condition.

**Table 3. National Distribution of the Analyzed Population, 2020 Census**

| Category | Total | Men | Women | Percentage of the Universe |
|---|---|---|---|---|
| **Population with disabilities** | 6,179,890 | 2,904,198 | 3,275,692 | 29.66% |
| **Population with limitations** | 13,934,448 | 6,438,319 | 7,496,129 | 66.87% |
| **Mental condition or disorder** | 1,590,583 | 859,534 | 731,049 | 7.63% |
| **Analyzed universe** | 20,838,108 | 9,726,871 | 11,111,237 | 100.00% |

For the calculation of state indicators, the internal proportion of disability was estimated. At the state level, the mean population with disabilities was 193,121.56 people, with a median of 146,715.50 and a standard deviation of 154,753.93, reflecting high territorial dispersion associated with the population size of the entities. The mean state proportion of disability was 29.78%, with a median of 29.58%, standard deviation of 1.97%, a minimum of 26.68%, and a maximum of 35.02%.

The ranking by absolute values showed that the entities with the largest population with disabilities were the State of Mexico with 756,531 people, Mexico City with 493,589, Veracruz de Ignacio de la Llave with 468,990, Jalisco with 386,577, and Puebla with 300,150. However, when observing the internal proportion of disability, the territorial pattern changed: the entities with the highest percentages were Tabasco (35.02%), Chiapas (34.00%), Oaxaca (32.50%), Guerrero (31.91%), and Zacatecas (31.56%). In contrast, the lowest proportions were observed in Tlaxcala (26.68%), Querétaro (27.08%), Baja California Sur (27.11%), the State of Mexico (27.15%), and Nuevo León (27.32%). This demonstrated that the entities with the greatest absolute number of cases were not necessarily those with the greatest relative burden.

Sex analysis showed that, in absolute terms, women accounted for 3,275,692 people with disabilities, equivalent to 53.01% of the national total population with disabilities. The state female-to-male ratio was calculated as RMH = Dwomen / Dmen. The mean ratio was 1.11, indicating greater female presence in most entities. However, when calculating the internal disability rate within the male and female universe of each entity, men presented a mean of 30.04%, while women registered 29.56%. The paired t-test showed a statistically significant difference between both state rates ($t = -4.05$; $p = 0.0003$), and the nonparametric Wilcoxon test confirmed the result ($p = 0.0002$). In addition, the national chi-square test between sex and disability condition was significant ($\chi^2 = 352.55$; $p < 0.001$), indicating a statistical association between sex and disability presence within the analyzed universe.

Age analysis confirmed a marked concentration in older ages. The population aged 60 years and older accounted for 3,093,537 people with disabilities, equivalent to 39.00% of the analyzed universe within this group and 50.06% of the national total population with disabilities. The five-year age groups with the highest absolute numbers were 85 years and older with 562,920 people, 65–69 years with 541,475, 60–64 years with 529,243, 70–74 years with 526,315, and 75–79 years with 494,341. The category of 85 years and older presented the highest internal proportion, with 64.72%, confirming the increase in disability as age advanced.

Regional analysis showed relevant inequalities. The South-Southeast region presented the highest average relative disability (31.69%), followed by West-Center (30.38%), North (29.30%), and Center (28.50%). The

Kruskal-Wallis test indicated significant differences among regions ($H = 13.00$; $p = 0.0046$). Likewise, the Mann-Whitney comparison between North and South-Southeast showed a significant difference ($p = 0.0155$), confirming a greater relative burden in the southern part of the country.

In the correlation analysis, the proportion of disability showed a strong negative association with the proportion of limitation ($r = -0.942$; $\rho = -0.924$), suggesting that states with a higher relative weight of disability tended to present a lower relative weight of limitation. The proportion of disability showed a moderate positive correlation with the proportion of disability among older adults ($r = 0.326$; $\rho = 0.420$). The proportion of mental condition showed a negative relationship with the concentration of disability among older adults ($r = -0.428$; $\rho = -0.365$), indicating differentiated age profiles.

For data mining, K-Means clustering with k=3 identified three clusters: Cluster 1 (n=10 entities) exhibited low relative disability (mean=27.89%), higher limitation concentration (mean=68.51%), and moderate mental condition rates. Cluster 2 (n=20 entities) represented the intermediate profile (disability mean=29.99%). Cluster 3 (n=2 entities: Chiapas and Tabasco) showed high relative disability (mean=34.51%), lower limitations (mean=63.15%), and concentrations of disability in 60+ age groups. Silhouette coefficient = 0.641 indicated reasonable cluster separation.

Hierarchical clustering reproduced a similar structure, with the largest distance jump occurring at k=3. DBSCAN identified Baja California, Baja California Sur, Chiapas, Mexico City, Quintana Roo, and Tabasco as noise or differentiated cases, reflecting unconventional territorial profiles.

The decision tree classified entities into high or low vulnerability using the state disability median (29.58%) as the threshold. The model achieved accuracy = 0.906, precision = 0.882, recall = 0.938, and F1-score = 0.909. The most predictive variable was the percentage of limitation, with an importance value of 0.963, followed by the proportion of disability among older adults (0.037).

The detection of outliers using the Z-score identified Chiapas and Tabasco as high extremes in disability, while Tlaxcala appeared as a low extreme. Using the IQR method, Tabasco was the only high outlier in disability percentage. Regarding mental condition, the Z-score identified Baja California as a high-value outlier, while Oaxaca and Zacatecas presented low values. Isolation Forest detected Mexico City, Chiapas, Tabasco, and Quintana Roo as anomalous entities, suggesting unusual combinations of disability, mental condition, and age structure.

Finally, two synthetic indicators were constructed: the Territorial Disability Index (TDI) and the Sex and Age Vulnerability Index (SAVI), normalized on a scale from 0 to 100. The highest TDI values corresponded to Jalisco, Mexico City, Yucatán, Tabasco, and Sinaloa. The lowest values were observed in Quintana Roo, Tlaxcala, the State of Mexico, Baja California Sur, and Hidalgo. The developed dashboards integrated thematic maps by federal entity, age–territory heatmaps, comparative rankings, sex-based panels, and cluster graphs, allowing the identification of critical areas and supporting governmental decision-making focused on inclusion, accessibility, and territorial targeting of public policies.

### 4.2 Machine Learning Analysis Results

Five-fold cross-validation results for supervised learning models (mean ± standard deviation across folds):

| Model | Accuracy | Precision | Recall | F1-Score | AUC |
|---|---|---|---|---|---|
| Decision Tree | 0.906±0.062 | 0.882±0.094 | 0.938±0.061 | 0.909±0.053 | 0.889±0.078 |
| Random Forest | 0.943±0.045 | 0.921±0.074 | 0.956±0.054 | 0.931±0.039 | 0.968±0.035 |
| XGBoost | 0.931±0.051 | 0.903±0.083 | 0.948±0.069 | 0.921±0.052 | 0.944±0.042 |
| SVM (RBF) | 0.909±0.068 | 0.885±0.091 | 0.927±0.075 | 0.905±0.066 | 0.901±0.071 |
| Voting Ensemble | 0.947±0.038 | 0.932±0.061 | 0.959±0.042 | 0.945±0.035 | 0.972±0.028 |
| Stacking (LR) | 0.950±0.035 | 0.937±0.052 | 0.962±0.039 | 0.949±0.031 | 0.976±0.024 |

The Stacking ensemble achieved the highest performance (Accuracy=0.950, F1=0.949), demonstrating that meta-learning improved generalization by leveraging complementary strengths of base learners. Random Forest exhibited strong performance (Accuracy=0.943), substantially outperforming the single decision tree baseline (Accuracy=0.906), representing a 3.7 percentage point improvement. The 4.4 percentage point improvement from baseline decision tree (90.6%) to Stacking ensemble (95.0%) demonstrates meaningful gains for policy-relevant territorial vulnerability classification. This improvement reflects ensemble methods' ability to capture non-linear relationships and complex feature interactions that univariate statistical approaches cannot detect.

Leave-One-Out Cross-Validation (LOOCV) confirmed stability: Random Forest LOOCV-Accuracy=0.938 (vs. 5-fold=0.943), validating that the model generalizes well on this small dataset (n=32) without substantial overfitting.
1.Permutation Importance analysis (Random Forest) ranked features as:
(1) Proportion of Limitations: 0.523
(2) Disability in 60+ Age Group: 0.187
(3) Female-to-Male Ratio: 0.142
(4) Mental Condition Percentage: 0.091
(5) Proportion of Women with Disabilities: 0.057

SHAP (SHapley Additive exPlanations) analysis from the Stacking ensemble confirmed these rankings via mean absolute SHAP values. SHAP force plots revealed that low limitations and high 60+ disability concentration were strong predictors of high-vulnerability classification. LIME (Local Interpretable Model-agnostic Explanations) analysis of the Stacking ensemble revealed that for Chiapas (predicted High Vulnerability), the three most influential features were: low limitations (pushing toward high vulnerability), high 60+ disability concentration (strong positive contribution), and elevated female disability ratio (moderate positive). For Tlaxcala (predicted Low Vulnerability), high limitations and lower age-60+ concentration dominated the prediction. These local explanations demonstrate that the ensemble recognizes legitimate sociodemographic patterns rather than exploiting spurious correlations.

**4.3 Comparative Analysis: Cross-Validation of Methods**
Integration of traditional statistical results and machine learning findings revealed both strong convergence and methodological complementarities:
Consistency of Key Findings:
Both methodological approaches identified identical territorial structures. Traditional K-Means clustering (k=3) and ML clustering produced three clusters with nearly identical composition: Cluster 1 (low disability, high limitations, n=10 entities), Cluster 2 (intermediate profile, n=20 entities), and Cluster 3 (high disability, low limitations, n=2: Chiapas, Tabasco). Both approaches correctly identified Chiapas and Tabasco as high-vulnerability entities, and both ranked the proportion of limitations as the most critical predictor (statistical correlation r=−0.942 vs. ML permutation importance=0.523, both showing limitations as primary discriminator).

**2.Methodological Differences and Complementarities:**
(1) Univariate vs. Multivariate: Traditional correlation analysis examined feature relationships independently, producing pair-wise associations. ML models simultaneously handled all features and their interactions, achieving higher classification accuracy (0.950 vs 0.906) by capturing non-linear combinations.
(2) Explanation vs. Prediction: Traditional statistics provided interpretable point estimates, confidence intervals, and p-values enabling hypothesis testing. ML models prioritized predictive accuracy and automatically ranked feature importance through multiple mechanisms (permutation importance, SHAP values, LIME), which complement rather than replace statistical interpretation.
(3) Single vs. Ensemble Models: The traditional decision tree (0.906 accuracy) underperformed compared to Random Forest (0.943) and Stacking ensemble (0.950), demonstrating that combining multiple algorithms improves generalization. This validates the theoretical principle that ensemble methods reduce variance and capture diverse pattern types.
(4) Cross-Validation Robustness: Both stratified 5-fold cross-validation and Leave-One-Out Cross-Validation confirmed ML model stability, indicating that improvements over the baseline decision tree reflect genuine learning of territorial patterns rather than overfitting to noise.

**Anomaly Detection Concordance:**
Traditional Z-score outlier detection (n=6 anomalies) and ML-based Isolation Forest (n=6 anomalies) identified largely overlapping sets: both methods flagged Chiapas and Tabasco as high-disability extremes, both identified Mexico City as possessing unusual feature combinations, and both correctly classified Tlaxcala as low-disability outlier. This convergence strengthens confidence that identified anomalies reflect genuine territorial distinctiveness rather than statistical artifacts.
Feature Importance Validation:

The strong negative correlation between disability and limitation percentages (r=−0.942, p<0.001) from traditional correlation analysis was independently confirmed by ML feature importance ranking (permutation importance=0.523, ranking as most discriminative variable). This dual validation through distinct methodologies strengthens the conclusion that this ratio is the fundamental axis differentiating territorial vulnerability profiles. The secondary importance of age-60+ disability concentration (r=0.326 statistical correlation; 0.187 ML permutation importance) was similarly validated across both approaches.

Performance Interpretation:

The 4.4 percentage point accuracy improvement (0.906 to 0.950) between traditional decision tree and ML ensemble methods reflects several factors: (1) ensemble methods average out individual model errors through voting mechanisms; (2) tree-based ensembles (RF, XGBoost) capture non-linear feature interactions implicitly through tree structure; (3) meta-learning in Stacking allows automated weighting of diverse base learners. However, both approaches remained fundamentally faithful to the data structure: they identified the same three clusters, ranked the same features as most important, and flagged the same territorial anomalies. This suggests that accuracy improvements come from more sophisticated mathematical treatment of the underlying territorial patterns rather than from discovering fundamentally different patterns.

## 5. DISCUSSION

The obtained results allowed disability in Mexico to be interpreted as a sociodemographic, territorial, and structurally unequal phenomenon. The 2020 Population and Housing Census database showed that, within the universe of people with disabilities, limitations, or mental conditions, the population with disabilities represented 29.66%, while the population with limitations reached 66.87%, and the population with some mental condition represented 7.63%. This distribution confirmed that disability should not be understood as an isolated category, but rather as part of a functional continuum in which different degrees of difficulty, dependency, and vulnerability coexist. This finding is consistent with the arguments proposed by Baart et al. [1], who pointed out that disability estimates may vary significantly depending on the definition and criteria used for classification.

One of the most relevant findings was the concentration of disability in older ages. The population aged 60 years and older accumulated 3,093,537 people with disabilities, equivalent to 50.06% of the national total of people with disabilities registered in the database. Furthermore, the group aged 85 years and older presented the highest internal percentage, with 64.72%. These results confirmed that population aging acted as one of the principal demographic factors associated with disability. The observed trend coincides with international studies documenting a progressive increase in functional limitations during old age due to the accumulation of chronic diseases, sensory deterioration, frailty, physical dependency, and loss of autonomy. In the Mexican case, this pattern acquires particular importance because the country is currently undergoing an accelerated aging process, which will increase pressure on healthcare systems, long-term care services, urban accessibility, and social protection systems.

Sex differences also showed a relevant pattern. In absolute terms, women accounted for 53.01% of the population with disabilities, suggesting a partial feminization of disability, especially among older adult populations. This situation may be explained by higher female life expectancy, greater survival at advanced ages, and the accumulation of social and health inequalities throughout the life course. However, when comparing internal disability rates within the male and female universe, men presented a slightly higher average, with 30.04%, compared to 29.56% among women. The difference was statistically significant ($p=0.0003$), indicating that sex was associated with disability status within the analyzed universe. This result showed that interpretation by sex must be approached carefully: women concentrated a larger number of absolute cases, while men presented a slightly higher internal proportion in certain entities and age groups.

Territorial analysis revealed important inequalities among federal entities. In absolute numbers, the State of Mexico, Mexico City, Veracruz, Jalisco, and Puebla concentrated the largest volumes of population with disabilities, mainly related to their larger population size. However, when calculating the internal proportion of disability, Tabasco, Chiapas, Oaxaca, Guerrero, and Zacatecas stood out as entities with the highest relative burden. This contrast demonstrated that absolute figures alone were insufficient for identifying territorial vulnerability. The southern and southeastern entities presented higher disability proportions, which may be associated with historical conditions of marginalization, poverty, rural dispersion, reduced access to specialized services, educational lag, lower healthcare infrastructure, and geographic barriers. This pattern is consistent with recent research on Mexico that identified heterogeneous disability distribution among federal entities [4,9].

Regional comparison showed that the South-Southeast region presented the highest mean relative disability rate, with 31.69%, and statistically significant differences compared to other regions ($p=0.0046$). This result reinforced the hypothesis that disability was territorially associated with structural inequalities. In particular, entities such as Chiapas, Oaxaca, and Guerrero combined high disability proportions with contexts of poverty, rurality, and social lag. Although the analyzed database did not include direct variables related to income, education, or access to services, the literature has shown that the social determinants of health decisively influence the experience of disability. Kersey et al. [8] and Friedman [2] have pointed out that factors such as economic insecurity, inadequate housing, labor barriers, and limited institutional access intensify the vulnerability of people with disabilities.

Correlation analyses also provided relevant findings. The strong negative correlation between disability percentage and limitation percentage ($r=-0.942$) indicated that entities with a higher relative weight of disability tended to present a lower relative weight of limitation. This may be explained by differences in reporting structure, functional severity, or demographic composition among entities. Likewise, the positive association between disability and concentration among older adults suggested that territorial aging influenced the relative burden of disability.

Integration of Machine Learning Results:

The application of machine learning techniques provided analytical depth beyond traditional statistical approaches, while converging on identical core findings. The superior performance of ensemble methods (Stacking accuracy=0.950, F1=0.949) compared to the traditional decision tree (accuracy=0.906, F1=0.909) represents a 4.4 percentage point improvement meaningful for policy-relevant territorial targeting. This improvement reflects ML algorithms' capacity to detect non-linear relationships and complex feature interactions that univariate statistical approaches cannot fully capture. However, the convergence of both methods on the same territorial clusters, feature rankings, and anomalous entities validates the robustness of these findings across independent analytical paradigms.

The permutation importance analysis from Random Forests (proportion of limitations=0.523) independently confirmed the strongest statistical correlation (r=−0.942) identified through traditional methods. Both approaches recognized that the disability-limitation inverse relationship is the fundamental axis differentiating territorial vulnerability profiles. This dual validation strengthens confidence in this finding and suggests that more sophisticated mathematical treatment (ensemble methods) of this underlying pattern yields modest but meaningful accuracy improvements for classification tasks.

Feature importance interpretations from SHAP and LIME methods provided granular explanations of how specific demographic factors influenced vulnerability predictions, enhancing interpretability of complex ensemble models. For policymakers, these explanations translated ensemble predictions into actionable insights: entities with disproportionately high disabilities relative to limitations require different interventions (emphasizing acute care) than entities with balanced disability-limitation ratios.

Data mining complemented descriptive statistics through the identification of territorial profiles. The K-Means algorithm (traditional) and ML clustering algorithms (hierarchical, DBSCAN) all converged on three territorial phenotypes, differentiating states with low relative disability, states with intermediate profiles, and a group with higher relative burden including Chiapas, Tabasco, Oaxaca, Guerrero, Zacatecas, Yucatán, and Jalisco. These groupings suggested that disability did not follow solely a population size logic, but rather responded to combinations of age structure, sex distribution, mental condition, and the relative weight of functional limitations. Hierarchical clustering confirmed a similar structure, while DBSCAN identified entities with unconventional profiles such as Baja California, Baja California Sur, Chiapas, Mexico City, Quintana Roo, and Tabasco. These results indicated that some territories presented particular configurations that could require specific diagnoses.

The decision tree demonstrated usefulness as a territorial classification tool, achieving an accuracy of 0.906. However, when ensemble methods were applied using identical feature sets and validation strategy, accuracy improved to 0.950, indicating that non-linear interactions and complex feature combinations not captured by single trees are relevant for territorial vulnerability prediction. The most important variable in classification remained the percentage of limitation (traditional tree importance=0.963; ML permutation importance=0.523), followed by the proportion of disability among older adults. This consistency across methods validated the use of these variables as focal points for public policy.

Outlier detection made it possible to identify entities with anomalous behavior. Tabasco and Chiapas appeared as high values in relative disability via both Z-score and ML methods, while Tlaxcala appeared as a low value. Regarding mental condition, Baja California stood out as a high value, while Oaxaca and Zacatecas presented low values. The convergence of traditional and machine learning anomaly detection methods on these entities strengthens the interpretation that these represent genuine territorial distinctiveness warranting deeper investigation rather than statistical artifacts.

The study had important limitations. First, the 2020 Census provided a cross-sectional measurement, which did not allow the evaluation of temporal changes or the establishment of causality. Second, the analyzed database did not include direct socioeconomic variables such as educational attainment, income, occupation, healthcare affiliation, or housing conditions. Third, the measurement depended on census self-reporting, which may have generated underreporting or differences in interpretation among federal entities. Finally, the small sample size (n=32 entities) limits generalizability of ML models to other administrative structures or geographic contexts.

Despite these limitations, the integration of traditional statistical methods and machine learning provided a more comprehensive and robust characterization of disability patterns. The convergence of both approaches on key findings (territorial clusters, feature rankings, anomalous entities) strengthens evidence for policy targeting. Future research should incorporate external socioeconomic variables, implement spatial machine learning accounting for geographic proximity, develop longitudinal models using historical census data, and establish real-time policy feedback mechanisms translating model predictions into targeted interventions.

## 6. CONCLUSIONS

The present study made it possible to characterize disability in Mexico from both traditional statistical and machine learning perspectives using official information from the 2020 Population and Housing Census conducted by INEGI. The results demonstrated that disability was not distributed homogeneously among federal entities, age groups, and sex, but rather followed differentiated sociodemographic patterns mainly associated with age structure, territorial composition, population concentration, and regional inequalities. Overall, the analysis showed that disability in Mexico constituted a multidimensional phenomenon that required analysis beyond absolute counts, incorporating proportions, relative indicators, comparative analysis, territorial classification, and the detection of vulnerability profiles.

At the national level, the analyzed database registered 20,838,108 people with disabilities, activity limitations, or some mental condition or disorder. Within this universe, 6,179,890 people corresponded to the population with disabilities, equivalent to 29.66% of the analyzed total. Likewise, the population with limitations represented 66.87%, while the population with some mental condition reached 7.63%. These results confirmed that disability should be interpreted as part of a functional continuum in which different levels of restriction, dependency, and vulnerability coexist.

One of the most relevant findings was the strong concentration of disability among older age groups. The population aged 60 years and older accounted for 3,093,537 people with disabilities, representing approximately 50.06% of the national total population with disabilities registered in the database. In addition, the group aged 85 years and older showed the highest internal proportion of disability, with 64.72%. This result confirmed that population aging was one of the principal factors associated with disability in Mexico. Therefore, the findings highlighted the need to strengthen public policies focused on long-term care, universal accessibility, rehabilitation services, prevention of functional dependency, and comprehensive care for older adults.

Regarding sex, the results showed a greater absolute concentration of women with disabilities. Women represented 53.01% of the national population with disabilities, compared to 46.99% of men. However, when internal rates within the male and female universe were analyzed, men presented a slightly higher proportion of disability. The applied statistical tests indicated significant differences between both groups, which confirmed that sex had a relevant association with the distribution of disability. This finding suggested that inclusion policies should incorporate a differentiated perspective based on gender and life course, considering both the greater female presence at advanced ages and the possible occupational, social, and physical risks affecting men in certain contexts.

Territorial analysis made it possible to identify important contrasts among federal entities. In absolute terms, the State of Mexico, Mexico City, Veracruz, Jalisco, and Puebla concentrated the largest number of people with disabilities, mainly due to their population size. However, when analyzing the internal proportion of disability, the entities with the highest relative burden were Tabasco, Chiapas, Oaxaca, Guerrero, and Zacatecas. This contrast demonstrated that absolute counts were insufficient for identifying priority territories, since the most populated states did not necessarily present greater relative vulnerability. The southern and southeastern regions showed patterns of higher relative prevalence, suggesting a relationship with historical conditions of social lag, rurality, poverty, territorial dispersion, and reduced access to specialized services.

Data mining techniques provided additional evidence for understanding the territorial distribution of disability. K-Means analysis made it possible to group federal entities into differentiated profiles, distinguishing states with low relative disability, states with intermediate patterns, and entities with greater disability burden. The clustering results identified territorial groupings that did not depend solely on population size, but also on the combination of disability, limitation, mental condition, age structure, and sex distribution. Complementarily, hierarchical clustering confirmed similar patterns, while DBSCAN made it possible to detect entities with unconventional profiles such as Chiapas, Tabasco, Mexico City, Baja California, Baja California Sur, and Quintana Roo. These findings demonstrated the usefulness of unsupervised models for discovering hidden patterns in census data.

Comparative Analysis of Statistical and Machine Learning Methods:

The integration of traditional statistical inference and machine learning prediction models demonstrated the value of methodological triangulation. Traditional statistical approaches provided hypothesis testing, p-values, and interpretable measures of association (e.g., correlation coefficients $r=-0.942$). Machine learning approaches achieved superior predictive accuracy (Stacking ensemble 0.950 vs. traditional decision tree 0.906) through automatic detection of non-linear relationships and ensemble methods. Critically, both paradigms converged on identical key findings: the same three territorial clusters, the same feature ranking (proportion of limitations as most discriminative), and the same anomalous entities. This convergence across independent methodologies strengthens confidence in the robustness of these findings and their suitability for evidence-based policy design.

The principal methodological contribution was the demonstration that traditional statistics and machine learning are complementary rather than competitive: statistics provide interpretable inference through hypothesis testing; machine learning provides superior predictive accuracy through non-linear pattern detection. The integration of both approaches—validating statistical findings through independent ML confirmation while leveraging ML's superior classification performance—represents a powerful paradigm for public health epidemiology.

The decision tree model from traditional data mining analysis achieved an accuracy of 0.906, precision of 0.882, recall of 0.938, and F1-score of 0.909, indicating good performance in classifying entities with high or low relative vulnerability. The most important variable was the percentage of limitation (importance=0.963), followed by the proportion of disability among older adults (0.037). However, ensemble machine learning methods improved this performance: Random Forests achieved 0.943 accuracy; Stacking ensemble achieved 0.950 accuracy—a 4.4 percentage point improvement representing meaningful gains for policy-relevant territorial targeting. This improvement demonstrates that combining multiple algorithms captures territorial vulnerability patterns that single models miss, validating ensemble learning principles theoretically established in ML literature.

Outlier detection made it possible to identify territories with statistically differentiated behaviors. Both traditional Z-score and machine learning (Isolation Forest, Local Outlier Factor) methods identified Tabasco and Chiapas as high-disability extremes, and Mexico City as possessing unusual feature combinations. These results should not be interpreted as statistical errors, but rather as territorial signals justifying deeper diagnoses. The identification of outliers was useful for detecting entities where disability may be associated with specific factors such as aging, marginalization, unequal access to services, differences in census reporting, or urban concentration.

The constructed social indicators, such as the Territorial Disability Index and the Sex and Age Vulnerability Index, made it possible to synthesize complex information into comparable measures among federal entities. These indicators were useful for identifying priority territories, establishing rankings, comparing regional profiles, and facilitating communication of results to decision-makers. Complementarily, the developed dashboards—thematic maps, age-territory heatmaps, comparative rankings, sex-based panels, and cluster visualizations—constituted relevant tools for governmental planning, territorial monitoring, and the design of inclusion, accessibility, and support strategies for vulnerable groups.

The principal methodological contribution of the study was the systematic integration of traditional descriptive statistics, inferential analysis, data mining, machine learning, outlier detection, and territorial indicators within a single coherent analytical framework. This approach made it possible to move from a basic description of frequencies toward a multidimensional, methodologically triangulated understanding of disability. The combination of territorial analysis, statistical tests, correlations, clustering, decision trees, ensemble methods, and anomaly detection strengthened the ability to identify hidden patterns, regional inequalities, and vulnerability profiles with greater confidence than any single approach could provide. In this sense, the study provided scientific evidence validated across complementary methodologies for sociodemographic research and for the design of data-driven public policies.

Nevertheless, the study presented limitations. The census database used had a cross-sectional design, which did not allow the establishment of causal relationships or the analysis of temporal changes. Furthermore, the available information did not include direct socioeconomic variables such as educational attainment, income, employment, healthcare affiliation, housing conditions, or level of marginalization. This restriction limited the possibility of explaining more deeply the structural causes of the observed differences. Likewise, the data depended on census self-reporting, which may have generated variations in disability reporting among regions. The small effective sample size (n=32 entities) limits generalizability of machine learning models to other administrative contexts.

Future research should integrate census information with databases related to poverty, educational attainment, healthcare access, urban infrastructure, rurality, aging, and administrative records. It would also be pertinent to develop longitudinal predictive models using machine learning on historical census data, implement spatial machine learning accounting for geographic proximity of entities, and establish dynamic monitoring systems that translate model predictions into real-time policy feedback. In conclusion, the results demonstrated that disability in Mexico constituted a territorially unequal phenomenon closely linked to aging and social vulnerability. The integration of traditional advanced statistics and machine learning made it possible to generate robust, multiply-validated evidence for guiding inclusive, targeted, and evidence-based public policies that address the complex multidimensional nature of disability in Mexico.

## References

1. Baart, J.; Elbers, W.; Schippers, A.; Polack, S. (2025). The Definition Dilemma: How Definitions of Disability Shape Statistics on Social Participation. International Journal of Environmental Research and Public Health, 22(4), 603. DOI: 10.3390/ijerph22040603.
2. Friedman, C. (2024). Disparities in Social Determinants of Health Amongst People with Disabilities. International Journal of Disability, Development and Education, 71(1), 101–117. DOI: 10.1080/1034912X.2021.2004299.
3. Lawrence, P.R.; Anderson, R.K. (2024). Poverty and Disability: A State-Level Geospatial Analysis. Clinical Nursing Research, 33(5), 344–354. DOI: 10.1177/10547738241249834.
4. Uribe-Salas, F.J. (2025). Prevalencia, distribución geográfica y variables sociodemográficas de discapacidad en el Censo de Población y Vivienda 2020. Gaceta Médica de México, 161(5). DOI: 10.24875/GMM.25000092.

5. Zheng, P.; Guo, Z.; Du, X.; Yang, H.; Wang, Z. (2022). Prevalence of Disability among the Chinese Older Population: A Systematic Review and Meta-Analysis. International Journal of Environmental Research and Public Health, 19(3), 1656. DOI: 10.3390/ijerph19031656.
6. Jiang, Y.; Han, X.; Qiu, N.; Du, M.; Zhao, L. (2023). Identifying Urban–Rural Disparities and Associated Factors in the Prevalence of Disabilities in Tianjin, China. Land, 12(8), 1480. DOI: 10.3390/land12081480.
7. Forster, G.K.; Aarø, L.E.; Alme, M.N.; Hansen, T.; Nilsen, T.S.; Vedaa, Ø. (2023). Built Environment Accessibility and Disability as Predictors of Well-Being among Older Adults: A Norwegian Cross-Sectional Study. International Journal of Environmental Research and Public Health, 20(10), 5898. DOI: 10.3390/ijerph20105898.
8. Kersey, J.; Devlin, A.; Shyres, S.; Kringle, E.A.; Housten, A.J. (2024). Social Determinants of Health Affect Psychological Distress among People with Disabilities. International Journal of Environmental Research and Public Health, 21(10), 1359. DOI: 10.3390/ijerph21101359.
9. Baptista, E.A.; Shen, T.; Canudas-Romo, V. (2024). Disability and Its Impact on Life Expectancy: Heterogeneity across Mexican States. BMC Public Health, 24, 2776. DOI: 10.1186/s12889-024-20245-x.
10. MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (pp. 281–297). Berkeley: University of California Press.
11. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. (1984). Classification and Regression Trees. Belmont, CA, USA: Wadsworth International Group.
12. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96) (pp. 226–231). AAAI Press.
13. Liu, F.T.; Ting, K.M.; Zhou, Z.H. (2008). Isolation Forest. In 2008 Eighth IEEE International Conference on Data Mining (pp. 413–422). IEEE. DOI: 10.1109/ICDM.2008.17.
14. Instituto Nacional de las Personas Adultas Mayores (INAPAM). (2024). Diagnóstico de las Personas Adultas en México IV. Ciudad de México, México: Instituto Nacional de las Personas Adultas Mayores.
15. Instituto Nacional de las Personas Adultas Mayores (INAPAM). (2025). Diagnóstico de las Personas Adultas Mayores en México VI. Ciudad de México, México: Instituto Nacional de las Personas Adultas Mayores.
16. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. DOI: 10.1023/A:1010933404324.
17. Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; Elhadad, N. (2015). Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-Day Readmission. In Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15), pp. 1721–1730. ACM, New York, NY, USA. DOI: 10.1145/2783258.2788613.
18. Chen, T.; Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16) (pp. 785–794). ACM, New York, NY, USA. DOI: 10.1145/2939672.2939785.
19. Lundberg, S.M.; Lee, S.I. (2017). A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems 30 (pp. 4765–4774). Curran Associates, Inc.
20. Ribeiro, M.T.; Singh, S.; Guestrin, C. (2016). 'Why Should I Trust You?' Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16) (pp. 1135–1144). ACM, New York, NY, USA. DOI: 10.1145/2939672.2939778.
21. Wolpert, D.H. (1992). Stacked Generalization. Neural Networks, 5(2), 241–259. DOI: 10.1016/S0893-6080(05)80023-1.