**Journal of Information Systems Engineering and Management**
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

# Analyzing Technology Access Inequality in California's Homeless Population Using Machine Learning Techniques

Md Refadul Hoque[1], Jannat Ara[2*], Lisa Chambugong[3], Arifa Siddiqua[4], Samina Ahmed[5], Iftekhar Hossain[6], Nasrin Akter Tohfa[7]

[1]Department of Management, St. Francis College, 179 Livingston St, Brooklyn, NY 11201, USA;

[2]Master's in Public Administration, Gannon University,109 University Square, Erie, PA 16541, USA;

[3]Master's of Management Sciences and Quantitative Methods, Gannon University,109 University Square, Erie, PA 16541, USA;

[4]Deputy Director (Administration), National Academy for Primary Education, Mymensingh, Bangladesh;

[5]Department of Computer Information Systems, New England College, 98 Bridge St, Henniker, NH 03242, USA;

[6]Master of Science in Cybersecurity, Washington University of Science and Technology, Alexandria, Virginia, USA;

[7]Master's in Information Systems Security, University of the Cumberlands, Kentucky, USA;

* Corresponding Author Email: Ara001@gannon.edu

**ARTICLE INFO**

Received: 02 Feb 2026

Revised: 01 March 2026

Accepted: 10 March 2026

**ABSTRACT**

The research analyzes technology access disparity on a county level throughout California through an interpretable machine-learning framework uniting supervised classification, explainable AI, and unsupervised clustering. With publicly available broadband and socioeconomic variables of the dataset Older 58 counties in California, using a dichotomous outcome model of lower and higher-access counties and evaluating several algorithms (logistic regression, SVM with RBF kernel, random forest, gradient boosting, and extra trees), the paper uses a dataset of publicly available broadband and socioeconomic variables entitled U.S. Broadband Availability. Results of comparative analysis indicate that logistic regression is the most effective one in general (accuracy about 0.611, F1-score of about 0.462), which also has a moderate discriminative ability (ROC-AUC about 0.59), with the best measure of precision-recall (average precision about 0.66). Reliability is also evaluated in this study by calculation and threshold-sensitivity analysis to bring out trade-offs pertinent to policy targeting. Explainability analyses (feature importance, SHAP, and partial dependence) reveal that structural socioeconomic aspects, particularly poverty, unemployment, population dynamics, and social assistance dependency, are core contributors to broadband exclusion, whereas direct household-level technology indicators have a secondary role. Last, the distinct profiles of the K-means clustering (K = 3) clearly distinguish broadband-access profiles (well-connected, transitional, and severely underserved), which then inform more customized intervention strategies.

**Keywords:** Digital divide, County-level analysis (California), Broadband access inequality, Socioeconomic vulnerability, Policy targeting.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

## Introduction

Connectivity to affordable and dependable broadband internet has already become a prerequisite to engage in the current economic, educational, and civic activities [1]. The concept of digital connectivity allows access to jobs, distance learning, medical services, and government services, among other factors, and is thus an important condition of social and economic integration [2, 3]. Although broadband infrastructure has been heavily invested in, there still exist enormous differences in access to technology by regions and populations in the United States [4-6]. These differences are commonly described as the digital divide, but are especially significant in regions with susceptibility to socioeconomic vulnerability, the instability of the labor market, and access to state resources [7, 8]. California, being among the largest and most economically diverse states, has had a high heterogeneity in terms of broadband access in counties [9, 10]. Although certain areas have good digital infrastructure and high adoption rates, others still have limited access to the same, based on a combination of economic disadvantage, demographics, and inequality in the adoption of digital infrastructure [11]. These disparities can be structured, and, therefore, understanding the underlying forces is crucial in developing specific and effective policy responses to decrease the disparity in technology access [12, 13].

Conventional methods of researching the inequality of access to broadband have depended much on descriptive statistics or regression analyses [14, 15]. These techniques work fine, but they tend to miss nonlinear associations, the interaction impacts, and nonhomogeneous patterns that may emerge because of the intricate interconnection of the socioeconomic circumstances and availability of infrastructure [16]. The recent developments in machine learning provide potent means to model the said complexity, but the way they are used in scenarios of interest to policymakers reads worryingly in terms of interpretability, reliability, and usability [17]. This research overcame these limitations by suggesting that an interpretable machine learning framework can be used to understand the inequality in access to technology in California counties [18]. The study goes beyond prediction by using supervised learning, explainable artificial intelligence, and unsupervised clustering to offer explanations for why some counties have low broadband access. The study focuses not only on the model performance, but also on the behavior of the model with various errors and threshold sensitivity regarding the parameters, which are of significant importance when the machine learning outputs are to be exploited to make policy decisions [8, 19, 20].

However, instead of referring to broadband access as an infrastructural issue, this research paper clearly looks at how other broader socioeconomic factors, like poverty, lack of employment, population dynamics, and dependence on state welfare, have contributed to the issue. In that way, it adds to a burgeoning literature on sustaining the idea that digital exclusion is rooted in structural unemployment and cannot be addressed only in terms of infrastructure expansion [21].

In this paper, the following are the contributions made:

- An end-to-end machine learning model for analyzing inequality in broadband access that fuses supervised classification, explainable AI methods, and unsupervised clustering.
- An interpretable assessment of socioeconomic determinants of low broadband adoption based on feature importance, SHAP values, and partial dependence plots.
- A sound assessment of the reliability of its models, reflecting trade-offs applicable to policy decision-making.
- Determination of the different profiles of broadband access at the county level by clustering, to highlight severely underserved areas, transitional areas, and well-connected areas.

The rest of this paper is structured in the following way: Section 2 explains the data sources, the pre-processing process, the model, and the evaluation process. Section 3 provides an elaboration of the empirical findings, namely predictive performance, explainability analyses, and clustering findings.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Section 4 gives a synthesis of prior research to outline that digital inequality is multidimensional, persistent, and motivated by intersecting socioeconomic, linguistic, and infrastructural factors as opposed to its access. Lastly, Section 5 wraps up the paper by summarizing contributions, providing limitations, and giving recommendations on further research activities [22].

## Methodology

Figure 1 gives a summary of the analytical model to be used in this research. The data used is processed and modeled based on supervised machine learning, explainable business strategies, and unsupervised groups to study technology access inequality among the counties in California. Four county-level socioeconomic and broadband-related variables are initially collected and processed by a definite preprocessing system. A series of machine learning models is subsequently trained to categorize counties as having broadband access available or not, and intensive performance metrics checks are subsequently conducted with complementary metrics. To be able to interpret the output of the model to guarantee transparency and the relevance of the policies, the feature importance analysis, SHAP explanations, and partial dependence plots are used. Lastly, the intrusive clustering is used to detect the latent broadband access profiles to offer extra information to the binary classification.

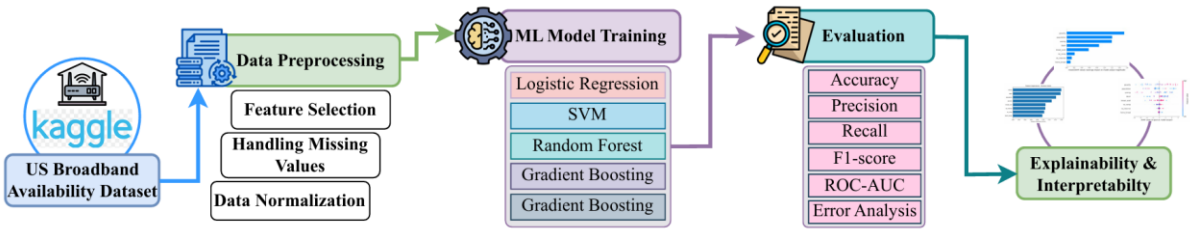



**Figure 1.** *Workflow of this research*

### 3.1 Data Source and Study Area

This research utilized publicly accessible broadband data and socioeconomic data in order to assess the inequality in access to technology in California counties. The "U.S. Broadband Availability" dataset is collected from Kaggle [23], which offers 3,142 county-wise observations in the United States. This dataset provides broadband access and infrastructure data, including broadband availability measures, mean download speed, and household broadband adoption, along with related demographic and socioeconomic factors. In the analysis, the counties in the state of California are confined to provide geographic consistency and relevance in issues of policy. The study sample will be comprised of 58 counties in California, which is the entire population of counties in the state after filtering the dataset by state. The data was filtered to eliminate counts that had no values on important variables before model estimation to retain the quality and strength of the data [24].

### 3.2 Feature Selection and Data Preprocessing

This study follows a systematic preprocessing pipeline that guarantees the quality of the data, the comparability of the features, and the strength of the machine learning models. Descriptions have been made of the preprocessing steps.

1. **Feature Selection [25, 26]:** A set of socioeconomic and broadband-related variables on the county level was selected as a way to reflect various dimensions of technology access inequality. The variables are the poverty rate, the unemployment rate, the percentage of households lacking a computer, the percentage of households lacking the internet, the household adoption rate to broadband, the markers of broadband availability, and the number of the population

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

within a county. The variables collectively represent economic susceptibility, access to digital infrastructure, and access requirements at the household level. Where the feature matrix is represented by

$$X = \{x_1, x_2, \ldots, x_p\}, \quad (1)$$

let there each row correspond to a county, and each column to a selected feature.

2. **Handling Missing Values [27]:** Listwise deletion was used to remove records with missing values in one of the chosen features. The filtered dataset can be defined as:

$$X' = \{x_i \in X | x_i \ contains \ no \ missing \ values\} \quad (2)$$

This is the best method to provide similarity among models and give a chance to avoid biases that might be brought with imputation due to a small portion of the missing observations.

3. **Feature Scaling and Standardization [28, 29]:** Support Vector Machines and Logistic Regression are feature-sensitive models, and it was necessary to normalize the features using the z-score to train these types of feature-sensitive models. Each of the features $x_j$ was normalized as:

$$x_j^{(scaled)} = \frac{x_j - \mu_j}{\sigma_j} \quad (3)$$

where $\mu_j$ and $\sigma_j$ represents the mean and standard deviation of the feature $j$, respectively. The models based on trees were trained using unscaled features because they are monotonic transformers of the variables being given as input.

4. **Target Variable Construction [30]:** The prediction problem is defined as a binary classification problem. Let $y$ denote the broadband access value for county $i$. The counties are identified on a median division of access to broadband:

$$y_i^* = \begin{cases} 1, & if \ y_i < median(y) \\ 0, & otherwise \end{cases} \quad (4)$$

where $y_i^* = 1$ refers to the mean low broadband access, and $y$ refers to the high broadband access.

5. **Train-Test Split:** Stratified sampling was used to divide the data into training and testing datasets to maintain the proportion of classes. Let $D$ denote the full dataset:

$$D = D_{train} \cup D_{test}, \ |D_{train}| = 0.7|D|, \ |D_{test}| = 0.3|D| \quad (5)$$

The stratification guarantees that both the subsets have a similar distribution of the target classes.

### 3.3 Model Development

In order to define counties with access inequality under technologies, this study utilizes various machine learning models that are currently found in diverse algorithmic families [31, 32]. With the help of different models, the analysis has an opportunity to include both the linear tendencies and nonlinear connections, as well as the complex interaction between the socioeconomic vulnerability and broadband access status. Each and every model is trained to estimate whether a county is part of the low broadband access population, as per the type of socioeconomic and infrastructure-related traits chosen.

1. **Logistic Regression (LR) [6, 33]:** LR serves as a baseline that measures the direct and linear relationship between socioeconomic variables and the risk of having low broadband access. The model offers a benchmark, which can be interpreted, since it estimates the impact of variations in variables like poverty rate, unemployment, or the absence of access to computers, on the likelihood of broadband exclusion.
Formally, the model estimates as:

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta^T x)}} \quad (6)$$

Where the feature vector is denoted by x, model coefficients by $\beta$, and the intercept by $\beta_0$.

2. **Support Vector Machine (RBF Kernel) [34]:** The Support Vector Machine (SVM) with a radial basis function (RBF) kernel is employed to model nonlinear boundaries between counties with high and low broadband access. This is particularly relevant for technology access inequality, where risk factors may interact in complex ways and threshold effects may exist (e.g., poverty levels beyond which access sharply declines). The RBF kernel is defined as:

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right) \quad (7)$$

By projecting counties into a higher-dimensional space, the SVM captures nonlinear separations that cannot be represented by linear models, thereby improving classification performance in heterogeneous regions.

3. **Random Forest (RF) [35, 36]:** RF is an ensemble algorithm that combines the predictions of several decision trees. Every tree is also trained on a bootstrap sample of the data, with node splits being chosen on a random subset of features. The majority is used to obtain the final prediction:

$$\hat{y} = mode\{T_1(x), T_2(x), \ldots, T_M(x)\} \quad (8)$$

with $T_m$ implying the $m$-th decision tree. This will minimize variance and enhance generalization.

4. **Gradient Boosting (GB) [37, 38]:** GB involves the building of weak learners sequentially, and with each new weak learner, the loss function of the previous ensemble becomes optimal. The update of the model at iteration $m$ can be defined as:

$$F_m(x) = F_{m-1}(x) + \eta h_m(x) \quad (9)$$

where $h_m$ is the newly trained weak learner, and $\eta$ is the learning rate. This model works well in representing complicated interactions between characteristics.

5. **Extra Trees [39]:** Extra Trees (Extremely Randomized Trees) classifier is an extension of the Random Forest, and extra randomness is added. The feature selection and the split thresholds are randomly chosen. This method also decreases variance further and frequently increases the performance on tabular data.

### 3.4 Evaluation Metrics

In order to offer a holistic measure of the model's performance, several evaluation metrics were adopted, which included predictive performance and class-specific performance.

1. **Accuracy [40, 41]:** The measures of accuracy give the general percentage of correct instances classified:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (10)$$

Accuracy is the overall accuracy of the model, but it can be deceptive when there is a class imbalance.

2. **Precision [42]:** Precision measures the proportion of low-access counties recognized to be low-access, relative to the number of counties forecasted to be low-access:

$$Precision = \frac{TP}{TP + FP} \quad (11)$$

When costs associated with false positives are greater, i.e., when well-connected counties may be labeled as underserved, precision is especially needed.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3. **Recall [43]:** The measure of the mass-produced model's correct identification of counties with low broadband access:

$$Recall = \frac{TP}{TP + FN} \quad (12)$$

The model with high recall is significant because it captures the majority of underserved counties, which is essential when it comes to policy intervention targeting.

4. **F1-score [43, 44]:** F1-score is defined as a harmonic mean between the precision and the recall; it is a more balanced metric of classification efficiency:

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (13)$$

F1-score is advantageous, especially in cases where there is class imbalance in the dataset, and causing a false positive and false negative occurrence is of consequence.

5. **ROC-AUC [45]:** Area under the Receiver Operating Characteristic curve, representing the trade-off between true positive rate and false positive rate.

6. **PR-AUC [46]:** Area under the Precision-Recall curve, particularly informative under class imbalance.

7. **Confusion Matrix [47]:** Confusion matrices were generated for each model to analyze misclassification patterns. ROC and Precision-Recall curves were plotted to enable direct visual comparison of discriminative performance across models.

### 3.5 Explainability and Interpretability

Since the policy is relevant to determining the counties with poor technology access, model interpretability is an essential element of the study. Instead of designing a predictive model that solely aims at predicting performance, the use of explainable machine learning methods is used to gain insights as to which socioeconomic and infrastructure predictors are used to predict low broadband access, as well as the effect of each predictor on the decisions made by the model [48, 49]. The best-performing tree-based model is analyzed to explain the model with the highest results, as determined with comparative evaluation metrics. The use of tree-based models is especially good when it comes to interpretability because it can capture nonlinear interactions, as well as enable post-hoc explanations [50].

#### 3.5.1 Feature Importance

The impurity-based measures to compute the feature importance scores are used to identify the most significant variables in the classification of counties into low and high broadband access [51]. Here, the importance of features shows how much a variable lowers uncertainty in differentiating underserved counties and better-connected ones. The importance of feature $j$ can be defined as:

$$FI_j = \sum_{t \in T} \Delta I_{t,j} \quad (14)$$

In which $\Delta I_{t,j}$ represents the reduction in impurity given to feature $j$ at a node $t$, aggregated across all trees in the ensemble.

#### 3.5.2 SHAP (SHapley Additive exPlanations)

In order to receive more specific and theoretically based explanations, SHAP (SHapley Additive exPlanations) [52, 53] values are used. SHAP computes these feature contribution values on individual prediction enabling the breakdown of the predicted risk of broadband access in a county in terms of the contribution of features. The SHAP value of feature $j$ is as follows:

$$\phi_j = \sum_{S \subseteq F\{j\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} [f(S \cup \{j\}) - f(S)] \quad (15)$$

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

where $F$ is the set of features. The influential variables and their directional impact on broadband access inequality were identified by means of SHAP summary and dependence plots.

### 3.5.3 Partial Dependence

Partial Dependence Plots (PDPs) are used to represent the marginal impact of individual features on the predicted probability of low broadband access, and averaged across the distribution of the remaining variables. The partial dependence of feature $x_j$ is given by the following:

$$PD_j(x_j) = E_{x_{-j}}[f(x_j, x_{-j})] \qquad (16)$$

This study benefitted by PDPs, especially to determine threshold effects, e.g., poverty or device-access levels after which the likelihood of broadband exclusion rises sharply. These insights help in the interpretation of model behavior in an aspect that is user-friendly to both policymakers and non-technical stakeholders.

## 3.6 Unsupervised Clustering

Besides the supervised learning, unsupervised clustering was also used in order to find the latent patterns between counties. Standardized features on the topics of poverty, internet access, broadband adoption, and availability were subject to the use of K-Means clustering [54]. The algorithm minimizes intra-cluster variance:

$$\min_{\{\mu_k\}} \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \qquad (17)$$

and $\mu_k$ represents the centroid of cluster $C_k$. The three-cluster solution ($K = 3$) was chosen to depict high-access counties, counties that are moderately underserved, and counties that are severely underserved. Summary statistics and visualizations were used to analyze cluster characteristics in order to place the results of supervised learning in perspective [55].

## Experimental Results

This section provides and discusses the empirical data of the research based on predictive performance measures, calibration analysis, error analysis, explainable machine learning outputs, and clustering results. The findings are structured in such a way that the performance of the models will be initially evaluated and compared, and then model reliability and the misclassification behavior will be studied in greater detail. Combined, those findings give a detailed picture of the technology access differences at the county level throughout California and the policy implications of these differences [56].

## 4.1 Comparative Performance of ML Models for Broadband Access Classification

The evaluation of the machine learning models applied to categorize California counties into low and high-access broadband will be a comparative assessment as outlined in Table 1. LR was one of the considered models, and overall performance was the highest, with the accuracy of 0.611 and the largest F1-score (0.462), which identified a fair trade-off between the precision and recall. This outcome implies that there are still significant linear associations between socioeconomic issues and broadband access in counties.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

**Table 1.** Classification performance comparison of ML models for identifying counties with low broadband access in California

| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Logistic Regression | 0.6111 | 0.75 | 0.3333 | 0.4615 |
| SVM (RBF) | 0.3889 | 0.3333 | 0.2222 | 0.2667 |
| Random Forest | 0.4444 | 0.4 | 0.2222 | 0.2857 |
| Gradient Boosting | 0.5556 | 0.6 | 0.3333 | 0.4286 |
| ExtraTrees | 0.5 | 0.5 | 0.2222 | 0.3077 |

The model performance in GB was competitive, as it provided an accuracy of 0.556 and an F1-score of 0.429. The fact that it is better recallable than most other nonlinear models implies that it has the capability of detecting underserved counties with more intensity, which is therefore the manifestation of the nonlinear interaction of the socioeconomic and infrastructure-related characteristics. Conversely, the SVM (RBF) and the RF showed lower predictive accuracy, especially in recall, suggesting that neither can easily predict low-access counties with limited sample size conditions.

The Extra Trees classifier was moderate in terms of accuracy and relatively low in terms of recall, indicating that though the ensemble-based model may model more intricate feature interactions, they might need more data as a whole or more fine-tuning to be generalized to such an environment effectively. All in all, these findings have shown that simpler and moderately complex models can be as good (or better) compared to more complex ensembles without comparison to county-level broadband access classification, which underlines the necessity of the interpretability and strength of models in policy-directed models.

### 4.2 Discriminative Performance Analysis Using ROC and Precision-Recall Curves

The Receiver Operating Characteristic (ROC) curves of all machine learning models to classify California counties depending on the status of broadband access are provided in Figure 2. In general, the ROC analysis indicates that there is a moderate level of discriminative power regarding models, which can be explained by the fact that it is difficult to predict the broadband access inequality on the basis of aggregated county-level data. LR has the greatest ROC-AUC (0.59) value, showing the best performance compared to ensemble and kernel-based models. This would imply that there is at least a linear separability between low-access and high-access counties and that more straightforward models can be used to predict prevailing trends in the data.

The GB model performs on average (ROC-AUC = 0.51), whereas the models of the Extra Trees, SVM, and RF show performance close to random. The findings suggest that more complicated nonlinear models do not always enhance discrimination in the context, probably because of the small sample size and structural homogeneity of counties at large levels of broadband access.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

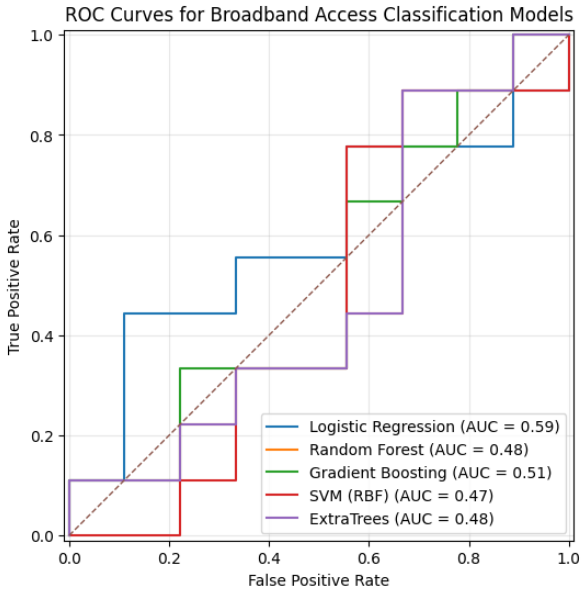



**Figure 2.** *ROC curves for ML models classifying broadband access across California counties*

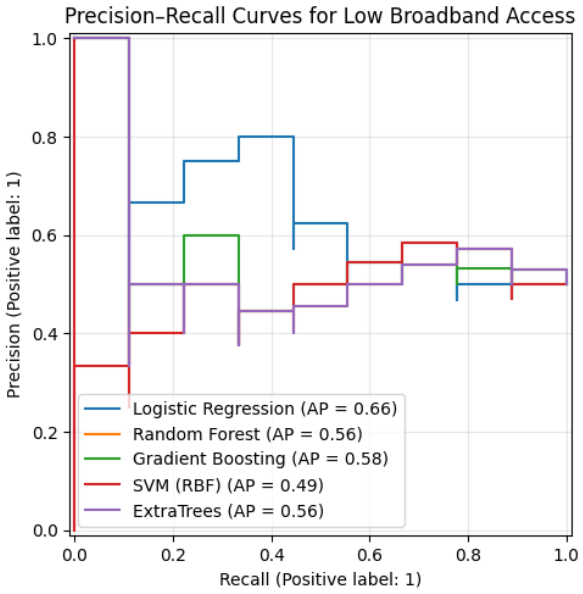



**Figure 3.** *Precision-Recall curves for low broadband access classification across ML models*

Figure 3 shows the Precision-Recall (PR) curves, which can be used as a better evaluative tool due to the conditions of class imbalance. In line with ROC findings, logistic regression has the best average precision (AP = 0.66), which means that LR can best identify low broadband access counties with a relatively high level of precision. GB and ensemble-based models have moderate average precision values as compared to the SVM, which has the lowest performance, especially at high recall levels.

Combined, the ROC and PR analyses support the conclusion that models with less complicated and easy-to-understand formats can provide competitive and, in many cases, optimal areas of discriminative performance on county-level broadband access classification. These results justify the application of the LR as a strong baseline model and reveal the necessity of balancing the complexity and interpretability of models used in policy-related purposes.

### 4.3 Calibration and Reliability of Predicted Broadband Access Risk

Figure 4 shows the calibration curves of all tested machine learning models, which depict the dependence on the predicted probabilities of a low probability of broadband access between the frequency of the observed results and the experimental results. On balance, the calibration analysis demonstrates a big difference in the probability reliability of the models, which is evident between discriminative performance and probabilistic accuracy.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

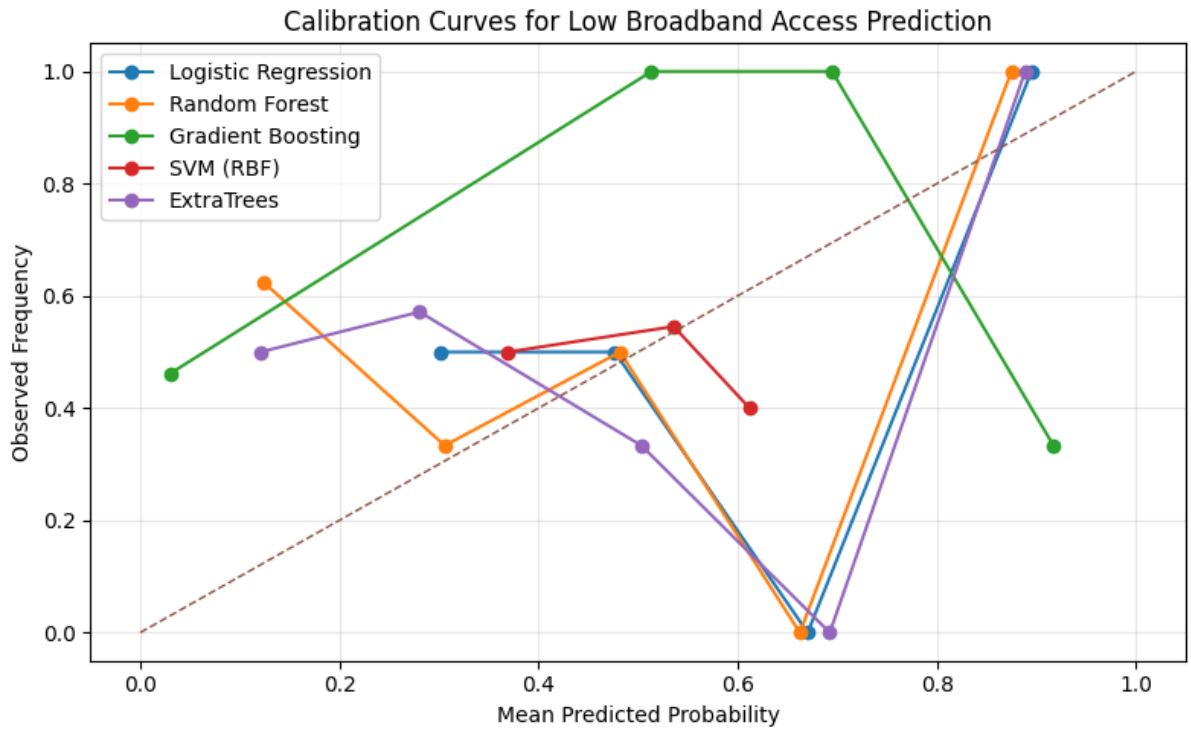



**Figure 4.** *Calibration curves for predicted low broadband access probabilities across ML models*

The calibration of the LR model is also better in nature, as the components' projected probabilities tend to fit the predicted line more in a series of probability bins. It means that the frequency of actual low broadband access (when the model offers moderate to high scores of risks) is fairly in accordance with the predictions. Conversely, the ensemble-based models, such as RF and Extra Trees, show much stronger deviations from the diagonal, indicating some trends towards overconfidence or underconfidence of some probability intervals.

The GB model exhibits a high degree of miscalibration, especially in the higher end of the prediction probability levels, whereby the outcomes observed deviate significantly from the risk as predicted. In the same fashion, the SVM model has an inconsistent calibration curve, which reduces its use in probabilistic interpretation. These results imply that not all models can obtain satisfactory model classification, but their probabilities are not always useful indicators of actual risk levels. Calibration is particularly relevant in terms of policy and decision support, whereby the predicted probability is quite often used to rank interventions. The findings demonstrate that simpler and correctly calibrated models can be more ideal in cases where probabilistic estimates are needed in order to inform the resource allocation to enhance access to broadband.

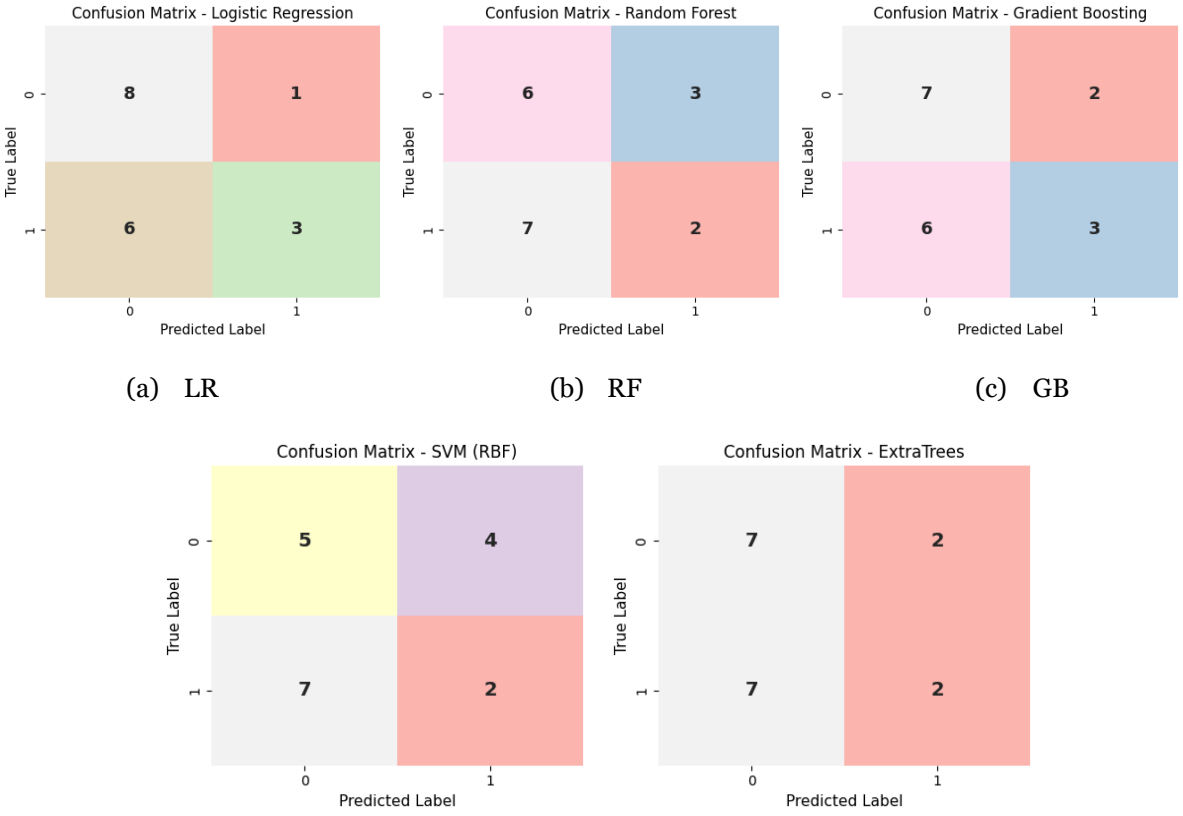
### 4.4 Error Analysis and Misclassification Patterns

Figure 5 provides a close-up of the products of misclassification details in the analysis of California counties based on broadband access coverage. Throughout the models, there is a general pattern where there is low recall of low broadband access class, which implies that it is challenging to identify underserved counties consistently. The error profile indicated by LR is the most balanced, making a greater number of low-access counties accurate than the other two models, although with a relatively lower number of false positives.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

On the contrary, the ensemble-based model (RF, GB, and Extra Trees) has a bias towards false negatives, where low-access counties are classified as high-access ones. Such a trend leads to the idea that as long as these models are able to identify key access trends, they would not find these marginal and borderline cases of broadband exclusion. The SVM is the lowest performing one, with high false positives and false negatives, which makes it even less suitable for this task.




(a) LR (b) RF (c) GB

(d) SVM (e) Extra Trees

**Figure 5.** *Confusion matrices for ML models classifying broadband access across California counties*

A deeper analysis of the best-performing model (LR) based on the distribution of predicted probabilities is also given in Figure 6. In the case of the correctly clustered high-access and low-access counties, there is partial separation in probability space, but those falsely identified are concentrated at the middle of the probability scale. These overlap points at intrinsic ambiguity in the conditions of county-level broadband access, especially of counties whose socioeconomic vulnerability is moderate or whose infrastructure features are mixed.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

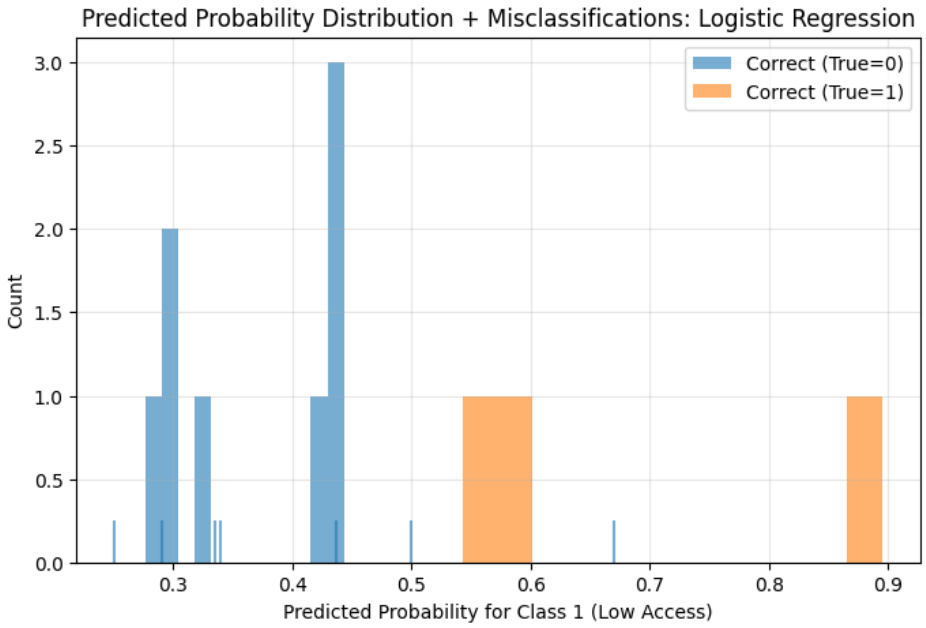



**Figure 6.** *Predicted probability distribution and misclassifications for the best-performing model*

Overall, the error analysis shows that the misclassifications are not scattered, but are concentrated between the counties having transitional access profiles. The findings emphasize that the inequality in access to technologies is quite complicated on the structural level and that the local inequality may be hidden when aggregating to the county level. Policy-wise, these borderline counties can be a subject of more qualified or sub-county research than binary classification can be.

### 4.5 Threshold Sensitivity Analysis for the Best-Performing Model

Further analysis of the decision behavior of the high-performing model (Logistic Regression) was done by a sensitivity threshold analysis by adjusting the classification probability threshold to consider counties considered low in terms of broadband access. Table 2 gives a summary of changes in precision, recall, and F1-score at a threshold point of 0.1 to 0.9. The findings indicate that there is a definite trade-off that exists between recall and precision. At these thresholds (i.e., 0.1-0.2), the model is very likely to be recalled, or rather, it identifies most of low access counties at the expense of substantially reduced precision, i.e., a higher number of false positives. The threshold decline of recall is high, with its precision increasing; this signifies a more conservative approach of classification to a higher degree of certainty than the coverage strategy.

**Table 2.** Precision, recall, and F1-score for the best-performing model under varying classification probability thresholds

| Threshold | Precision | Recall | F1-score |
|---|---|---|---|
| 0.1 | 0.5 | 0.888889 | 0.64 |
| 0.2 | 0.5 | 0.666667 | 0.571429 |
| 0.3 | 0.375 | 0.333333 | 0.352941 |
| 0.4 | 0.4 | 0.222222 | 0.285714 |
| 0.5 | 0.5 | 0.222222 | 0.307692 |
| 0.6 | 0.5 | 0.111111 | 0.181818 |

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

| | | | |
|---|---|---|---|
| 0.7 | 1 | 0.111111 | 0.2 |
| 0.8 | 1 | 0.111111 | 0.2 |
| 0.9 | 0 | 0 | 0 |

The peak F1-score is achieved at a small threshold (0.1), indicating that the overall performance of the maximization of recalls in this application is the most balanced. Nonetheless, a cutoff beyond 0.5 leads to significant declines in recall, which means that numerous underserved counties would be missed in the event of the implementation of default probability cutoffs. The model will only identify a small proportion of low-access counties at very high thresholds (>=0.7), which makes its application to locate at-risk areas practically ineffective. In regard to policy, this discussion has also illustrated how the threshold should be set based on the intervention goals. In some instances, the smaller thresholds may be appreciated where one aims at ensuring that the underserved counties are not overlooked, even at the expense of more false positives. On the contrary, more advanced thresholds can be relevant in cases where the resources are scarce, and the interventions need to focus on a set scope. This flexibility has helped to underscore the importance of probabilistic outputs as far as determinate binary decisions are concerned in the planning of broadband access.

### 4.6 Feature Importance Analysis of Socioeconomic and Infrastructure Drivers

The feature importance rankings based on the RF and Extra Trees models are given in Figure 7 and Figure 8, respectively. In both the ensemble techniques, unemployment rate stands out as the most predictive variable of low broadband access rates, which means that the labor market factor factors in as a key influencer to the technology access inequalities at the county level. This observation underscores the high level of correlation between economic instability and low digital connectivity.

The population size and enrollment in social assistance services (SNAP) are also among the leading contributors in both models, which is indicative of the fact that population-wide structural and demographic processes are major determinants of broadband access outcomes. Those counties that have more citizens or are with increased socioeconomic vulnerability seem to have different access patterns, possibly due to differences in infrastructure investment and access to services.

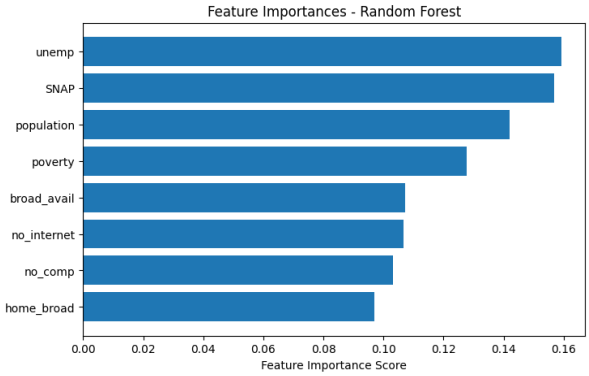



**Figure 7.** *Feature importance rankings from the RF model for broadband access classification*

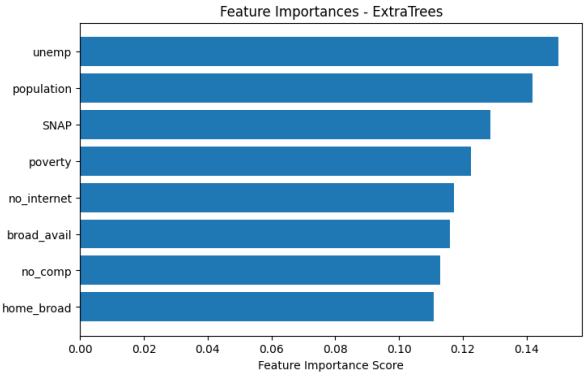



**Figure 8.** *Feature importance rankings from the Extra Trees model for broadband access classification*

Variables related to the digital infrastructure and access to homes directly, including broadband access, no access to the Internet, and no access to a computer, reflect moderate yet stable values in both models. Although these factors are intuitively connected with access to technology, their rather low position in the ranking with the others (unemployment and population) indicates that the structural socioeconomic

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

factors can be even stronger in predicting the broadband exclusion beyond the infrastructure access only.

The consistency in feature rankings in both Random Forest and Extra Trees also lends support to the findings that the stated factors that drive technology access inequality are not specific to a particular model. Altogether, the analysis of feature importance supports the conclusion that technological infrastructure barriers are entrenched in more extensive socioeconomic factors, and not necessarily the reason why people experience gaps in broadband access.

### 4.7 SHAP-Based Interpretation of Drivers of Broadband Access Inequality

Figure 9 and Figure 10 show SHAP-based explanations of the best-performing tree-based model, which includes global and local information about the factors that contribute to the prediction of low access to broadband in California counties. The SHAP summary plot (Figure 9) will show the contribution of the different features per county, whereas the SHAP bar plot (Figure 10) will rank the features in terms of the average contribution of the absolute values to the model output.

In both visualizations, poverty is the most significant variable, and the high values of poverty are always connected to the positive values of SHAP, which reveals that there is the likelihood of low broadband storage. This observation underscores the importance of economic disadvantage in affecting inequality in access of technology. In a similar manner, the population size and unemployment rate also demonstrate strong contributions, indicating that both the magnitude of the population and the situation in the labor market are major determinants of the broadband access.

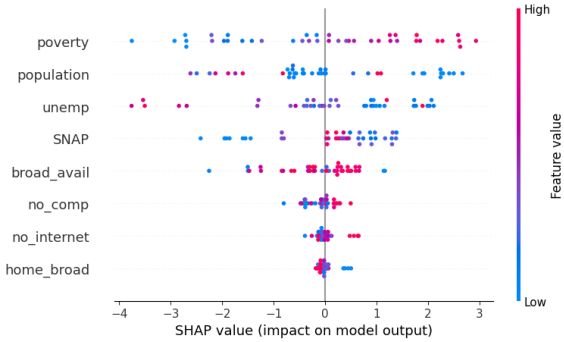



**Figure 9.** *SHAP summary plot showing feature contributions to low broadband access predictions*

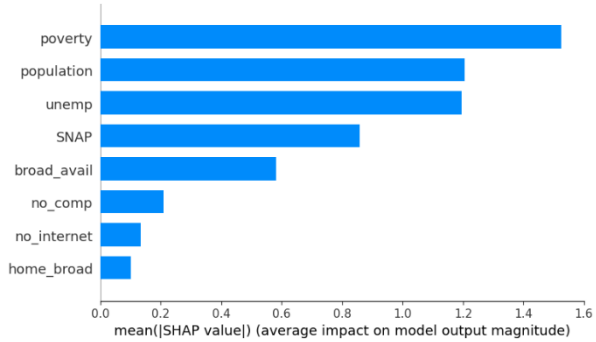



**Figure 10.** *Global feature importance based on mean absolute SHAP values*

There is also a significant change in participation in social assistance programs (SNAP), and it supports the correlation between the general socioeconomic vulnerability and the lack of connectivity. By contrast, the variables that are immediately dependent on the household level technology access, including the absence of computer ownership and the absence of internet access, have smaller average effects, even though their impact is directional. The more these indicators are high in a county, the higher the likelihood of a high risk of broadband exclusion.

The SHAP summary plot also shows us that the effects of the features are heterogeneous, with counties showing a strong positive or negative contribution based on their socioeconomic profile. This variation implies that there is no single cause of inequality in access to technology, and instead, the structural conditions that vary between regions will create inequality in access. In general, the SHAP analysis reveals transparent model-consistent evidence of disparities in access to broadband being mainly based on socioeconomic factors, where the variables related to infrastructure have a second significant but supportive influence.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

### 4.8 Partial Dependence Analysis of Socioeconomic Threshold Effects

Figure 11 provides Partial Dependence Plots (PDPs) of the major socioeconomic and access-related variables to demonstrate how individual features had a marginal impact on the predicted probability of lacking broadband access with all other variables kept constant. These plots can give an understanding of the relationships and threshold effects that are not reflected by the global feature importance metrics.

A nonlinear rising risk of predicted broadband exclusion rises strongly above moderate poverty rates in the PDP of poverty rate. Counties in which the poverty rate is above the mid-range have a drastically increasing partial dependence score, which means that the probability of low broadband access is significantly higher. This implies that there is a poverty line past which the inequality of access to technology gets strikingly more acute. On the same note, there is a nonlinear trend in the unemployment rate where the risk is predicted to be high at low-to-moderate as well as higher unemployment levels, and then there is a steep increase at a critical level. This trend means structural labor market vulnerability and amplifies it on broadband access, especially in counties that have faced chronic employment instability.

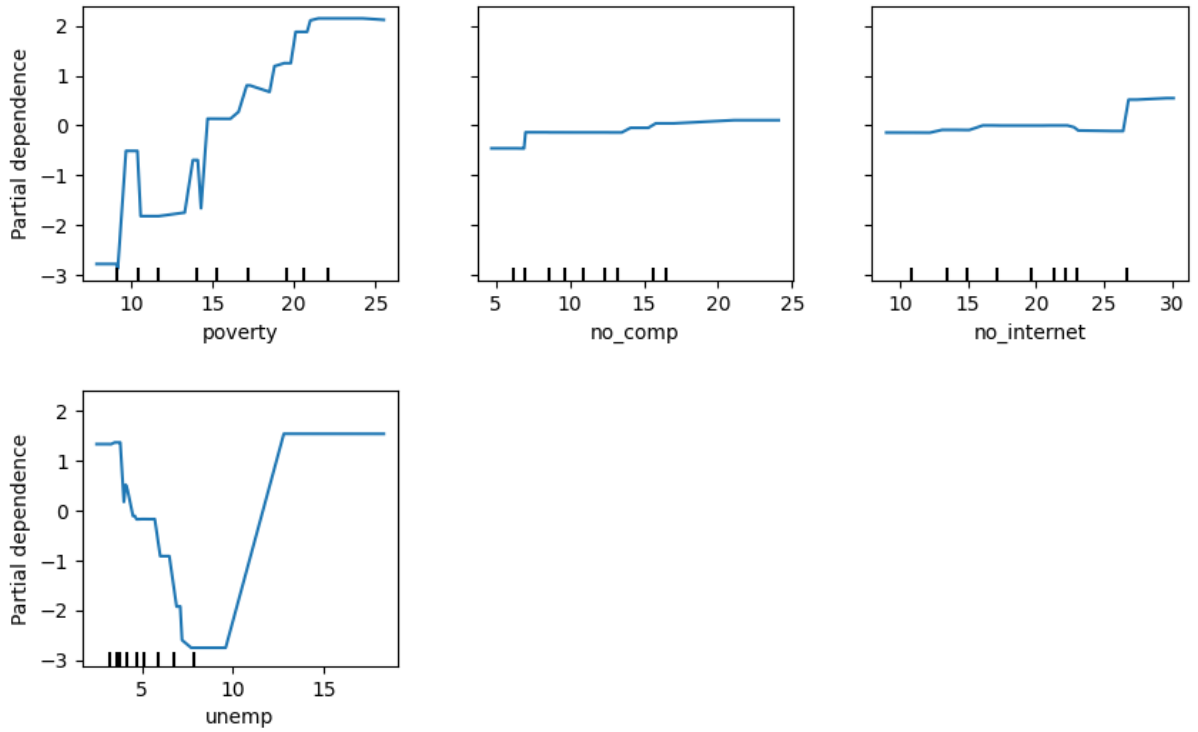



**Figure 11.** *Partial dependence plots showing the marginal effects of key socioeconomic variables on low broadband access predictions*

By comparison, the relationships between the PDPs of absence of computer ownership and absence of internet access will be rather flatter, which implies that their impact on the risk of broadband exclusion will be less quick. Though increased values of these indicators are linked to greater predicted risk, their impact on the risk is not significant in relation to more extensive socioeconomic variables. This result supports the conclusion that structural economic factors are predominant, with deficits on household levels of technology as reinforcing and not major causes.

In general, the analysis of partial dependence points to the idea that inequality in access to technology would also be the phenomenon that is defined by the threshold-based socioeconomic impact, and in this context, the gradual improvements might not be adequate for the counties that are already above

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

the levels of critical vulnerability. Such insights can be used as a guideline for specific policy interventions to reduce broadband exclusion.

### 4.9 Unsupervised Clustering of Broadband Access Profiles

The standardized cluster centers calculated using K-Means clustering analysis are shown in Figure 12, and it shows that California counties possess very different broadband access and socioeconomic patterns. The homogenous grouping solution with three clusters shows that there is some substantial segmentation based on the inequality of the homogenous access to technology that cannot be represented by supervised classification alone.

Cluster 0 is a group with high levels of poverty, high rates of households lacking internet access, and significantly low broadband penetration and household broadband take-up. The cluster is the most underserved county in which the socioeconomic disadvantage, together with the infrastructure constraints, becomes the source of broadband exclusion.

Cluster 1, in its turn, has lower poverty rates, less internet deprivation, and comparatively high household broadband occupation rates, as well as moderate levels of broadband availability. This cluster of counties can be explained as well-interconnected areas, in which the economic and access infrastructure has enabled increased levels of digital inclusion.

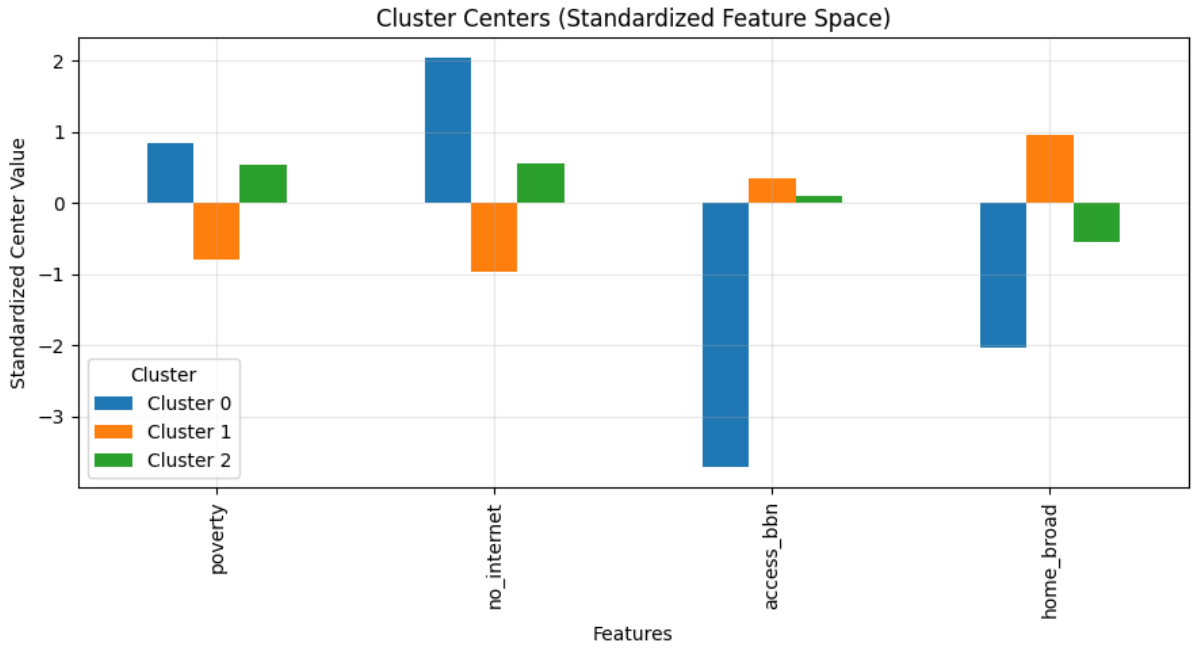



**Figure 12.** *Cluster centers in standardized feature space illustrating broadband access profiles across California counties*

Cluster 2 is the mean one, having moderate poverty and internet deprivation, near-average access to broadband, and a slightly lower broadband adoption among households. This segment is representative of the transitional countries, in which the conditions of access fall between being acutely limiting and being completely satisfying, which means that they may be susceptible to instances of digital disenfranchisement in the future.

All in all, the advantages of the clustering analysis to complement the results of the supervised learning include the demonstration of a clear access profile to structure across the counties. These results indicate that the issue of inequality in access to broadband has multidimensional characteristics and

imply that the key to addressing it should be to target the particular needs of each cluster instead of a blanket approach.

## Discussion

Research came to a number of basic inequality patterns. Constant digital divides also existed even in the face of overall increases in access, with inequalities becoming concentrated in and between already disadvantaged groups [57]. The 81.5% rate of access to the Internet created large gaps, 30% of which had no significant access [58]. Access was not the only source of telehealth disparities. Patients with limited English proficiency had less utilization (4.8% versus 12.3% post-adjustment of the internet access), suggesting multifaceted barriers such as digital literacy and content in their native language [59]. Surprising results showed that the use of telehealth meant more emergency department visits among all the patients. The disparity between broadband quality was also quite significant, with the rural areas having a speed 2.5 times less than that of the urban ones [60]. FCC data most strongly overstated the rates within rural and low-income communities and can be misleading to policy-making [61].

Families in K-12 experienced mixed development. The access to the devices was boosted significantly, and the share of broadband adoption rose by 86% to 93% between 2019 and 2021, and the share of smartphone-only households decreased to 10% and 4% [62]. Nonetheless, the gap in the quality and dependability of the connections persisted with low-income families. The access to broadband was still based on the parental payment capacity, and education, frequency of use, and breadth were the strongest predictors of eHealth literacy inequality [63]. There were great disparities in the use of assistive technology according to age and race, ethnicity, education, income, and disability type [64]. There was an emphasis on policy recommendations focusing on specific strategies. Expanding access in rural communities and increasing the proportion of poor urban people enrolling in services were complementary measures [65]. There were language choices of learning management systems and multilingual digital literacy training that met the needs of K-12 [66]. The personalization of the interventions to less-educated and internet-experienced persons filled the eHealth literacy gaps. Specific attention was paid to the dimension of telehealth equity, which is limited English proficiency.

The identified barriers were not limited to infrastructure and cost, and these barriers also fell under the categories of digital literacy disparities, language barriers, and unconventional means of care. The facilitators were English proficiency, higher education, and end-to-end provision, connectivity, and training. The limitations of the study were self-reported data, the lack of evaluation of all the telehealth modalities, and the possible bias of crowdsourced data. The studies incorporated reported similar patterns of inequality in access to technologies on various levels, whereas they focused on different populations, technologies, and time spans. The evidence shows that inequality acts in many interlocking pathways that exacerbate disadvantage, as opposed to contradictory findings.

## Conclusion

This research introduces a detailed ML-driven model to examine the issue of inequality in technology access in the counties of California. The combination of supervised classification, explainable artificial intelligence models, and unsupervised clustering contributes to the development of the study in the context of revealing how socioeconomic and infrastructural variables together determine the disparity of broadband access. The results show that structural socioeconomic factors, especially poverty, unemployment, demographics, and dependence on social support programs, are key determinants of broadband exclusion, and, in many cases, infrastructure access is not significant. By utilizing interpretable models and explainability piping like SHAP and partial dependence plots, the research gives clear, understandable findings relevant to policymakers that are not based purely on predictive models to determine actionable elements of digital inequality.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

In spite of such contributions, there are a number of limitations that should be recognized. The data used to perform the analysis is based on county-level aggregate data, which can cause sub-county disparity in broadband access and socioeconomic vulnerability to go unnoticed. The sample size is rather small and restricts the potential usefulness of more complex machine learning models, and could restrict extrapolation outside of the area under study. Also, the measures of broadband availability do not comprehensively depict such dimensions as affordability, quality of services, or actual usage, which are essential components of digital inclusion. These constraints imply that the findings are supposed to be viewed as reflective of structural patterns instead of likely causal connections.

Future studies can be expanded to include smaller geographic scales, including census tract data, and include longitudinal data to study temporal patterns in the disparity in access to broadband. The analysis could be further enriched with the expansion of the feature sets that comprise the affordability, pricing, and indicators of digital literacy. Methodologically, subsequent research might be done on methods of causal inference or hybrid models that can integrate machine learning with the econometric in order to make more effective policy assessment. Comprehensively, the framework suggested in this study offers a viable and meaningful basis for research and evidence-based policymaking towards eliminating technology access inequality in the future.

## References

[1] M. Alenezi, S. Wardat, and M. Akour, "The Need of Integrating Digital Education in Higher Education: Challenges and Opportunities," *Sustainability*, vol. 15, no. 6, p. 4782doi: 10.3390/su15064782.

[2] M. de Clercq, M. D'Haese, and J. Buysse, "Economic growth and broadband access: The European urban-rural digital divide," *Telecommunications Policy,* vol. 47, no. 6, p. 102579, 2023/07/01/ 2023, doi: https://doi.org/10.1016/j.telpol.2023.102579.

[3] J. Chen and Z. Xu, "The Impact of the Digital Divide on Labor Mobility and Sustainable Development in the Digital Economy," *Sustainability*, vol. 16, no. 22, p. 9944doi: 10.3390/su16229944.

[4] T. H. Grubesic and E. Helderop, "California's digital divide and the specter of data uncertainty for evaluating broadband coverage," *Telematics and Informatics,* vol. 71, p. 101837, 2022/07/01/ 2022, doi: https://doi.org/10.1016/j.tele.2022.101837.

[5] R. Yao, W. Zhang, R. Evans, G. Cao, T. Rui, and L. Shen, "Inequities in Health Care Services Caused by the Adoption of Digital Health Technologies: Scoping Review," (in eng), *J Med Internet Res,* vol. 24, no. 3, p. e34144, Mar 21 2022, doi: 10.2196/34144.

[6] S. M. Asad, M. A. Hussain, R. M. Monim, and K. Islam, "Application of Machine Learning for Early Disease Diagnosis in Healthcare," *Cuestiones de Fisioterapia,* vol. 51, no. 3, pp. 332-355, 2022.

[7] A. Afzal, D. Firdousi, A. Waqar, and M. Awais, "The Influence of Internet Penetration on Poverty and Income Inequality," *SAGE Open,* vol. 12, p. 215824402211161, 08/17 2022, doi: 10.1177/21582440221116104.

[8] K. Amarasinghe, K. T. Rodolfa, H. Lamba, and R. Ghani, "Explainable machine learning for public policy: Use cases, gaps, and research directions," *Data & Policy,* vol. 5, p. e5, 2023, Art no. e5, doi: 10.1017/dap.2023.2.

[9] C. Fan, J. Xu, B. Y. Natarajan, and A. Mostafavi, "Interpretable machine learning learns complex interactions of urban features to understand socio-economic inequality," *Computer-Aided Civil and Infrastructure Engineering,* vol. 38, no. 14, pp. 2013-2029, 2023/09/01 2023, doi: https://doi.org/10.1111/mice.12972.

[10] P. Dhal and C. Azad, "A comprehensive survey on feature selection in the various fields of machine learning," *Applied Intelligence,* vol. 52, no. 4, pp. 4543-4581, 2022/03/01 2022, doi: 10.1007/s10489-021-02550-9.

[11] A. Palanivinayagam and R. Damaševičius, "Effective Handling of Missing Values in Datasets for Classification Using Machine Learning Methods," *Information*, vol. 14, no. 2, p. 92doi: 10.3390/info14020092.

[12] D. U. Ozsahin, M. T. Mustapha, A. S. Mubarak, Z. S. Ameen, and B. Uzun, "Impact of feature scaling on machine learning models for the diagnosis of diabetes," in *2022 International Conference on Artificial Intelligence in Everything (AIE)*, 2-4 Aug. 2022 2022, pp. 87-94, doi: 10.1109/AIE57029.2022.00024.
[13] H. M. Sozib *et al.*, "Cloud Computing in Business: Leveraging SaaS, IaaS, and PaaS for Growth," *Journal of Applied Research,* p. 38.
[14] J. R. Wilson, K. A. Lorenz, and L. P. Selby, "Introduction to Binary Logistic Regression," in
[15] A. P. Gopi, R. N. S. Jyothi, V. L. Narayana, and K. S. Sandeep, "Classification of tweets data based on polarity using improved RBF kernel of SVM," *International Journal of Information Technology,* vol. 15, no. 2, pp. 965-980, 2023/02/01 2023, doi: 10.1007/s41870-019-00409-4.
[16] A. Hassaan, M. M. Jamshaid, M. N. Siddique, Z. Akbar, and S. Niaz, "ETHICAL ANALYTICS & DIGITAL TRANSFORMATION IN THE AGE OF AI: EMBEDDING PRIVACY, FAIRNESS, AND TRANSPARENCY TO DRIVE INNOVATION AND STAKEHOLDER TRUST," *Contemporary Journal of Social Science Review,* vol. 1, no. 04, pp. 1-18, 2023.
[17] M. R. Abbasniya, S. A. Sheikholeslamzadeh, H. Nasiri, and S. Emami, "Classification of Breast Tumors Based on Histopathology Images Using Deep Features and Ensemble of Gradient Boosting Methods," *Computers and Electrical Engineering,* vol. 103, p. 108382, 2022/10/01/ 2022, doi: https://doi.org/10.1016/j.compeleceng.2022.108382.
[18] A. Pagliaro, "Forecasting Significant Stock Market Price Changes Using Machine Learning: Extra Trees Classifier Leads," *Electronics*, vol. 12, no. 21, p. 4551doi: 10.3390/electronics12214551.
[19] Y. Deng, M. R. Eden, and S. Cremaschi, "Metrics for Evaluating Machine Learning Models Prediction Accuracy and Uncertainty," in *Computer Aided Chemical Engineering*, vol. 52, A. C. Kokossis, M. C. Georgiadis, and E. Pistikopoulos Eds.: Elsevier, 2023, pp. 1325-1330.
[20] G. Naidu, T. Zuva, and E. M. Sibanda, "A Review of Evaluation Metrics in Machine Learning Algorithms," in *Artificial Intelligence Application in Networks and Systems*, Cham, R. Silhavy and P. Silhavy, Eds., 2023// 2023: Springer International Publishing, pp. 15-25.
[21] D. Chicco and G. Jurman, "The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification," *BioData Mining,* vol. 16, no. 1, p. 4, 2023/02/17 2023, doi: 10.1186/s13040-023-00322-4.
[22] S. Liu *et al.*, "Comparison of evaluation metrics of deep learning for imbalanced imaging data in osteoarthritis studies," (in eng), *Osteoarthritis Cartilage,* vol. 31, no. 9, pp. 1242-1248, Sep 2023, doi: 10.1016/j.joca.2023.05.006.
[23] M. Mijwil and M. Aljanabi, "A Comparative Analysis of Machine Learning Algorithms for Classification of Diabetes Utilizing Confusion Matrix Analysis," *Baghdad Science Journal,* vol. 21, 10/20 2023, doi: 10.21123/bsj.2023.9010.
[24] Y. Yang, Y. Yuan, Z. Han, and G. Liu, "Interpretability analysis for thermal sensation machine learning models: An exploration based on the SHAP approach," *Indoor Air,* vol. 32, no. 2, p. e12984, 2022/02/01 2022, doi: https://doi.org/10.1111/ina.12984.
[25] Z. Akbar, A. Hassaan, M. M. Jamshaid, M. N. Siddique, and S. Niaz, "Leveraging Data and Artificial Intelligence for Sustained Competitive Advantage in Firms and Organizations," *Journal of Innovative Computing and Emerging Technologies,* vol. 3, no. 1, 2023.
[26] K. Purohit, S. Vats, R. Saklani, V. Kukreja, V. Sharma, and S. P. Yadav, "Improvement in K-Means Clustering for Information Retrieval," in *2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 6-8 July 2023 2023, pp. 1239-1245, doi: 10.1109/ICESC57686.2023.10193031.
[27] E. Arenas-Arroyo, D. Fernández-Kranz, and N. Nollenberger, "High speed internet and the widening gender gap in adolescent mental health: Evidence from hospital records," IZA Discussion Papers, 2022.
[28] R. Dantu, "Assessing the Foundations: A Critical Review of the Public Advocates Office's Report on Broadband Pricing Trends in California."
[29] J. Dine and J. Kane, "The state of US broadband in 2022: Reassessing the whole picture," Information Technology and Innovation Foundation, 2022.
[30] N. Bell, P. Hung, A. Lòpez-De Fede, and S. A. Adams, "Broadband access within Medically Underserved Areas and its implication for telehealth utilization," (in eng), *J Rural Health,* vol. 39, no. 3, pp. 625-635, Jun 2023, doi: 10.1111/jrh.12738.