EduGap: A Machine Learning Powered Analysis of Digital Readiness and Effectiveness of Intervention Efforts

Fakiha Tariq

1. Abstract:

The digital divide remains a persistent and evolving barrier to equity in the modern age, particularly in rural and semi-rural populations. While most research highlights disparities in broadband access and device availability, fewer studies explore skill-based inequity in digital readiness across socioeconomic and demographic groups and evaluation of intervention efforts. This study introduces a machine learning driven framework designed to predict and analyze digital skill readiness and skill gains after interventions across three core areas: basic computer knowledge, internet, and mobile literacy. Using a synthetic dataset representative of populations in underserved regions, multiple models were evaluated to classify individuals as above or below "average" in each skill and skill gain, though for simplicity purposes the terms above and below average are used even though the median is used as the splitting point, not the mean. The median is used in order to improve equity and balance between the classes. A final pipeline using a MultiOutputClassifier with a RandomForest base achieved a testing accuracy of around .6, much greater from initial accuracies near .35, while avoiding overfitting. The model was able to reveal overarching trends through consistently identifying individuals with lower or no formal education, youth, early-career workers, unemployed or informally employed, and low and high income groups as the most underserved. Some of these same groups, however, demonstrated above-average skill gains, highlighting strong potential for targeted interventions, while other groups continued to fall below average, indicating groups requiring additional support. This combination of model-based predictions and group-level analysis demonstrates how machine learning can complement traditional digital divide research, serving as a foundation for more scalable, data-driven approaches to improving digital readiness and tailor future programs and policies.

2. Introduction

The digital divide is a phenomenon which has been commonly characterized by an increasing gap between groups of people due to the unequal access to telecommunication devices. Though, in recent years, the concept of the digital divide has been revised to consider not only access to technologies but also the ability to properly utilize them through having the proper skills and knowledge (Echauri, 2024).

With the fast paced adoption and emergence of new technologies, those who lack the opportunities to learn how to work their way around these technologies can be left behind, leading to them being at a disadvantage that the rest of the population. According to the research report released by The National Skills Coalition and the Federal Reserve Bank in Atlanta, 92% of jobs require workers to have digital skills but one-third of Americans lack the ground-level digital skills needed for these jobs. The report further stated that replacing a job that requires no digital skills to at least three can increase worker's pay by an average of 45% which can ultimately help contribute to a growth of state and federal tax revenue (National 2023). Furthermore, it's been found that some level of digital proficiency is needed for over 80% of middle-skill jobs (Kloza 2023). Without the proper amount of proficiency in digital skills, these individuals can ultimately miss out on these benefits and opportunities causing them to be further left behind.

Previous studies have been conducted on the topic of the digital divide and have revealed trends and patterns. Research by Sanders and Scanlon (2021) highlights how the lack of access to high-speed internet continues to disproportionately affect low income, rural, older, and other minority populations in the US. It stresses the idea that digital inequity is systematic and having access to the internet is a basic human right. Timotheou et al. (2022) discusses the acceleration of digitalization after COVID-19 but this leads to increased learning gaps and inequality in underprepared schools. They discuss how digital transformation is more than simple access or performance but a systematic process also about infrastructure, policies, funding, etc. Gallardo (2022) introduces the Digital Divide Index (DDI) which reveals that rural, low-income, veteran, and disabled communities are disproportionately represented in counties with high digital inequity. It highlights income and geography as persistent barriers to broadband access and device availability.

Unlike the past studies which have focused more on access to technologies and mainly income based groups, this paper focuses on pre-existing inequities in digital literacy/readiness and the effectiveness of digital literacy programs in order to inform targeted improvements through utilizing machine learning in

order to analyze which socioeconomic and demographic groups are disadvantaged throughout different digital skills and who performed well through the training and who may need more support. This deviates from traditional statistical studies which often examine correlations through linear models as it utilizes interactions across multiple variables to uncover complex, nonlinear relationships between socioeconomic and demographic factors and digital readiness. This study also focuses on more socioeconomic groups than just low income and age which are common factors observed in previous studies. I created two models, one for analyzing and predicting pre-training inequities throughout groups and one for post-training analysis for program effectiveness. I utilized a supervised learning model, which is when a model is trained through labeled datasets, data where each input is paired with an output which the model uses to learn, in order to make predictions. The prediction performance of the model is evaluated through different metrics, such as accuracy, F1 score, precision, and recall. Though, more importantly, the patterns of inequity are revealed through various graphs and scores illustrating who to target for inclusivity efforts and how to improve these interventions. I hypothesized that low income groups, low or no education, unemployed, and seniors would be more at a disadvantage.

3. Methods

3.1 Dataset and Features

For this study, I utilized a synthetic dataset of 1,000 users found on Kaggle which was generated to reflect statistical patterns over real-world digital readiness for rural and semi-rural populations. Even though it's not drawn from actual individuals, the data was constructed to stimulate realistic distributions, enabling safe exploration of inequity patterns using machine learning. Table 1 lists the columns from the dataset I used for the features and outputs. It includes socioeconomic, demographic, behavioral, and engagement data along with digital skill scores prior to and post training. For my first model for pre-existing digital inequity, I utilized solely socioeconomic or demographic features including education levels, household income, employment status, and ages. For my labels, I used 3 digital skills to focus on: basic computer knowledge score, internet usage score, and mobile literacy score. For my second model for program evaluation, I used the same features as my first model but with the addition of behavioral and engagement features such as modules completed, engagement level, session count, and average time per module. For the labels, in order to measure growth of skill and knowledge, a new column was created by finding the difference between post training and pre training score columns.

Some features that seem to be significant for finding deeper insights have some limitations to them in being features for model-prediction due to data leakage and possibilities of self-reporting. Instead, they were analyzed as individual variables in terms of visuals and statistics to work as supporting information for analysis purposes. These variables are adaptability score, skill application, and overall literacy score,

and they work to add more depth to my results and analysis. Employment impact was left out from being one of the supporting variables due to concerns about interpretability and potential confounding variables.

Table 1. List of features used for the model from the Kaggle dataset. Table created by student researcher through Google Docs, 2025. Data source: Digital Literacy Education Dataset (Ziya, 2025).

Column Name	Description
Age	Age of user in years
Education_Level	Highest level of formal education completed (e.g., Primary, High School)
Employment_Status	Current employment situation (e.g., Student, Farmer)
Household_Income	Approximate income level (e.g., Low, Medium)
Basic_Computer_Knowledge_Score	Score reflecting basic computer literacy skills
Internet_Usage_Score	Score reflecting basic internet usage skills
Mobile_Literacy_Score	Score reflecting competence with mobile devices and apps
Post_Training_Basic_Computer_Knowledge_Score	Score reflecting basic computer literacy skills after training
Post_Training_Internet_Usage_Score	Score reflecting basic ability with internet skills after training
Post_Training_Mobile_Literacy_Score	Score reflecting competence with mobile devices and apps
Modules_Completed	Number of educational modules completed in the program
Average_Time_Per_Module	Average time spent per module in minutes
Session_Count	Number of sessions attended
Engagement_Level	Categorical indicator of engagement (e.g., Low, Medium, High)

3.2 Data Preprocessing

I began by checking for null values and I found my education level column had some. I reviewed my dataset to find all the values for “None” were being translated as null so I filled those “null” values with “No School” which gave me a clean dataset to work with.

Next, I turned my categorical features into numerical features through the process of feature mapping. Feature mapping allows for encoding the categorical values so it can be analyzed by the machine learning model and allows for the model to capture complex relationships by enhancing underlying patterns. For example, for education level, “No School” was encoded as 0, “Primary” was encoded as 1, “Secondary” was encoded as 2, and so on. This allowed the model to interpret relative educational attainment as a feature and helped maintain a consistent structure across the dataset. I used the bucketing process for the age feature, which divides continuous data into bins in order to create categorical data, creating age groups prior to mapping in order to make it easier to evaluate inequity in the analysis process. Bins of 0-24, 24-40, 40-60, 60+ were created to create “Youth”, “Early Career”, “Midlife”, and “Senior” groups.

In order to effectively train the model on the categorical features for predicting digital readiness levels, a binary classification label was created based on participants’ scores. Individuals whose scores fell below or at the median were labeled as “Below Average” or class 0 and those above the median were labeled as “Above Average” or class 1. The median was chosen as the splitting point in order to have a clear and fair way of defining underserved and reflecting real-world skill gaps instead of the mean which can be greatly affected by outliers in the data. The terms “Below Average” and “Above Average” were used for simplicity purposes. The same was done for my second model with defining “Low Growth”, class 0, versus “High Growth”, class 1. Other quantile based splits were explored in order to improve prediction abilities for the underserved group, class 0, which will be discussed further on. Label generation occurred prior to model training in order to avoid information leakage and served as the target variable during supervised learning.

Class balance was checked prior to testing and the classes weren’t very imbalanced which allowed me to skip using Synthetic Minority Oversampling Technique or any similar class balancing method. Results of the class distribution are shown in Table 2. Multicollinearity, or the occurrence of two or more explanatory variables being very highly linearly related which can lead to misleading conclusions, was checked through using Variance Inflation Factor (VIF). If any features return a VIF value greater than 5, it

suggests high multicollinearity present but this wasn't an issue as none of features returned a value higher than 2.87.

Table 2 Class distribution values per label. Table created by student researcher through Google Docs, 2025.

	Computer_Skill_Label class label distribution	Internet_Usage_Label class label distribution	Mobile_Literacy_Label class label distribution	Computer_Skill_Gain_Label class label distribution	Internet_Gain_Label class label distribution	Mobile_Gain_Label class label distribution
Class 0	.506	.51	.504	.506	.528	.505
Class 1	.494	.49	.496	.494	.472	.495

3.3 Model Development and Optimization

To train my model, the dataset was split randomly into two categories. 80% of the data was used for training while the other 20% was used for testing. The training set was used to train the model while the testing dataset was set aside for an unbiased assessment of the model's performance.

Building the model involved an iterative process of training, evaluating, and refining a machine learning pipeline to predict whether individuals were below or above average in digital readiness and skill growth. Initially, I used ordinal encoding for categorical features and more basic classifiers, with location type as one of my features instead of employment status. Multiple models including GradientBoostingRegressor, KNeighborsClassifier, GaussianNB, HistGradientBoosting, LogisticRegression, CatBoost, were tested for improving prediction accuracy and performance though they were underfitting as both training and testing accuracy was low. The accuracy started from as low as .35 for the testing data.

To improve performance, I reassessed feature selection and encoding. Location type was removed due to its lack of significance and replaced with employment status, which showed higher feature importance. I also transitioned from ordinal encoding to feature mapping, allowing the model to more meaningfully interpret categorical inputs.

Ultimately, the best-performing configuration used a MultiOutputClassifier with a RandomForestClassifier as a base estimator. MultiOutputClassifier works to evaluate the three digital

skills and skill gains through creating separate models using all the features (Geeks 2017). RandomForestClassifier creates a more accurate and stable prediction by combining predictions from multiple decision trees (Geeks 2020). It was utilized as the base estimator helps to prevent overfitting and doesn't require normalizing the features through StandardScaler because it has the ability to handle unscaled data. After hyperparameter tuning, including setting n_estimators = 200 and max_depth = 15 among others shown on Table 3, model accuracy improved significantly and was not overfitting, indicating better generalization. The accuracy after optimization increased to around .6 for the testing data.

Table 3 Hyperparameters for RandomForestClassifier. Table created by student researched through Google Docs, 2025.

Parameters	**Value**
n_estimators	200
max_depth	15
min_samples_leaf	4
min_samples_split	5
class_weight	'balanced'
max_features	'sqrt'
random_state	42
n_jobs	-1

Accuracy is only one of the multiple metrics that can be used to evaluate classification models and by itself it can be misleading. For a more nuanced evaluation of my classification model, it was evaluated consistently using precision, recall, and F1 score along with accuracy in order to assess not only performance but also class balance and fairness in order to not look over class 0, or the underserved individuals, the focal point of this study. This made recall especially important to maximize in order to ensure the model correctly captured as many truly underserved individuals as possible, critical in an equity-focused study. Two label definitions were used in order to do so. To fairly identify through the patterns the model found which socioeconomic or demographic groups were most digitally underserved,

binary labels were generated using median digital skill scores for its ability to be resistant to outliers, unlike mean, in order to split individuals into “below average” and “above average” groups, these terms are used for simplicity. Though, another label was created for improving classification accuracy and model predictive performance through using quantiles which allowed for the model to better identify the two classes, especially class 0. These metrics helped assess class balance and avoid neglecting the group the model was intended to identify and support.

Accuracy uses both true positives and true negatives in order to measure the proportion of correctly predicted instances. It’s easy to interpret and understand. (Geeks, 2023)

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision is the proportion of true positive predictions, which are correct positive predictions) out of all the positive instances. It gives insight into how well the model can predict positives accurately. (Geeks, 2023)

Precision = TP / (TP + FP)

Recall measures how correctly a model can identify all positive instances through proportion of true positive predictions, which are positive instances that were predicted correctly, out of all the positive instances. (Geeks, 2023)

Recall = TP / TP + FN

F1 score is a single value that combines both precision and recall in order to provide a balanced assessment of how well a model performs. (Geeks, 2023)

F1 Score = 2x ((Precision * Recall) / (Precision + Recall))

3.3 Underserved Group and Program Effectiveness Identification

After the model was trained to classify individuals, the next step was to apply those predictions to uncover patterns of inequity across socioeconomic groups in the classification. In order to have a way to evaluate the model’s accuracy of the trends it found, a custom function was developed that aggregated raw data by group categories such as income level, employment status, etc, and calculated the proportion of individuals predicted to be below and above average within each group. This allowed for the identification of groups that lacked basic digital skills the most vs the least and groups that had below vs above average skill gain.

Another function was developed in order to visualize which demographic subgroups were most affected by low digital readiness as predicted by the models which calculated and displayed the proportion of individuals predicted to be below average throughout the different categorical features. The function iterates through selected socioeconomic/demographic columns and computes the percentage of "below average" classifications within each subgroup using normalized value counts. For each socioeconomic/demographic feature, the function selects the top n (in my case 2) subgroups with the highest proportions of predicted below average individuals. These are then visualized in a bar chart where each bar represents a subgroup and the y-axis shows the proportion of "below average" predictions within each group. This visualization aids in identifying the most underserved populations as recognized by the model and helps prioritize which groups may need the most targeted support. It also works for identifying and visualizing below average gain allowing for finding groups that may need extra support.

Lastly, feature importance analysis was conducted to quantify how much each input variable contributed to the prediction of digital readiness outcomes. These scores were visualized using a horizontal bar chart where longer bars represent features with greater influence. This provides a clear visualization of which features have the highest importance which suggests they played the most significant role in determining whether an individual was predicted to be below or above average in digital readiness and skill growth. In contrast, features with lower importance contributed less to the model's decisions. This form of visualization adds interpretability to the model, helping to reveal which socioeconomic/demographic factors most strongly shape digital inequity patterns, a crucial step in guiding real world inclusivity efforts.

3.4 Supporting Variable

The dataset included variables that wouldn't work well as features for the model but still have the potential for contributing insightful information for the study and its purposes. These variables were utilized through grouping them with each of the main features and visualizing them with bar graphs. The bar graphs provided insights through graphing the average value of the variable within each of the subgroups, like low, medium, and high income, and giving the opportunity to compare the average value across the subgroups.

4. Results

4.1 Model Accuracy and Fairness

Table 4 and 5 displays the performance of the two models, pre-training skills and skill growth respectively, on the training set through precision, recall, F1, and accuracy score for each label. For basic computer skills, the overall accuracy was .56 with class 0 having a recall score of .60 and precision score

of .64. For internet usage skills, the overall accuracy was .625 with class 0 having a recall score of .66 and precision score of .66. For mobile literacy skill, the overall accuracy was .615 with class 0 having a recall score of .65 and precision score of .74. The moderately strong recall rate shows that the model is doing well at catching underserved individuals and the moderately strong precision scores show that it's doing well at not mislabeling people as well.

For computer skill gain, the overall accuracy was .53 with class 0 having a recall score of .60 and precision score of .57. For internet usage skill gain, the overall accuracy was .52 with class 0 having a recall score of .60 and precision score of .55. For mobile literacy skill gain, the overall accuracy was .51 with class 0 having a recall score of .56 and precision score of .59. While overall accuracy of the model remains modest, the focus on underserved individuals, or class 0, the model achieved with reasonable success.

Table 4 Results of metrics for all three skill labels for pre-training score model. Table created by student researcher through Google Docs, 2025.

	Basic Computer Skills		Internet Usage		Mobile Literacy	
	Class 0	Class 1	Class 0	Class 1	Class 0	Class 1
Recall	.64	.46	.66	.58	.74	.44
Precision	.60	.51	.66	.58	.65	.54
F1 Score	.62	.48	.66	.58	.69	.48
Overall Accuracy	.56		.625		.615	

Table 5 Results of metrics for all three skill labels for skill gain model. Table created by student researcher through Google Docs, 2025.

	Computer Skill Gain		Internet Usage Gain		Mobile Literacy Gain	
	Class 0	Class 1	Class 0	Class 1	Class 0	Class 1
Recall	.57	.47	.55	.48	.59	.42
Precision	.60	.44	.60	.43	.56	.45
F1 Score	.59	.45	.57	.45	.57	.43
Overall Accuracy	.53		.52		.51	

4.2 Feature Importance

Figure 1 and 2 illustrates the significance of the different features involved in the models. The most significant from the pre-training underserved groups model were employment status then education level with income level being the least important. In the skill gain model, the most significant features were average time per module then session count.

Figure 1 Graph of features importance for pre-training skill model. Labels from top to bottom: Education_Level, Household_Income, Employment_Status, Age_Group. Graph created by student researcher using Python (matplotlib), 2025.

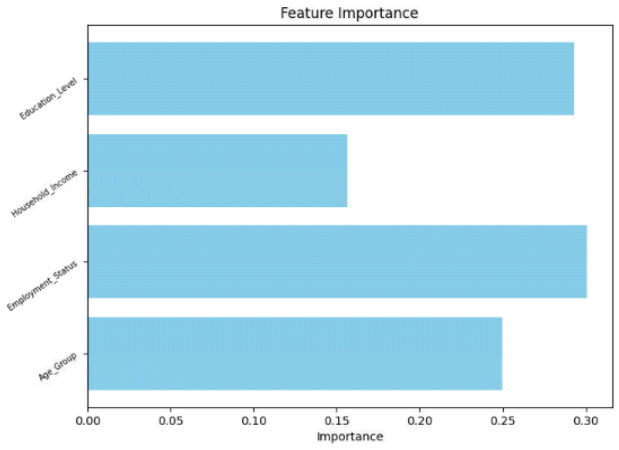
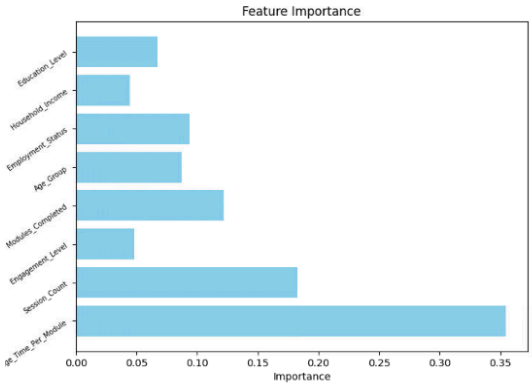


Figure 2 Graph of features importance for skill gains model. Labels from top to bottom: Education_Level, Household_Income, Employment_Status, Age_Group, Modules_Completed, Engagement_Level, Session_Count, Average_Time_Per_Module. Graph created by student researcher using Python (matplotlib), 2025.



4.3 Identifying the Pre-Training Underserved Groups

Figure 3, 4, 5 illustrate the top 2 sub-groups predicted to have the least digital literacy in basic computer knowledge, internet usage, and mobile literacy skills respectively. For pre-training underserved groups, the overall most common groups predicted to be the most below average across skills for education level are no school and primary. For income, they are low and high income. In employment status, it's unemployed, self-employed, and farmers. For age group, they are youth and midlife. Based on raw data calculations, the actual most common groups with the highest amount below average across skills for education level are primary and high school. For income, they are low and high income. In employment status, it's unemployed, self-employed, and farmers. For age group, they are youth and early career.

Figure 3 Graph of two most below average subgroups predicted by model for basic computer skills. Labels are from left to right: Education_Label: Primary, Education_Label: Secondary, Income_Label: Low, Income_Label: High, Employment_Level: Unemployed, Employment_Level: Self_Employed, Age_Label: Youth, Age_Label: Early Career. Graph created by student researcher using Python (matplotlib), 2025.

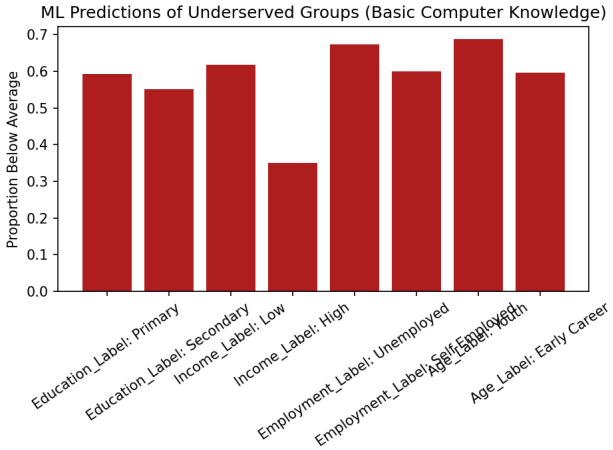


Figure 4 Graph of two most below average subgroups predicted by model for internet usage skills. Labels are from left to right: Education_Label: No School, Education_Label: Primary, Income_Label: High, Income_Label: Low, Employment_Level: Self-Employed, Employment_Level: Farmer, Age_Label: Senior, Age_Label: Midlife. Graph created by student researcher using Python (matplotlib), 2025.

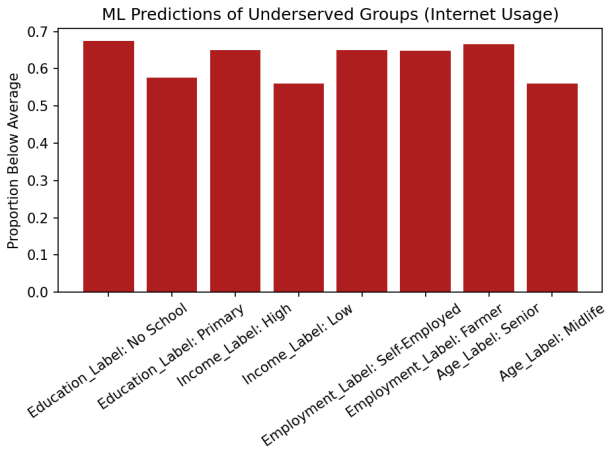
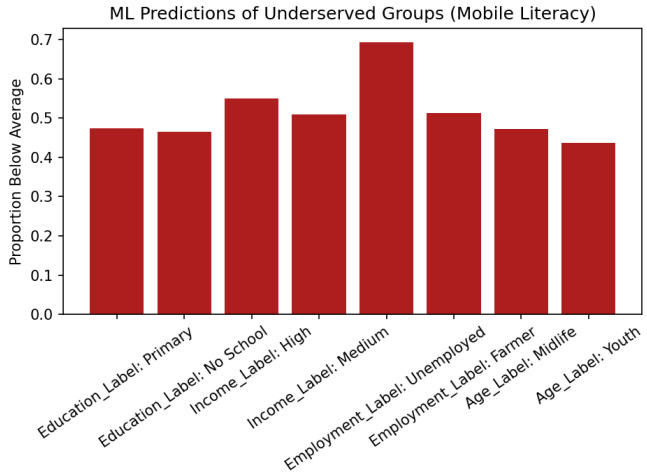


Figure 5 Graph of two most below average subgroups predicted by model for mobile literacy skills. Labels are from left to right: Education_Label: Primary, Education_Label: No School, Income_Label: High, Income_Label: Medium, Employment_Level: Unemployed, Employment_Level: Farmer, Age_Label: Midlife, Age_Label: Youth. Graph created by student researcher using Python (matplotlib), 2025.



4.4 Effect of Intervention Efforts on Groups

Figures 6, 7, 8 illustrate the top 2 sub-groups with the least digital literacy growth after training efforts in basic computer knowledge, internet usage, and mobile literacy skills respectively. For post-training underserved groups, the overall most common groups predicted to be the most below average across skill gain for education level is no school. For income, they are low and medium income. In employment status, they are unemployed and farmers. For age, they are youth and senior career. Based on raw data

calculations, the actual most common groups with the highest amount below average across skills for education level are primary and high school. For income, they are low and high income. In employment status, it's unemployed, self-employed, and farmers. For age group, they are youth and early career.

Figure 6 Graph of two most below average subgroups predicted by model for basic computer skill gain. Labels are from left to right: Education_Label: Secondary, Education_Label: No School, Income_Label: High, Income_Label: Medium, Employment_Level: Unemployed, Employment_Level: Self_Employed, Age_Label: Midlife, Age_Label: Senior. Graph created by student researcher using Python (matplotlib), 2025.

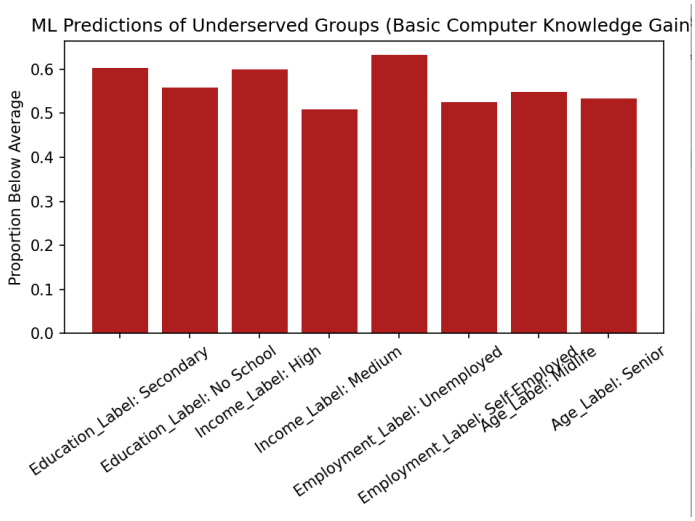


Figure 7 Graph of two most below average subgroups predicted by model for internet usage skill gain. Labels are from left to right: Education_Label: High School, Education_Label: No School, Income_Label: Low, Income_Label: Medium, Employment_Level: Farmer, Employment_Level: Unemployed, Age_Label: Youth, Age_Label: Senior. Graph created by student researcher using Python (matplotlib), 2025.

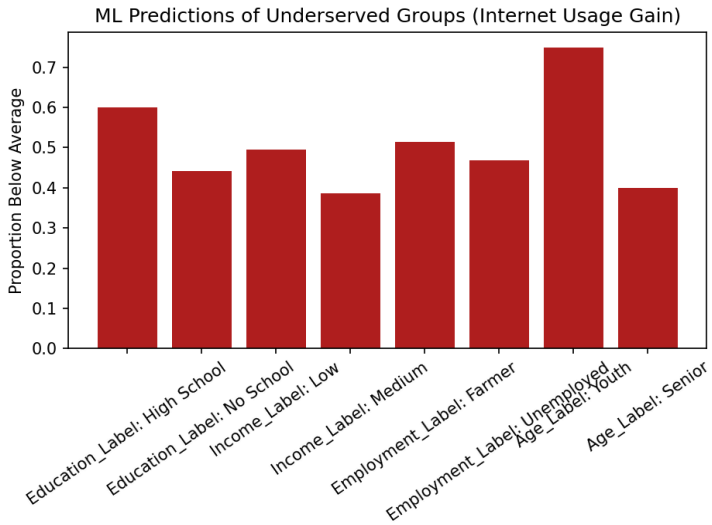
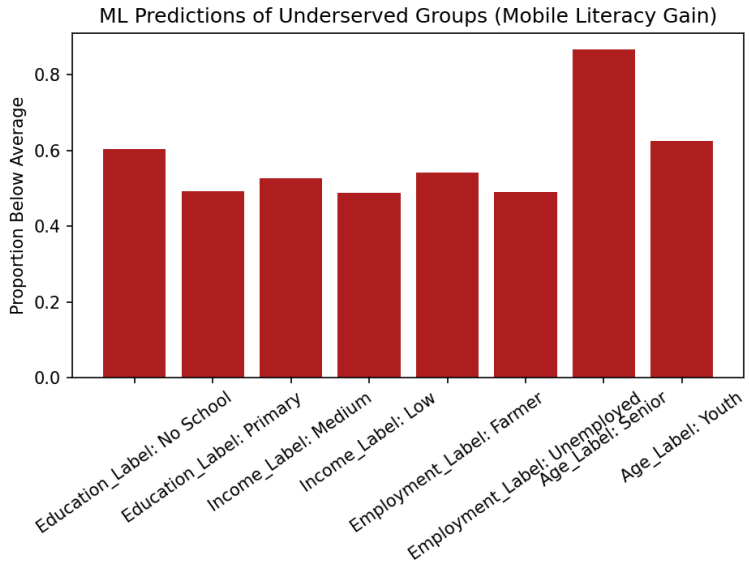


Figure 8 Graph of two most below average subgroups predicted by model for mobile literacy skill gain. Labels are from left to right: Education_Label: No School, Education_Label: Primary, Income_Label: Medium, Income_Label: Low, Employment_Level: Farmer, Employment_Level: Unemployed, Age_Label: Senior, Age_Label: Youth. Graph created by student researcher using Python (matplotlib), 2025.



5. Discussion

5.1 Interpreting Model Performance

The classification model trained using a quantile based labeling approach was used for maximizing accuracy, recall, and precision for class 0, individuals below average, in order to enhance its predictive abilities while reducing bias than the median based labeling approach model used for having a standard way of comparing inequity patterns across the three digital literacy outputs.

Furthermore, this analysis focuses primarily on the model's ability to identify individuals in class 0, reflecting the study's central aim of uncovering inequities and targeting support toward participants with the lowest digital readiness or gains.

Overall testing accuracy was moderately strong for the pre-training skill model, trained to predict skill scores prior to training, suggesting that it can be somewhat reliable in predicting individuals' baseline digital readiness based on demographic and socioeconomic features. While not highly precise, this level of performance is reasonable given the complexity of digital readiness, potential noise, or other limitations discussed in section 5. Nonetheless, the model does well in identifying groups who are more

likely to begin with overall lower digital skills, making it a strong foundation for future inequity determining models.

In the pre-training skill model, class 0, which represents individuals below average, consistently achieved recall scores above .64 and precision scores above .6. This indicates the model is especially effective at identifying vulnerable individuals who are more likely to need additional support due to their lack of digital readiness, the focus of this study.

For the skill gain model, trained to predict participants' improvement after training, it achieved a lower overall testing accuracy as compared to the pre-training model. This could be expected due to the complexity of modeling human outcomes which are influenced by a variety of unmeasured features such as motivation and learning environment. Despite this, the model consistently identified low-gain individuals, class 0, moderately well, suggesting that the model is directionally useful for potentially flagging individual participants who may not benefit as strongly from current training programs. Although the model's predictive power is more limited, it remains valuable for identifying growth potential and its performance highlights that digital skill gains may be more influenced by complex, often unobserved factors.

5.2 Equity Gaps Across Groups + Feature Importance Review

The equity analysis models used a median-based threshold to define "below average" and "above average" performance. Although median splits may be less sensitive in skewed distributions, they offer a highly interpretable and standardized reference point. The median doesn't rely on assumptions about shape distribution and provides a consistent benchmark across groups, making it a useful tool for identifying equity gaps, while using quantiles would be changing/defining what is considered "below average."

In employment status, based on the results, unemployed, self-employed, and farmers were the most common groups to fall the most below average across the three skills. The model was able to determine similar findings as it also revealed self-employed, farmers, and unemployed struggled the most. Also considering this is the most significant factor in model prediction of pre-training digital skills, this indicates that being unemployed or working a job with limited access to technology, digital training opportunities, and environments can highly lead to people to fall behind in terms of digital literacy. On the other hand, a student or someone working in a professional workplace has access to structured training and digital upskill opportunities allowing them to have a greater level of digital readiness.

Education levels resulted in no school and secondary being the most common groups to fall the most below average. The model predicted primary and no school were considered as underserved groups, successfully identifying the no school group but failing to catch the secondary group. The model underestimated the primary education level group, potentially because some primary-educated individuals have more access to digital literacy through school opportunities to learn and use digital skills, factors which weren't explicitly captured in the dataset. This portrays how unmeasured contextual variables can lead to discrepancies between predicted and observed readiness. Educational level is also the second most important predictor for pre-training digital skills, highlighting the important role gaining an education can have in shaping an individual's ability to access and develop digital skills. Since lower education levels fell behind, this serves to show that higher education allows for opportunities for students to gain foundational knowledge and experience with tools and technologies to encourage higher levels of digital literacy. People who spent less time gaining an education, and in result have lower educational attainment and had less digital literacy opportunities, gained less digital literacy skills than their higher educated counterparts.

Age, the third most important predictor of digital skill baseline, had youth and early career as its most common groups to be most below average. The model successfully predicted youth groups being underserved but incorrectly assumed senior groups were also underserved. This could reflect limited representation of seniors in the dataset or the inclusion of seniors actively using technology and maintain high digital readiness. This suggests the model is sensitive to sample distribution and may require additional features to better capture generational differences in technology use. Generally, it's considered that younger participants are more exposed to digital tools in comparison to older adults who may have a more difficult time adapting to the new technologies. Youth being one of the underserved groups could indicate that the younger generation might struggle with practical digital literacy skills such as internet usage and basic computer skills while being better at using mobile since they were not the most below average group for mobile literacy. This indicates that initiatives need to be taken to target the youth in order to teach them applied, career-oriented skill development. Being in the earlier part of a career can indicate not having much exposure to advanced digital literacy opportunities or training which can cause them to be at more of a disadvantage as compared to midlife or seniors who might've been exposed to these training and technologies in earlier times in their life.

Lastly, income was the least important predicting factor with, unexpectedly, both low and high income groups as the most common below average groups. The model was able to accurately predict these groups

as underserved. Due to this feature being least influential in the model's predictions, this could reflect that within high-income brackets factors such as educational level or employment type could have led to lower digital readiness levels since the previously discussed factors tend to have a much stronger tendency to determine baseline digital skills causing the appearance of both low and high income as disadvantaged groups.

Overall, my results indicated that employment status and education levels are significant factors to determine where to target intervention efforts since being in an environment where structured digital literacy learning or growing opportunities aren't present can cause individuals to be more prone to being left behind and struggling to learn these digital literacy skills. This also reveals machine learning's ability to identify general trends across the dataset. Any mispredictions highlight the areas where future data collection and feature design can better reflect real-world conditions. Understanding these discrepancies is crucial for designing targeted interventions and improving predictive fairness, ensuring that underserved populations are accurately identified.

5.3 Evaluating Intervention Efforts and Outcomes

The skill gains model follows the same median based split for evaluating "below average" and "above average" gains. This analysis looks to evaluate groups that were able to do well with the intervention results from the dataset and groups that still need some extra help in order to reach the same level as other groups.

Reference Table 5 for a breakdown of individual underserved groups and their general supporting variable results. Using the calculations for the amount below and above average skill gains, from the disadvantaged groups identified, secondary schooling, farmers (to an extent), and early career groups were able to have some of the highest gains in their according digital literacy skill showing how they went from some of the most below average starting scores to having some of the greatest above average growth. The model predicted farmers to have one of the most low gains, underestimating their growth. Though, the model successfully not underestimating growth for secondary and early career groups portrays a critical opportunity that underserved populations, when engaged effectively, are capable of meaningful digital advancement and responsiveness to training despite their low starting points. Though, considering the supporting variable bar graphs these groups still have room for literacy improvement in comparison to other groups which have higher levels of digital literacy by the end of the program. These groups also tend to be better at adjusting to new digital tools and skills, showing their potential for more digital challenges and growth and how these programs can grow the individuals in these groups, but they

tend to be worse at applying the skills they learn indicating future programs may need to start integrating more real-world simulations or practical tasks.

The model successfully identified unemployed, youth, low income, and no schooling to have the lowest gains which parallels the observed patterns in the data proportions. This depicts the continued need for more and better interventions due to the barriers that limit these groups that started more behind from having deeper or sustained growth. Continued efforts to target these groups who seem to struggle more in improving these digital skills can help empower them to be on the same level as groups that aren't as disadvantaged and gain more opportunities. These groups tend to be better at applying than adapting which could be due to feeling overwhelmed with a lot of new tools or skills or the training style may have been too task-specific so they couldn't have built broader digital problem solving skills. Future training should teach more conceptual understanding and pattern-recognizing across tools in order to boost adaptability and overall digital literacy scores.

Table 5 Breakdown of the underserved groups and their general adaptation, application, and overall literacy results. Table created by student researcher through Google Docs, 2025.

	Adapting	Applying	Overall Literacy
Unemployed	Low	High	Low
Low-Income	High	Low	Low
Youth	Low	High	Low
No School	High	High	Low
Secondary Schooling	Low	Low	Low
Farmer	High	Low	Low
Early Career	High	Low	Low

It's also important to address that another reason these groups consistently exhibit lower average scores in adaptability, skill application, and overall literacy can be due to the fact that these groups started with lower digital literacy baselines and not because they aren't making progress. Groups with higher prior exposure to technology, often correlated with income, employment, and education, may find it easier to navigate and apply digital skills causing them to have higher post-training scores. It's necessary to

recognize that these groups are getting benefit and progress through these programs, as shown from the lack of drastically different average score values between subgroups, but due to them starting at more of a disadvantage, these groups specifically are going to need some more extra, targeted support to reach the same level and strengthen their skills.

Though the model may not accurately predict every underserved group within each skill, it has the ability to reveal overall trends as it does well in identifying groups that most commonly tend to start or stay behind across the different skillsets. These results work to affirm the value of machine learning in surfacing nuanced inequity patterns that may not be clear through traditional statistical summaries. By combining model predictions and focusing on not only pre-training digital inequities in literacy levels but also gain-based analysis, this study provides a scalable, data-driven approach to identifying not only who is underserved, but the ability for these groups to go through improvements through these interventions and which groups may need some extra support. These findings can inform the design and evaluation of future programs, ensuring resources are allocated where they are both most needed and most impactful.

6. Limitations

While this study offers meaningful insights into patterns of digital inequity using machine learning, there are several limitations that should be acknowledged.

First, the dataset used was synthetic and not collected from real participants along with focused solely on rural and semi rural population. Although the dataset was designed to mirror realistic demographic and behavioral distributions, its simulated nature may limit the generalizability of the model's predictions to real-world populations, especially urban populations though the focus on rural and semi rural groups was good for the goals of this study. Future work incorporating datasets with real-life participants' data would provide stronger external validity and allow for deeper evaluation of intervention outcomes.

Second, the model's predictive accuracy, even though it was improved through iterations, remained moderate. This reflects the complexity of the task and it also suggests that some patterns of digital readiness and growth may not be fully captured by the available input features, especially for skill growth which may require more environmental and engagement features.

Third, the use of quantile thresholds to define "above average" or "below average" was efficient for my dataset with a lack of pre-labels for the digital literacy skills and model goals but it may oversimplify the

spectrum of digital skill progression. Cut-off based labeling holds the possibility of distorting some group comparisons.

Lastly, while the model revealed underserved groups and relative skill gains, it could not account for external variables such as quality of training, local infrastructure, language barriers, or motivations. These are additional factors that can have a significant impact on digital literacy outcomes. These elements must be considered in future studies to develop a more holistic view of digital equity.

7. Conclusion

This study used machine learning to portray its potential to identify and analyze nuanced patterns of digital inequity across socioeconomic and demographic groups. By classifying individuals below or above average in digital readiness and skill gain, the model successfully revealed that individuals with lower levels or no education, unemployment or informal employment, younger, and low and high income groups are most significantly underserved. Though, it tended to underestimate midlife and senior groups.

Though the predictions also accurately indicated that multiple of these same groups, like early career and secondary schooling, showed high potential for improvement by not underestimating their growth. This suggests that targeted interventions have the possibilities for producing meaningful impact if properly implemented. Though, the model was able to reveal groups like unemployed, youth, low income, and no schooling need to continue to be targeted when it comes to intervention efforts because they, having the potential to grow too, still struggle with lower literacy levels. The combination of model based prediction and group level analysis allowed a scalable method for identifying inequity.

While limitations such as synthetic data and modest accuracy need to be acknowledged, this research lays important groundwork for future real-world applications. It demonstrates how machine learning can support data-driven policy decisions, continuous assessment, and targeted program development aimed at narrowing the digital divide. With further refinement and real-world validation, this project can become a powerful asset in the pursuit of digital equity.

8. Bibliography

Echauri, G. (2024). *The Evolution of the Digital Divide: New Dimensions of Digital Inequality | Platypus*. Castac.org. https://blog.castac.org/2024/11/the-evolution-of-the-digital-divide-new-dimensions-of-digital-inequality/

Gallardo, R. (2022, August 17). *The State of the Digital Divide in the United States – Purdue Center for Regional Development*. Purdue Center for Regional Development – Purdue Center for Regional Development. https://pcrd.purdue.edu/the-state-of-the-digital-divide-in-the-united-states/

GeeksforGeeks. (2023, October 16). *Classification Metrics using Sklearn*. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/sklearn-classification-metrics/

GeeksforGeeks. (2017, July 20). *Multiclass classification using scikitlearn*. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/multiclass-classification-using-scikit-learn/

GeeksforGeeks. (2020, September 4). *Random Forest Classifier using Scikitlearn*. GeeksforGeeks. https://www.geeksforgeeks.org/dsa/random-forest-classifier-using-scikit-learn/

Kloza, B. (2023, February 27). *Impact of the Digital Divide: Economic, Social, and Educational Consequences - Connecting the Unconnected*. Connecting the Unconnected. https://ctu.ieee.org/blog/2023/02/27/impact-of-the-digital-divide-economic-social-and-educational-consequences/

National Skills Coalition. (2023, February 6). *New report: 92% of Jobs Require Digital skills, one-third of Workers Have Low or No Digital Skills Due to Historic underinvestment, Structural Inequities*. National Skills Coalition; National Skills Coalition. https://nationalskillscoalition.org/news/press-releases/new-report-92-of-jobs-require-digital-skills-one-third-of-workers-have-low-or-no-digital-skills-due-to-historic-underinvestment-structural-inequities/

Sanders, C. K., & Scanlon, E. (2021). The digital divide is a human rights issue: Advancing social inclusion through social work advocacy. *Journal of Human Rights and Social Work*, *6*(2), 130–143. https://doi.org/10.1007/s41134-020-00147-9

Timotheou, S., Miliou, O., Dimitriadis, Y., Sobrino, S. V., Giannoutsou, N., Cachia, R., Monés, A. M., & Ioannou, A. (2022). Impacts of digital technologies on education and factors influencing schools'

digital capacity and transformation: A literature review. *Education and Information Technologies*, *28*(28), 6695–6726. https://doi.org/10.1007/s10639-022-11431-8

Ziya. (2025, January 16). *Digital Literacy Education Dataset*. Kaggle. https://www.kaggle.com/datasets/ziya07/digital-literacy-education-dataset