ANALYZING THE DIGITAL DISTRESS IN ILLINOIS REGION

by

Saiprasanna Cheedepudi

B.Tech., Andhra University, 2018

A Thesis
Submitted in Partial Fulfillment of the Requirements for
the Master of Science Degree

School of Computing
in the Graduate School
Southern Illinois University Carbondale
December 2022

THESIS APPROVAL

ANALYZING THE DIGITAL DISTRESS IN ILLINOIS REGION

by

Saiprasanna Cheedepudi

A Thesis Submitted in Partial

Fulfillment of the Requirements

for the Degree of

Master of Science

in the field of Computer Science

Approved by:

Dr. Koushik Sinha, Chair

Dr. Bidyut Gupta

Dr. Henry Hexmoor

Graduate School
Southern Illinois University Carbondale
November 9, 2022

AN ABSTRACT OF THE THESIS OF

Saiprasanna Cheedepudi, for the Master of Science degree in Computer Science, presented on November 9, 2022, at Southern Illinois University Carbondale.

TITLE: ANALYZING THE DIGITAL DISTRESS IN ILLINOIS REGION

MAJOR PROFESSOR: Dr. Koushik Sinha

As the digital economy and society continue to grow, communities and individuals have a shot at improving their quality of life. Unfortunately, not everyone is being able to participate at the same functional level in this knowledge-based economy. Those on the wrong side of the digital divide are being left behind. This has prompted the focus on creation of strategies to ensure everybody can reap the benefits of the Internet revolution. For creation of such strategies, it is essential to understand the important factors that influence the adoption of the digital ecosystem. We have chosen the state of Illinois as a case study for this MS Thesis. We have explored various factors affecting digital inclusion in the context of two broad indicators: broadband internet subscription (or lack of), and device ownership (or lack of). For the specific goals of the MS thesis, namely, understanding the current state of broadband in southern Illinois, barriers to adoption, and developing a digital-divide elimination plan, we have utilized data published by the US Census Bureau's American Community Survey (ACS), Office of Broadband of the state of Illinois, and other publicly available datasets. Our approach uses a combination of data fusion and statistical as well as machine learning analysis to gain fine-grained insights into factors affecting broadband availability and broadband adoption in Illinois at both the County-level and county-tract level within counties. Our results clearly show that there is significant digital divide in Illinois and the rural regions suffers from significant deficiency in access to computing device and high-speed broadband Internet due to various demographic and economic factors.

Keywords: Digital Distress, Device Deficiency, Internet Insufficiency, Correlation ,Multi-class Machine Learning Classification Model, Tuning Hyperparameters, Cross Validation

ACKNOWLEDGMENTS

I would like to thank my supervisor Dr. Koushik Sinha for his supervision and mentorship. I would like to send my gratitude to the School of Computing. Additionally, I would like to express gratitude to committee members, Dr. Bidyut Gupta, and Dr. Henry Hexmoor.

DEDICATION

I dedicate this thesis to my father EdukondalaRao Cheedepudi and my mother KanakaDurga Cheedepudi.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE PAGE

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

Today, the Internet is increasingly making its presence felt, not only playing an important role in research and education but also serving as a catalyst to a country’s socio–economic, cultural, and political development. It is therefore not a surprise that the Internet has become a development of the highest significance.

The importance we attach to Internet development has naturally led us to notice widening gaps in technology between industrialized and developing nations, exacerbating an already significant moral and practical problem. The uneven adoption of modern information communication technologies (ICT) will further widen the division of "the information rich and information poor."

While the industrialized nations are pressing ahead with the Internet development, some of the less fortunate countries have yet to taste the fruits of this new technological invention. This is especially true for many regions where "universal access to basic communication and information services still remains a distant dream".

As the digital economy and society continue to grow, communities and individuals have a shot at improving their quality of life. However, not everybody is being able to participate in this digital age. Those on the wrong side of the divide are being left behind, prompting the creation of strategies to ensure everybody can reap the benefits of this new age.

Digital inclusion is one such strategy. The Office of the Comptroller of the Currency (OCC), when talking about bank financing for broadband initiatives in a recently released report, defined digital inclusion as "*the adoption of broadband technologies and its meaningful use for social and economic benefits*".

Being able to identify areas in dire need of digital inclusion efforts is an important step to take to ensure everybody participates and benefits of the digital age.

Multiple analyses exist that focus primarily on access to broadband of at least 25 Megabits per second download and 3 Megabits per second upload, or 25/3 for short, current broadband definition per the Federal Communications Commission. While useful to jump start conversations, these analyses miss an entire dimension of digital inclusion: how the technology is leveraged, or not.

Now — thanks to data from the U.S. Census Bureau American Community Survey (ACS) — a look at this dimension of digital inclusion is explored through looking at two broad indicators: type of internet subscription (or lack of) and device ownership (or lack of).

What is Digital Distress?

Digital distress areas have a harder time using and leveraging the internet to improve their quality of life due to the type of internet subscription or devices owned.

Digital distress is defined here as census tracts (neighborhoods) that had a 1) high percent of homes not subscribing to the internet or subscribing only through a cellular data plan and a 2) high percent of homes with no computing devices or relying only on mobile devices.

Four indicators were taken from the 2013–2017 U.S. Census Bureau ACS to measure digital distress for the nation's 72,400+ Census tracts:

- percent of homes with a cellular data only subscription
- percent of homes with no internet access (not subscribing)
- percent of homes relying only on mobile devices

- percent of homes not owning a computing device

Homes relying solely on cellular data or not subscribing to the internet are not benefiting from digital applications due to limited data plans and restricting access to outside the home. On the other hand, homes relying on mobile devices only or not owning computing devices make it harder to leverage digital applications due to smaller screens or no screens at all placing the home in digital distress.

Given that the subscription and computing device variables are not mutually exclusive, simply adding up all homes to obtain a final percentage was not possible. For this reason, variables 1 & 2 (see list above) were compiled into one percentage called subscriptions while variables 3 & 4 were compiled into one percentage called devices. Z-scores were then calculated for both aggregated percentages (subscriptions and devices), added together, and normalized to a score from 0 to 100, where a higher number denotes a higher digital distress. Tracts were considered digitally distressed if their score was greater than 50.

Figure below shows the breakdown for the U.S. and digitally distressed tracts for each of the four variables utilized. As expected, tracts in digital distress (orange bar) had a higher share of homes in each of these indicators. Almost half or 45.8 percent of homes in digitally distressed tracts did not access the internet versus 17.6 percent in the U.S. overall. Similar pattern can be seen with not owning computing devices where slightly more than one-third (34.4 percent) of homes in digitally distressed areas did not own a computing device compared to 12.8 percent in the U.S. overall.

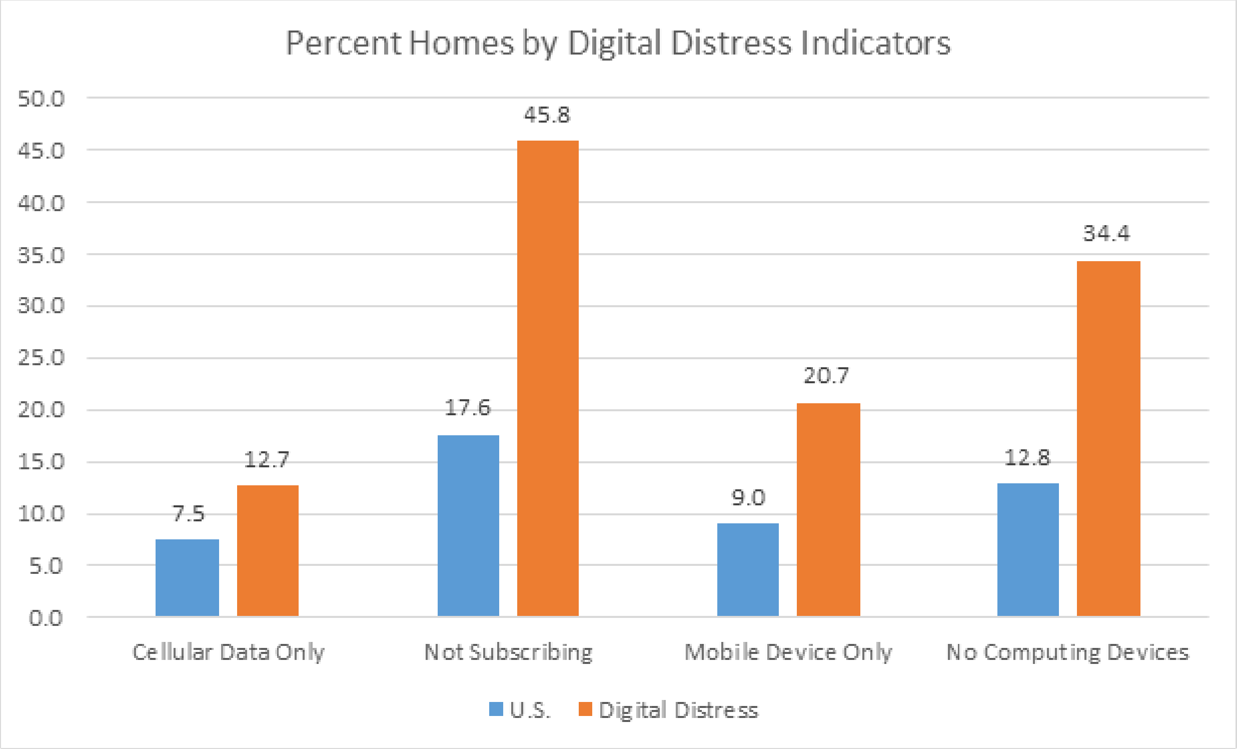


Fig.1 Percent Homes by Digital Distress Indicators

To calculate some of the parameters on the digital indicator dashboard, we defined some the parameters as follows:

1. **Connected with Device**: Cellular Internet only + Satellite + Broadband such as cable, fiber optic or DSL + Dial-up only. Our rationale for this definition stems from the fact that each of these different internet access methods require a personal device to access the Internet.
2. **Fully Disconnected**: Percent of Households with No Internet Access.
3. **Internet Insufficient**: Dial-up Internet only + Cellular data plan only + Without an Internet subscription. We have defined internet insufficiency as either the lack of Internet subscription, or access through only low data-rate dial-up connectivity or having only a cellular data subscription plan. Our rationale for including only cellular data subscription

plan in this definition is that cellular data plans are mostly associated with accessing the internet from a mobile device, which can turn out to be restrictive in terms of a exploring the full potential/experience of the Internet. For example, many websites do not render properly on a smartphone; certain tasks can be accomplished with far greater efficiency on a desktop or laptop. Moreover, most cellular data subscription plans are comparatively more expensive than fixed broadband subscription plans of similar data-volume/speed, and therefore a significant number of users use a metered cellular data plan that leads to a self-imposed restriction on their amount of time and type of Internet access per day (or month).

4. **Device Deficient**: Desktop or laptop with no other type of computing device + Smartphone with no other type of computing device + Tablet or other portable wireless device with no other type of computing device + No computing device. We have defined device deficiency as the lack of any Internet-access capable computing device or having only one such device. Our rationale for this definition is that only one type of device constrains a person from being able to carry out the full range of digital and/or online activities, whenever and wherever.
5. **Digital Distress**: The description of digital distress given by PCRD, Purdue University states "households with cellular data only or no internet as well as mobile only or no computing devices". Noting that most people with cellular data plan also use mobile devices for accessing the Internet, we first compute the percentage of the joint scenario: "People accessing the Internet using cellular data plan only and using mobile device only". To calculate this, we have defined a new measure called "Access Internet using cellular plan only, using mobile device only" as follows:

LetP_{mobile} = Mobile device only (smartphones, tablets and other portable devices), and P_{cell} = Cellular data plan only. Then, our new measure Access Internet using cellular plan only, on mobile device only is given by $P_{cell}^{mobile} = \frac{P_{cell}*P_{mobile}}{P_{cell}+ P_{mobile}}$.

In terms of specific columns from the ACS 2019 data, Access Internet using cellular plan only, on mobile device only = Cellular data plan only * (Smartphone with no other type of computing device + Tablet or other portable wireless computer with no other type of computing device) / (Cellular data plan only + Smartphone with no other type of computing device + Tablet or other portable wireless computer with no other type of computing device).

Using the above definition of the new measure “Access Internet using cellular plan only, on mobile device only” we then calculate our Digital Distress parameter as: Digital Distress = Access Internet using cellular plan only, on mobile device only + Dial-up with no other type of Internet subscription + Without an Internet subscription + No computer from the ACS 2019 data.

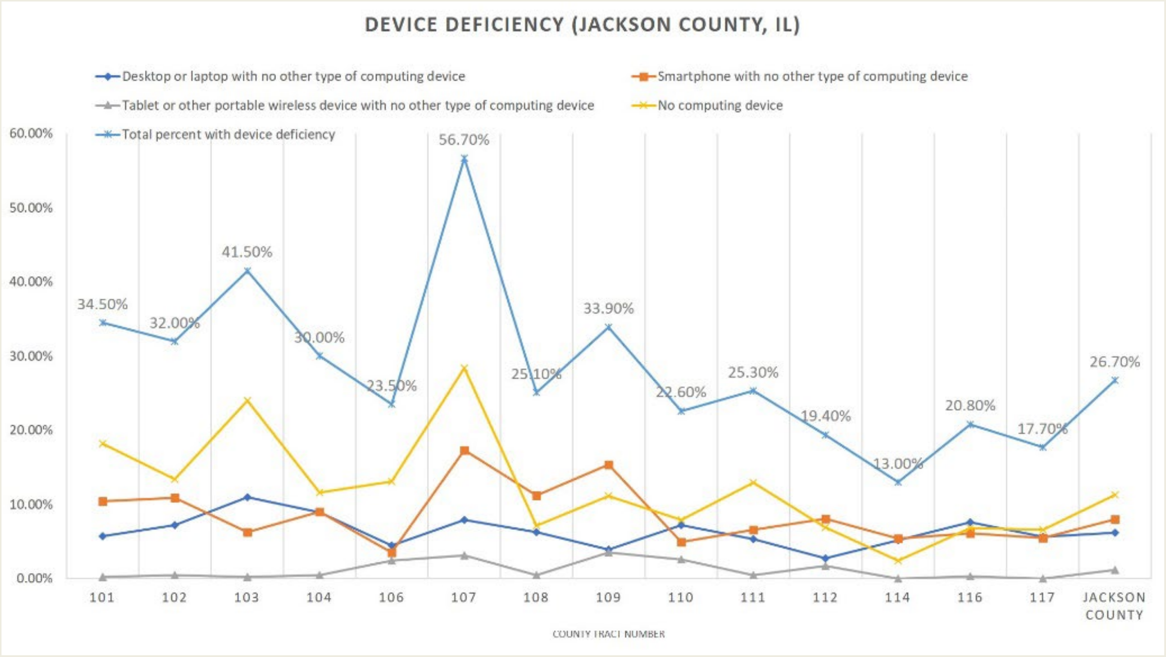


Fig.3 Device deficiency statistics at county-tract level, Jackson County

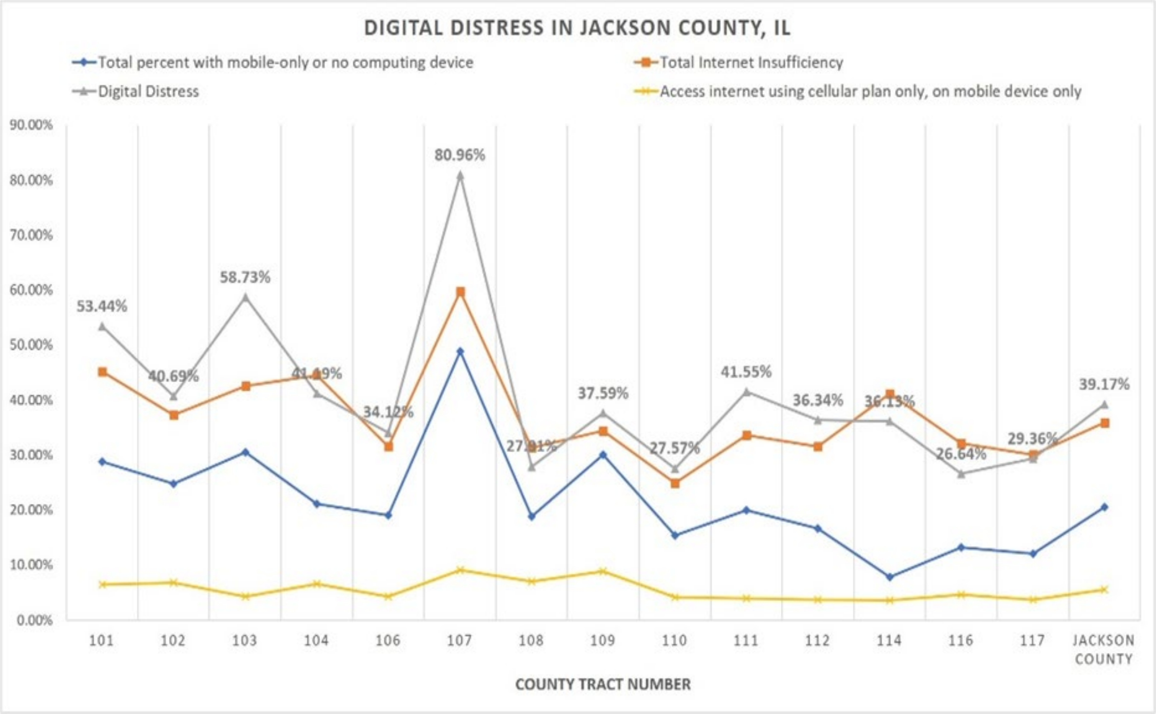


Fig.2 Digital Distress statistics at county-tract level, Jackson County

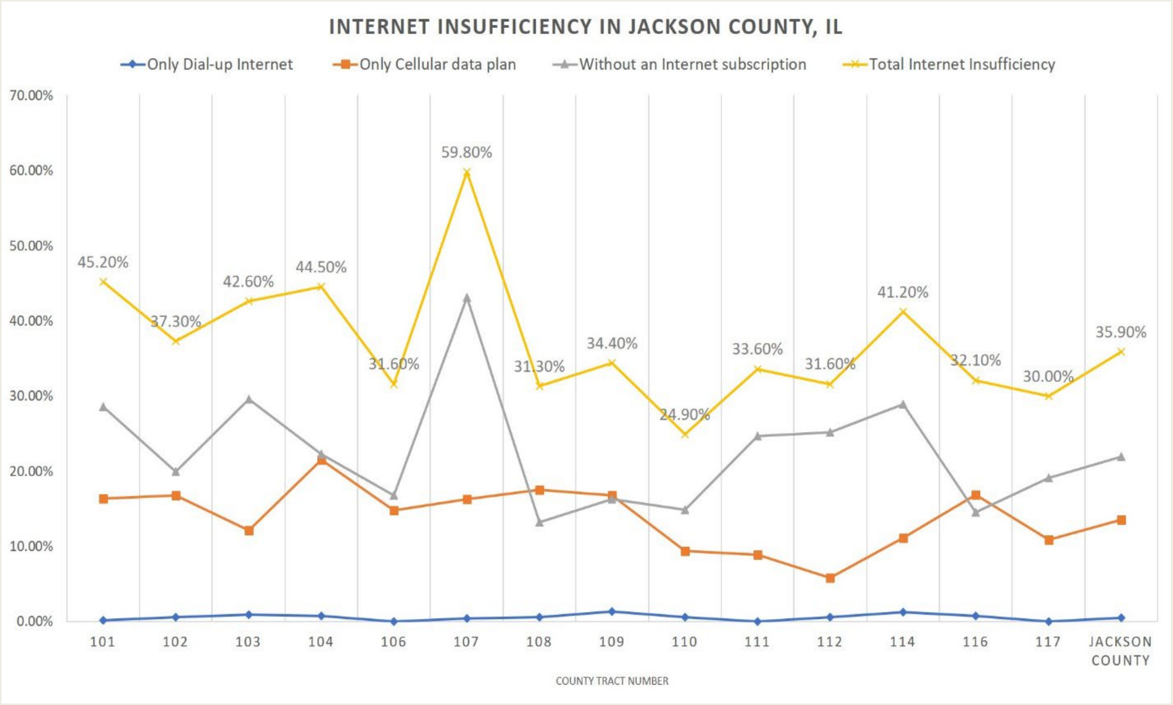
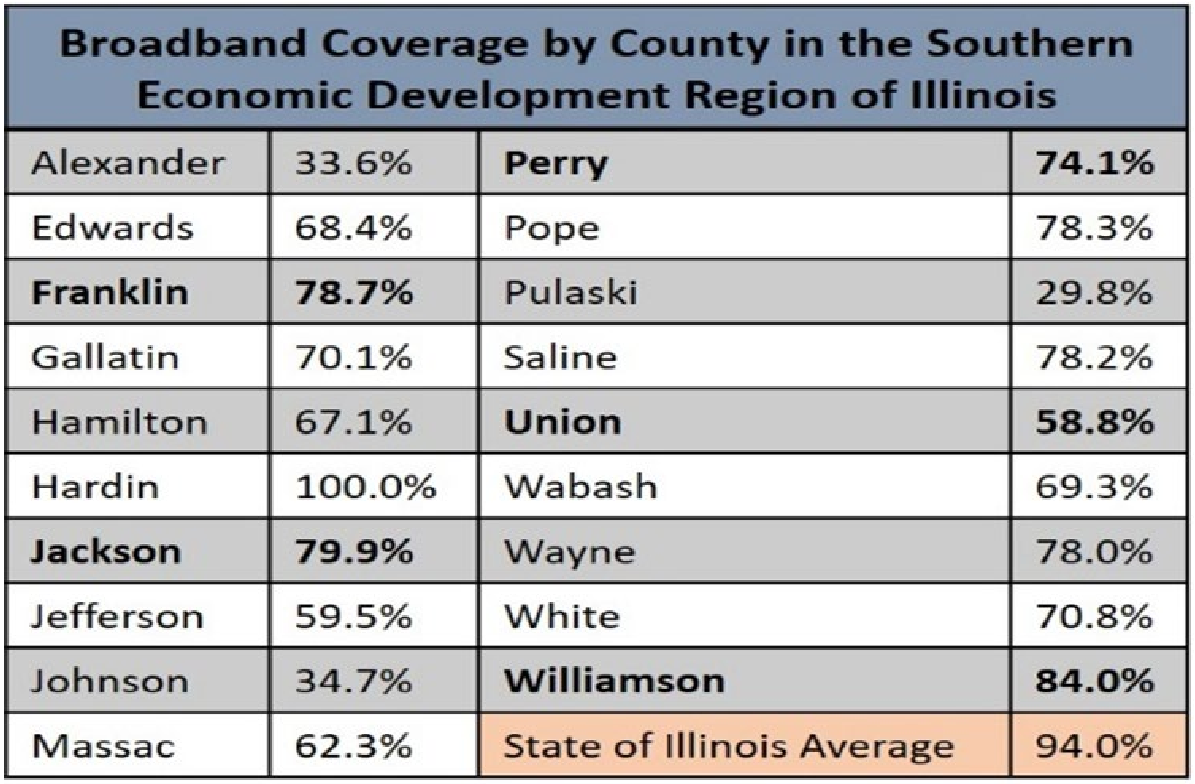


Fig.4 Internet insufficiency statistics at county-tract level, Jackson County



Broadband Coverage by County in the Southern Economic Development Region of Illinois

Alexander	33.6%	**Perry**	**74.1%**
Edwards	68.4%	Pope	78.3%
Franklin	**78.7%**	Pulaski	29.8%
Gallatin	70.1%	Saline	78.2%
Hamilton	67.1%	**Union**	**58.8%**
Hardin	100.0%	Wabash	69.3%
Jackson	**79.9%**	Wayne	78.0%
Jefferson	59.5%	White	70.8%
Johnson	34.7%	**Williamson**	**84.0%**
Massac	62.3%	State of Illinois Average	94.0%

Table:1 Broadband Coverage by County in Southern Illinois Region

1.2 AIMS AND OBJECTIVES

Group all IL households into 4 categories (4.866M households):

- Connected, with Device (CD) 72%
- Device Deficient (DD) 4%
- Internet Insufficient (II) 10%
- Fully Disconnected (FD) 14%

Special objective of this project is to minimize the percent values of Device Deficient (DD), Internet Insufficient (II) and Fully Disconnected (FD) and to maximize Connected (CD) values by configuring the factors affecting them.

1.3 CONTRIBUTION

The thesis explains factors affecting Digital Distress. We analyzed many factors that may have considerable effect on Digital Distress considering ZIP code areas all over Illinois. In near future, we can figure out other factors which may affect Digital Distress.

1.4 THESIS ORGANIZATION

This thesis includes five chapters. Chapter 1 explained the Background and the main goals and objectives. Chapter 2 explained the System Model Chapter 3 explains the Methodology & 4 shows the best Machine Learning Classification Models and Chapter 5 contains Conclusion and Future work.

CHAPTER 2

SYSTEM DESIGN

2.1 DATA SELECTION

To better analyze the digital inclusion-exclusion picture of a region, it is important to first understand the socio-economic indicators of the region. They are important for a digital inclusion narrative for two reasons. First, some socio-economic indicators impact technology adoption, meaning people in those groups are more or less likely to use technology. Second, socio-economic indicators can also impact access to online information and services. Now, let us consider Southern Illinois and therefore, we present below some socio-economic indicators for the Southern Illinois region.

Data from U.S. American Community Survey shows that 59.7% of the population in this region are in the age range of 18-64 years, and 19.2% are 65 years or above. 17.7% of the population is living below poverty level. 27.5% of the household have one or more children. In terms of total population and number of households, we see that the southern Illinois is predominantly a sparsely populated region, with wide variations among the individual counties of the Southern Illinois region. Only 20.2% hold a bachelor's or higher degree, compared to the state average of 34.7%. Several counties have a significantly higher percentage who have an education-level equivalency of less than high school or just high school. Another important characteristic of the Southern Illinois region is that a large proportion of the counties have a significantly higher than the state average percentage of education-level of "some college". Data from 2013 Illinois DCOE survey [2] showed that the median household income in the SEDR was $39,730, significantly lower than $57,792 for Illinois. We can depict the various county-level socio-economic statistics for the SEDR region in terms of age, race/ethnicity, households with children, education-level, and poverty-level.

As expected, educational attainment was lower in digitally distressed areas. A little more than 30 percent of residents did not finish high school, compared to less than 13 percent in the U.S. overall. Likewise, less than 10 percent of digitally distressed residents had a bachelor's degree or higher compared to almost 31 percent in the U.S. overall.

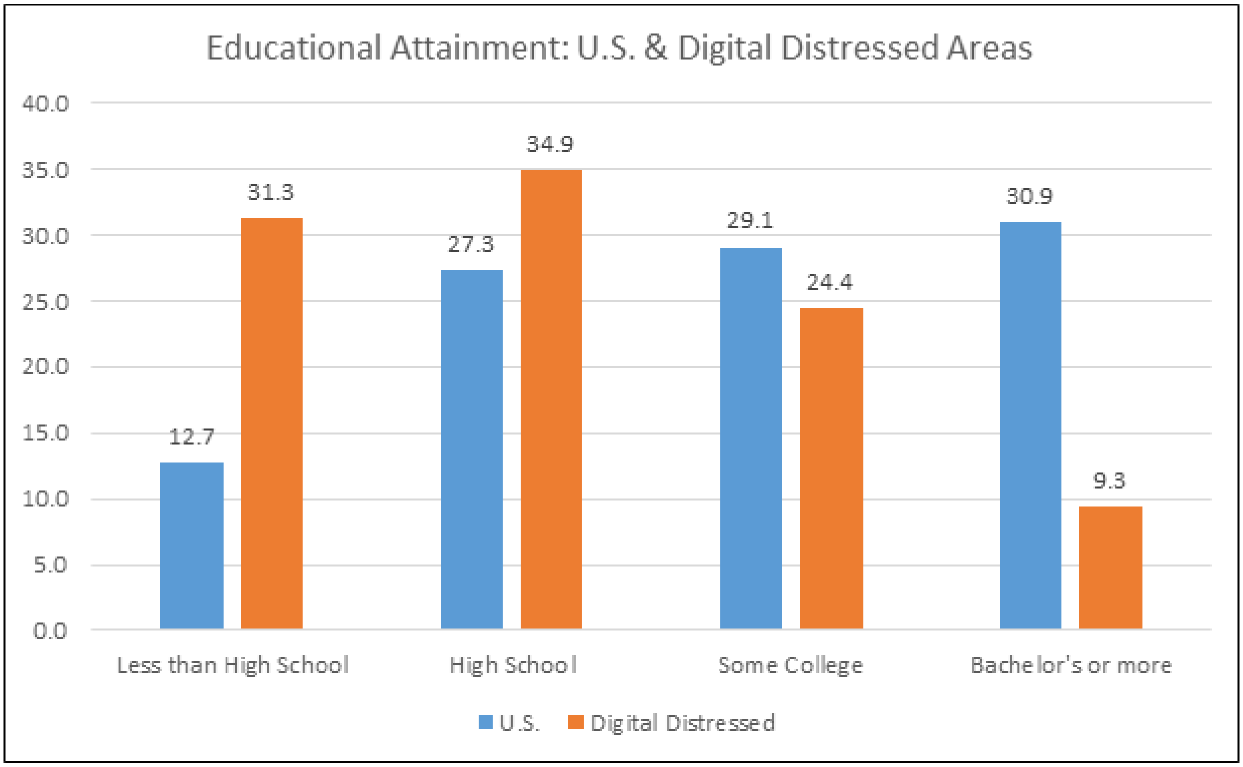


Fig.5 Education Attainment: U.S. & Digital Distressed Areas

Individual poverty was twice as high in digitally distressed areas compared to the U.S. average (14.6 versus 36.2 percent). Likewise, the share of population with disabilities was also higher in digitally distressed neighborhoods compared to the nation.

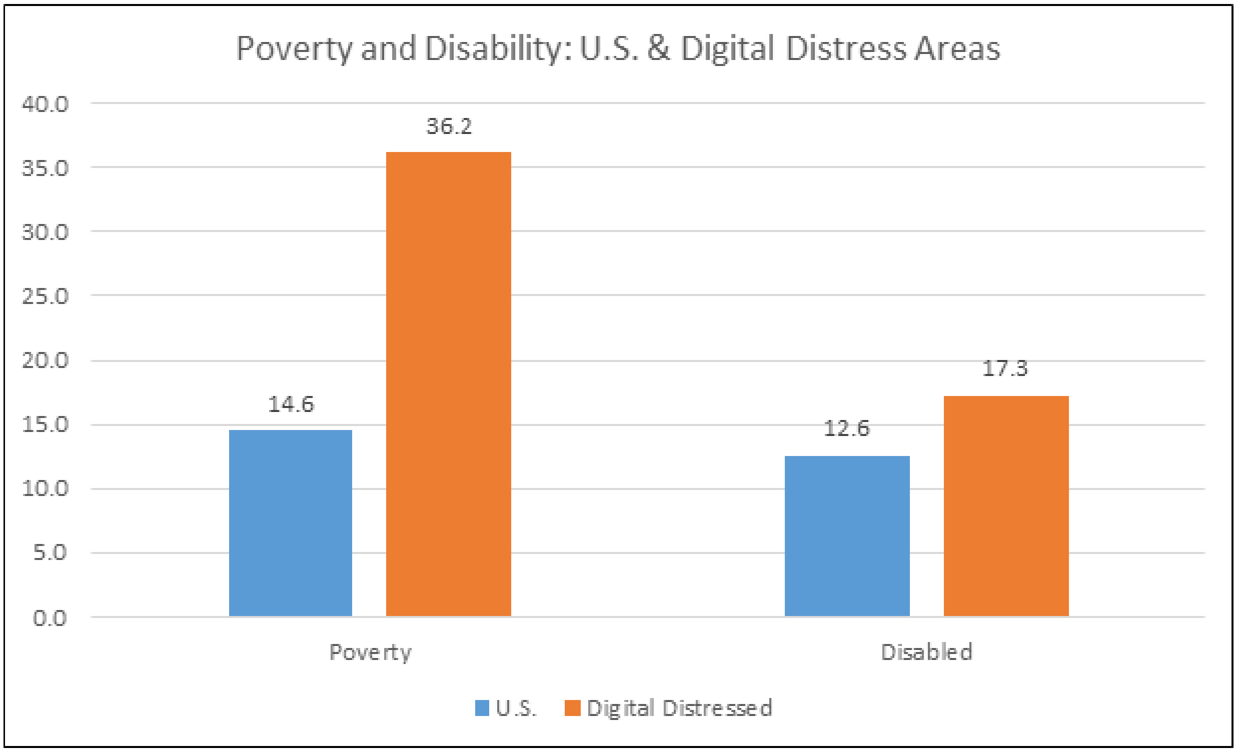


Fig.6 Poverty and Disability: U.S. & Digital Distressed Areas

A similar, if not identical pattern, is seen regarding median household income. The figure below shows that the median household income in digitally distressed areas was half of the U.S. median income.

Again, what is causing what could be argued both ways: low-income leads to digital distress or digital distress leads to low-income. Regardless, having a harder time leveraging digital applications that in turn could lead to improvements in education or employability does impact the residents' ability to earn more.

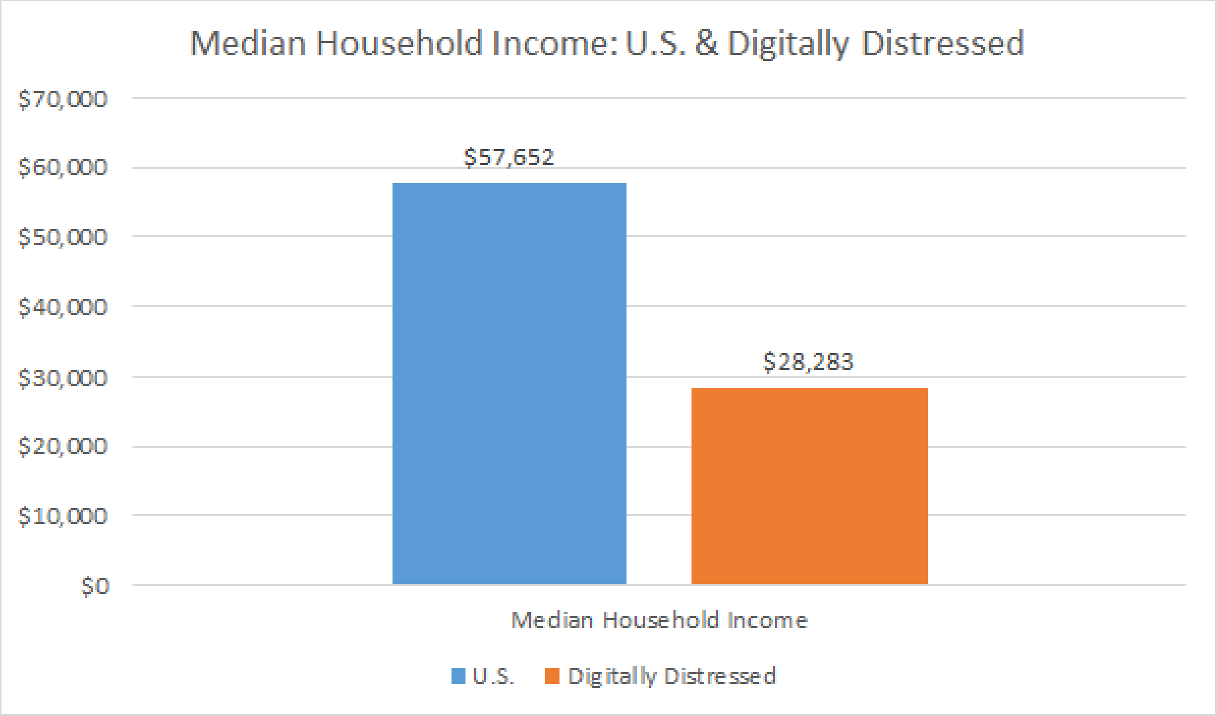


Fig.7 Median Household Income: U.S. & Digital Distressed Areas

Lastly and regarding race & ethnicity, digitally distressed areas had a higher share of minorities compared to the nation, as shown in the figure below. The share of white non-Hispanics in digitally distressed areas was less than half the national average (24.8 percent in digitally distressed areas versus 61.5 percent overall) while the share of black non-Hispanic was three times the national share (12.3 versus 36.4 percent). Also note, the share of Hispanics in digitally distressed areas was almost double the share found in the nation overall (17.6 versus 32.6 percent).

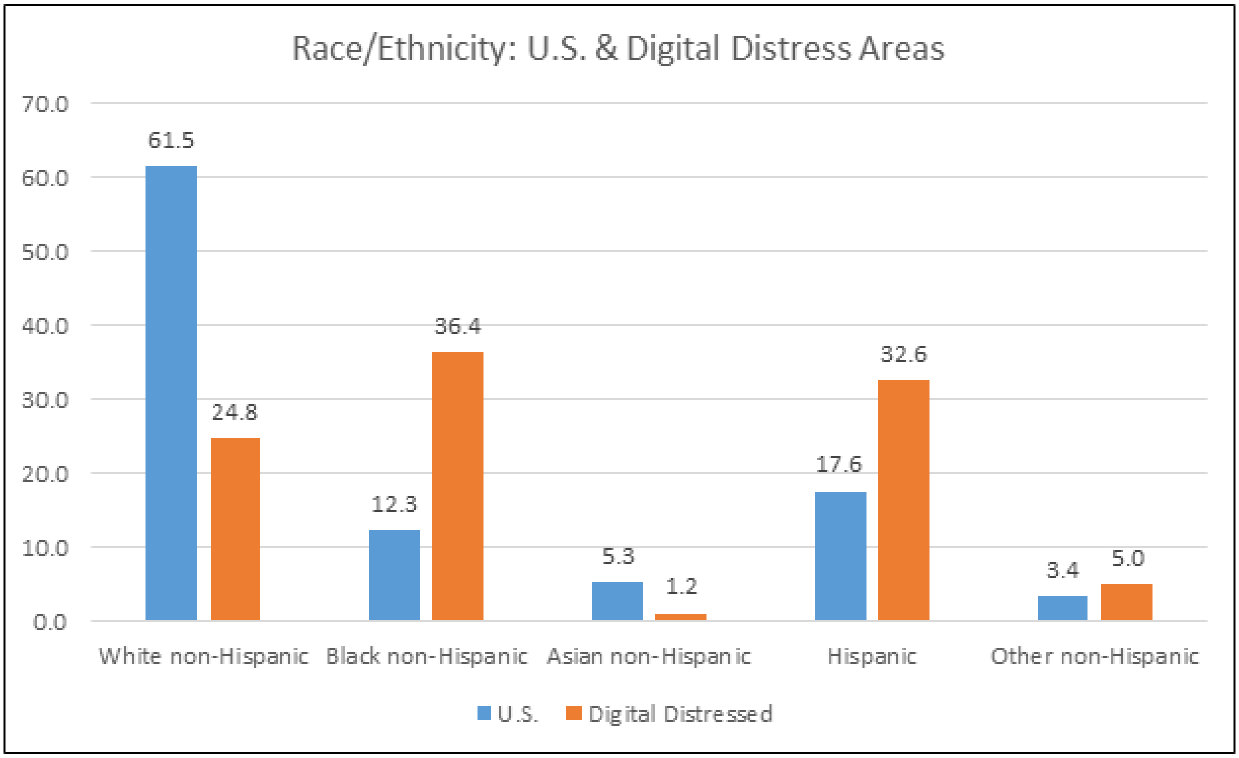


Fig.8 Race/Ethnicity: U.S. & Digital Distressed Areas

To conclude, there are multiple points worth discussing. First, digital distress is a phenomenon that exists putting families and children at a disadvantage. This phenomenon affects 4.4 percent of the total population and cuts across urban and rural areas.

Second, the socioeconomic characteristics of those in digital distress denote a higher share of minorities, less educated, poorer, and younger residents. Ironically, these same groups could benefit greatly from digital applications to improve their quality of life. However, being in digital distress places them at a disadvantage.

2.2 DATA PREPARATION

- By considering all the factors discussed in "DATA SELECTION" section, collected data from 4 different tables in U.S. Census Bureau American Community Survey (ACS).
  - Table -1: Table with Income levels, device counts and Internet subscriptions (Households)

  - Table-2: Table with Age criteria (Population)
  - Table-3: Table with Education Attainment criteria (Population)
  - Table-4: Table with Race & Ethnicity criteria (Population)

- Combined data from these 4 tables to a single table.
- Each row in the table represents data of unique ZIP code from Illinois.
- Total data of 1384 in the data set.
- Total 20 features in data set and the target column is Digital Distress.
- Data is continuous.

Features in data set

- Device Deficient
- Internet Insufficient
- Less than $20,000
- $20,000 to $74,999
- $75,000 or more
- White alone
- Black or African American alone
- American Indian and Alaska Native alone
- Asian alone
- Native Hawaiian and Other Pacific Islander alone
- Below poverty level
- At or above the poverty level
- Less than high school graduate
- High school graduate (includes equivalency)

- Some college or associate degree
- Bachelor's degree or higher
- 45 to 54 years
- 55 to 59 years
- Under 18 years
- 18 to 24 years
- 15 to 44 years
- 60 years and over
- Digital distress

Digital Distress is the target column.

2.3 DIGITAL DISTRESS COLUMN

Target column Digital Distress numerical values are account to form string values to make the model it a classification model to achieve better results and better understanding of the content. So, for this conversion, outliers are used.

What are Outliers?

An outlier is that datapoint or observation which behaves very differently from the rest of the data.

Calculated 5 outliers of Digital Distress data.

1. Lowest
2. Q1
3. Q2
4. Q3
5. Highest

With calculations, obtained data as follows.

0		Lowest
7.36		Q1
22.09		Q2
86.2975		Q3
1826.75		Highest

Categorial values used for the digital distress are Low, Low moderate, Moderate and High.

These values are defined as

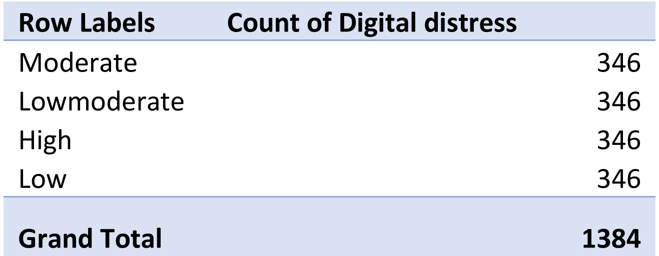
Low - =IF (value<7.36)

Low moderate - =IF (value(A2>=7.36 AND A2<22.09))

Moderate - =IF((A2>=22.09 AND A2<86.2975))

High - =IF((A4>=86.2975 AND A4<1826.75))

With these calculations, the values of Digital Distress column can be summarized as follows.



Row Labels	**Count of Digital distress**
Moderate	346
Lowmoderate	346
High	346
Low	346
Grand Total	**1384**

Table:2 Digital Distress values

CHAPTER 3

METHODOLOGY

3.1 STATISTICAL DATA ANALYSIS

With the data we collected we can draw the following conclusions.

1. Below poverty level accounts for High Digital Distress.

Table:3 Data Analysis: Below poverty level & Digital Distress

Row Labels	Sum of Below poverty level
High	684814
moderate	96482
lowmoderate	26628
Low	9831
Grand Total	**817755**

Represented in pivot chart:

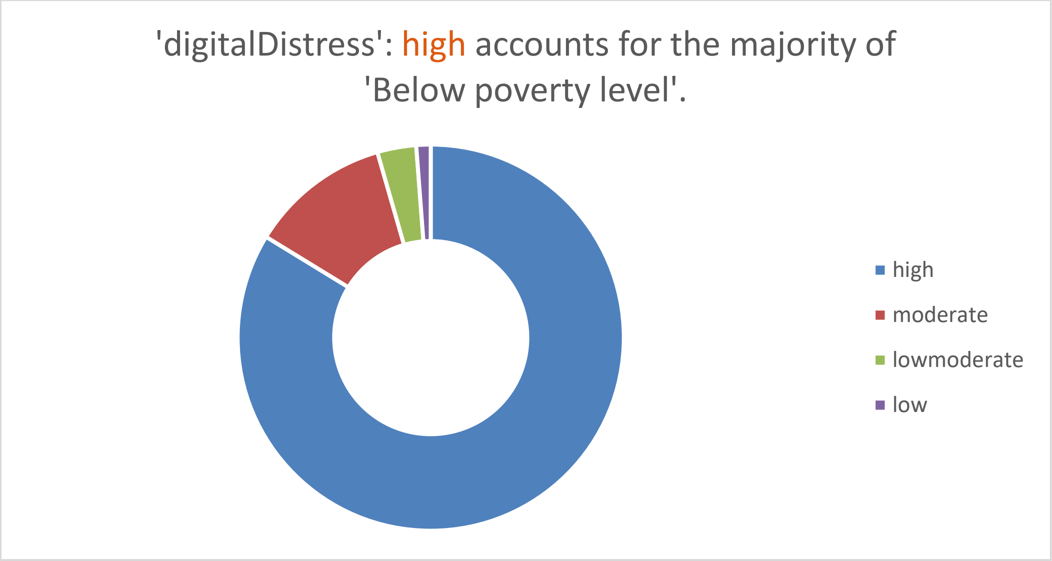


Fig.9 Data Analysis: Below poverty level & Digital Distress

2. Age criteria especially Below 18 years accounts for High Digital Distress.

Table:4 Data Analysis: Age Criteria: Under 18 years & Digital Distress

Row Labels	Sum of Under 18 years
High	2215379
moderate	473153
lowmoderate	125007
Low	42216
Grand Total	**2855755**

Represented in pivot chart:

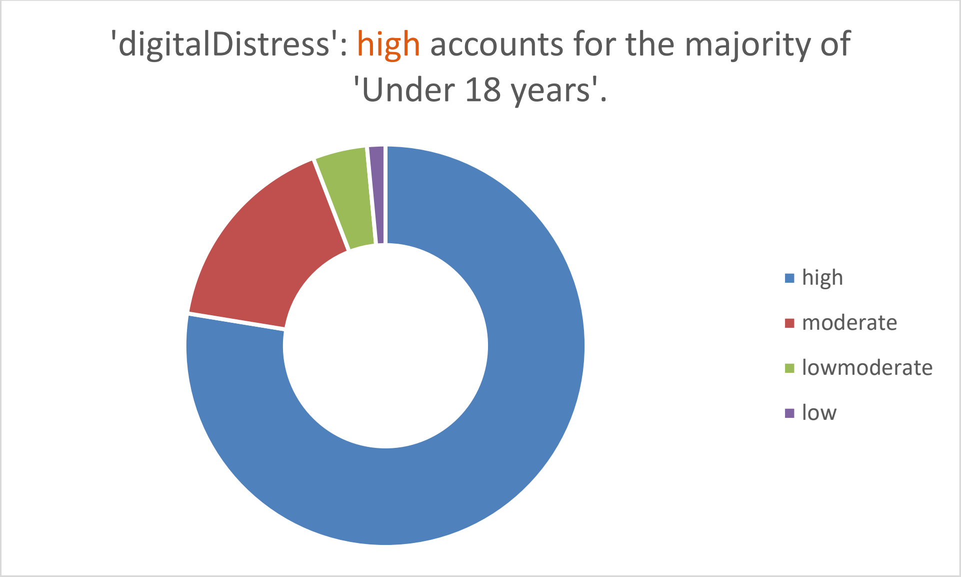


Fig:10 Data Analysis: Age Criteria: Under 18 years & Digital Distress

3. Education Attainment: Less than high school graduate accounts for High Digital Distress

Table:5 Data Analysis: Education Attainment: Less than high school graduate & Digital Distress

Row Labels	Sum of Less than high school graduate
High	543017
moderate	55697
lowmoderate	16910
Low	5353
Grand Total	**620977**

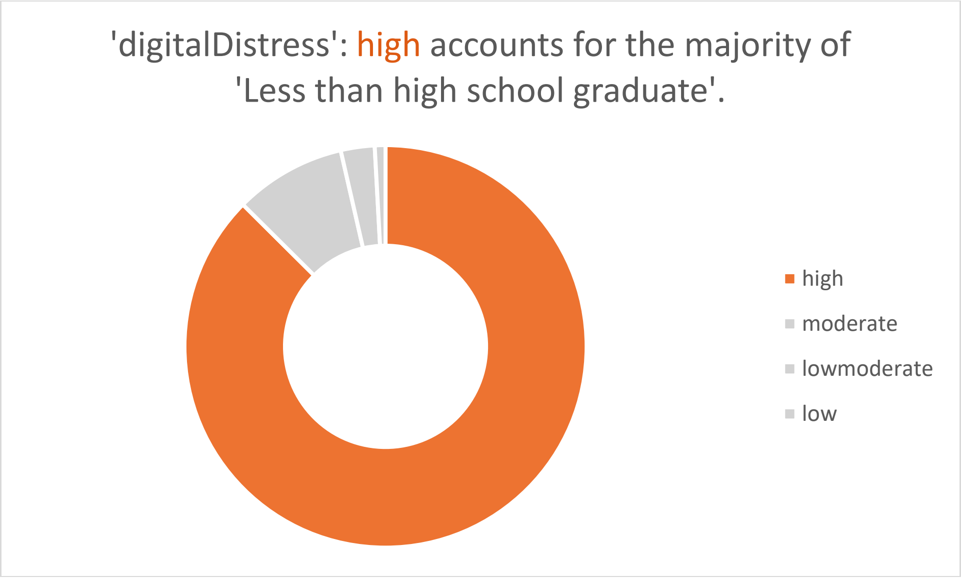
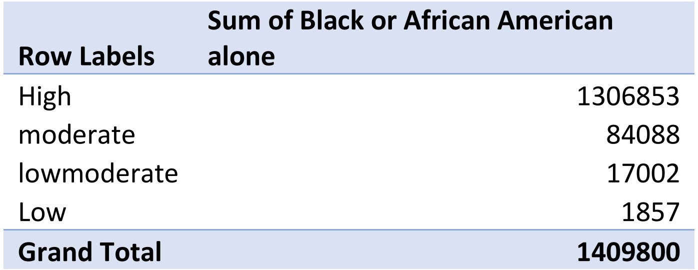


Fig.11 Data Analysis: Education Attainment: Less than high school graduate & Digital Distress

4. Race & Ethnicity: Black or African American Alone accounts for High Digital Distress

Table:6 Data Analysis: Race & Ethnicity: Black or African American Alone & Digital Distress



Row Labels	**Sum of Black or African American alone**
High	1306853
moderate	84088
lowmoderate	17002
Low	1857
Grand Total	**1409800**

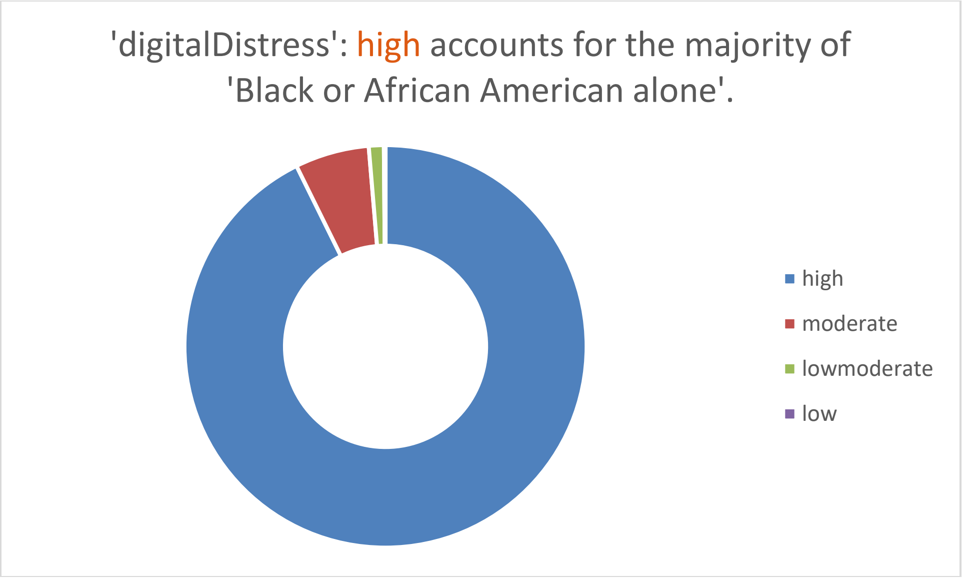
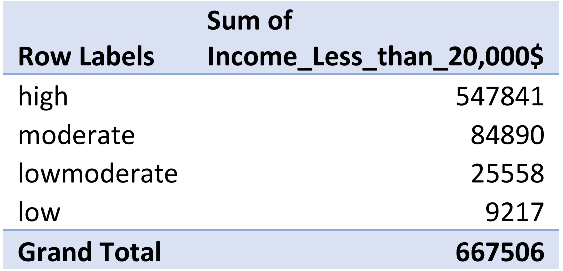


Fig.12 Data Analysis: Race & Ethnicity: Black or African American Alone & Digital Distress

5. Income: Less than 20,000$ accounts for High Digital Distress.

Table:7 Data Analysis: Income: Less than 20,000$ & Digital Distress



Row Labels	Sum of Income_Less_than_20,000$
high	547841
moderate	84890
lowmoderate	25558
low	9217
Grand Total	**667506**

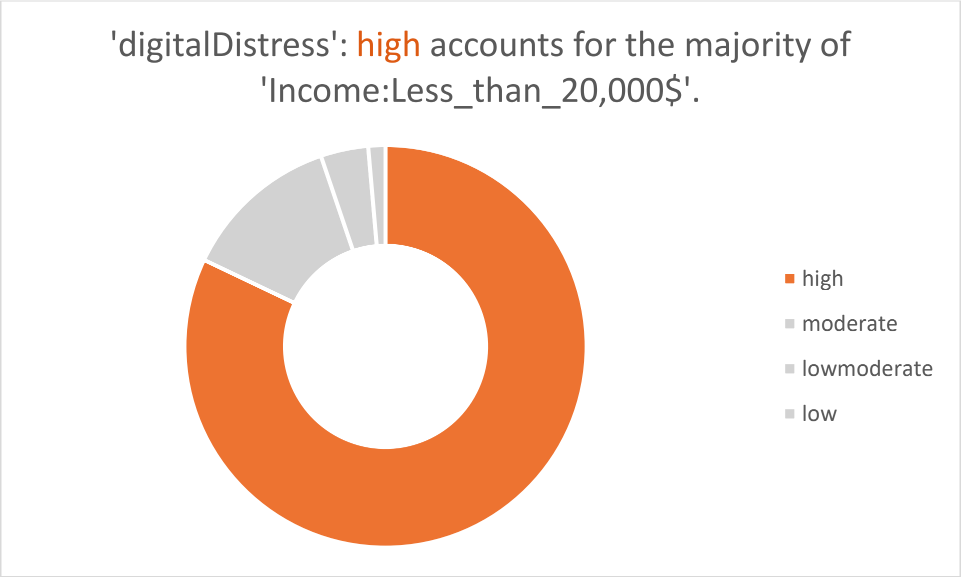


Fig.13 Data Analysis: Income: Less than 20,000$ & Digital Distress

6. The factors which accounts for High Digital Distress are Dependent on each other. Income: Less than $20,000 and Education Attainment: Less than high school graduate both accounts for High Digital Distress, this is obtained from the above analysis.

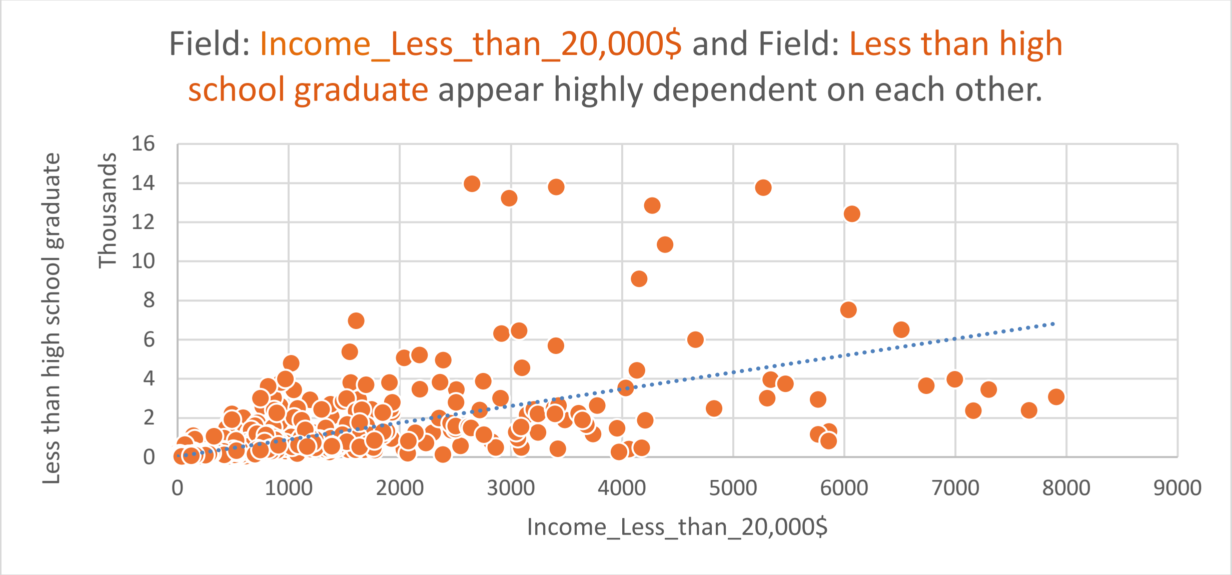


Fig.14 Data Analysis: Income: Less than 20,000$ & Education Attainment: Less than high school graduate

3.2 DIGITAL DISTRESS CORRELATION COEFFICIENT

What is correlation?

Correlation is a bi-variate analysis that measures the strength of association between two variables and the direction of the relationship. In terms of the strength of relationship, the value of the correlation coefficient varies between +1 and -1. A value of ± 1 indicates a perfect degree of association between the two variables. As the correlation coefficient value goes towards 0, the relationship between the two variables will be weaker. The direction of the relationship is indicated by the sign of the coefficient; a + sign indicates a positive relationship and a - sign indicates a negative relationship.

Usually, in statistics, we measure four types of correlations:

- Pearson correlation
- Kendall rank correlation
- Spearman correlation
- Point-Biserial correlation.

Pearson R Correlation

As the title suggests, we'll only cover Pearson correlation coefficient. I'll keep this short but very informative so you can go ahead and do this on your own. Pearson correlation coefficient is a measure of the strength of a linear association between two variables — denoted by r. You'll come across Pearson r correlation

Questions a Pearson correlation answers

- Is there a statistically significant relationship between age and height?
- Is there a relationship between temperature and ice cream sales?
- Is there a relationship among job satisfaction, productivity, and income?

- Which two variables have the strongest co-relation between age, height, weight, size of family and family income?

Coefficient correlation with Digital Distress on all the factors.

Coefficient correlation with Digital Distress	
Device deficiency	0.941905
Internet insufficiency	0.936914
Less than $20,000:	0.879079
$20,000 to $74,999:	0.895473
$75,000 or more:	0.578197
White alone	0.593048
Black or African American alone	0.718632
American Indian and Alaska Native alone	0.635942
Asian alone	0.380462
Native Hawaiian and Other Pacific Islander alone	0.248950
Below poverty level	0.865343
At or above the poverty level	0.766341
Less than high school graduate	0.813912
High school graduate (includes equivalency)	0.885736
Some college or associate degree	0.850139
Bachelor's degree or higher	0.475764
45 to 54 years	0.783510
55 to 59 years	0.779779
Under 18 years	0.829878
18 to 24 years	0.717530
15 to 44 years	0.798971
60 years and over	0.786252

Table:8 Correlation between Digital Distress and some demographic and socio-economic factors

3.3 BROADBAND ACCESS CORRELATION COEFFICIENT

Socio-economic factors have a strong influence on the adoption of digital technologies since they often describe the underlying story behind the use of computing and Internet in occupational choices and daily life. Therefore, as a first step towards understanding the bigger picture on adoption in the southern Illinois region, we attempted to correlate socio-economic factors with broadband usage. We calculated the Pearson correlation value for the broadband usage percentage in the 19-county southern Illinois region with different socio-economic/demographic parameters. The data used in the analysis was obtained from ACS 2019 survey results, and from http://www.arcgis.com .

We see that availability Internet access, and more specifically higher download speeds have a positive influence on the broadband adoption or usage in the southern Illinois region. Similarly, access to computing devices such as laptops and desktops in households would have a strong positive impact on broadband usage. However, it is interesting to note that the degree of correlation of access to computing device with broadband usage is considerably weaker than that observed with broadband access. One possible explanation for this observation could be that a portion of the population uses only smartphones (device deficient) for accessing the Internet via Wi-Fi or cellular data plans.

Demographic / Socio-economic parameter	**Rural Population Percent**	**Percent of Population below Poverty Level**	**Median Household Income**	**Percent of Households with No Internet Access**	**Percent of Households with no Computing Device**	**Ookla Median Download Speed (Mbps)**	**Broadband Access in Household**
Broadband-Usage Pearson Correlation Coefficient	-0.46570	-0.17133	0.29780	-0.71970	-0.70449	0.68281	0.73442

Table:9 Correlation between broadband usage and some demographic factors

We see that there is good correlation of broadband usage with higher percentage of people having an education level of bachelor’s degree or higher. This might potentially reflect the effect of education on occupational and household dynamics.

Socio-economic parameter	Race/Ethnicity factor		Age factor			Education factor	
	White, non-Hispanic	Black, non-Hispanic	25-44 Age	18-64 Age	65 or older	Up to High school degree	Bachelor’s degree or higher
Broadband-Usage Pearson Correlation Coefficient	0.40396	-0.50089	0.42453	0.48765	-0.30968	-0.59150	0.54839

Table:10 Correlation between broadband usage and some socio-economic factors

CHAPTER 4

MACHINE LEARNING CLASSIFICATION MODELS

4.1 MULTI CLASS ML CLASSIFICATION MODEL

Machine learning is a field of study and is concerned with algorithms that learn from examples.

Classification is a task that requires the use of machine learning algorithms that learn how to assign a class label to examples from the problem domain. An easy-to-understand example is classifying emails as "*spam*" or "*not spam*."

There are many different types of classification tasks that you may encounter in machine learning and specialized approaches to modeling that may be used for each.

There are perhaps four main types of classification tasks that you may encounter; they are:

- Binary Classification
- Multi-Class Classification
- Multi-Label Classification
- Imbalanced Classification

Multi-Class Classification comes into picture for the Data set being used.

Multi-class classification refers to those classification tasks that have more than two class labels. Unlike binary classification, multi-class classification does not have the notion of normal and abnormal outcomes. Instead, examples are classified as belonging to one among a range of known classes.

Many algorithms used for binary classification can be used for multi-class classification.

Popular algorithms that can be used for multi-class classification include:

- k-Nearest Neighbors.
- Decision Trees.

- Naive Bayes.
- Random Forest

Algorithms that are designed for binary classification can be adapted for use for multi-class problems.

This involves using a strategy of fitting multiple binary classification models for each class vs. all other classes (called one-vs-rest) or one model for each pair of classes (called one-vs-one).

- **One-vs-Rest**: Fit one binary classification model for each class vs. all other classes.
- **One-vs-One**: Fit one binary classification model for each pair of classes.

Binary classification algorithms that can use these strategies for multi-class classification include:

- Logistic Regression.
- Support Vector Machine.

4.2 TRAINING DATA

By training these models on the data, the following accuracy values are obtained.

	Model	**Accuracy**
0	Random Forest	0.735577
1	Decision Tree	0.661058
2	Naive Bayes	0.605769
3	SVC	0.653846
4	KNN	0.706731
5	Logistic	0.569712

Table:11 Accuracy Values of ML models

Confusion matrix for the above models.

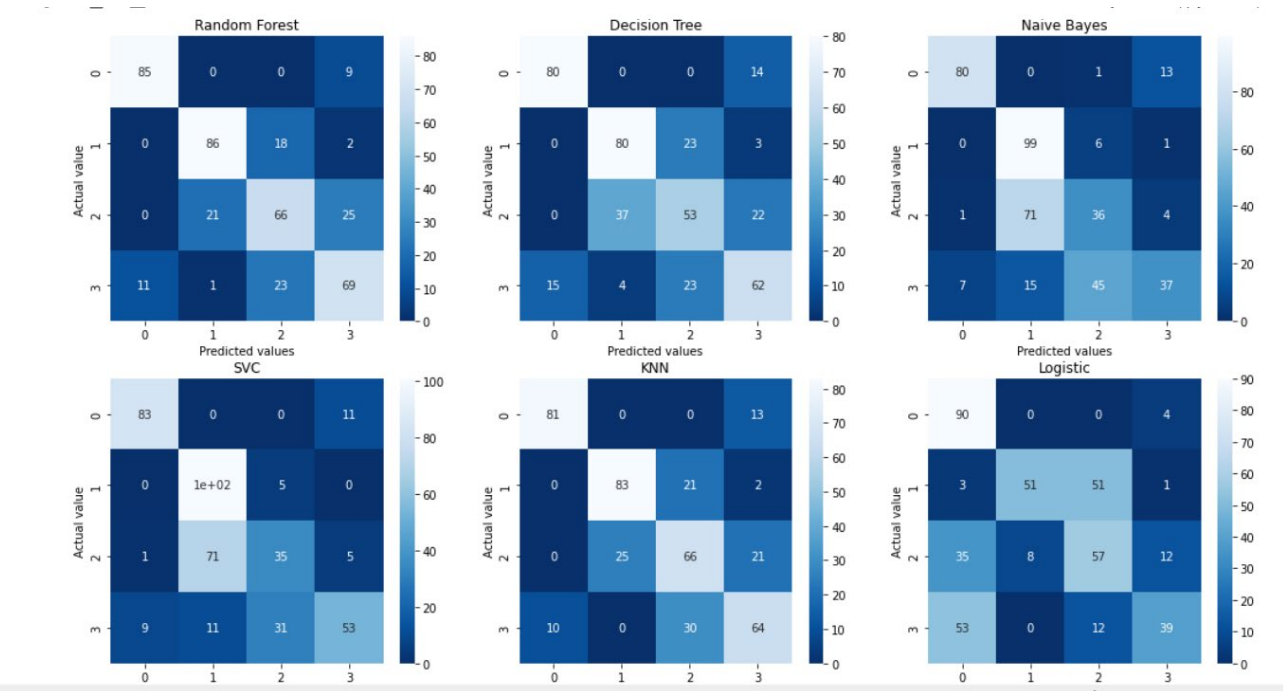


Fig.15 Confusion Matrix of ML models

4.3 CROSS VALIDATION

GridSearchCV is the process of performing hyperparameter tuning to determine the optimal values for a given model. As mentioned above, the performance of a model significantly depends on the value of hyperparameters. Note that there is no way to know in advance the best values for hyperparameters so ideally, we need to try all possible values to know the optimal values. Doing this manually could take a considerable amount of time and resources and thus we use GridSearchCV to automate the tuning of hyperparameters.

GridSearchCV is a function that comes in Scikit-learns (or SK-learn) model selection package. So an important point here to note is that we need to have the Scikit learn library installed on the computer. This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set. So, in the end, we can select the best parameters from the listed hyperparameters.

With GridSearchCV on Decision Tree Classification model,

The accuracy is 0.7331730769230769

4.4 TUNING HYPERPARAMETERS

The best way to think about hyperparameters is like the settings of an algorithm that can be adjusted to optimize performance. While model *parameters* are learned during training — such as the slope and intercept in a linear regression — *hyperparameters* must be set by the data scientist before training. In the case of a random forest, hyperparameters include the number of decision trees in the forest and the number of features considered by each tree when splitting a node. (The parameters of a random forest are the variables and thresholds used to split each node learned during training). Scikit-Learn implements a set of sensible default hyperparameters_for all models, but these are not guaranteed to be optimal for a problem. The best hyperparameters are usually impossible to determine ahead of time, and tuning a model is where machine learning turns from a science into trial-and-error based engineering.

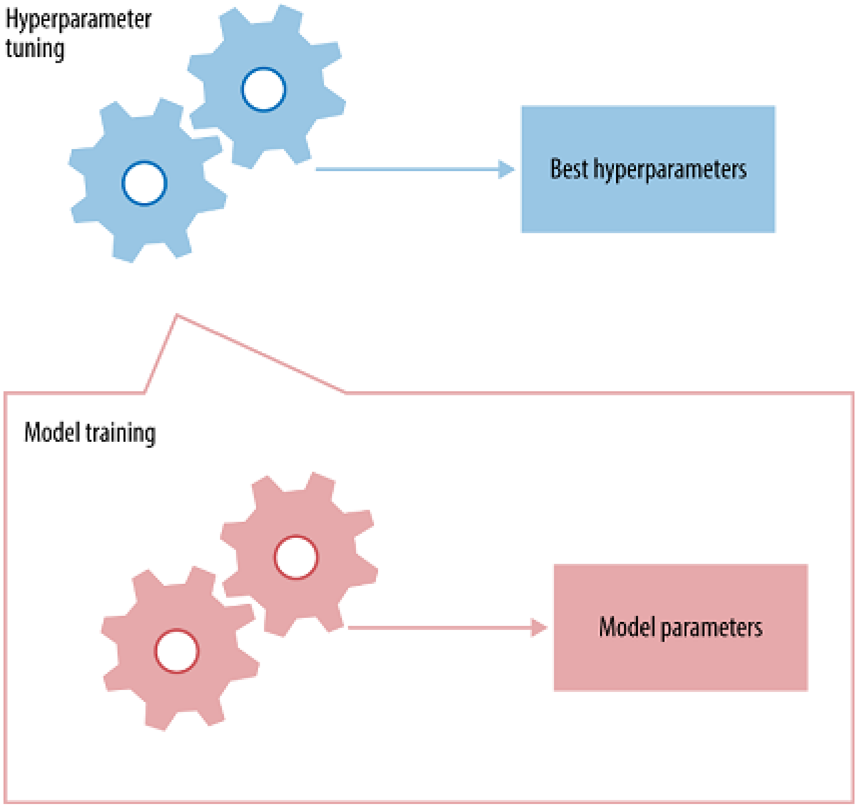


Fig.16 Hyperparameters

4.4.1 Decision Tree Classification

Hyperparameter used in Decision Tree Classification.

DecisionTreeClassifier (criterion='entropy',

max_depth=11,

max_features=0.8301024279155249,

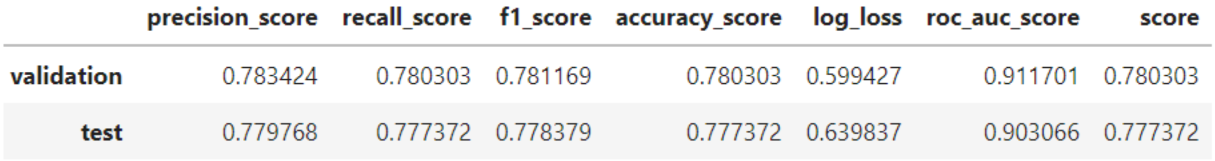
min_samples_leaf=0.2277036762124418,

min_samples_split=0.2009520698298776,

random_state=542470820)

By tuning hyperparameters on Decision Tree Classification, the results are as follows.

Table:12 Results on Decision Tree Classification



	precision_score	recall_score	f1_score	accuracy_score	log_loss	roc_auc_score	score
validation	0.783424	0.780303	0.781169	0.780303	0.599427	0.911701	0.780303
test	0.779768	0.777372	0.778379	0.777372	0.639837	0.903066	0.777372

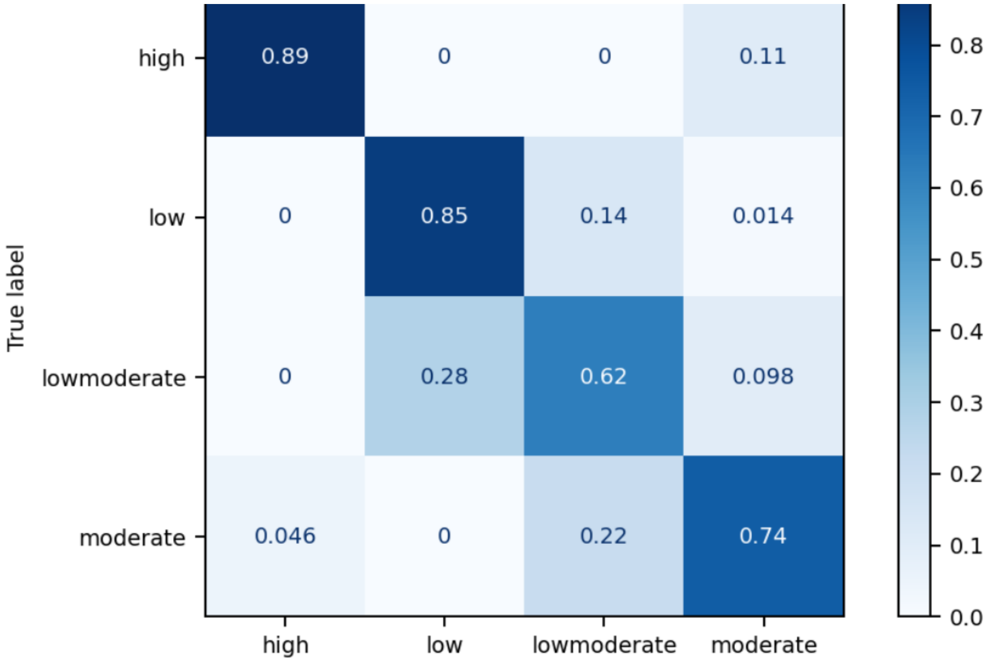


Fig.17 Confusion Matrix of Decision Tree Classification

4.4.2 Random Forest Classification

Hyperparameter used in Random Forest Classification.

RandomForestClassifier (criterion='entropy',

max_depth=5,

max_features=0.6772485796634158,

min_samples_leaf=0.03707910745988508,

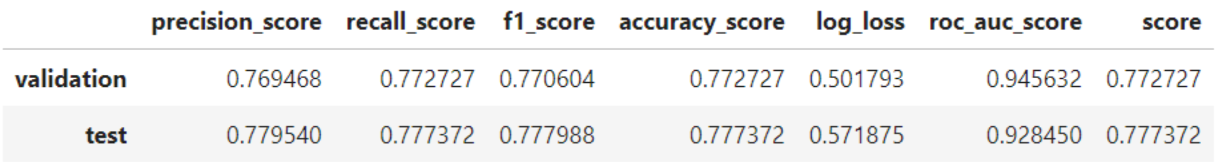
min_samples_split=0.04600774138805236,

n_estimators=177,

random_state=542470820)

By tuning hyperparameters on Random Forest Classification, the results are as follows.

Table:13 Results on Random Forest Classification



	precision_score	recall_score	f1_score	accuracy_score	log_loss	roc_auc_score	score
validation	0.769468	0.772727	0.770604	0.772727	0.501793	0.945632	0.772727
test	0.779540	0.777372	0.777988	0.777372	0.571875	0.928450	0.777372

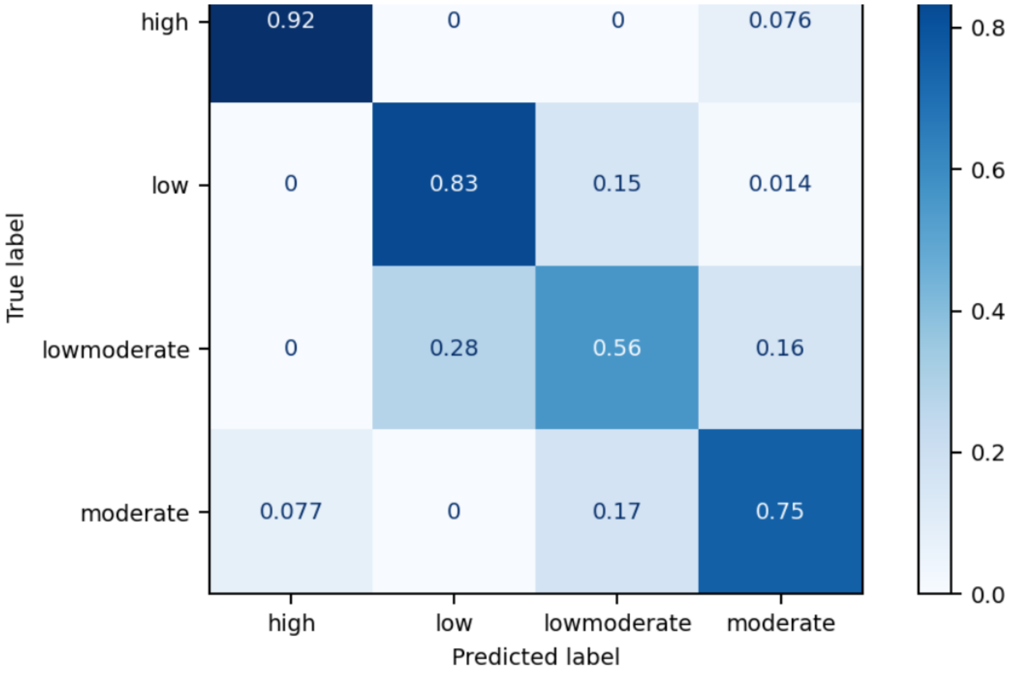


Fig.18 Confusion Matrix of Random Forest Classification

CHAPTER 5

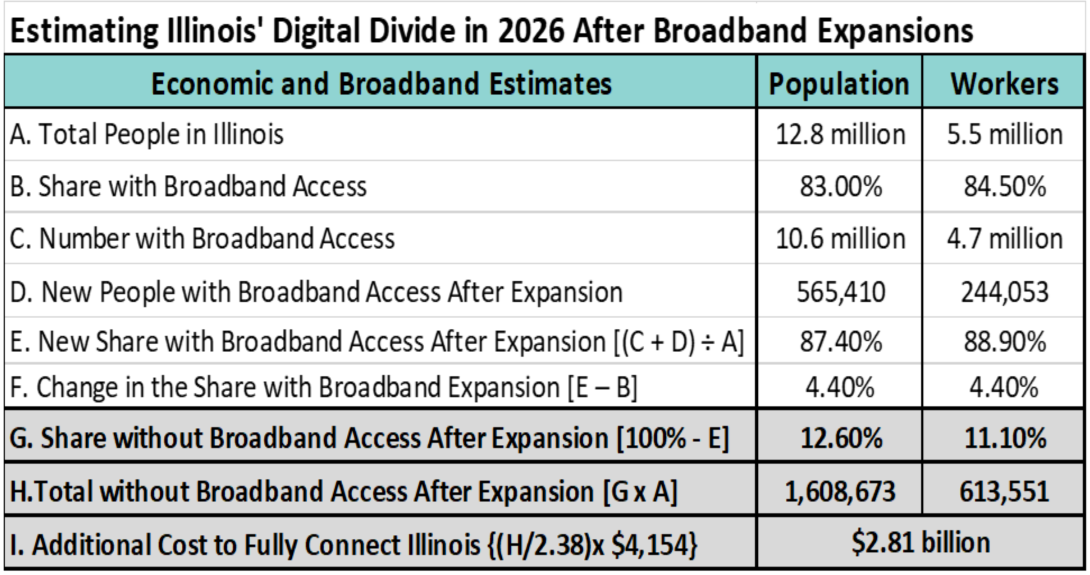
CONCLUSION AND FUTUREWORK

In this thesis, we have stated various factors that may affect the Digital Distress in Illinois Region. We have come up with an efficient Machine Learning model to predict the Digital Distress at any region. Here, we have used various models by tuning their hyper-parameters to build best prediction models. Future works in this direction can be done using Federated learning which may further add any factors to the model or may remove from the model for better training of data. And the data can be collected all over the United States or starting from Mid-west countries and then expanding.

To achieve the objective of overcoming the problem of Digital Distress in Illinois:

- Need to educate the people to stay CONNECTED.

Based on U.S. Census Bureau figures, researchers estimate that 13% of residents will still be left without reliable high-speed internet access after 2026. Closing this gap would require additional public and private investments in broadband infrastructure amounting to $2.8 billion.



Estimating Illinois' Digital Divide in 2026 After Broadband Expansions

Economic and Broadband Estimates	Population	Workers
A. Total People in Illinois	12.8 million	5.5 million
B. Share with Broadband Access	83.00%	84.50%
C. Number with Broadband Access	10.6 million	4.7 million
D. New People with Broadband Access After Expansion	565,410	244,053
E. New Share with Broadband Access After Expansion [(C + D) ÷ A]	87.40%	88.90%
F. Change in the Share with Broadband Expansion [E – B]	4.40%	4.40%
G. Share without Broadband Access After Expansion [100% - E]	**12.60%**	**11.10%**
H.Total without Broadband Access After Expansion [G x A]	**1,608,673**	**613,551**
I. Additional Cost to Fully Connect Illinois {(H/2.38)x $4,154}	**$2.81 billion**	

Table:14 Estimating Illinois Digital Divide in 2026 After Broadband Expansions

REFERENCES

[1]US Census Bureau Data, https://data.census.gov/cedsci/.

[2]https://www.illinoisworknet.com/DownloadPrint/Southern%20EDR%20Data%20Packet_Page.pdf.

[3]Broadband Internet in Illinois, BroadbandNow, https://broadbandnow.com/Illinois.

[4]https://medium.com/design-and-tech-co/digital-distress-what-is-it-and-who-does-it-affect

[5]https://illinoisupdate.com/2022/06/01/broadband-study/

[6]Scikit Learn, https://scikit-learn.org/

[7]GridSearchCV, https://www.mygreatlearning.com/blog/gridsearchcv/

[8]Correlation,https://towardsdatascience.com/pearson-coefficient-of-correlation-explained-369991d93404

[9]Hyperparameters,https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74

[10]M-Lab Statistics by US County, https://datastudio.google.com/u/0/reporting/e5be593c-da8f-489c-a1d5-73b28fb82acc/page/gxXiB?s=rM4_hclXQKc.

[11]Purdue Center for Regional Development, Purdue University, https://pcrd.purdue.edu/data-analysis/.

[12]Broadband Adoption in Illinois, http://www.broadbandillinois.org/uploads/cms/documents/broadband_adoption_in_illinois.11.09-edsb.pdf, 2012.

[13]Survey of Illinois Households Examines Broadband Usage, Benefits, Barriers, Broadband Illinois,http://www.broadbandillinois.org/Research/Internal-Research/Adoption-and-Use.html, 2012.

[14]https://www.nado.org/report-connect-si-regional-prosperity-through-collaborations-in-southern-illinois/.

[15]The Illinois Broadband Map, https://gis.connectednation.org/portal/apps/webappviewer/index.html?id=caedfe7ce8924660a4ce62de6a75a7fd, 2021.

[16]U.S. Department of Agriculture. 2019. A Case for Rural Broadband. American Broadband Initiative. Available at https://www.usda.gov/sites/default/files/documents/case-for-rural-broadband.pdf.

[17]LoPiccalo, K. 2021. Impact of Broadband Penetration on U.S. Farm Productivity. OEA Working Paper 50, Federal Communications Commission. Available at https://docs.fcc.gov/public/attachments/DOC-368773A1.pdf.

[18]Heartland Forward, https://heartlandforward.org/case-study/connecting-the-heartland-how-heartland-forward-is-doing-our-part-to-bridge-the-digital-divide/, 2021.

[19]Connect Illinois Computer Equity Network, https://www2.illinois.gov/dceo/ConnectIllinois/Pages/PCsForPeople.aspx, 2021.

[20]https://ltcillinois.org/wp-content/uploads/2020/12/2020-IL-School-District-Technology-Survey.pdf, 2020.

[21]Open Data Network, Jackson County, IL, https://www.opendatanetwork.com/entity/0500000US17077/Jackson_County_IL/demographics.population.count?ref=search-entity&year=2018, 2021.

[22]Data USA, Jackson County Health, IL https://datausa.io/profile/geo/jackson-county-il#health, 2021.

APPENDIX I

```
from __future__ import print_function
import os
import numpy as np
import pandas as pd
from pandas import DataFrame,Series
import matplotlib
import matplotlib.pyplot as plt
%matplotlib inline
from sklearn.model_selection import train_test_split
import warnings
warnings.filterwarnings('ignore')
dataset = os.sep.join( ['CategorialData.csv'])
df = pd.read_csv(dataset)


x_new = [x for x in df.columns if x!='digitalDistress']
X = df[x_new]
y = df['digitalDistress']
X_train, X_test, Y_train, Y_test = train_test_split(X,y, test_size=0.3, random_state=42)



from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn.naive_bayes import GaussianNB
from sklearn.neighbors import KNeighborsClassifier
from sklearn.linear_model import LogisticRegression

model_pipeline = []
model_pipeline.append(RandomForestClassifier())
model_pipeline.append(DecisionTreeClassifier())
model_pipeline.append(GaussianNB())
model_pipeline.append(SVC())
model_pipeline.append(KNeighborsClassifier())
model_pipeline.append(LogisticRegression(solver="liblinear"))

from sklearn import metrics
from sklearn.metrics import classification_report, confusion_matrix

model_list = ['Random Forest','Decision Tree','Naive Bayes','SVC','KNN','Logistic']
acc_list =[]
auc_list = []
cm_list = []
```

```
for model in model_pipeline:
    model.fit(X_train,Y_train)
    y_pred = model.predict(X_test)
    acc_list.append(metrics.accuracy_score(Y_test,y_pred))
    cm_list.append(confusion_matrix(Y_test,y_pred))
```

```
import seaborn as sns
fig = plt.figure(figsize=(18,10))
for i in range(len(cm_list)):
    cm = cm_list[i]
    model = model_list[i]
    sub = fig.add_subplot(2,3,i+1).set_title(model)
    cm_plot = sns.heatmap(cm,annot=True,cmap = 'Blues_r')
    cm_plot.set_xlabel('Predicted values')
    cm_plot.set_ylabel('Actual value')
```

```
result_df = pd.DataFrame({'Model':model_list, 'Accuracy': acc_list})
result_df
```

For GridSearch CV

```
from sklearn.model_selection import GridSearchCV

# Write your code to set parameters (max_depth and feature importancr) for cross validation
params =
{'max_depth':range(1,DTC.tree_.max_depth),'max_features':range(1,len(DTC.feature_importances_)+1)}

# Write your code to apply Grid search cv using the paramters
grid_search_CV = GridSearchCV(DTC,params,scoring = 'accuracy',cv=3)

#write your code to fit the model
grid_search_CV.fit(X_train,Y_train )
```

```
GridSearchCV(cv=3,
             estimator=DecisionTreeClassifier(criterion='entropy', max_depth=11,
                                              random_state=542470820,
                                              splitter='random'),
             param_grid={'max_depth': range(1, 11),
                         'max_features': range(1, 23)},
            scoring='accuracy')
grid_search_CV.n_features_in_

22


print('Number of nodes: ', grid_search_CV.best_estimator_.tree_.node_count)
print('Depth of model:  ', grid_search_CV.best_estimator_.tree_.max_depth)

Number of nodes:  29
Depth of model:   4

y_test_predict =  grid_search_CV.predict(X_test)
print(model_accuracy(Y_test,y_test_predict))

0.7331730769230769
```

APPENDIX II

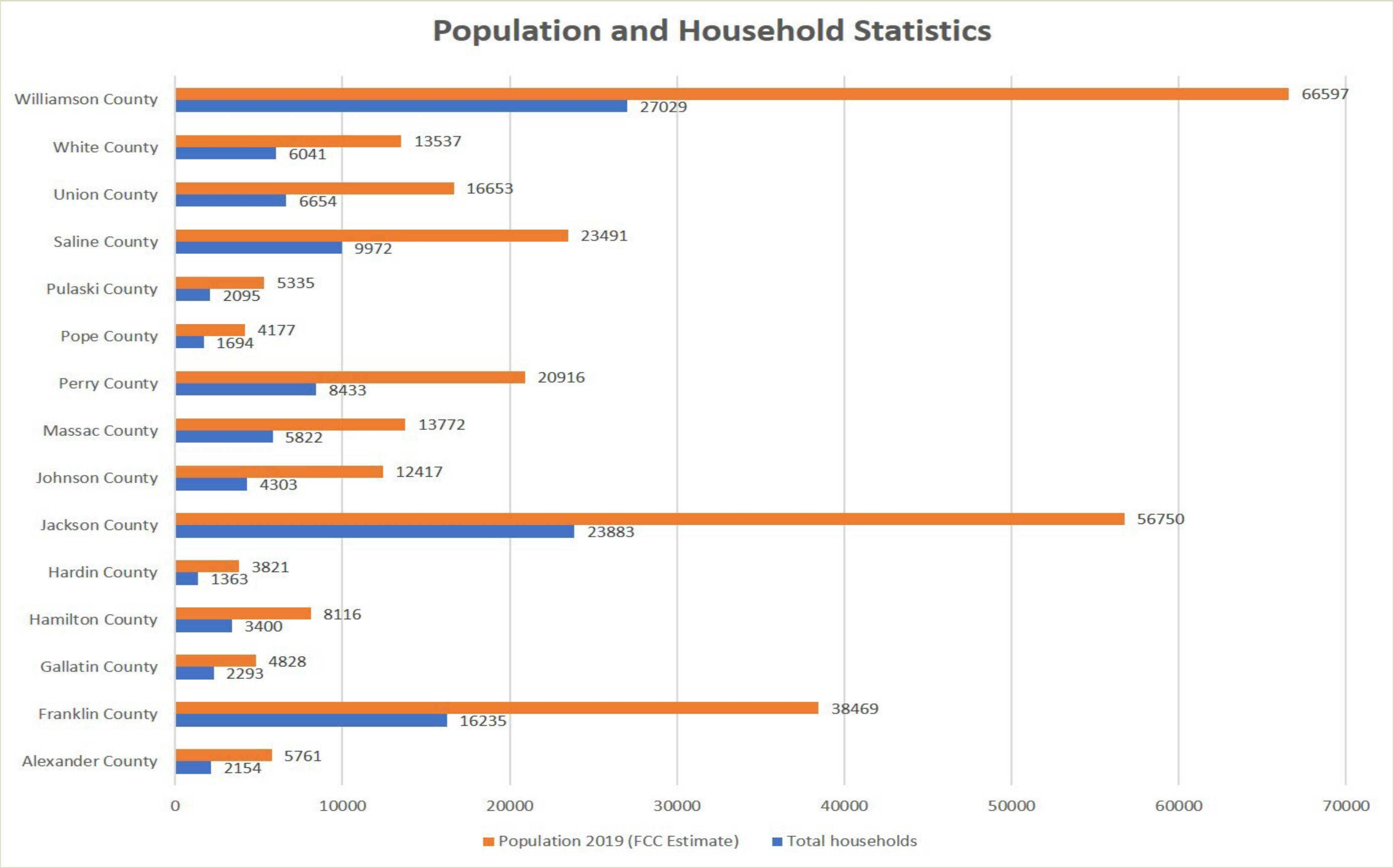


Fig.19 Population and households in SEDR counties

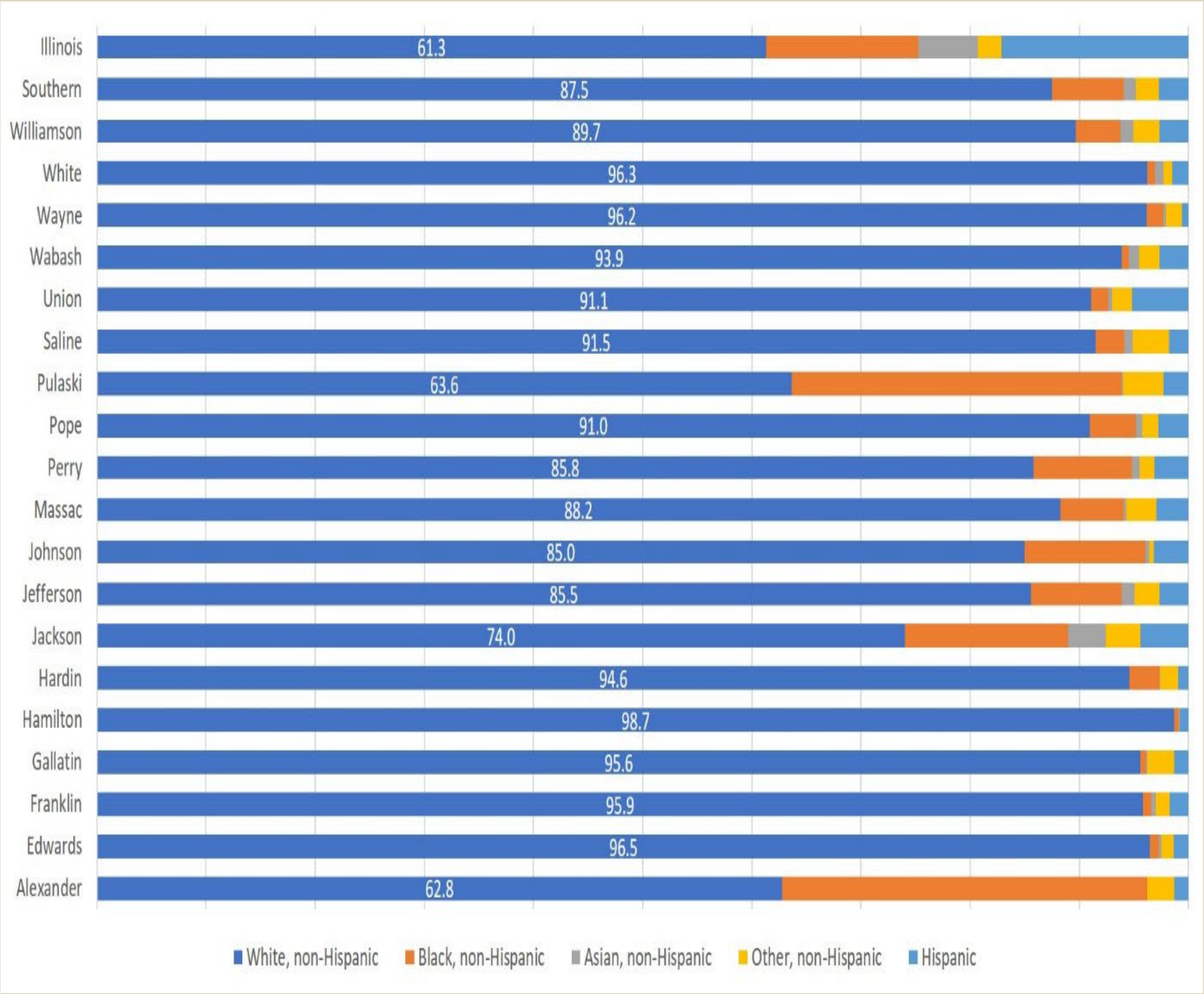


Fig.20 Race/ethnicity breakdown of the SEDR counties

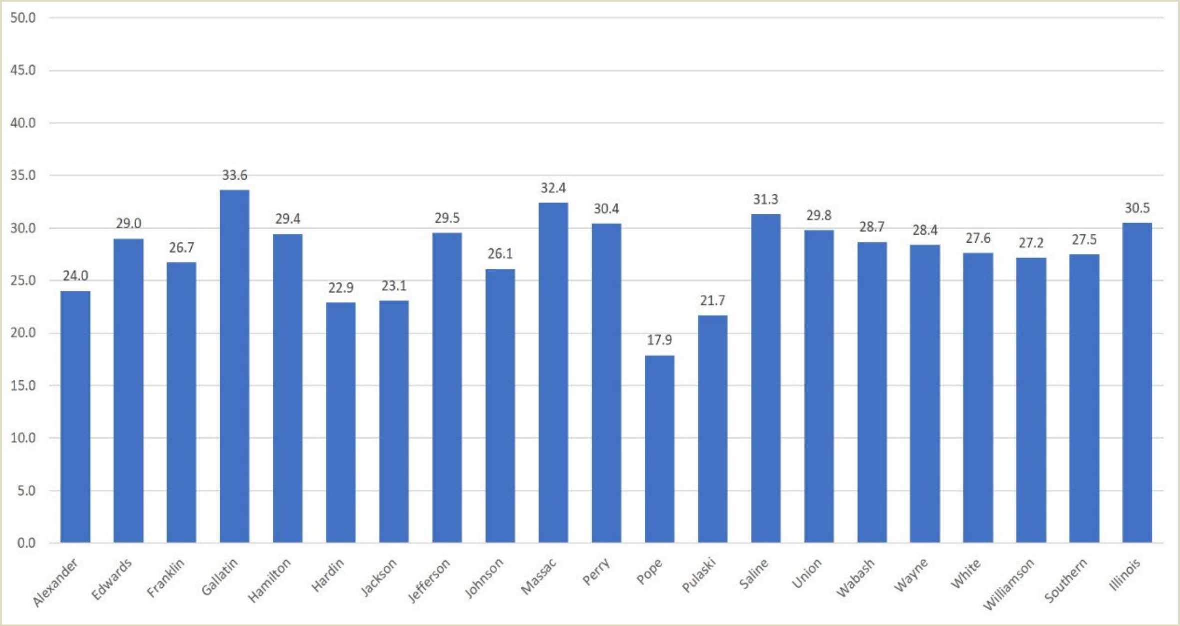


Fig.21 Percentage of households with children in SEDR counties

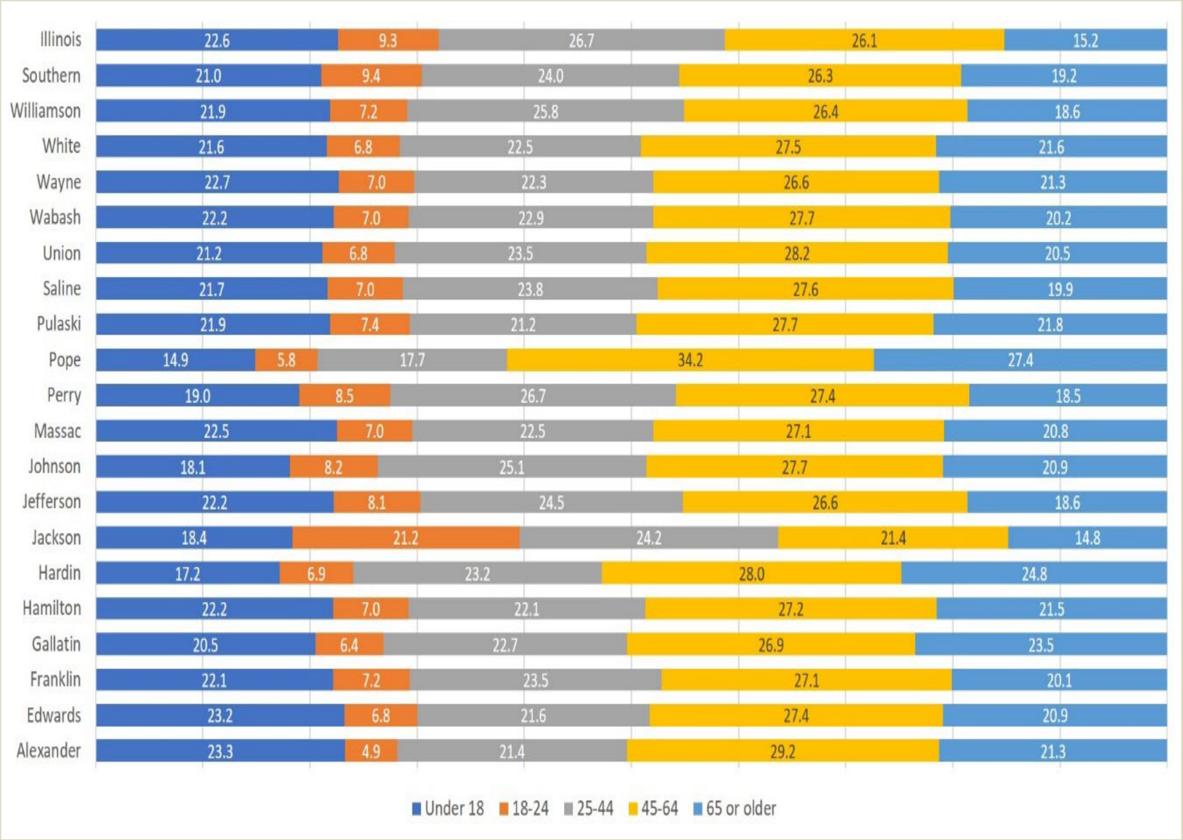


Fig.22 Age breakdown of the SEDR counties

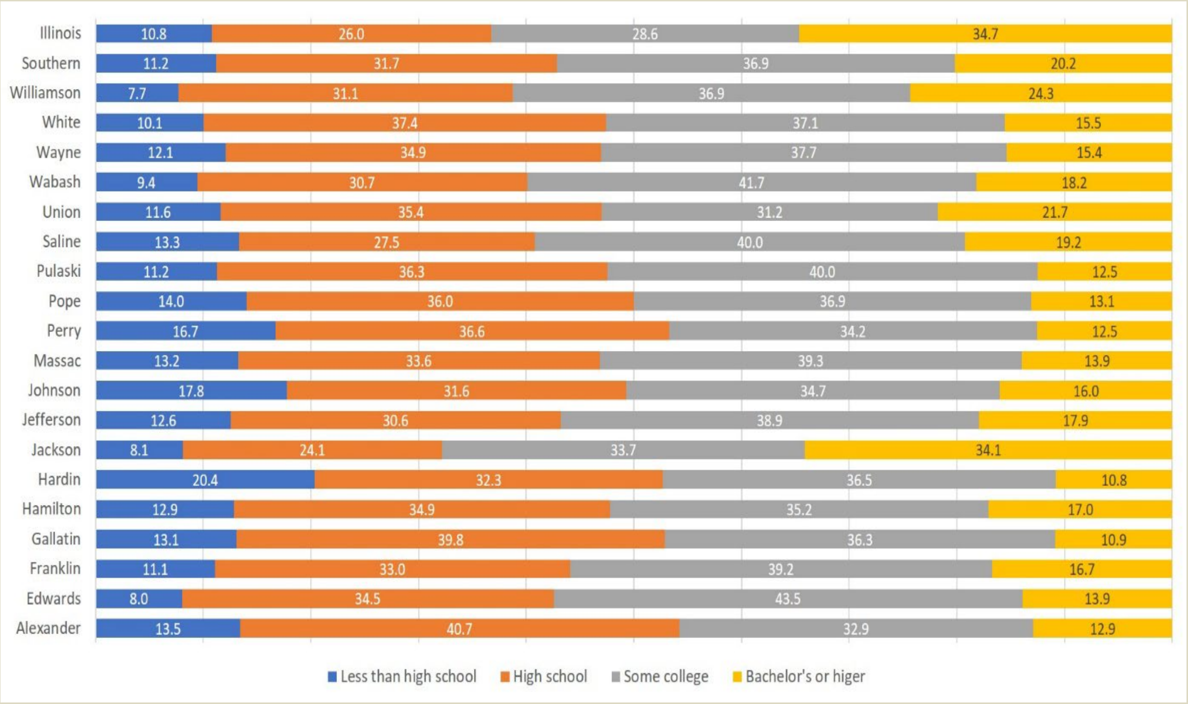


Fig.23 25+ years of age education-level breakdown in the SEDR counties

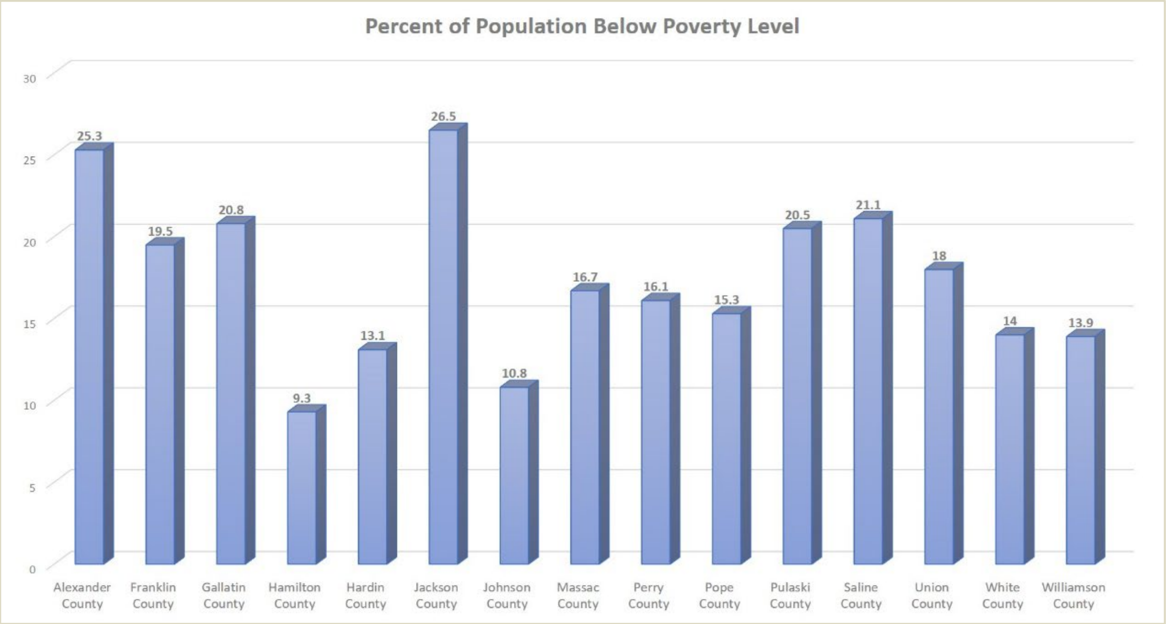


Fig.24 Poverty rate in the SEDR region

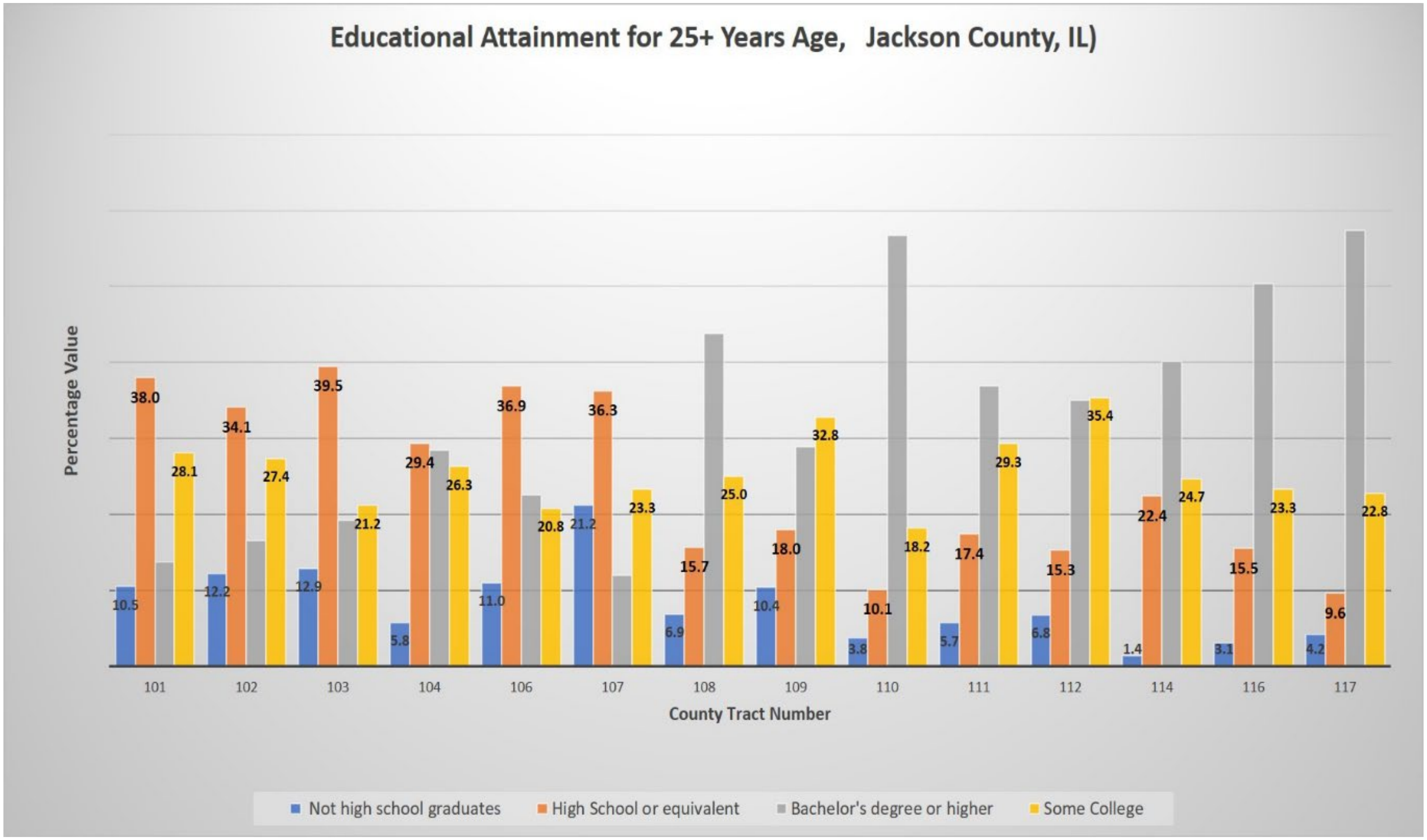
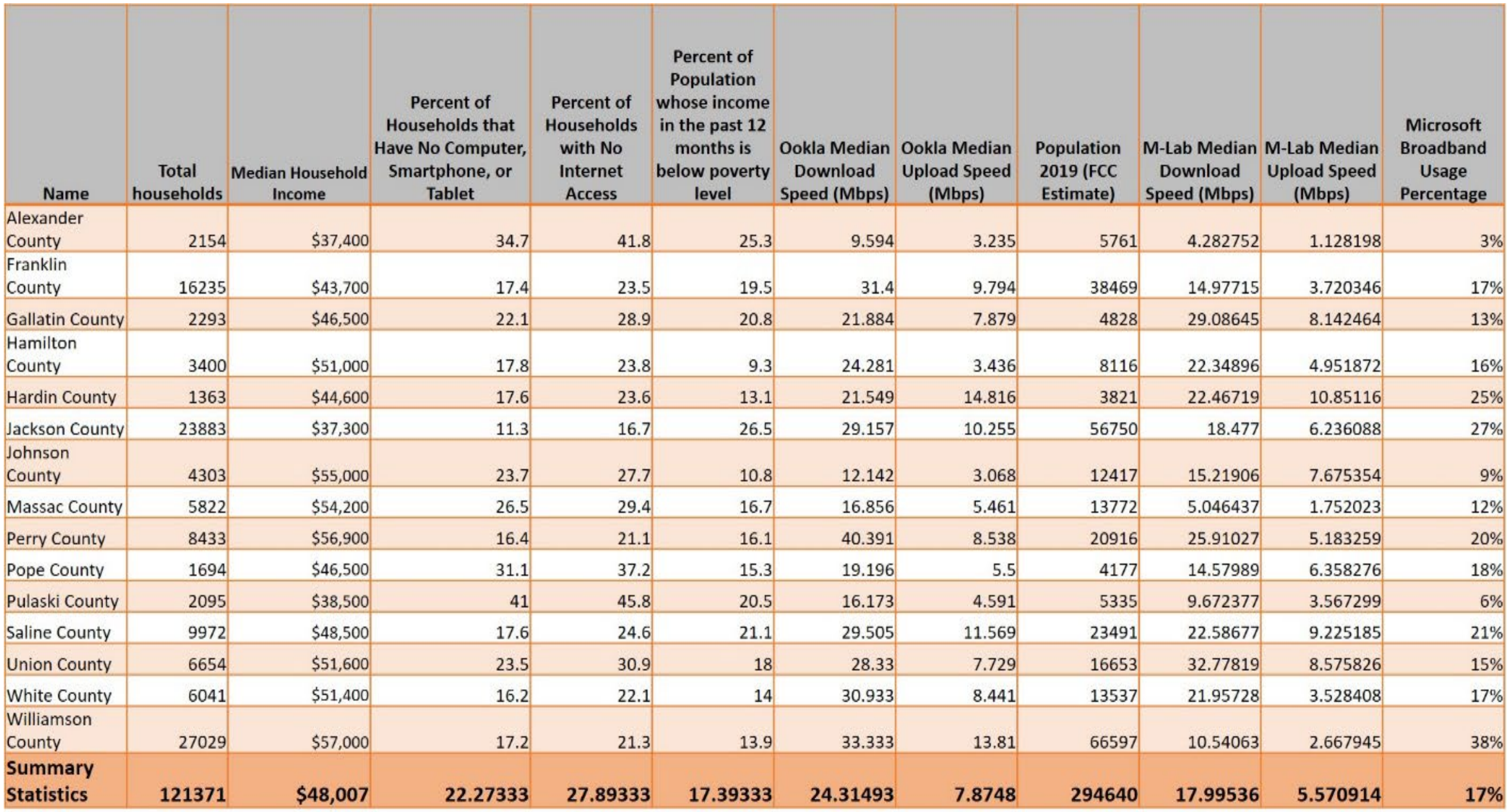


Fig.25 Education level at county tract level, Jackson County



Name	Total households	Median Household Income	Percent of Households that Have No Computer, Smartphone, or Tablet	Percent of Households with No Internet Access	Percent of Population whose income in the past 12 months is below poverty level	Ookla Median Download Speed (Mbps)	Ookla Median Upload Speed (Mbps)	Population 2019 (FCC Estimate)	M-Lab Median Download Speed (Mbps)	M-Lab Median Upload Speed (Mbps)	Microsoft Broadband Usage Percentage
Alexander County	2154	$37,400	34.7	41.8	25.3	9.594	3.235	5761	4.282752	1.128198	3%
Franklin County	16235	$43,700	17.4	23.5	19.5	31.4	9.794	38469	14.97715	3.720346	17%
Gallatin County	2293	$46,500	22.1	28.9	20.8	21.884	7.879	4828	29.08645	8.142464	13%
Hamilton County	3400	$51,000	17.8	23.8	9.3	24.281	3.436	8116	22.34896	4.951872	16%
Hardin County	1363	$44,600	17.6	23.6	13.1	21.549	14.816	3821	22.46719	10.85116	25%
Jackson County	23883	$37,300	11.3	16.7	26.5	29.157	10.255	56750	18.477	6.236088	27%
Johnson County	4303	$55,000	23.7	27.7	10.8	12.142	3.068	12417	15.21906	7.675354	9%
Massac County	5822	$54,200	26.5	29.4	16.7	16.856	5.461	13772	5.046437	1.752023	12%
Perry County	8433	$56,900	16.4	21.1	16.1	40.391	8.538	20916	25.91027	5.183259	20%
Pope County	1694	$46,500	31.1	37.2	15.3	19.196	5.5	4177	14.57989	6.358276	18%
Pulaski County	2095	$38,500	41	45.8	20.5	16.173	4.591	5335	9.672377	3.567299	6%
Saline County	9972	$48,500	17.6	24.6	21.1	29.505	11.569	23491	22.58677	9.225185	21%
Union County	6654	$51,600	23.5	30.9	18	28.33	7.729	16653	32.77819	8.575826	15%
White County	6041	$51,400	16.2	22.1	14	30.933	8.441	13537	21.95728	3.528408	17%
Williamson County	27029	$57,000	17.2	21.3	13.9	33.333	13.81	66597	10.54063	2.667945	38%
Summary Statistics	**121371**	**$48,007**	**22.27333**	**27.89333**	**17.39333**	**24.31493**	**7.8748**	**294640**	**17.99536**	**5.570914**	**17%**

Fig.26. Digital demographics data for southern Illinois region

VITA

Graduate School
Southern Illinois University Carbondale

Saiprasanna Cheedepudi

saiprasanna.cheedepudi@siu.edu

Andhra University, India
Bachelor of Technology, Computer Science Engineering, May 2018

Thesis Paper Title:
Analyzing the Digital Distress in Illinois Region

Major Professor: Dr. Koushik Sinha