

THE EFFECT OF SCHOOL FUNDING DECISIONS ON INCOME INEQUALITY: AN  
EXPLORATION OF MACHINE LEARNING FOR CAUSAL ANALYSIS

by

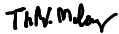
Benvin Fan Lozada

A Senior Honors Thesis Submitted to the Faculty of  
The University of Utah  
In Partial Fulfillment of the Requirements for the  
Honors Degree in Bachelor of Science

In

Economics

Approved:



Thomas N. Maloney (Aug 2, 2023 20:49 MDT)

---

Thomas Maloney, PhD  
Thesis Faculty Supervisor  
Chair, Department of Economics  
Honors Faculty Advisor *Locum Tenens*

---

Monisha Pasupathi, PhD  
Interim Dean, Honors College

August 2023  
Copyright © 2023  
All Rights Reserved

THE EFFECT OF SCHOOL FUNDING DECISIONS ON INCOME INEQUALITY: AN  
EXPLORATION OF MACHINE LEARNING FOR CAUSAL ANALYSIS

by


Benvin Fan Lozada

A Senior Honors Thesis Submitted to the Faculty of  
The University of Utah  
In Partial Fulfillment of the Requirements for the  
Honors Degree in Bachelor of Science

In

Mathematics

Approved:



---

Jing Yi Zhu, PhD  
Thesis Faculty Supervisor



---

Kevin Wortman, PhD  
Honors Faculty Advisor, Department of  
Mathematics



---

Tommaso De Fernex, PhD  
Chair, Department of Mathematics

---

Monisha Pasupathi, PhD  
Interim Dean, Honors College

August 2023  
Copyright © 2023  
All Rights Reserved

## ACKNOWLEDGEMENTS

Special thanks to:

Dr. Thomas Maloney

Dr. Jing Yi Zhu

Dr. Eunice Han

Dr. Gabriel Lozada

Dr. Jessie Fan

Benning Lozada

## ABSTRACT

The effects of public K-12 education funding disparities on student outcomes remain one of the most contentious issues in the realm of education. While many studies approach this problem through the analysis of discrepancies in short-term results such as test scores, no study has yet attempted to analyze the effect of discrepancies on long-run economic outcomes; this study attempts to fill this divide. To do so, we investigate the long-run economic health of children born from 1978-1983 and draw comparisons with school funding statistics from the 1991-1992 school year. We conduct this analysis using an instrumental variable approach combined with the deployment of machine learning regression algorithms in hopes to accurately model the causal impact of disparities in school funding. We find that machine learning models are more effective at modeling the causal relationship between school funding and income at age 35 than a standard linear regression model, using state fiscal neutrality scores as an instrument. We conclude that increases in school funding in the school district where a child grew up are causally linked to that child's outcome at age 35 and demonstrate that increasing school funding could be one potential solution to help remedy income inequality in the United States.

## TABLE OF CONTENTS

ABSTRACT	iii
1: INTRODUCTION	1
1.1: Literature Review	8
2: METHODOLOGY	11
2.1: Supervised Learning	13
2.2: Instrumental Variables	20
2.3 Data	23
3: FINDINGS	27
3.1: Descriptive Statistics	27
3.2: Instrumental Variable Analysis	30
4: CONCLUSION	39
REFERENCES	42
APPENDICES	48
Appendix A: Datasets	48
Appendix B: Figures	49
Appendix C: Code	55

## 1: INTRODUCTION

Since the early 1980s, the United States has experienced a dramatic increase in income inequality (Pew Research Center 2020). While a significant part of this trend has been caused by an increase in income at the top of the distribution, those at the bottom of the distribution have paid the price. Lower-income communities have faced consequences ranging from waning political power (Gielns 2012) to slipping health outcomes (Bor, Cohen, & Galea 2017), and these issues can lead to both damaged growth prospects nationwide (Furman & Stiglitz 1997) and an increase in social and political upheaval and violence (Muller, 1985). The impact of this inequality on our society is not just a theoretical concept; we can observe its effects through the unrest and division that have caused significant disruptions to the political, economic, and social climate of the 2010s and 2020s, with few indications of progress thus far. Therefore, it is crucial to identify effective solutions to address the escalating income inequality in our nation and safeguard stability and equity throughout the United States.

Income inequality is also tied to socioeconomic mobility, which is the ability of people to move up in their socioeconomic status. The ideas of mobility and self-determination underlie the fundamental promise of the American Dream, which provides hope that challenging work can allow people to pull themselves up by their bootstraps and build better lives for themselves. Such ideals, however, often omit the reality of opportunity available to those who most need it. In the United States, the likelihood of poor children staying poor and being unable to move up the socioeconomic ladder is higher than it has been in decades and is lower than almost

all other developed countries (The White House, 2013). Such limitations to intergenerational mobility are further complicated by further racial segmentation, with significant barriers existing creating gaps in socioeconomic mobility between African Americans, Hispanics, and whites (Collins & Wanamaker, 2017; Zan, Fan, & Lozada, 2023)<sup>1</sup>. This research highlights not only the instability of the present income inequality situation in the United States but also the inefficacy of current measures in narrowing this gap. Consequently, policymakers need to closely focus on this matter and explore potential solutions, with education being a key consideration among them.

Education as a means of socioeconomic mobility has been enshrined in the American consciousness; the opportunity to gain additional knowledge and skills has only become more important to succeed in an ever more competitive world, and as such the allure of educational attainment represents a ticket to a brighter future for many aspiring students. This is not simply an abstract benefit, but an actual one: when controlling for socio-demographic variables, it is estimated that men with bachelor's degrees earn \$655,000 more in median lifetime earnings than their non-degree holding counterparts, while women earn \$450,000 more in median lifetime earnings (Social Security Administration, 2015). This is not limited just to the college level, but applies to the high school level as well, where non-graduates have median weekly earnings of \$606 compared to \$749 for high school graduates (Bureau of Labor Statistics, 2019). These statistics are also coupled with the multiple positive societal externalities of education. People who receive an

---

<sup>1</sup> This study was conducted with assistance from the author.

education are less likely to commit crimes (Machin, Marie, & Vujić 2011), have better health outcomes in terms of quality and length of life (Eide & Showalter, 2011), and are even less likely to experience deaths of despair, morbidity, and emotional distress (Deaton & Case, 2022). The overwhelming appeal of these statistics has led to a wide promotion of education as one of the best solutions to the current economic mobility crisis the United States is currently experiencing; however, it is not without its unique struggles.

One of the main barriers to the promotion of education is that educational opportunities are not uniformly distributed across our society, but rather are influenced by a variety of social and economic factors. Economically, children whose parents are in the top 1% of the income distribution are 77 times more likely to attend an Ivy League university than those whose parents lie in the bottom 20% (Chetty, Friedman, Saez, Turner, & Yagan, 2017). Educational attainment also differs across demographic groups; a study conducted over 26 years found that dropout rates were least among whites, females, and those whose parents received a postsecondary education or owned a home. Factors associated with higher rates of dropping out included being non-white, male, residing in a large city, and having a female head of household (Hauser, Simmons, & Pager, 2000). These disparities may exist for a variety of reasons: studies have found issues ranging from a lack of quality information regarding a student's outcomes and opportunities among low-income communities (Hoxby & Turner, 2015) to a lack of English proficiency among some immigrant communities (Lutz 2007), but one of the most critical and most



studied barriers to educational attainment has been the funding of K-12 schools throughout the United States.

Funding of K-12 schools in the United States has typically been seen as a state issue, with little in the way of federal intervention; only 7.9% of funding for public K-12 schools comes from the federal government, with states having broad discretion over how, and how much, to fund school districts within their territory (Education Data Initiative, 2022). This broad discretion has led to a large variety of funding methods, the most prominent being the property tax funding model. The usage of a property tax to fund universal public schooling has its roots in the Reformation and has developed to become one of the most longstanding Anglo-Saxon traditions regarding education. This movement towards compulsory education constituted one of the most significant contributions of the Reformation, and the property tax has become inseparable from the movement for educational attainment (Walker 1984).

However, a critical problem arises with the utilization of a local property tax to fund schools: it leads to significant disparities in funding between low and high-income neighborhoods. Specifically, neighborhoods with substantial tax bases are more capable of providing extensive support to the schools in their vicinity compared to neighborhoods with smaller tax bases. This discrepancy can be attributed to the fact that neighborhoods with large property tax bases often correspond to areas where residents are more affluent. Consequently, this system of funding exacerbates educational inequality to a considerable extent; schools in financially prosperous neighborhoods have access to more resources, including

better facilities, highly qualified teachers, and a wider range of extracurricular activities. Conversely, schools located in neighborhoods with smaller tax bases face significant challenges in adequately supporting their students. These schools struggle to secure sufficient funding for essential resources, resulting in outdated textbooks, limited technological infrastructure, overcrowded classrooms, and a reduced range of educational opportunities. As a result, students in these schools are at a severe disadvantage, hindering their academic progress and limiting their potential for future success.

Such a system not only perpetuates educational disparities but also perpetuates broader socioeconomic inequality. By granting greater advantages to already advantaged communities, the local property tax-based funding system widens the gap between the rich and the poor, further entrenching social stratification. This cycle of inequality becomes self-reinforcing, as individuals from underprivileged backgrounds are more likely to face barriers to educational achievement, limiting their ability to improve their economic circumstances. This relationship has been documented as existing in the United States since the 19<sup>th</sup> century (Vollrath, 2013), but was most notably challenged in the Supreme Court case *San Antonio Independent School District v. Rodriguez*, where a 5-4 majority of the Court ruled that “education ...[is] not within the limited category of rights recognized by this Court as guaranteed by the Constitution”, and that the property tax funding method “[did] not disadvantage any suspect class” (San Antonio Independent School District v. Rodriguez, 1973). *San Antonio* both effectively authorized discrimination against the poor and stated that education was not a

basic human right protected by the Constitution; as such, the decision spurred immense criticism both on its legal and moral grounds (Gammon, 1976; Walsh, 2011) and led to a national movement to reform school funding policies.

This national movement to reform school funding provides an ideal window to investigate the true effect of school funding on economic mobility and forms the basis of this research study. Throughout this paper, we analyze the effect of school funding levels and state finance policies on the long-run income of students. We focus on students born between 1978 and 1983, as these students were born only a few years after *San Antonio* and were likely to be attending public schools during the school funding reformation movement. To gain accurate representations for school funding and state policies, we utilize the year 1992 due to two main reasons. First, this is the only year wherein all students in our target group would be in early adolescence, and thus in a key period of learning, growth, and neurobiological development (Dahl, Allen, Wilbrecht, & Suleiman 2018). Secondly, this period is where we see some of the most significant variation in school funding policies; many reform policies were put in place before 1992<sup>2</sup>, but significant variation still existed in school funding policies nationwide<sup>3</sup>.

This study stands to contribute to the literature in two distinct ways. First, this study aims to fill a gap in studying the long-term effect of school funding on students; the majority of studies focus on shorter-term outcomes, which are slightly

---

<sup>2</sup> Examples include the policies adjustments to the school funding policies mandated in *Serrano v. Priest* in California in 1977 and the implementation of Abbott Districts in New Jersey in 1985.

<sup>3</sup> Inequitable school funding policies were in effect in Arizona and Texas but would be struck down in Texas (*Edgewood Independent School District v. Kirby*, 1993) and Arizona (*Roosevelt Elementary School District v. Bishop*, 1994) soon after.

easier to measure but are not necessarily indicative of future career or personal success. The second notable addition to the literature is related to the innovative methodology utilized in this study. This study deploys an instrumental variable approach to approximate the causal effect of differences in school funding on student outcomes but does so using a unique expansion of the standard implementation of instrumental variables typically done in econometrics: machine learning. Machine learning and artificial intelligence algorithms have an incredible potential to find relationships that are more nuanced than the standard regression models often utilized in instrumental variable analysis, but have almost never been utilized alongside instruments. As such, this study provides one of the first implementations of machine learning into the framework of instrumental variables, which can provide a new innovative method to find causal relationships between variables in a non-controlled setting using real-world datasets and outcomes.

Our study aims to answer the following question: is there a significant link between the school district where a child grew up and their long-run economic outcomes? To do so, we will begin with a literature review of the existing research regarding educational funding and its relationships to intergenerational mobility. We then discuss the innovative methodology of our study, using a combination of machine learning models and causality methods to accurately ascertain the nature of the relationship in question. We will then demonstrate the study's findings, including comparisons of several different analysis methodologies to conduct a survey of analysis methods. Finally, we will summarize our findings and discuss their implications for both education and statistical research projects in the future.

## 1.1 Literature Review

Previous studies on the effectiveness of school funding have focused on a variety of educational outcomes; however, no study has yet been conducted to link school funding to real-world long term student outcomes<sup>4</sup>. In terms of the existing literature, increases in school funding have been linked to increases in student academic achievement, often measured by performances on standardized tests (Tow, 2006; Sebold & Dato, 1981; Neymotin, 2010). Furthermore, increases in funding for K-12 institutions have been tied to reduced dropout and increased graduation rates, as well as increases in college enrollment numbers (Kreisman & Steinberg 2019). These positive results are related to a variety of factors, including that schools with better funding are more likely to have better staffing ratios, class sizes, and teacher wages (Baker, Farrie, & Sciarra 2016). These impacts have been found on both broad and granular levels: increases in targeted school special education funding have been found to increase the number of students with disabilities meeting educational standards (Cruz, Lee, Aylward, Kramarczuk, & Voulgarides 2022), further demonstrating the value of funding and resources to achieving educational goals.

However, the literature also highlights significant disparities in school funding across the United States. Funding gaps between high and low-poverty school districts can reach as high as over \$1,000 per student nationwide, as can the gap between white and non-white students (Carey, 2004). Further studies have shown the existence of specific disadvantages among multiple racial and ethnic

---

<sup>4</sup> Lovenheim and Willen (2019) examined teacher collective bargaining and long run student outcomes such as annual earnings but did not study school expenditures specifically.

groups. One study demonstrated that adequate spending to achieve national average outcomes was less common in districts with large proportions of Latinx students across the United States, while another showed that school districts with large populations of nonwhite students, economically disadvantaged students, or English Language learners were likely to receive less funding per student than less diverse districts, even after adjusting for the additional state and federal government funds made accessible to them (Baker, Srikanth, Cotto Jr, & Green 2020; Weiss 2020). These persistent gaps have led to a substantial body of work dedicated to providing actionable solutions to these disparities. While the details of this research lie beyond the scope of this review, example solutions include changing school funding models to better reflect principles of equity and evaluating the payment structures of staff to gain more well-qualified and experienced teachers (Roza & Miles, 2002; Adamson & Darling-Hammond 2012).

However, these conclusions are still debated: some summaries of the literature have claimed that there is not a strong or consistent relationship between school funding and performance after accounting for variations in family inputs (Hanushek, 1997), and think tanks from the conservative-leaning Heritage Foundation to the liberal-leaning Center for American Progress have duelled over whether race-based disparities in school funding are myth or reality (Richwine, 2011; Epstein & Miller, 2011). In the face of this continual uncertainty, clarity is needed on the definitive effects of education funding to cut above the noise and provide a clearer signal as to the issues and remedies of the current school funding situation in the United States.

However, the quality of an education is a difficult metric to ascertain; simple measures such as test scores are deeply problematic in their attempts to measure talent and intelligence (Rooney & Schaeffer 1998; Croizet & Dutrévis 2004), and long-term data from students, sorted by school district of origin, are hard to come by. In this, we turn to a neighboring literature on economic inequality and geographic location, particularly the work of Dr. Raj Chetty and the Opportunity Insights project based at Harvard University. Using unique access to data from the US Census Bureau and Internal Revenue Service, Dr. Chetty's team assembled a novel dataset connecting people's current income to the census tract where they were born and raised. Research borne from this project has detailed the effects of geographic neighborhoods on future life outcomes and shown that where a child grows up is the single biggest predictor of their future success (Chetty and Hendren, 2018). Given the impact of schooling on young children, and the geography-based nature of school districts, school funding may be a major factor in this perceived predictor. As such, we concentrate on the gap in the literature between school funding results and future income calculations and conduct this study to bridge these two divides.

To summarize, previous literature has found numerous positive educational benefits that are associated with an increase in school funding, including higher performance on standardized tests. However, no studies have connected school funding to long-run student economic outcomes, and some studies dispute if any relationship exists. To resolve this, we focus on datasets linking people's income at age 35 to their census tract of origin to help provide context to this issue.

## 2: METHODOLOGY

The methodology that we will utilize to analyze this issue centers on the intersection between two different, yet connected, methods in mathematics and statistics: machine learning models and instrumental variables. Machine learning is a method, or more accurately a set of methods, which leverage computational algorithms to analyze and extract patterns from large datasets. It encompasses various techniques designed to enable computers to “learn” from data and make predictions or decisions without being explicitly programmed and is often used to find unique patterns and relationships within datasets. Instrumental variables, on the other hand, are a statistical technique designed to allow causal inferences to be made in non-controlled experimental designs. This usefulness applies to a considerable number of fields, including statistics, econometrics, and epidemiology, where controlled experiments are difficult to conduct (Lousdal, 2018).

We begin by detailing a key limitation on our deployment of machine learning. Our primary focus in this study lies on supervised learning models, rather than unsupervised learning models. Supervised learning involves training models with labeled data, where each data point is associated with a known outcome or value. The primary objective of supervised models is to map input data to corresponding output data accurately. By learning from this labeled data, these models can recognize patterns and associations within the data and can subsequently make precise predictions or classifications for new, unseen instances. On the other hand, unsupervised learning deals with unlabeled data, where no explicit outcome or value is assigned to each data point. Instead of predicting



specific outcomes, unsupervised learning algorithms focus on discovering underlying structures, patterns, or clusters within the data. This makes unsupervised learning valuable for tasks such as data exploration, anomaly detection, and clustering, but not well-fit for predictive uses.

In our research, the central goal is to investigate the impact of numerous factors, such as school funding statistics, on long-run income. To achieve this, it is essential to establish meaningful connections between the independent variables (e.g., school funding) and the dependent variable (e.g., long-run income). Therefore, we require models capable of accurately predicting or classifying the long-run income based on the input variables, which aligns perfectly with the objectives of supervised learning. Unsupervised learning is primarily concerned only with discovering patterns and lacks the ability to make specific predictions or classifications; thus, it would not provide us with the necessary insights to establish cause-and-effect relationships between the variables of interest.

Secondly, this study will focus on a few key basic supervised machine learning models. In addition to standard linear regressions and their variations, the focus will be on the k-Nearest Neighbors (k-NN), Decision Trees, and Feedforward Neural Networks methods. These are three of the main supervised learning methods designed for regression (i.e., predicting a continuous variable) rather than classification (i.e., predicting a discrete variable). Other methods, such as Naïve Bayes or Support Vector Machines (Support Vector Regression, an alternative usage, is suitable for regression), are primarily utilized for classifying data points into discrete groups and thus do not fall into the scope of analysis detailed in this study.

## 2.1 Supervised Learning

As detailed above, supervised learning models are methods for predicting an outcome group or value based on a set of inputs. A subsection of supervised learning models are focused on regression, which is the aim of accurately predicting continuous variables. However, multiple variations on supervised learning models exist, and their limitations are important to understand when drawing conclusions regarding their appropriateness in deployment. This section details the main mechanisms that underlie supervised learning models and the resulting limitations. For our framework, we mainly follow the textbook *Introduction to Machine Learning* by Alex Smola and S.V.N Vishwanathan (2008), with non-substantive deviations for simplicity and readability.

A basic supervised machine learning model implementation proceeds as follows. We begin with a dataset of  $n$  independent and identically distributed (iid) observations, each with a feature vector  $x_n$  representing the input characteristics of the observation  $n$  and a feature vector  $y_n$  representing the output characteristics of the observation  $n$ . For supervised machine learning algorithms in their implementation in this study, vector  $y_n$  represents the value of the output variable from the observation in the dataset  $n$ . These vectors are first paired by observation, generating a set of values of the form  $\{(x_1, y_1), (x_2, y_2), \dots, (x_{n-1}, y_{n-1}), (x_n, y_n)\}$ . The observations are then often split into two groups, the training group and the testing group. There is no one “best” split due to the bias-variance dilemma, which will be discussed later in this section; rather, the percentage of observations placed in the training group vs. the testing group is selected at the initialization of the machine

learning model. This is to allow for a model to be evaluated to ensure that the model retains a certain degree of accuracy on data on which it was not trained.

The training group is then used to train the machine learning model, whose final goal is to generate a function  $f: X \rightarrow Y$  that maps the input space  $X$  to the output space  $Y$ . Each machine learning method has its own method of generating and optimizing the function  $f$ , but its primary target is to minimize the difference between the predicted output and the actual output for each pair in the training dataset. Mathematically, this is often termed as the minimization of the empirical risk of the function  $R(f)$ , which is itself modeled as the expected value of the loss function of  $f$ , modeled as  $L(y_i, f(x_i))$ . As such,  $R(f) = \mathbf{E}[L(y_i, f(x_i))] = \int L(y_i, f(x_i)) dP(x, y)$ , where  $P(x, y)$  represents the joint probability distribution of  $X$  and  $Y$ . The loss function  $L$  models the difference between the value of the output variable predicted by the function  $f$  and the actual value of the function, using methods such as Mean Square Error. A detailed examination of loss functions is beyond the scope of this paper but can be found in Smola and Vishwanathan's text as well as other areas in the literature (Wang, Ma, Zhao & Tian 2020). The function that minimizes the empirical risk function  $R(f)$  is the final product of the model and is selected for use.

While the final function  $f$  has been selected, it has not yet been assessed for its accuracy on data on which it was not trained. This is where the testing dataset comes in; due to the split between the training and the testing dataset, the model has never encountered the observations in the testing subsection of the dataset, and thus will allow us to gauge the ability of the function to generalize to new observations. The performance of the model on these testing observations can thus

provide insight into the true effectiveness of the model. This is important due to one of the most critical issues that plagues machine learning algorithms: overfitting.

Overfitting is a term for when a model is too well-fitted and tailored to the dataset that it was trained on and cannot be effectively generalized to data points outside of its original training set. This occurs when the machine learning model detects patterns in the dataset that do not represent underlying trends but rather the underlying randomness and noise in the data; this means that the model has a much higher accuracy on its own training set, where these idiosyncrasies exist, than on other sets where they do not. This contrasts with an underfitted model, where the model cannot capture either the underlying noise, or the actual fundamental underlying patterns in the data, which means that this model also cannot effectively be generalized to new datasets. This leads to a delicate balance that needs to be struck when creating machine learning models: models need to be complex enough to effectively capture important patterns within the dataset, but not so complex that they begin to account for fabricated patterns caused by noise in the data. This is an example of the bias-variance dilemma, which affects all supervised learning models.

The bias-variance dilemma states that there is a tradeoff between bias and variance; if you try to reduce one source of error, you inevitably increase the other source. When applied to overfitting, we see the exact tradeoff; an overfitted model will result in high variance because the model has specified itself to the training dataset and cannot accurately predict new datapoints, while an underfitted model will result in high bias due to the model being unable to pick up on the fundamental trends in the data. To counter the bias-variance tradeoff, several methods exist; for

the purposes of our analysis, we will use cross-validation techniques to do so, although it is also possible to penalize complex functions during model selection.

We conclude this section with a summary of the exact methods of machine learning that will be utilized throughout the course of this study: Linear Regressions, Decision Trees and Random Forests, K-Nearest Neighbors, and Feedforward Neural Networks.<sup>5</sup> Each method is a supervised machine learning model suitable for the analysis of continuous dependent variables.

Linear regression is widely considered one of the most fundamental, and thus most widely utilized, forms of machine learning. Linear regression attempts to model a linear relationship between the input variables and the output value through finding a best-fit line that minimizes the difference between the model's predicted output values and the actual output values. A simple one-variable linear regression takes the form of  $\hat{y} = \beta_0 + \beta_1 x_1 + \varepsilon$ , where  $\hat{y}$  represents the predicted value of the output,  $\beta_0$  represents the y-intercept of the regression line (the predicted value of  $y$  when  $x = 0$ ),  $\beta_1$  represents the change in the predicted value of the output for a 1-unit increase in the value of the input variable,  $x_1$  represents the value of the input variable, and  $\varepsilon$  represents the value of the error term of the regression equation. Such a model can be expanded to include more variables, non-linear relationships between variables, or even interactions between variables, but the basic principle underlying the model remains the same. However, one of the most notable features of most linear regression models is their usage of the least

---

<sup>5</sup> Information on these methods is gathered collectively from the work of Smola and Vishwanathan, and course notes from Math 4100: Introduction to Data Science at the University of Utah by Dr. Anna Little and Dr. Alexander Lex (Spring 2021)

square error, or the LSE, to develop their regression model. The least square error is used to develop a line of best fit through the dataset by attempting to minimize the sum of the squared residuals between each data point and the line of best fit. The method works through the adjustment of the  $\beta_0$  and  $\beta_1$  parameters to minimize the residual sum of squares (RSS), modeled as  $\sum_1^n (y - \hat{y})^2$ . This optimized result gives the line of best fit under least-squares, which allows for the easy interpretation of linear regression predictions. Given this ease and their ubiquity, we will utilize the basic linear regression framework as a starting point for the purposes of our study.

Decision trees, another method of supervised machine learning, are multi-tiered models that repeatedly recursively divide the input feature space based on the values of the input features. Each tier consists of a series of decision nodes, which each represent a decision to be made depending on the value of one or more input features. These decisions select a path that either leads to further decision nodes which further divide the feature space, or to a “leaf” node which corresponds to a final prediction on the value of the input feature. Decision trees provide a clear and simple method for interpretation, as any prediction can be boiled down to a sequence of individual decisions based on the input features. However, decision trees are often prone to overfitting datasets, and can thus prove difficult to generalize. As such, multiple decision trees are sometimes combined in an ensemble learning method known as Random Forests; in Random Forests, multiple decision trees are trained on different random subsets of the data using bootstrapping, and then averaged to create a final prediction model. This means that Random Forests, while being slightly harder to interpret, can improve the generalization

performance of Decision Trees. While Decision Trees will be discussed throughout the analysis of different machine learning models, Random Forests will be the primary method used in the study due to its applicability for general prediction.

Similar to Decision Trees, K-Nearest Neighbors (k-NN) is a non-parametric method for predicting outcomes based on a set of input values; however, instead of using the multi-tiered approach of Decision Trees, KNN predicts an output for a data point based on the averaging of the  $k$  nearest data points in the feature space. The main control for overfitting in k-NN is the selection of  $k$ , which determines the model's balance between complexity and flexibility. To determine the  $k$  nearest data points, k-NN relies on calculations of distance in a  $d$ -dimensional space, where  $d$  represents the number of input features; this leads to some complexity regarding measurements of distance for continuous and discrete variables. Continuous variable measurements often rely on Euclidian distance in introductory settings, but there are a wide range of metric spaces that can be utilized to run K-Nearest Neighbors algorithms. Examples of such include the Manhattan distance (where distances are calculated using the sum of the absolute difference between their coordinates, similar to how someone would move across city blocks) and other metrics from the broad family of Minkowski distances, although others exist as well. The choice of the distance function is a critical element of K-NN to ensure that the algorithm is well-prepared to model the relationships in the data. As such, given the relative simplicity of our planned analysis and the widespread use of the Euclidian metric for univariate modeling, we choose to utilize it moving forward to demonstrate a suitable and standard implementation of the K-NN algorithm.

While significantly less easy to interpret than other machine learning methods, neural networks are some of the most powerful models available for data analysis. Neural networks are designed to mimic a human brain through a collection of layered interconnected artificial neurons, which can each process information and signal other neurons throughout the network. Through this, neural networks can very effectively “learn” patterns in large datasets and uncover unique patterns that other algorithms may not always be able to find. This characteristic, especially when using multiple layers of neurons (a method known as deep learning), makes them useful for a variety of tasks, including common deployment in image and speech recognition; however, these same traits make them desirable for their potential usage in regression models. Feedforward neural networks, which we will be focusing on throughout this study, are the simplest model within the neural network family, yet retain much of their computational abilities. Feedforward neural networks, also known as multilayer perceptrons (MLPs), consist of input, hidden, and output layers of interconnected nodes (neurons). The input layer takes each feature value as an input and allows the hidden layers’ nodes to apply a complex series of non-linear transformations to the weighted sum of its inputs, which is finally outputted as a regression result. As a result of this multi-layer design, neural networks with multiple hidden layers are capable of learning incredibly nuanced and complex relationships between variables, often more so than the aforementioned machine learning models. As such, controlling the number of layers and nodes within the network is critical to reduce overfitting and ensuring the generalizability of the model and the meaningfulness of the network’s results.



## 2.2 Instrumental Variables

Instrumental Variables are variables utilized to attempt to infer causality from data that was not produced in a controlled or experimental setting. Given the various confounding issues that can occur with uncontrolled data, the instrumental variable approach specifically endeavors to address endogeneity in regression models. Endogeneity is the phenomenon when an input variable is correlated with the error term in a regression, which can lead to an inability for the model to accurately model relationships between variables. This can occur for two main reasons: Omitted Variable Bias, when a significant variable was not included in the model, or Simultaneity Bias, when the outcome and input variables influence each other (Lynch & Brown, 2011)

Omitted Variable Bias (OVB) can interfere with causal analysis through a failure to model the accurate nature of the relationship between the input and output variables, which can lead to confounding effects and flawed predictions of the true relationship between the variables. Simultaneity Bias can interfere with causation due to an obstruction of the true relationship between the input and output variables; regression models are only designed to predict the output given a certain feature space, and errors can occur if the output affects the input feature space itself, which can lead to elements of reverse causality (where the output actually causes changes in the input feature space rather than vice versa) or generate a feedback loop (where the input space and the output space both exert influence on each other). Given the extensive time span that we study, we have little concern for simultaneity bias in our model; this form of bias cannot typically occur

with an output that occurs significantly after the input is applied, as is the case with education and long-run income figures; while it is possible for the past to influence the future, it is impossible for the future to influence the past. Thus, we turn our attention to the most significant source of interference; omitted variable bias.

OVV poses a significant issue in our study; given the considerable time gap between our input feature space and our output, there are a nearly unfathomable number of omitted variables to be considered when attempting to draw causal results from our dataset. As such, we assume that any standard model that we utilize regressing school funding on long-run outcomes will have some form of endogeneity caused by OVB. This does not mean that such an exploration is not valuable; we will conduct these regressions as well, but mainly in an exploratory capacity. For causal inference, however, we will need to use the instrumental variables method to control for this issue.

To understand how instrumental variables address endogeneity, we begin with a basic instrumental variable model using simple one-variable linear regression. Assume we have a one variable regression model of the form  $\hat{Y} = \beta X + \varepsilon$ ; we can model endogeneity through the formula  $cov(X, \varepsilon) \neq 0$ . We want to find an instrumental variable  $Z$  that is correlated with the endogenous variable  $X$  ( $cov(X, Z) \neq 0$ ) while being independent of the error term ( $cov(Z, \varepsilon) = 0$ ). Thus,  $Z$  must comply with a few key assumptions. The first is the exogeneity assumption, which assumes that the instrumental variable  $Z$  is uncorrelated with the error term  $\varepsilon$  in the structural equation ( $cov(Z, \varepsilon | X) = 0$ ) to ensure that the instrument affects the endogenous variable  $X$  only through its correlation with  $X$ . The second assumption

is relevance, which states that the instrument must be correlated with the endogenous variable ( $\text{cov}(Z, X) \neq 0$ ) to capture some of the variation in  $X$ . The last assumption is the Exclusion Restriction Assumption, which states that the only effect that the instrumental variable can have on the outcome value is through its effect on the endogenous variable ( $\text{cov}(Z, Y|X) = 0$ ). This is essential to ensure that the instrument isolates the effect of the endogenous variable on the outcome value.

The primary estimation method of instrumental variables is conducted in two stages. In the first stage, the instrument  $Z$  is regressed on the endogenous variable  $X$  to gain the predicted values of the endogenous variable based on the instrument, modeled by  $\hat{x} = f(z, \varepsilon_1)$ . The second stage consists of using the predicted values of the endogenous variables to then estimate the causal effect of the endogenous variable on the outcome value, modeled by  $\hat{y} = f(\hat{x}, \varepsilon_2)$ . This process is most commonly using variants of linear regression, in which case it is referred to as a 2-Stage Least Squares model. The literature has outlined the appropriateness and benefits of integrating instrumental variables with machine learning in an abstract setting, but almost no studies have been done to apply these theoretical benefits into the analysis of real-world datasets. (Pech & Laloe 1997, Xu, Chen, Srinivasan, de Freitas, Doucet, & Gretton 2021). This study aims to fill this gap; through deploying a practical implementation of machine learning prediction models in concert with an instrumental variable analysis, we hope to both demonstrate the untapped potential of machine learning regression in noncontrolled experiments and provide a more reliable and insightful investigation into the true causal relationship between school funding and long-run income.

## 2.3 Data

To assemble the necessary information to conduct a long-term analysis of school funding effects, we procure data from four sources: the United States Census Bureau, the National Center for Education Statistics (NCES), the United States General Accounting Office (now called the Government Accountability or GAO), and the Opportunity Insights Project at Harvard University. From the United States Census Bureau, we utilize two datasets. The first dataset contains information from the U.S. Census Bureau and NCES's combined 1992 Annual Survey of School System Finances (Form F-33), which contains extensive information on school district revenues and expenses for school districts across the United States. For each school district, we supplement the above data with racial demographic information for all children enrolled in public schools within the district; however, no publicly accessible data exists for racial composition on any sufficiently granular level (census tract or school district) during this period. To provide a loose approximation for these levels, we use information from the inaugural American Community Survey in 2005, which provides such information; however, due to the 13-year time gap, this data is more so intended for dataset exploration than prediction. Lastly, we utilize data from the 1980 US Census Bureau regarding census tract poverty rates, which are sourced from the Public Health Disparities Geocoding Project at Harvard University. To standardize tracts, we link this information to 2010 census tracts using the Longitudinal Tract Database (LTDB) developed by researchers at Brown University.

The second source of data that we utilize originates from the June 1998 report from the General Accounting Office entitled “*State Efforts to Equalize Funding Between Wealthy and Poor School Districts*” (General Accounting Office 1998). The GAO conducted an extensive investigation of equity in school funding across states, with a focus on tax base targeting efforts and income gaps between schools. The investigation contains data on all 50 states for the 1991-1992 school year, which is the same time span as our other datasets encompass.

Our third source of data is the Opportunity Insights Project at Harvard University. The organization’s mission is to “conduct scientific research using “big data” on how to improve upward mobility and work collaboratively with local stakeholders to translate these research findings into policy change” (Opportunity Insights, 2023). The specific dataset utilized in the study is termed the “Opportunity Atlas,” which is a collaborative effort between Opportunity Insights and the Census Bureau to release national social mobility data. This data contains detailed information about children across the US born from 1978-1983 and their long-run economic outcomes, including their eventual income information in adulthood, which was gathered from the Internal Revenue Service. The dataset aggregates the information for individual children by the census tract where they grew up; as such, we have the need to connect tracts to the school districts serving them.

This is where the final data source of note comes in; the School District Geographic Relationship Files from the National Center for Education Statistics. Given the various sets of geographical units in standard usage throughout the US Government, the Geographic Relationship Files (GRF) are a set of files intended to

store the relationship between school districts and other sets of units, including census tracts. As such, we use these files to merge information regarding individual school districts with information regarding student outcomes. We utilize the 2013 GRF due to the format of the dataset from Opportunity Insights to ensure that both datasets utilize tracts from the 2010 Census. To concatenate information between school districts in 1992 and 2013, we utilize each district's unique NCES Local Education Agency ID (LEAID) numbers, which are time-invariant.

In our final cleaned dataset, we have a total of 56,408 valid Census tracts for which all our major variables under consideration were available. Each tract is measured on up to 254 variables; some tracts are missing minor information on some variables within the larger feature set (such as missing income projections for a certain race or gender), but each tract had information on school district expenditures and revenue, state fiscal neutrality scores, projected income at age 35 and poverty percentage in 1980, which constitute the main features of the set. For Census tract information, we experience multiple Census tracts which contain the same values for school district funding; this is because multiple Census tracts are often serviced by one school district. We keep the data in this format to avoid an overrepresentation of small school districts and an underrepresentation of large districts, as the set population of census tracts allows us to appropriately weigh larger districts through the inclusion of additional district datapoints. Of the 254 variables, 123 of them are variables relating to different school funding amounts, 108 are income characteristics at age 35 disaggregated by race, gender, and parental income bracket, 19 are school district demographic characteristics as determined by

the American Community Survey, one is the Fiscal Neutrality Score as determined by the GAO, one is the percentage poverty rate in the Census tract during the 1980 Census, and the remaining seven (7) are informational variables containing information such as state names.

In our data cleaning stage, we make 3 notable changes to our existing dataset. First, we adjust for the split between 1980 and 2010 Census tracts through averaging the poverty rates of separated 1980 tracts to provide unified and accurate 1980 poverty rates for 2010 tracts. Second, we exclude the state of Hawaii from our study; in the year 1992, Hawaii had only one school district, making it impractical for district-by-district comparisons (GAO, 1997). As such, including Hawaii in the analysis could potentially introduce inaccuracies in the results. Lastly, we create a subsection of the existing dataset that focused on districts in high-poverty areas. Specifically, we identified 2010 census tracts with a 1980 poverty rate above 13.9875%, which represented the 75<sup>th</sup> poverty rate percentile of census tracts in the sample. By defining high-poverty areas in this manner, we hope to focus on districts with a history of higher poverty rates, allowing for a targeted examination of the relationship between school funding disparities and socioeconomic conditions.

In our data cleaning, feature engineering, and analysis processes, we utilized the Python programming language, along with the matplotlib, pandas, and sklearn libraries. All datasets that this study utilized are publicly available; while every effort has been made to ensure the accuracy of the data, no guarantee is provided regarding their correctness. All datasets and the code that support the findings of this study are available in Appendices A and C, respectively.

## 4: FINDINGS

### 4.1 Descriptive Statistics

Of the 254 variables in our dataset, we focus on 6 to gain an estimate of the broader characteristics of the data: *Whitepct*, *Expenditure\_Per\_Student*, *Revenue\_Per\_Student*, *Fiscal\_Neutrality\_Scores*, *income\_allrace\_allgender\_pall*, and *povertypct*. We create summary statistics for this information on two datasets, one for all districts *Final\_Data* and one for lower-income neighborhoods *Poor\_Data*.

For our main dataset we naturally have a total of 56,408 observations. The variable *Whitepct*, which represents the percentage of the white population in each school district among school-age children enrolled in public schools (and as such is equal to 1- the percentage of nonwhite students), has an average of approximately 51.14% and a standard deviation of around 28.98%. The minimum value recorded is 0%, while the maximum value is 100%. These extreme values are likely linked to small school districts in rural areas, which may only service a small population with relative racial homogeneity. We also investigate each school district's per-pupil expenditures "*Expenditure\_Per\_Student*" and revenues "*Revenue\_Per\_Student*". For expenditures, we find that the average value of expenditures is approximately 5.92 thousand dollars, with a standard deviation of approximately 1.89 thousand dollars. The minimum and maximum were 2.51 thousand and 33.92 thousand dollars, respectively. For each school district's per-pupil revenue, we see that the mean revenue per student is approximately 5.79 thousand dollars, with a standard deviation of around 1.76 thousand dollars. The minimum and maximum revenue per student values are 1.68 and 29.23 thousand dollars, respectively.



The variable *Fiscal\_Neutrality\_Scores* is a measure of the fiscal neutrality of a state's funding policies with respect to school districts, measured using the elasticity of district total (state and local) funding relative to district wealth. The variable had an average score of approximately 0.15 and a standard deviation of about 0.14. The range of scores goes from -0.556 to 0.469. For the variable *income\_allrace\_allgender\_pall*, which references the average income of students who grew up in a particular census tract at age 35, irrespective of their race, gender, or parental income bracket, we see a mean average income of around \$30,459.18, with a standard deviation of approximately \$6,618.85. The minimum average income recorded is \$8,487, and the maximum average income is \$56,587. Lastly, for the *povertypct* variable, which measures the percentage of the population living in poverty in the census tract where the students grew up, we find an average poverty percentage of approximately 11.23%, with a standard deviation of about 9.63%. The minimum and maximum poverty percentages are 0% and 100% due to outliers in the data, with the 25<sup>th</sup> percentile being 4.8565% and the 75<sup>th</sup> percentile being 13.9421%. We use this 75<sup>th</sup> percentile to divide the dataset, with districts with poverty rates above this value being considered districts with high rates of poverty.

As such, we then form summary statistics with the subsection of districts considered as high-poverty and compare their statistics to the overall dataset. We consider 13,625 districts that are in this category. For *Whitepct*, we see a substantially higher mean for our overall population of districts at 51.14% when compared to high-poverty districts at 35.03%, indicating that there is a large delta wherein lower-income districts have a higher percentage of minority students than

in all school districts nationwide; this is consistent with the race-based disparities in school funding noted in the literature. Both datasets have similar means for "Expenditure\_Per\_Student" and "Revenue\_Per\_Student," with poor districts having slightly lower values (5.915549 vs 5.793566 thousand dollars for expenditures, 5.786576 vs 5.697586 793566 thousand dollars for revenues). However, revenues and expenditures have smaller standard deviations in the dataset focused on low-income districts, indicating less variation in per-pupil expenditure and revenue.

For *Fiscal\_Neutrality\_Scores*, we see a mean of 0.158261 in low-income districts, which is higher than the 0.149733 observed in all districts; this indicates that as an average, low-income districts are in states with more inequitable funding policies. However, both datasets have a similar standard deviation, which suggests that both datasets have similar fiscal neutrality score characteristics in terms of variability. Lastly, for *income\_allrace\_allgender\_pall* and *povertypct*, we see results consistent with our expectations. High-poverty districts have lower average income projections relative to all districts, meaning that students who grew up in high-poverty areas were more likely to stay in poverty; this matches results in the literature regarding the economic immobility that currently exists in the United States. By construction, our *povertypct* variable has a higher mean in the high-poverty district set; however, it also has a higher standard deviation, indicating more relative variance in the poverty rate among low-income areas. Overall, we find significant differences between low-income districts and all districts, confirming elements of the existing literature and setting the stage for our analysis. Full tables and summary statistics are located in Appendix B.

## 4.2 Instrumental Variable Analysis

To conduct our instrumental variable analysis, we attempt to analyze the relationship between school district expenditures per pupil in low-income census tracts (the *Expenditure\_Per\_Student* variable of the *Poor\_Data* subsection of the dataset) and the average income for all races and all genders from all parental income brackets from the census tract (*income\_allrace\_allgender\_pall*). For our instrument, we use the GAO's Fiscal Neutrality Scores for the 1991-1992 school year (*Fiscal\_Neutrality\_Scores*). A state's Fiscal Neutrality Score, also the tax base elasticity of total funding, is a measure of the funding gap between rich and poor districts within a state. A positive Fiscal Neutrality Score means that per-pupil funding increases in higher-income neighborhoods, and vice versa; larger values signify more inequity in school funding.

A state's Fiscal Neutrality Score plays a critical role in assessing the impact of school funding on long-term economic mobility in impoverished districts. However, for this neutrality score to be considered a reliable instrument for studying the effects of school funding in disadvantaged districts, it must not violate the exogeneity assumption, the relevance assumption, or the Exclusion Restriction assumption. The exogeneity assumption requires us to ensure that there is little correlation between the error term of the endogenous variable equation and the instrumental variable. Fortunately, the tight-knit relationship between the Fiscal Neutrality Score and funding in low-income school districts provides confidence in meeting this assumption. The strong and clear relationship between these two elements also helps fulfill the relevance assumption through insuring that the

instrument is impactful on the endogenous variable. Lastly, by relying on the Fiscal Neutrality Score as a proxy for the disparity between high and low-income school districts, we can confidently assume that there are no other systemic factors influencing both the instrument and the endogenous variable, thus satisfying the Exclusion Restriction assumption.

This choice of instrument is made with the acknowledgement that the two-stage model will restrict the amount of district-by-district variation we can encompass in our final model. This is because the Fiscal Neutrality Score is a state-level statistic, while expenditures per student is a district-level statistic; as such, predictions of district expenditures based merely on the score will only vary at the state level. However, the nature of the Fiscal Neutrality Score as an aggregate state statistic of district-by-district variations in funding means that much of the district inequality caused by school funding policies will still be accounted for in the instrument. Additionally, given our focus on high-poverty districts in our model, the Fiscal Neutrality Score only predicts the average expenditures of districts in high-poverty areas. An analog can be drawn to a usage of the state-level Gini coefficient as an instrument for the income of people in the bottom 20% of the income distribution; while individual data regarding income may be lost, the aggregate nature of the Gini coefficient ensures that the instrument still carries information about the income of a subsection of the overall income distribution. In a similar way, the Fiscal Neutrality Score can capture some aspects of the expenditures of districts in high-poverty areas, as it is designed to model funding inequities between school districts. This instrument also helps ensure that we do not run into endogeneity

concerns regarding neighborhood effects; the statewide nature of the Fiscal Neutrality Score helps to alleviate concerns that an instrument on a district or tract level would be correlated with positive or negative neighborhood characteristics aside from school funding policies. While the loss of individual district-by-district data is a significant sacrifice, the instrumental variable approach is critical to the elimination of omitted variable bias and ensuring the causal validity of our analysis.

We run the instrumental variable analysis in a 2-stage model using the above specification. We begin in the first stage by selecting the best-fitting regression model for the instrumental variable using our predetermined suite of machine learning models. For each model, we conduct hyperparameter tuning using a grid search when appropriate (linear regressions do not have hyperparameters to tune), which is the process of evaluating a series of different model configuration settings to ensure that the model is performing at its most optimal level. We then select the best hyperparameters for each model as measured by the model's coefficient of determination or r-squared score. The r-squared value indicates how well the model explains the variance in the instrumental variable and can be used to signify which model performs the best at predicting the true value of the outcome variable. Using the  $r^2$  values, we then compared the best performing model of each type together and selected the best model's prediction of the values of our endogenous variable for usage in the second stage of our model. In the second stage of our model, we utilize the values of the endogenous variable gathered using the best-performing model in the first stage to predict the causal effect of the endogenous variable in the second stage using each of the five models. Each model then reported a variety of

summary statistics regarding the model's performance, including r-squared values, for us to determine the effectiveness of each model on the testing dataset.

When examining the results of each machine learning model when predicting *Expenditure\_Per\_Student* using *Fiscal\_Neutrality\_Scores* in our first stage, we find substantial variance in the performance of different regressors. The linear regression model reported the lowest r-squared value of only 0.0064993, which is indicative of a model that is almost completely inept at approximating the variation of the predicted variable. Next, the MLP neural network reported the second-lowest value of 0.16354, which indicates a substantial improvement over the linear regression model but an underperformance when compared to our other three models. . The poor performance of linear regression and MLP neural networks relative to other models is likely related to the fact that the neural network implementation utilized relied on a linear activation, which means that the output of each neuron in the network is a simple weighted sum of the neuron's inputs, similar to the way a linear regression model computes its output as a weighted sum of the input features. Thus, if the patterns in the data are almost entirely non-linear (characteristics that this dataset is demonstrating), both the neural network and the linear regression will underperform other models with more appropriate designs. This hypothesis is likely also to be true because of the time that it took the neural network to converge to a specific model; a standard implementation of an MLP neural network would often struggle to converge in our study, and modifying the implementation of the model to a significantly more computationally intensive

approach was the only way in which the neural network would consistently converge; this indicates that the MLP was a poor fit for the dataset under study.

When compared to the performance of neural networks and linear regressions, the remaining three models performed significantly better. The K-Nearest Neighbors algorithm had an r-squared of 0.22195, which is a notable improvement from the linear regression and neural network models. However, the Decision Trees and Random Forests algorithms performed best of all, with r-squared values of 0.50183 and 0.50119, respectively; these represented a 2.25x advantage over KNN, 3x advantage over MLP neural networks, and a 77x improvement over the standard linear regression model. This dramatic increase in performance is likely due to the benefits that Decision Trees and Random Forests have when it comes to capturing non-linear relationships due to their nodal construction; as such, they could dramatically outperform the previous two models due to their ability to better model patterns in the data. In particular, the biggest success is the performance of Decision Trees; while Random Forests typically will outperform Decision Trees due to being an ideal combination of multiple Decision Trees, the relative simplicity of the dataset means that the Decision Trees algorithm was able to capture patterns in the dataset efficiently without any advantage to be gained from the more computationally complex method. As a result, we opt to use the predictions of Decision Trees to continue to the second stage of our model.

After selecting the decision trees model and gaining the predicted values of *Expenditure\_Per\_Student*, we proceed to run the second stage of the model, where we utilize newly specified models to predict the income at age 35 of students who

grew up impoverished school districts *income\_allrace\_allgender\_pall*. As in the first stage, we include hyperparameter tuning and evaluate the appropriateness of each model utilizing the coefficients of determination.

In this second stage, the pattern that we detected regarding nonlinear trends within the first stage continues; the MLP neural network yielded an r-squared value of 0.05627 and the linear regression noted a r-squared value of 0.05626. This, combined with the continual errors in the convergence of the neural network (the network would often fail to converge with 200 iterations), suggests that this relationship is once again very poorly modeled by linear relationships. However, both models significantly outperform K-Nearest Neighbors, which scored the lowest with an r-squared value of 0.01776. The difference in performance and KNN can be attributed to how KNN measures the closest neighbors of a point; KNN relies on the Euclidean distance to find nearby points in the feature space. However, this approach can encounter limitations in our dataset because we only have 50 state fiscal neutrality scores; as a result, all the school districts within the same state tend to have similar feature values. This similarity within states can lead to challenges for KNN, as it may prioritize neighboring points within the same state, even if they belong to different classes. On the other hand, we expect some algorithms, such as decision trees, to perform better in such cases as they can capture non-linear relationships between features and the target variable. The decision tree's ability to identify relevant features and create splits based on their importance allows it to make better predictions despite the data's similarity within states. This theory is confirmed when looking at the performance of the decision trees and random



forests model, which outperform the rest of the field with  $r^2$  values of 0.081997 and 0.081956, respectively. These results again see the standard decision tree outperforming the more robust random forests model, serving to indicate that the data is very well-modeled by a decision tree and that the model is sufficient to capture many of the nuances in the dataset; this is likely an indicator that the simplicity of our dataset is such that a decision tree can effectively model it.

Overall, random forests appear to be the most effective model in predicting the causal impact of disparities in school funding in poor districts on the long-run economic outcomes of students from those districts. However, given the black-box nature of machine learning, we cannot effectively evaluate the actual relationship that the decision tree is modeling without gaining additional information. This is the main disadvantage of machine learning models; while they can effectively model incredibly complex relationships, they are less interpretable than models typically used in econometrics such as linear regression. However, some recent literature in the field of agricultural economics has begun attempting to find ways to approximate the functional form of a dataset from a machine learning model (Storm, Baylis, & Heckeley 2020). This can be done through analyzing the input-output pairs of an artificially created dataset. This method can be deployed at its simplest with univariate regression; given that machine learning models are also trained to predict complex interactions with multiple features, capturing the effects of changes in a multivariate model is possible, although it increases the complexity of the issue significantly. This method to approximate the effect encapsulated in a machine

learning model has been recently discussed in relation to agricultural economics, but its methodology still applies to analysis of non-agricultural datasets.

To conduct this analysis, we create a dataset modeling artificial school funding per pupil amounts ranging from 2.509895 (the minimum value of the dataset to 10.914362 (3 standard deviations above the mean) thousand dollars of school expenditures per pupil, iterating in increments of \$10. These endpoints were selected to encompass as much of the sample as possible while controlling for the extreme outliers that existed in the upper bound of our dataset; as such, we felt as though the approximation of a value being 3 standard deviations above the mean as being an outlier was appropriate. When applying the same function to the lower range of the dataset, we gain a lower bound value of 0.67277. However, to avoid extrapolation, we utilize the minimum number in the dataset instead. Each artificial dataset was then fed into the various predictive models that we utilized to attempt to ascertain the shape and general trends of each model. The finished output was then examined to detect patterns in the data that may give us insight into the causal relationship between our variables of investigation. The completed graphs for all five (5) regression algorithms can be found in Appendix B, Figures 4-13.

Investigating the trends that we can see in each model, we can draw a few key conclusions regarding the nature of the relationship. First, all 5 models predict a mostly consistent upwards trend in values of *income\_allrace\_allgender\_pallas Expenditures\_Per\_Pupil* increases, as can be seen by models 5, 7, 9, 11, and 13; however, this relationship is not linear in nature, as demonstrated by the poor second-stage performance of the MLP neural network and linear regression. Rather,

the most successful models predict a sharp drop-off around \$9,000 of expenditures, but one that makes a fast recovery as funding approaches \$10,000. This could be caused by overfitting, as machine learning models tend to conform to the training dataset and can sometimes exhibit anomalies of the dataset that are not truly reflective of the relationship between the two variables. This may also be the cause for the variability in KNN, Decision Trees, and Random Forest predictions between \$4,000 and \$7,000 in per-pupil expenditures. Nevertheless, these findings provide valuable information regarding the functional form of the relationship between per-pupil-expenditures and long-run income.

Turning our attention to 8 and 10, we observe a rhombus-like pattern exhibited by both top-performing models, outlining the predicted income bounds corresponding to various levels of expenditures. As expenditure increases, the upper bound of predicted income rises until reaching around \$6500, where it levels off. Meanwhile, the lower bound experiences modest growth until approximately \$6000, beyond which it accelerates significantly. These insights shed light on the potential outcomes for low-income districts; in the best-case scenario, increasing expenditure levels to at least \$6500 yields tangible, positive effects on long-term student income. Moreover, even beyond this threshold, additional expenditures further enhance the lower bound of predicted income, reducing variability and increasing the likelihood of students reaching this upper income bound. This emphasizes the importance of providing additional resources to schools in low-income areas; by supporting schools with the funding they need, we can help America's most disadvantaged groups fulfill the promises of the American Dream.

## 5: CONCLUSION

This study analyzes the effect of school funding disparities on long-run income for students growing up in high-poverty areas. Data was gathered from a variety of different public sources, including government agencies such as the U.S. Census Bureau and National Center for Education Statistics as well as databases from research groups such as Opportunity Insights based at Harvard University. The data was then analyzed through an instrumental variable approach; however, in lieu of the standard 2-Stage Least Squares linear regression model, we utilize four other different machine learning models (K-Nearest Neighbors, Decision Trees, Random Forests, and Multi-Layer Perceptron Neural Networks) to provide a more robust analysis of the causal relationship between school district per-pupil expenditures and the average income at age 35 of students who grew up in the school district in question. Our study finds that the utilization of machine learning models to attempt to predict the causal relationship using instrumental variables yields noticeable and important benefits over utilizing standard linear regressions. When deployed at the two stages of data analysis, machine learning models, most notably decision trees, were able to outperform linear regression by a significant margin in terms of being able to explain the variation in the target variable, and as such were more effective methods of predicting changes in long-run income based on changes in school expenditures. In terms of the relationship, the machine learning models indicated a significant positive causal relationship between school funding and income, meaning that increases in school funding have a direct beneficial impact on the long-run income of students in the school districts experiencing the increase.

As such, this work establishes two major additions to the existing literature. First, the analysis demonstrates the benefits of machine learning models in adapting to complex patterns, even in univariate datasets, and highlights the usefulness of implementing large learning models into economic research. This is bolstered through the combination of instrumental variables and machine learning, which allows the study to find a significant nonlinear causal relationship between school expenditures per pupil and students' income at age 35. This provides the second contribution of our study; through establishing the causal link between educational expenditure and long-run outcomes, this study quantitatively demonstrates the tangible real-world benefits to increasing funding to educational institutions. This adds to the existing literature by expanding on studies linking funding to academic outcomes; as such, we demonstrate that the link between funding and outcomes extends not just to test scores and grades, but also to real-world results. As such, we conclude that increasing funding to schools in low-income areas is critical to ensuring that students attending schools in those regions have the highest chances of being economically mobile and having full access to the opportunities that may be available to them.

Unfortunately, there remain multiple limitations to our study. First, the restrictions on the amount and quality of information that is publicly available regarding school district funding and outcomes means that our study is restricted to the district level, among other limitations; access to more granular information would help us to strengthen our analysis through refining our model specifications and information, thus helping us to better analyze the relationship under study.

Second, given the black-box approach of machine learning models, we are unable to put together meaningful estimates of the statistical significance of different parameters. However, this limitation is not of great concern; standard models for statistical significance, such as the utilization of p-values and hypothesis tests, are not necessary for us to draw determinate conclusions regarding the impact of the variables in our dataset because the purpose of p-values is to estimate a population statistic from an unbiased sample. However, as we have a census of the dataset, we do not need to provide estimates for population parameters; our datasets will yield the relationship within the population itself, not a sample of it. As such, irrespective of p-values, we can be confident that a significant causal relationship exists based on our study. Lastly, this study provided only a basic survey into learning algorithms and did not encompass many more complex machine learning methods or significant alterations to existing functions. This means that while we were able to produce models that outperform standard linear regressions, we were not able to produce a definitive survey of the implementations of different model types.

Although these limitations establish a foundation for future work, this study still offers a significant contribution to the existing literature. Leveraging machine learning algorithms opens possibilities for uncovering novel insights in the realm of social sciences, which in turn could lead to meaningful policy changes in areas like addressing school funding gaps. By engaging in thorough analysis and dedicated advocacy, we have the power to narrow the disparities in accessing quality education within the US; by doing so, we can truly uphold our commitment to being a land of where every individual is afforded a fair opportunity at success.

## REFERENCES

- Adamson, F., & Darling-Hammond, L. (2012). Funding disparities and the inequitable distribution of teachers: Evaluating sources and solutions. *education policy analysis archives*, 20(37), n37.
- Baker, B. D., Farrie, D., & Sciarra, D. G. (2016). Mind the gap: 20 years of progress and retrenchment in school funding and achievement gaps. ETS Research Report Series, 2016(1), 1-37.
- Baker, B. D., Srikanth, A., Cotto Jr, R., & Green III, P. C. (2020). School Funding Disparities and the Plight of Latinx Children. *Education policy analysis archives*, 28(135), n135.
- Bor, J., Cohen, G. H., & Galea, S. (2017). Population health in an era of rising income inequality: USA, 1980–2015. *The Lancet*, 389(10077), 1475-1490.
- Carey, K. (2004). The funding gap 2004. Education Trust (Fall 2004). The full report can be downloaded at <http://www2.edtrust.org/NR/rdonlyres/30B3C1B3>.
- Case, A., & Deaton, A. (2022). The great divide: education, despair, and death. *Annual Review of Economics*, 14, 1-21.
- Chetty, R., & Hendren, N. (2018). The impacts of neighborhoods on intergenerational mobility I: Childhood exposure effects. *The Quarterly Journal of Economics*, 133(3), 1107-1162.
- Chetty, R., Friedman, J. N., Saez, E., Turner, N., & Yagan, D. (2017). Mobility report cards: The role of colleges in intergenerational mobility (No. w23618). national bureau of economic research.

- Collins, W. J., & Wanamaker, M. H. (2017). African American intergenerational economic mobility since 1880 (No. w23395). National Bureau of Economic Research.
- Croizet, J. C., & Dutrévis, M. (2004). Socioeconomic status and intelligence: Why test scores do not equal merit. *Journal of Poverty*, 8(3), 91-107.
- Cruz, R. A., Lee, J. H., Aylward, A. G., & Kramarczuk Voulgarides, C. (2022). The effect of school funding on opportunity gaps for students with disabilities: Policy and context in a diverse urban district. *Journal of Disability Policy Studies*, 33(1), 3-14.
- Dahl, R. E., Allen, N. B., Wilbrecht, L., & Suleiman, A. B. (2018). Importance of investing in adolescence from a developmental science perspective. *Nature*, 554(7693), 441-450.
- Eide, E. R., & Showalter, M. H. (2011). Estimating the relation between health and education: What do we know and what do we need to know? *Economics of Education Review*, 30(5), 778-791.
- Furman, J., & Stiglitz, J. E. (1998). Economic consequences of income inequality. *Income Inequality: Issues and Policy Options*, Jackson Hole, Wyoming, Federal Reserve Bank of Kansas City.
- Gammon, T. E. (1976). Equal protection of the law and San Antonio Independent School District V. Rodriguez. *Val. UL Rev.*, 11, 435.
- Gilens, M. (2012). *Affluence and influence: Economic inequality and political power in America*. Princeton University Press.



- Hauser, R. M., Simmons, S. J., & Pager, D. I. (2000). High School Dropout, Race-Ethnicity, and Social Background from the 1970s to the 1990s.
- Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: An update. *Educational evaluation and policy analysis*, 19(2), 141-164
- Hanson, Melanie. "U.S. Public Education Spending Statistics" EducationData.org, June 15, 2022, <https://educationdata.org/public-education-spending-statistics>
- Harvard University. (2023, July 24). Policy solutions to the American Dream. Opportunity Insights. <https://opportunityinsights.org/>
- Hoxby, Caroline M., and Sarah Turner. 2015. "What High-Achieving Low-Income Students Know about College." *American Economic Review*, 105 (5): 514-17.
- Kreisman, D., & Steinberg, M. P. (2019). The effect of increased funding on student achievement: Evidence from Texas's small district adjustment. *Journal of Public Economics*, 176, 118-141.
- Leachman, M., Masterson, K., & Figueroa, E. (2017). A punishing decade for school funding. Center on Budget and Policy Priorities, 29.
- Little, A & Lex, A (2021). Math 4100: Introduction to Data Science at the University of Utah [Course Notes]. <https://datasciencecourse.net/2021/index.html>
- Lousdal, Mette Lise. An introduction to instrumental variable assumptions, validation and estimation. *Emerging Themes in Epidemiology* 15.1 (2018): 1-7.
- Lovenheim, M. F., & Willén, A. (2019). The long-run effects of teacher collective bargaining. *American Economic Journal: Economic Policy*, 11(3), 292-324.

- Lutz, A. (2007). Barriers to high-school completion among immigrant and later-generation Latinos in the USA: Language, ethnicity, and socioeconomic status. *Ethnicities*, 7(3), 323-342.
- Lynch, S. M., & Brown, J. S. (2011). Stratification and inequality over the life course. In *Handbook of aging and the social sciences* (pp. 105-117). Academic Press.
- Machin, S., Marie, O., & Vujić, S. (2011). The crime reducing effect of education. *The Economic Journal*, 121(552), 463-484.
- Muller, E. N. (1985). Income inequality, regime repressiveness, and political violence. *American sociological review*, 47-61.
- Mitchell, T. (2020, January 9). 1. trends in income and wealth inequality. Pew Research Center's Social & Demographic Trends Project.  
<https://www.pewresearch.org/social-trends/2020/01/09/trends-in-income-and-wealth-inequality/>
- Neymotin, F. (2010). The relationship between school funding and student achievement in Kansas public schools. *Journal of Education Finance*, 88-108.
- Pech, N and Laloë, F, Use of Principal Component Analysis with Instrumental Variables (PCAIV) to analyse fisheries catch data, *ICES Journal of Marine Science*, Volume 54, Issue 1, February 1997, Pages 32-47
- Richwine, J. (2011). The Myth of Racial Disparities in Public School Funding. Background. No. 2548. Heritage Foundation.
- Rooney, C., & Schaeffer, B. (1998). Test Scores Do Not Equal Merit: Enhancing Equity & Excellence in College Admissions by Deemphasizing SAT and ACT Results.

- Roza, M., & Miles, K. H. (2002). Moving toward Equity in School Funding within Districts: A Comparison of Traditional Funding Policies and More Equitable Formulas.
- Sebold, F. D., & Dato, W. (1981). School funding and student achievement: An empirical analysis. *Public Finance Quarterly*, 9(1), 91-105.
- Smola, A., & Vishwanathan, S. V. N. (2008). *Introduction to Machine Learning*. Cambridge University Press.
- Social Security Administration. (2015, November). Education and Lifetime Earnings. Research Summary: Education and Lifetime Earnings.  
<https://www.ssa.gov/policy/docs/research-summaries/education-earnings.html>
- Storm, H., Baylis, K., & Heckelei, T. (2020). Machine learning in agricultural and applied economics. *European Review of Agricultural Economics*, 47(3), 849-892.
- Tow, C. (2006). The Effects of School Funding on Student Academic Achievement. Unpublished undergraduate thesis). University of California, Berkeley, Berkeley, CA.
- U.S. Bureau of Labor Statistics. (2019). Median weekly earnings \$606 for high school dropouts, \$1,559 for advanced degree holders. U.S. Bureau of Labor Statistics. <https://www.bls.gov/opub/ted/2019/median-weekly-earnings-606-for-high-school-dropouts-1559-for-advanced-degree-holders.htm#:~:text=Those%20without%20a%20high%20school,college%20or%20an%20associate%20degree.>

- Vandivier, D. (2022). What is the Great Gatsby Curve? National Archives and Records Administration. Retrieved July 2, 2022, from <https://obamawhitehouse.archives.gov/blog/2013/06/11/what-great-gatsby-curve#:~:text=The%20Great%20Gatsby%20Curve%20illustrates,ladder%20compared%20to%20their%20parents.>
- Vollrath, D. (2013). Inequality and school funding in the rural United States, 1890. *Explorations in Economic History*, 50(2), 267-284.
- Wang, Q., Ma, Y., Zhao, K., & Tian, Y. (2020). A comprehensive survey of loss functions in machine learning. *Annals of Data Science*, 1-26.
- Walsh, C. (2011). Erasing Race, Dismissing Class: San Antonio Independent School District v. Rodriguez. *Berkeley La Raza LJ*, 21, 133.
- Walker, B. D. (1984). The local property tax for public schools: Some historical perspectives. *Journal of Education Finance*, 9(3), 265-288.
- Weiss, A. H. (2020). School funding inequalities in the Texas panhandle related to the racial, socio-economic, and linguistic composition of school districts. *Journal of Education Finance*, 46(1), 20-46.
- Xu, Chen, Srinivasan, de Freitas, Doucet, Gretton. Learning Deep Features in Instrumental Variable Regression. Published as a conference paper at ICLR 2021
- Zan, H., Fan, J. X., & Lozada, B. (2023). The economic disparity between Hispanic and non-Hispanic White households: An analysis of middle-class achievement. *American Journal of Economics and Sociology*.

## APPENDIX A: DATASETS

Opportunity Insights. (2020). The Opportunity Atlas.

<https://www.opportunityatlas.org/>

United States Census Bureau (2022, August 16). Annual Survey of School System Finances Tables. Census.gov. <https://www.census.gov/programs-surveys/school-finances/data/tables.html>

Chunyu, M. (n.d.). Census geography: Bridging data for census tracts across time. Diversity and disparities.

<https://s4.ad.brown.edu/projects/diversity/Researcher/Bridging.html>

National Center for Education Statistics. (n.d.). Education Demographic and Geographic Estimates. Index.

<https://nces.ed.gov/programs/edge/tableviewer/acsProfile/2009>

Johnson, E. L., Fastrup, J. C., & Billingshurst, B. (1998). State Efforts to Equalize Funding Between Wealthy and Poor School Districts. General Accounting Office, Health, Education, and Human Services Division.

<https://www.gao.gov/assets/hehs-98-92.pdf>

Harvard T.H. Chan School of Public Health. (2021, February 9). U.S. Census Tract Poverty Data. The Public Health Disparities Geocoding Project.

<https://www.hsph.harvard.edu/thegeocodingproject/u-s-census-tract-poverty-data/>

National Center for Education Statistics. (n.d.). School District Geographic Relationship Files.

<https://nces.ed.gov/programs/edge/Geographic/RelationshipFiles>

## APPENDIX B: FIGURES

Figure 1: Summary Table for *Final\_Data*

<i>Final_Data</i>	count	mean	std	min	25%	50%	75%	max
Whitepct	56408	51.14042	29	0	23.9	52.2	77.1	100
Expenditure_Distance	56408	-2.82E-17	1.25	-3.83	-0.64	-0.174	0.405	28.277
Revenue_Distance	56408	1.36E-17	1.14	-4.25	-0.66	-0.126	0.432	20.987
Expenditure_Per_Student	56408	5.915549	1.85	2.51	4.674	5.4772	6.682	33.917
Revenue_Per_Student	56408	5.786576	1.76	1.68	4.609	5.4409	6.566	29.231
Fiscal_Neutrality_Scores	56408	0.149733	0.14	-0.56	0.055	0.141	0.25	0.469
income_allrace_allgender_pall	56408	30459.18	6619	8487	25811	30333	35001	56587
povertypct	54790	0.112255	0.1	0	0.049	0.082	0.139	1

Figure 2: Summary Table for *Poor\_Data*

<i>Poor_Data</i>	count	mean	std	min	25%	50%	75%	max
Whitepct	13635	35.032	26.5	0	12.8	27.3	55	100
Expenditure_Distance	13635	-0.0286	0.96	-3.32	-0.59	-0.141	0.351	10.031
Revenue_Distance	13635	0.0062	0.94	-3.29	-0.57	-0.064	0.442	9.9821
Expenditure_Per_Student	13635	5.7936	1.71	2.51	4.58	5.4378	6.662	18.148
Revenue_Per_Student	13635	5.6976	1.6	2.133	4.6	5.5034	6.566	15.023
Fiscal_Neutrality_Scores	13635	0.1583	0.14	-0.56	0.071	0.153	0.25	0.469
income_allrace_allgender_pall	13635	24652	5423	8487	20740	24258	27909	53074
povertypct	13635	0.2461	0.1	0.14	0.169	0.2141	0.298	1

Figure 3: Summary Table of Regression Model Performance

	Stage 1 R^2	Stage 2 R^2	Stage 2 MSE
Linear Regression	<b>0.0064994</b>	0.056266924	40968682.21
K-Nearest Neighbors	0.22195846	0.017766506	42640035.49
Decision Trees	<b>0.5018339</b>	<b>0.08199798</b>	<b>39851663.4</b>
Random Forests	0.50119441	0.081956185	39853477.92
MLP Neural Network	0.16354953	<b>0.05627524</b>	<b>40968321.3</b>

*Note for figures 4-13: The keys for graphs are using slightly incorrect terminology: The models are two-stage models, but not two-stage least squares (2SLS) models. The incorrect labeling arises from errors in the initial terminology given to the models but has no effect on the accuracy or quality of the models themselves.*

Figure 4 and 5: Linear Regression 2-Stage Prediction on Test and Artificial Data

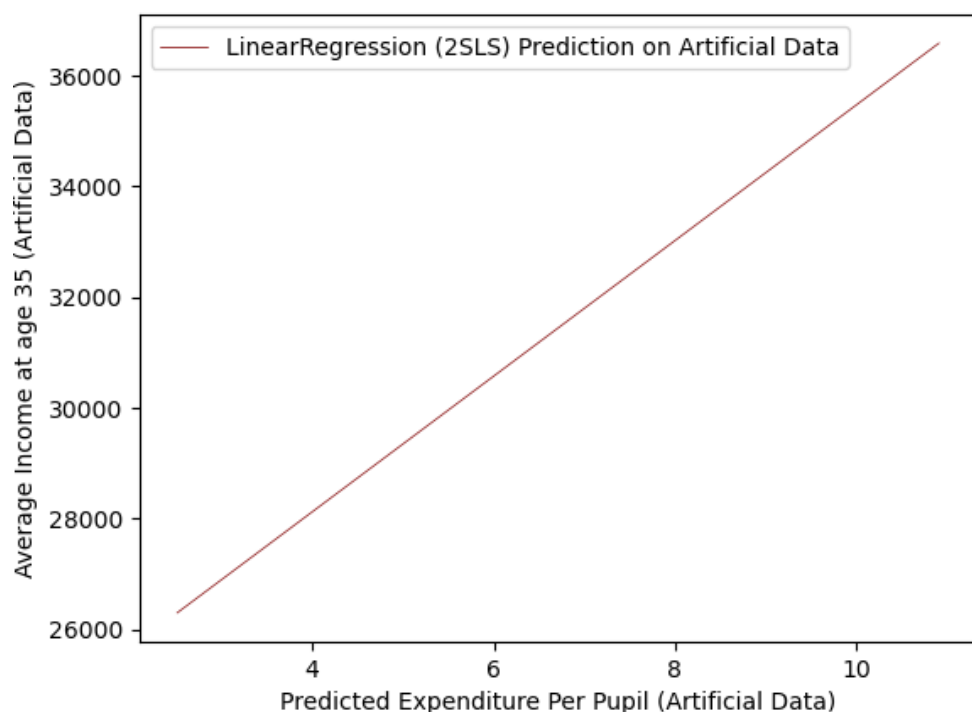
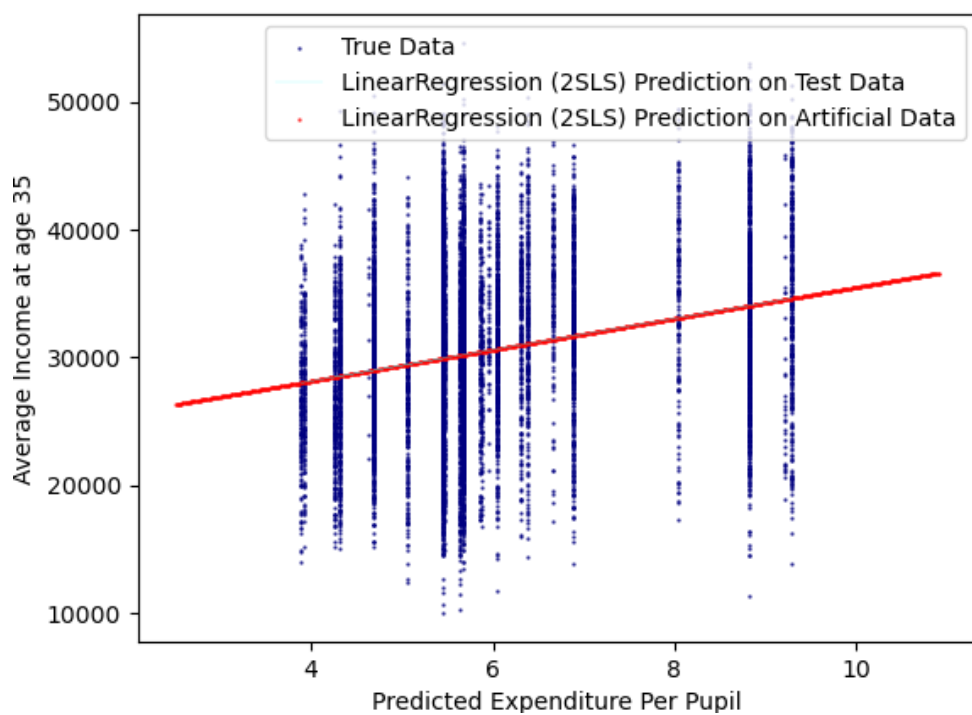


Figure 6 and 7: Linear Regression 2-Stage Prediction on Test and Artificial Data

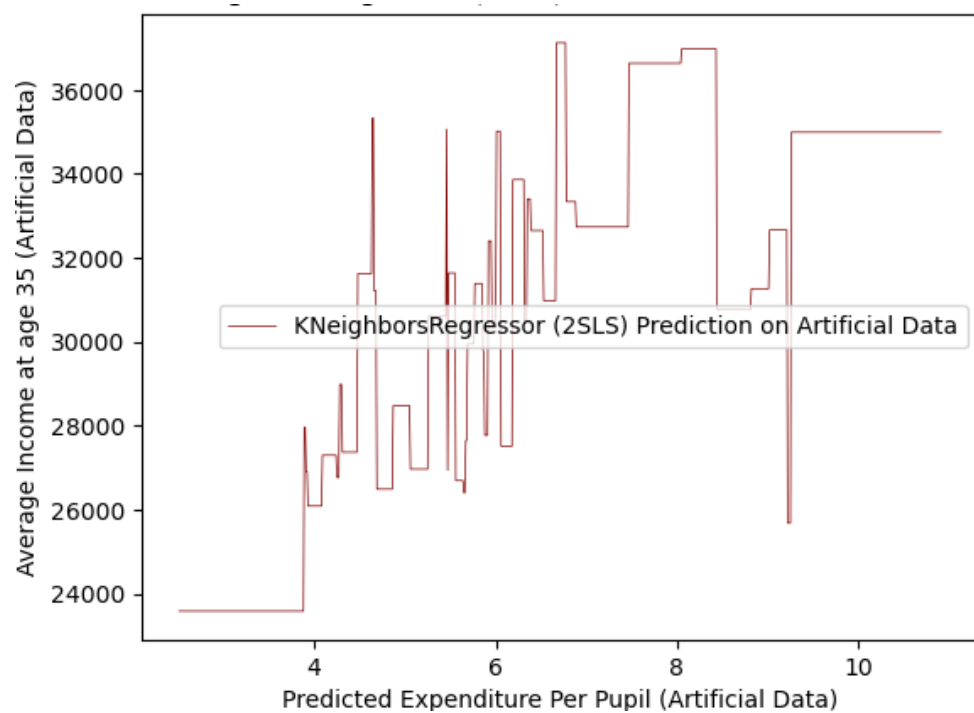
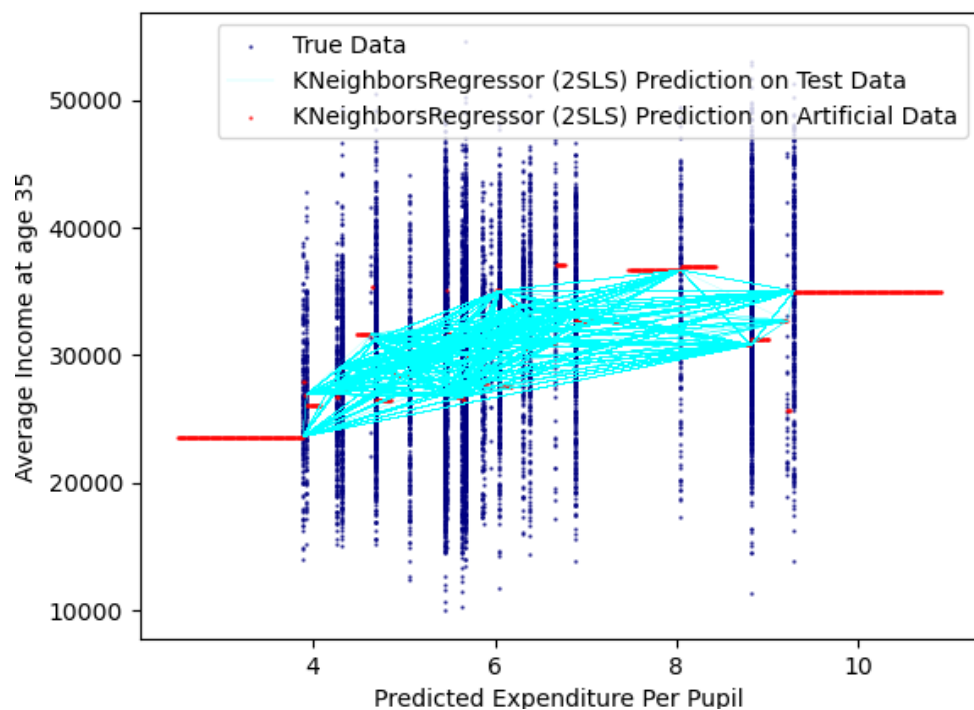




Figure 8 and 9: Decision Trees 2-Stage Prediction on Test and Artificial Data

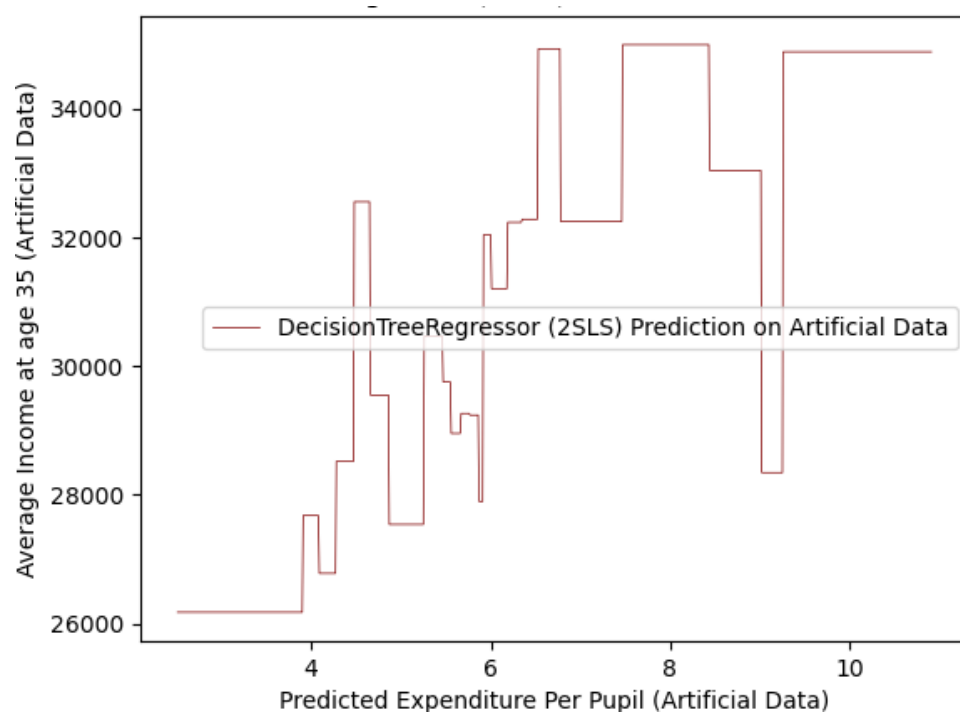
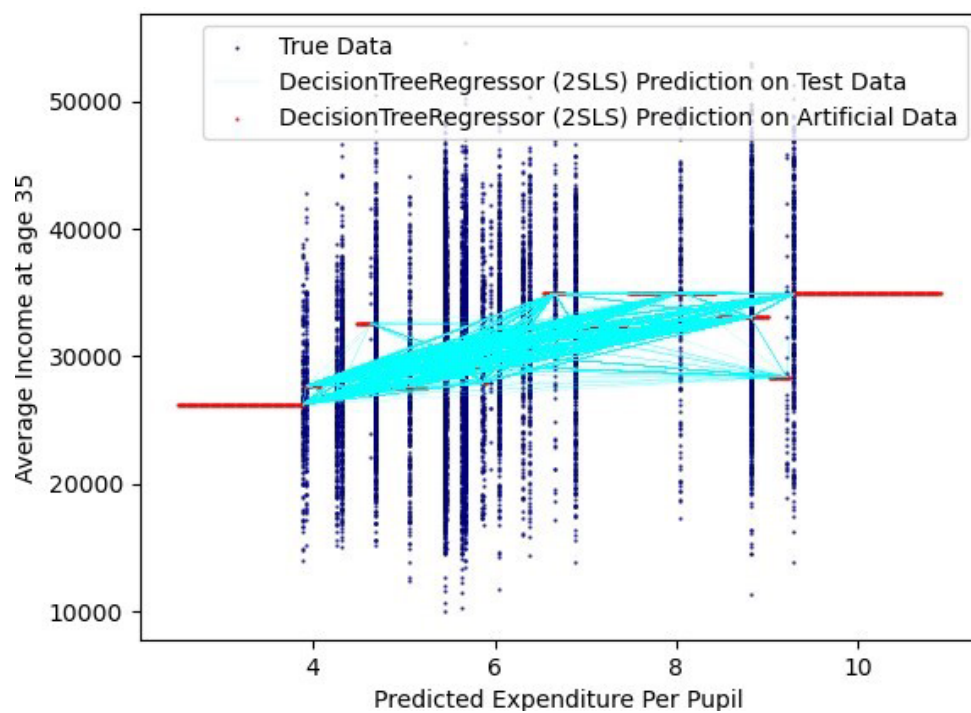


Figure 10 and 11: Random Forests 2-Stage Prediction on Test and Artificial Data

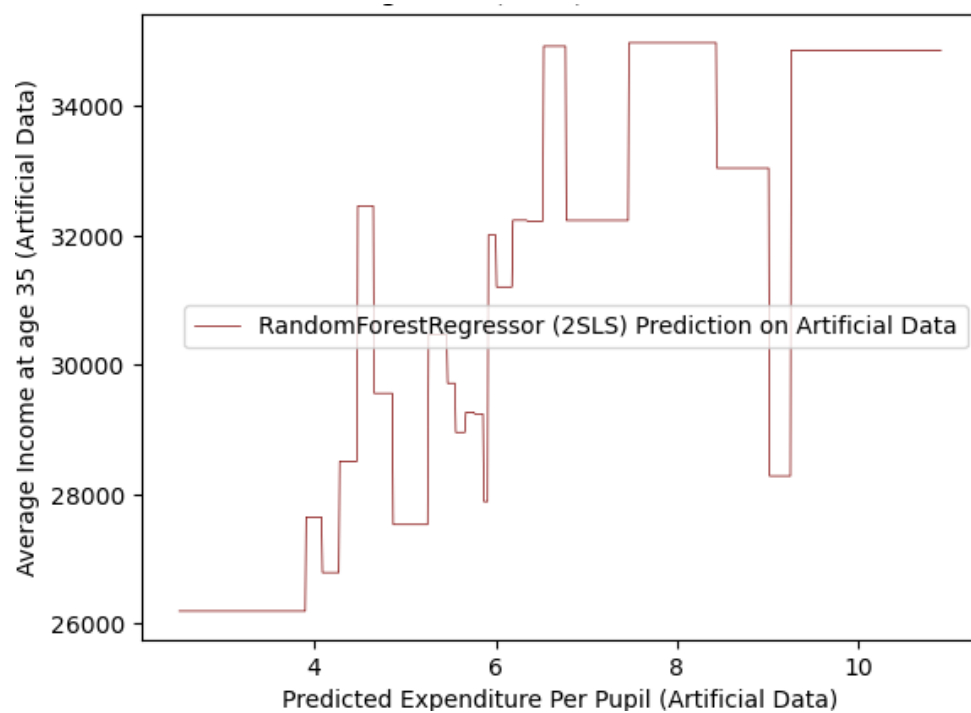
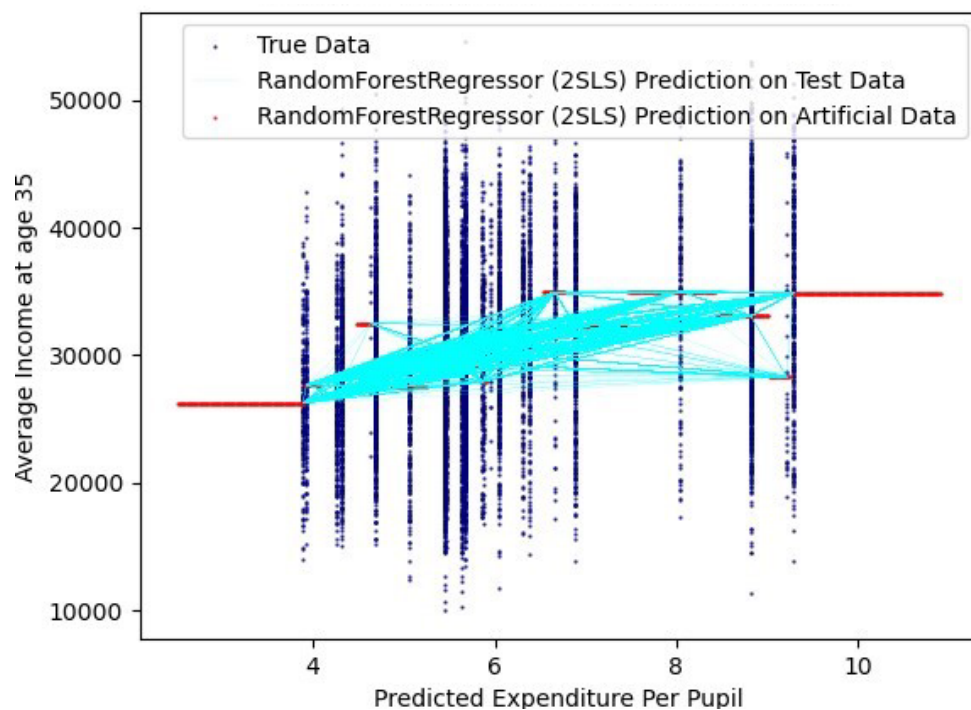
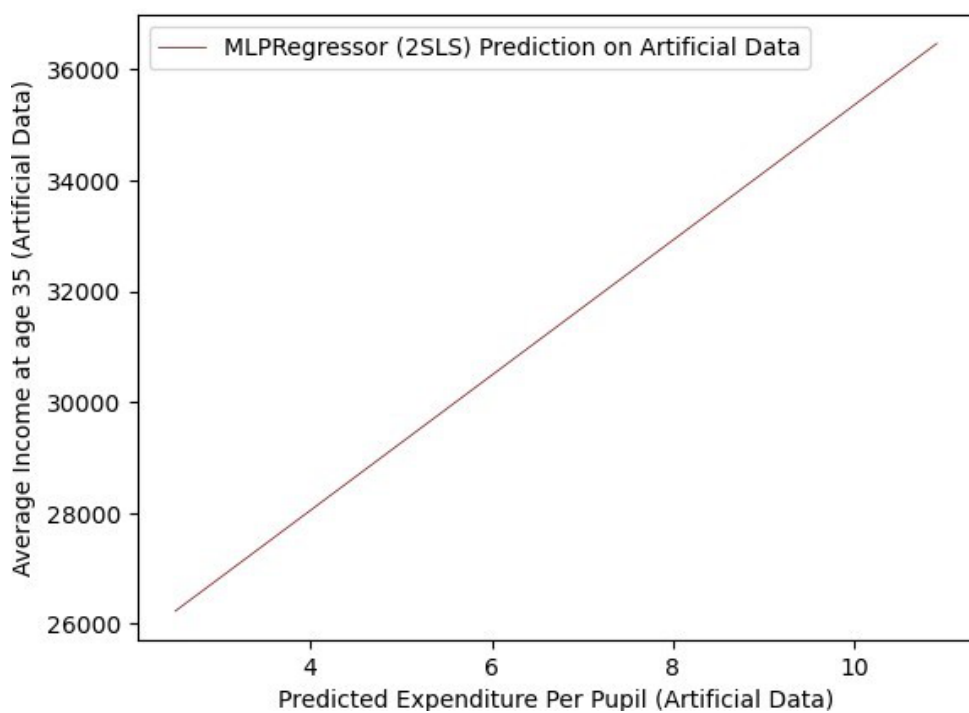
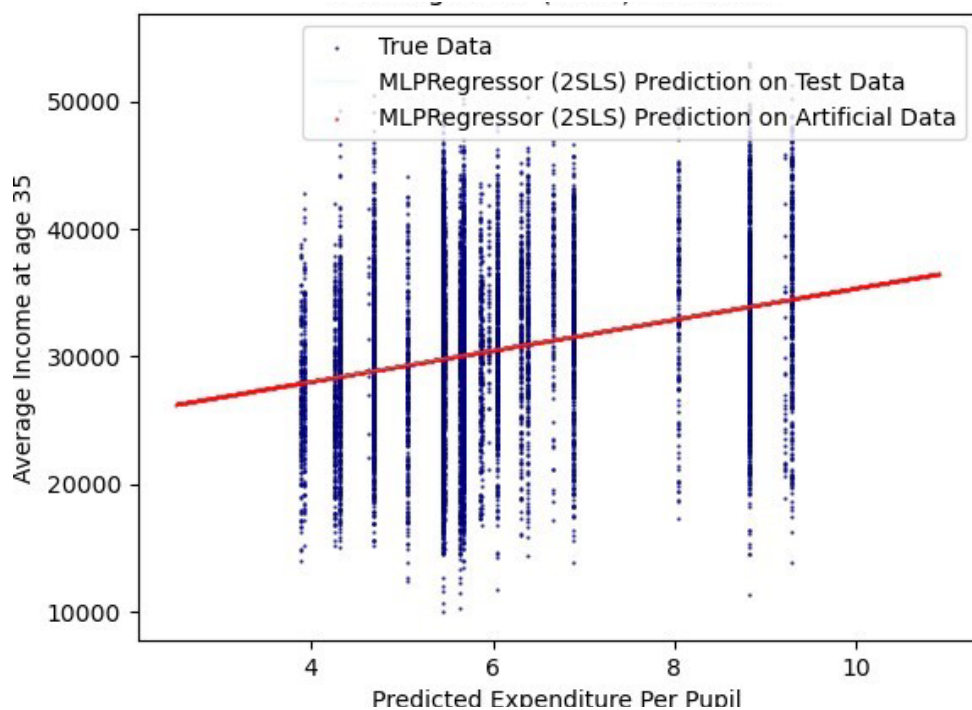


Figure 12 and 13: Neural Network 2-Stage Prediction on Test and Artificial Data



## APPENDIX C: CODE

All code was implemented in Python 3 using Jupyter Notebook and the numpy, matplotlib, pandas, and sklearn libraries.

```
In [1]:import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.neighbors import KNeighborsRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.neural_network import MLPRegressor
from sklearn.metrics import mean_squared_error
from sklearn.metrics import mean_absolute_error, r2_score,
mean_squared_error,
explained_variance_score, median_absolute_error
from sklearn.metrics import mean_squared_error, mean_absolute_error,
r2_score,
explained_variance_score, median_absolute_error
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import GridSearchCV
from sklearn.exceptions import ConvergenceWarning
from sklearn import linear_model
import warnings

In [2]:District_Tract = pd.read_csv(r'District_Tract.csv', low_memory=False)
Income_Tract = pd.read_csv(r'IndividualIncome_Tract.csv',
low_memory=False)
Race_District = pd.read_csv(r'Race_District.csv', low_memory=False)
Fund_District = pd.read_csv(r'Fund_District.csv', low_memory=False)
Poverty_1980Tract = pd.read_csv(r'Poverty_1980Tract.csv',
low_memory=False)
Tract_Tract = pd.read_csv(r'1980Tract_2010Tract.csv', low_memory=False)
District_Tract = District_Tract.rename(columns={'TRACT': 'tract'})
District_Tract = District_Tract.rename(columns={'NAME_LEA19': 'name'})
District_Tract = District_Tract.rename(columns={'LEAID': 'NCESID'})
Race_District = Race_District.rename(columns={'LEAID': 'NCESID'})
District_Tract = District_Tract.drop(['COUNT', 'LANDAREA', 'WATERAREA'],
axis=1)
```

```

Fiscal_Neutrality_Scores = {'statename': ['Alabama', 'Alaska', 'Arizona',
'Arkansas', 'California', 'Colorado', 'Connecticut', 'Delaware',
'Florida', 'Georgia', 'Hawaii', 'Idaho', 'Illinois', 'Indiana', 'Iowa',
'Kansas', 'Kentucky', 'Louisiana', 'Maine', 'Maryland', 'Massachusetts',
'Michigan', 'Minnesota', 'Mississippi', 'Missouri', 'Montana', 'Nebraska',
'Nevada', 'New Hampshire', 'New Jersey', 'New Mexico', 'New York', 'North
Carolina', 'North Dakota', 'Ohio', 'Oklahoma', 'Oregon', 'Pennsylvania', 'Rhode
Island', 'South Carolina', 'South Dakota', 'Tennessee', 'Texas', 'Utah', 'Vermont',
'Virginia', 'Washington', 'West Virginia', 'Wisconsin',
'Wyoming'], 'Fiscal_Neutrality_Scores': [+0.290, -0.272, +0.141, +0.220, +0.073,
+0.154, +0.241, +0.072, +0.239, +0.323, +0.0, +0.247, +0.338, +0.153, +0.031, +0.014, +0.126,
+0.216, +0.176, +0.469, +0.447, +0.290, +0.113, +0.007, +0.362, +0.393, +0.154, -0.556,
+0.238, +0.168, +0.004, +0.370, +0.250, +0.236, +0.315, -0.053, +0.166, +0.300, +0.274,
+0.150, +0.367, +0.242, +0.003, +0.036, +0.176, +0.377, +0.055, +0.071, +0.129, -0.196]}
Neutrality_State = pd.DataFrame(Fiscal_Neutrality_Scores)
Neutrality_State.reset_index(drop=True, inplace=True)
Neutrality_State.insert(0, 'STATE', range(1, len(Neutrality_State) + 1))
Income_District = pd.merge(Income_Tract, District_Tract, on='tract',
how='outer')
Income_District = Income_District.dropna(subset=['kir_rP_gP_pall'])
Income_District = Income_District.dropna(subset=['NCESID'])
Fund_District = Fund_District.loc[(((Fund_District["NCESID"]) != '') &
((Fund_District["NCESID"]) != ' ') & ((Fund_District["NCESID"]) != '0 0'))]
Income_District["NCESID"] = Income_District.NCESID.astype(int)
Fund_District["NCESID"] = Fund_District.NCESID.astype(int)
Fund_Income = pd.merge(Fund_District, Income_District, on='NCESID')
Fund_Income["Revenue_Per_Student"] =
Fund_Income["TOTALREV"]/Fund_Income["V33"]
Fund_Income["Expenditure_Per_Student"] =
Fund_Income["TOTALEXP"]/Fund_Income["V33"]
Race_Income = pd.merge(Fund_Income, Race_District, on='NCESID')
Dataset = pd.merge(Race_Income, Neutrality_State, on='STATE')
Dataset = Dataset[~np.isinf(Dataset['Revenue_Per_Student'])]
variable_list = Dataset.columns.tolist()
filtered_variable_list = [variable for variable in variable_list if not
variable.startswith('CDP') and not variable.endswith('_I')]
variables_to_remove =
['YRDATDEP', 'GeoId', 'STATE', 'ID', 'SUPID', 'Geography', 'name_x',
'Iteration', 'name_y', 'FIPS', 'SCHLEV', 'WEIGHT', 'YRDATIND']
Filtered_Data = Dataset[filtered_variable_list]

```

```

Filtered_Data = Filtered_Data.drop(columns=variables_to_remove)
mapping_dict = {'kir': 'income', 'rA': 'asian', 'rW': 'white', 'rNA': 'native', 'rB':
'black', 'rH': 'hispanic', 'rP': 'allrace', 'gP': 'allgender', 'gF': 'female', 'gM': 'male',}
columns_to_rename = [col for col in Filtered_Data.columns if any(substring in
col for substring in mapping_dict.keys())]
new_column_titles = [col for col in Filtered_Data.columns]
for old_substring, new_substring in mapping_dict.items():
    new_column_titles = [col.replace(old_substring, new_substring) for col
in new_column_titles]
Filtered_Data.columns = new_column_titles
Poverty_1980Tract.rename(columns={'01001020100 ': '1980tract', '.':
'povertypct'},
inplace=True)
Poverty_1980Tract['povertypct'] =
pd.to_numeric(Poverty_1980Tract['povertypct'].replace('.', float('nan'))
Tract_Tract.rename(columns={'trtid80': '1980tract', 'trtid10': '2010tract'},
inplace=True)
Tract_Tract = Tract_Tract.drop(['placefp10', 'cbsa10',
'metdiv10', 'ccflag10', "weight"], axis=1)
Poverty_2010Tract = pd.merge(Poverty_1980Tract, Tract_Tract,
on='1980tract')
Filtered_Data.rename(columns={'tract': '2010tract'}, inplace=True)
Final_Filter = pd.merge(Poverty_2010Tract, Filtered_Data, on='2010tract')
avgFinal_Filter = Final_Filter.groupby('2010tract').agg({'povertypct': 'mean',
**{col: 'first' for col in Final_Filter.columns if col != 'povertypct'}})
Final_Data = avgFinal_Filter.drop(['1980tract', '2010tract'], axis=1)
Final_Data = Final_Data.dropna(subset=['NCESID'])
Final_Data = Final_Data[Final_Data['statename'] != "Hawaii"]
state_avg_expenditure = Final_Data.groupby('statename')
['Expenditure_Per_Student'].mean()
state_avg_revenue =
Final_Data.groupby('statename')['Revenue_Per_Student'].mean()
Final_Data = Final_Data.merge(state_avg_expenditure, on='statename',
suffixes=(',', '_state_avg'))
Final_Data = Final_Data.merge(state_avg_revenue, on='statename',
suffixes=(',', '_state_avg'))
Final_Data['Expenditure_Distance'] = (Final_Data['Expenditure_Per_Student']
-Final_Data['Expenditure_Per_Student_state_avg'])
Final_Data['Revenue_Distance'] = (Final_Data['Revenue_Per_Student'] -
Final_Data['Revenue_Per_Student_state_avg'])

```

```

Poor_Data = Final_Data[Final_Data['povertypct'] >= 0.139875]
SumData = Final_Data[['statename', 'NAME', 'Whitepct',
'Expenditure_Distance', 'Revenue_Distance',
'Expenditure_Per_Student', 'Revenue_Per_Student',
'Fiscal_Neutrality_Scores',
'income_allrace_allgender_pall', 'povertypct']]
PoorSum = Poor_Data[['statename', 'NAME', 'Whitepct',
'Expenditure_Distance', 'Revenue_Distance',
'Expenditure_Per_Student', 'Revenue_Per_Student',
'Fiscal_Neutrality_Scores',
'income_allrace_allgender_pall', 'povertypct']]
display(SumData.describe())
display(PoorSum.describe())
In [3]: Z = Final_Data.Fiscal_Neutrality_Scores
X = Final_Data.Expenditure_Per_Student
y = Final_Data.income_allrace_allgender_pall
X = X.values.reshape(-1, 1)
Z = Z.values.reshape(-1, 1)
models_first_stage = [(LinearRegression(), {'fit_intercept': [True, False]}),
(KNeighborsRegressor(), {'n_neighbors': [3, 5, 7]}),
(DecisionTreeRegressor(), {'max_depth': [5, 10, 15]}),
(RandomForestRegressor(), {'n_estimators': [50, 100, 150], 'max_depth': [5,
10, 15]}),
(MLPRegressor(max_iter=200), {'hidden_layer_sizes': [(50,), (100,), (50,
50)]})]
best_r2 = -1
best_first_stage_model = None
mlp_converged = True
mlp_fail_count = 0
for model, params in models_first_stage:
    if model.__class__.__name__ == "MLPRegressor" and not
mlp_converged:
        print("MLPRegressor did not converge during the first stage.
        Skipping it.")
        continue
    try:
        with warnings.catch_warnings():
            warnings.simplefilter("ignore",
            category=ConvergenceWarning)
        grid_search = GridSearchCV(model, params, scoring='r2')

```

```

        grid_search.fit(Z, X.ravel())
except ConvergenceWarning:
    print(f"{model.__class__.__name__} did not converge.")
    if model.__class__.__name__ == "MLPRegressor" and
    mlp_converged:
        mlp_fail_count += 1
        if mlp_fail_count == 1:
            print("MLPRegressor failed to converge for the
            first time. Skipping it from now on.")
            mlp_converged = False
        continue
X_rf_estimated = grid_search.best_estimator_.predict(Z).reshape(-1, 1)
r2 = r2_score(X, X_rf_estimated)
print(f"{model.__class__.__name__} R-squared: {r2}")
if r2 > best_r2:
    best_r2 = r2
X_rf_estimated = best_first_stage_model.predict(Z).reshape(-1, 1)
X_train_rf, X_test_rf, y_train, y_test = train_test_split(X_rf_estimated, y,
test_size=0.2)
print( ('End first stage'))
expenditures_list = []
expenditure_value = 2.509895
increment = 0.01
while expenditure_value <= 10.914362:
    expenditures_list.append(expenditure_value)
    expenditure_value += increment
artset = np.array(expenditures_list).reshape(-1, 1)
artset_predict = None
def evaluate_model(model, X_train, y_train, X_test, y_test, model_name,
artificial_data):
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    artset_predict = model.predict(artset)
    mse = mean_squared_error(y_test, y_pred)
    r2 = r2_score(y_test, y_pred)
    print(f"{model_name} Metrics:")
    print(f"R-squared: {r2}")
    print(f"MSE: {mse}")
    print("\n")

```



```

plt.scatter(X_test, y_test, color='navy', label='True Data', alpha=0.7,
s=0.5)
plt.plot(X_test, y_pred, color='cyan', linewidth=1,
label=f'{model_name} Prediction on Test Data')
plt.scatter(artset, artset_predict, color='red', label=f'{model_name}
Prediction on Artificial Data', alpha=0.7, s=0.5)
plt.xlabel('Predicted Expenditure Per Pupil')
plt.ylabel('Average Income at age 35')
plt.legend()
plt.title(f'{model_name} Prediction')
plt.show()
plt.plot(artset, artset_predict, color='maroon', linewidth=.5,
label=f'{model_name} Prediction on Artificial Data')
plt.xlabel('Predicted Expenditure Per Pupil (Artificial Data)')
plt.ylabel('Average Income at age 35 (Artificial Data)')
plt.legend()
plt.title(f'{model_name} Prediction on Artificial Data')
plt.show()
X_train_rf, X_test_rf, y_train, y_test = train_test_split(X_rf_estimated, y,
test_size = 0.2)
models_second_stage = [(LinearRegression(), {'fit_intercept': [True,
False]}),(KNeighborsRegressor(), {'n_neighbors': [3, 5,
7]}),(DecisionTreeRegressor(), {'max_depth': [5, 10,
15]}),(RandomForestRegressor(), {'n_estimators': [50, 100, 150], 'max_depth':
[5,10,15]})]
best_r2_second_stage = -1
best_second_stage_model = None
if mlp_converged:
    with warnings.catch_warnings():
        warnings.simplefilter("ignore", category=ConvergenceWarning)
        grid_search_mlp = GridSearchCV(MLPRegressor(max_iter=1000),
{'hidden_layer_sizes': [(50,), (100,), (50, 50)]}, scoring='r2')
        grid_search_mlp.fit(Z, X.ravel())
        if 'grid_search_mlp' in locals():
            models_second_stage.append((MLPRegressor(max_iter=1000),
{'hidden_layer_sizes': [(50,), (100,), (50, 50)]}))
for model, params in models_second_stage:
    grid_search = GridSearchCV(model, params, scoring='r2')
    grid_search.fit(X_train_rf, y_train)
    best_model = grid_search.best_estimator_

```

```
    r2 = grid_search.best_score_  
evaluate_model(best_model, X_train_rf, y_train, X_test_rf, y_test,  
f"{model.__class__.__name__} (2SLS)", artset)  
if r2 > best_r2_second_stage:  
    best_r2_second_stage = r2  
    best_second_stage_model = best_model  
    global artificial_data  
    artificial_data = artset  
    global artificial_predictions  
    artificial_predictions = artset_predict
```

Name of Candidate: Benvin Fan Lozada

Date of Submission: August 7, 2023