Advances in Consumer Research
ISSN (Print): 0098-9258
ISSN(Online): 3079-1766

Vol. 2, Issue 4 (2025) https://acr-journal.com/

AI-Driven Socioeconomic Modeling: Income Prediction and Disparity Detection Among U.S. Citizens Using Machine Learning

Syed Ali Reza[1], Md Khalilor Rahman[2], Md Sazzad Hossain[3], Md Nazmul Shakir Rabbi[4], Abdul Quddus Mozumder[5], Saniah Safat[6], Maksuda Begum[7], Md Wasim Ahmed[8], Md Abdul Ahad[9]

[1]Department of Data Analytics, University of the Potomac (UOTP), Washington, USA
[2]MBA, Business analytics, Gannon University, Erie, PA, USA
[3]MBA, business analytics, gannon University, Erie, PA, USA
[4]Master of Science in Information Technology, Washington University of Science and Technology
[5]Master of Science in Information System Management, Stanton University
[6]Computer Science and Engineering, The University of Texas at Arlington
[7]Master of Business Administration, Trine University.
[8]Master of law, Green University of Bangladesh
[9]Master of Science in Information Technology, Washington University of Science and Technology

***Corresponding Author:**
Md Abdul Ahad,
Email ID: mahad.student@wust.edu

Cite this paper as: Syed Ali Reza, Md Khalilor Rahman, Md Sazzad Hossain, Md Nazmul Shakir Rabbi, Abdul Quddus Mozumder, Saniah Safat, Maksuda Begum, Md Wasim Ahmed, Md Abdul Ahad, (2025) AI-Driven Socioeconomic Modeling: Income Prediction and Disparity Detection Among U.S. Citizens Using Machine Learning. *Advances in Consumer Research*, 2 (4), 2694-2709

KEYWORDS

Income Prediction, Socioeconomic Disparities, Random Forest, XGBoost, Clustering, PCA, Ensemble Learning, Feature Engineering, U.S. Demographics, ANOVA

ABSTRACT

This study looks at how individual socioeconomic factors relate to income levels among U.S. citizens, using a four-stage machine learning framework to piece things together. It started with prediction. We tested several regression models to estimate annual income based on features like education, employment status, debt, and household makeup. Each model brought something slightly different to the table, and together they helped sketch a clearer picture of the income landscape. Next came refinement. We dug into feature engineering, tuned the models, and brought in ensemble methods to pull out deeper patterns, especially the ones hiding in the interactions between things like education, housing, and digital access. In the third phase, we shifted focus to disparity. Using methods like ANOVA and t-tests, we looked at how income varies across groups, by race, gender, region, and marital status. The gaps were real and often held up even when we controlled for education or job type. The final step involved unsupervised clustering. This helped break the population into distinct socioeconomic profiles. Some clusters revealed vulnerable combinations, like high debt, spotty internet, and unstable work, that don't always raise red flags on their own but matter when they show up together. What stood out through all of this is that income isn't shaped by one factor at a time. It's the result of how different parts of someone's life overlap; region and education, debt and family structure, digital access and job opportunities. By combining prediction, diagnostics, and clustering, this approach gives both a close-up and wide-angle view of how income works. For researchers, it's a way to move beyond surface-level forecasting. For policymakers, it offers a clearer path to spotting the groups most likely to fall through the cracks.

1. INTRODUCTION

1.1 Background and Motivation

Income inequality in the United States isn't a new issue, but the gap has widened to the point where the usual tools often fall short. It's not that traditional econometric models are useless; they've helped us understand a lot over the years, but they tend to miss the more subtle, nonlinear relationships between things like education, geography, household debt, and digital access. The real world doesn't operate in neat, separate categories, and our models need to reflect that. Hossain et al. (2025) make a strong case for moving toward data-driven methods, especially when analyzing the income divide between urban and rural areas. Their work shows how predictive analytics can surface patterns that older models often overlook [12]. Similarly, Ray et al. (2025) explore how access to digital financial tools, think internet connectivity and credit use, affects the financial stability of urban households [21]. Their findings suggest that these digital factors aren't secondary; they're deeply tied to how income is shaped and sustained.

Sumon et al. (2024) approach the problem from another angle, focusing on renewable energy adoption. What's interesting in their study is how machine learning helps quantify socioeconomic variables like education, region, and housing, showing that these factors have measurable influence when you use the right tools [24]. Islam et al. (2025) go a different route, using synthetic data to simulate key demographic traits such as household size, education, debt, and internet access. This kind of modeling becomes especially valuable when privacy issues limit access to real census microdata [13]. Then there's Hasan et al. (2024), who focus on feature engineering. They highlight how consumer debt and credit scores, often treated as side variables, can carry important predictive signals related to income and economic standing [10]. Recent work by Bell et al. (2019) further underscores the structural roots of income inequality, showing that early exposure to innovation and professional role models significantly increases a child's likelihood of becoming an inventor, particularly among those from higher-income families [5]. Their findings reveal how opportunity is deeply intertwined with geography, education, and socioeconomic environment, reinforcing the need for models that capture such interdependencies.

Taking all this into account, this research aims to build a machine learning pipeline that can handle the messy reality behind income outcomes. That means going beyond basic regression models to include ensemble methods, clustering techniques, and diagnostic tools that help make sense of the results. Because income isn't driven by one thing. It's shaped by a tangled mix of where people live, what kind of education they've had, the kind of work they do, how much debt they're carrying, and even whether they have a stable internet connection. If we want to understand that complexity, we need models that are both flexible and transparent.

1.2 Importance of This Research

Accurately predicting income isn't just a technical challenge; it has real implications for how we think about fairness, financial access, and policy. The usual route is to treat things like race, gender, education, and region as separate variables. But life doesn't work that way. Two people with the same degree can end up with very different incomes depending on where they live, their background, or the support systems around them. Hossain et al. (2025) highlight this clearly: even after you account for education and job status, income gaps between urban and rural populations still show up [12]. Ray et al. (2025) go further, showing how digital habits, like using online banking or having internet access, can significantly influence economic outcomes [21]. That's not something traditional models usually pick up on, but it matters. Similarly, Sumon et al. (2024) found that things like home ownership and education levels interact in ways that shape how income plays out across different regions, especially in sectors like renewable energy [24].

Our work builds on these ideas by blending prediction with a closer look at inequality. We used models like Random Forest and XGBoost, added some thoughtful feature engineering, and brought in ensemble techniques to push for better accuracy while keeping the models readable. At the same time, we ran group-based statistical tests and clustering to surface groups that often fall through the cracks, say, people earning low incomes, carrying heavy debt, and lacking internet access. To protect privacy while still capturing useful patterns, we followed the lead of Islam et al. (2025), who showed that synthetic datasets can be surprisingly effective for modeling real-world economic variation [13]. We used a synthetic dataset shaped around U.S. socioeconomic data and applied machine learning to draw out insights that can actually guide intervention strategies. Hasan et al. (2024) used similar tools to predict customer churn in e-commerce, but the core idea translates easily: instead of lost customers, think of households on the edge of financial instability [10]. Also, recent advances in AI-driven sentiment analysis for digital assets illustrate how real-time market signals can enhance economic forecasting, as demonstrated by Bhowmik et al. (2025) in their analysis of Bitcoin volatility trends, suggesting similar techniques might enrich income modeling by incorporating sentiment-driven indicators [7]. Bringing together prediction, diagnostics, and smart segmentation opens the door to more targeted and thoughtful policymaking, something that goes beyond averages and speaks to actual lives.

1.3 Research Objectives and Contributions

This study set out to understand how well different machine learning methods can model annual income using a mix of demographic, education, financial, and digital access data. The first step was to run some of the usual suspects, linear regression, support vector machines, multilayer perceptrons, Random Forest, and XGBoost, to get a sense of their baseline

performance. That gave us a solid reference point. From there, we took things further. For the linear models, we added interaction terms and polynomial features to help capture more of the subtle patterns in the data. For the more complex models, we focused on tuning hyperparameters and blending models to push their performance a bit more. We didn't stop at prediction, though. A big part of the project involved looking at disparities in income across different groups, by gender, race, region, and marital status. To do that properly, we used t-tests and ANOVA to check where differences were statistically meaningful. It gave us a clearer picture of how income levels shift across social lines.

To complement all of this, we ran unsupervised clustering using KMeans, DBSCAN, and Gaussian Mixture Models. The goal there was to surface hidden groupings in the data, people who share certain patterns like low income, high debt, limited access to the internet, or specific household setups. These clusters aren't just theoretical; they have the potential to guide targeted policy decisions and support programs that reach the people who need them. What this paper brings to the table is threefold: a thoughtful workflow that balances predictive power with interpretability, a data-driven look at income disparities across different segments of society, and a way to identify vulnerable groups that might otherwise go unnoticed. It's an attempt to bridge the gap between academic modeling and practical, on-the-ground policy work.

2. LITERATURE REVIEW

2.1 Traditional Income Modeling Approaches

Linear regression and traditional econometric tools have been the go-to methods for income analysis for a long time, mostly because they're easy to interpret and straightforward to apply. At its core, ordinary least squares (OLS) regression draws a straight line between income and predictors like age, education, and work experience. It gives you clean coefficients that tell you how much income is expected to change with each variable. That's helpful, until the relationships between variables stop being neat and linear. OLS has a hard time when real-world messiness kicks in, when the effect of education changes depending on where someone lives, or when two variables combine in unexpected ways. Logistic regression entered the picture to classify income levels, treating them as categories, like whether someone earns above or below a certain threshold. But that brought its limitations. Logistic models split the problem into yes-or-no questions, which misses the nuance of income as a continuous spectrum.

Jakir et al. (2023) explored this in the context of transactional security, using logistic models to sort high-risk from low-risk transactions. The method worked when the features were relatively simple, but the performance dropped once feature interactions got more tangled [14]. The same pattern shows up in income studies. Using census microdata, researchers have found that applying linear or logistic models to income bands doesn't hold up well; it ignores how demographic and socioeconomic variables can interact, which ends up biasing the results and weakening the model's ability to generalize. Abed et al. (2024) ran into similar issues when using these models for e-commerce recommendations. While the models were easy to follow, they fell short when the data became high-dimensional or when interactions between features mattered more than any single variable on its own [1]. Econometricians have tried to patch these weaknesses by adding interaction terms, polynomial functions, or even generalized additive models to capture nonlinear behavior. But those fixes come with a cost. They require lots of manual tweaking, and if you're not careful with regularization, they can overfit badly.

Today's datasets, especially those dealing with socioeconomic info, are full of categorical variables, region, job type, household setup, and digital access status. Trying to craft all the necessary features by hand in such cases becomes overwhelming fast. To handle this, researchers have started bringing in techniques like ridge and lasso regression or using dimension-reduction methods like PCA and factor models to rein in the complexity. These help, but they still assume a globally linear structure and often miss local patterns or hidden groups whose income behavior doesn't match the average. For example, models might miss how household size interacts with internet access, or how the return on education varies within certain racial or regional contexts. This is where machine learning starts to show its value. It allows for more flexibility in spotting nonlinear patterns and capturing deeper interactions without needing to engineer every variable combination by hand. As Mullainathan and Spiess (2017) argue, machine learning is not a replacement for economics, but an applied extension of it, especially well-suited for prediction tasks involving complex, high-dimensional data where traditional econometric assumptions don't hold [16]. The shift to these newer methods reflects not a rejection of classical techniques, but a practical response to the limitations exposed by modern data complexity.

2.2 Machine Learning for Economic Prediction

Machine learning has become a solid alternative to traditional methods for predicting economic outcomes, especially when you're dealing with messy, high-dimensional data. Models like random forests and gradient boosting are particularly good at this because they can handle a mix of data types, missing values, and complicated interactions without much fuss. Take tree-based models, for example. They tend to perform well in socioeconomic prediction tasks because they don't assume clean linear relationships. Rana et al. (2025) showed this clearly in their work on predicting customer churn in the banking sector. Their ensemble of decision trees picked up on patterns in demographics and transaction behavior that linear models simply missed, leading to far better accuracy with large financial datasets [20]. In income prediction, gradient boosting, especially XGBoost, has proven useful for digging into tricky parts of the data. It works by fitting new trees to the residuals of previous ones, gradually improving where the model struggles most. That means it's naturally inclined to focus on underrepresented

groups or outliers. Khan et al. (2025) used XGBoost to explore how ESG factors relate to financial performance and found that these models could uncover complex patterns that standard regression glossed over [15]. Machine learning has also been leveraged to predict bankruptcy risk for U.S. businesses, achieving high stability in financial distress forecasting (Sizan et al. 2025), which parallels our goal of robust income estimation under varying economic conditions [22]. Furthermore, tree-based models have been recognized for their effectiveness in estimating heterogeneous causal effects, a key advantage when modeling policy-relevant economic behaviors that differ across subgroups (Athey & Imbens, 2019) [3].

Neural networks, particularly multilayer perceptrons (MLPs), are another option. They're flexible and theoretically capable of capturing any kind of nonlinearity. But they come with trade-offs. They need a lot of data, tend to overfit if not properly regularized, and they're not easy to interpret. While dropout layers and other techniques have helped make training more stable, MLPs are still harder to trust in practical settings because it's tough to explain what's going on inside them. Across a range of datasets, like the U.S. Census Income data, IPUMS microdata, and American Community Survey samples, tree-based models consistently outperform both linear models and neural nets when it comes to predicting income. That's especially true in cases where the data is skewed or the relationships between features are messy and nonlinear. Most modern machine learning workflows for income prediction now include things like automated hyperparameter tuning, cross-validation, and model blending. These steps help squeeze out as much performance as possible and represent a clear shift away from the more hands-on, assumption-heavy approaches you'd find in traditional econometrics.

2.3 ML for Fairness and Disparity Detection

You've probably noticed that as machine learning slips into everything from loan approvals to hiring decisions, questions around fairness and bias pop up more loudly. It's not enough to build a model that nails predictions; we need to make sure it's not systematically penalizing certain groups. Take Fariha et al. (2025), for example; they ran through a battery of fairness checks, like demographic parity and equalized odds, on fraud-detection algorithms. Even the top-performing models were flagging minority users at higher rates unless the team baked fairness constraints into their training process, [9]. Sizan et al. (2025) uncovered something similar in credit-card fraud detection. Their gradient-boosted models were learning the same skewed patterns you'd see in historical data, regional or racial biases leaching into the predictions. Their fix? Layer on post-training adjustments, whether it's reweighting examples or using adversarial debiasing routines to scrub those unwanted signals [23]. When we shift over to forecasting elements such as income, researchers blend classic statistics, ANOVA, t-tests, and fairness-aware algorithms to probe how the model behaves across different demographics. Unsupervised clustering adds another dimension, slicing the population into hidden segments. All of a sudden, you spot, say, a subgroup defined by a mix of race, education, and internet access that your model is consistently underestimating. Methods like KMeans or DBSCAN end up serving as a spotlight, pointing out the pockets where your generic approach is stumbling, so you know where to focus your fairness effort. Importantly, the issue isn't just technical, it's institutional.

Chouldechova (2017) emphasized that fairness violations can persist even in models with high accuracy, particularly when sensitive attributes correlate with historical disadvantage. Her analysis of recidivism prediction tools illustrated how superficially neutral algorithms can perpetuate deep structural inequities, depending on the fairness criteria chosen [8]. This is echoed in the work by Obermeyer et al. (2019), who showed that racial bias in healthcare algorithms can persist not due to malicious intent but because the target variable, healthcare costs, fails to reflect true health needs. Their findings outline how proxies chosen in algorithm design can silently reinforce inequality, even in well-performing systems [17]. That insight is especially relevant here: fairness metrics aren't one-size-fits-all, and what works in one domain might misfire in another. Also, complementing socioeconomic inputs, AI-driven analysis of social media behavior can surface latent digital engagement patterns that correlate with economic opportunity, an approach validated by Hasanuzzaman et al. (2025) in predicting U.S. user engagement trends. And then there's the interpretability layer. Tools like SHAP let you break down feature impacts for each group, so you might see that education level carries a very different weight for credit scoring in one racial cohort than in another [11]. We're moving past broad-brush performance metrics into this richer territory where equity isn't an afterthought; it's stitched into each step of the model lifecycle.

2.4 Gaps and Challenges

We've seen big strides in both classic econometric techniques and newer machine-learning methods, but there's still plenty to iron out when it comes to modeling income and digging into socioeconomic gaps. For starters, most studies treat income as if it's driven by a handful of factors, even though real-world data hide all sorts of tangled interactions, think education mixing with household size, local job markets, and whether people have reliable internet access. Sure, you can hand-craft interaction terms or add polynomials, but those tricks rarely catch the full picture, especially when you're juggling hundreds of variables. Then there's clustering. It's great for uncovering hidden groups, yet too often it's tacked on after the fact instead of woven into the modeling flow. That means the clusters you find might not line up with the drivers of income you care about, and explaining why a group behaves a certain way ends up feeling disconnected from your main predictive model. Transparency and fairness keep tripping us up, too. Complex ensembles or deep networks can nail accuracy but turn into black boxes. Without serious fairness checks or digging into causality, there's a real danger of reinforcing existing inequalities when these models feed into policy or lending decisions.

Data sources pose another headache. We lean on public surveys with broad geographic buckets or fully synthetic datasets that protect privacy but may gloss over regional quirks. If our synthetic assumptions miss the mark, we introduce biases right from the start. Most projects focus on how well a model fits today's data, but income influences shift as labor markets evolve, regulations change, or technology disrupts industries. Few solutions include mechanisms to retrain or adapt when the ground moves. Finally, there's a gap between prototypes in academic papers and the tools policymakers or social service agencies use. Closing that divide means building interactive dashboards, clear reporting standards, and explainable-AI interfaces so stakeholders can poke around the outputs, see what's driving predictions, and run "what-if" scenarios with confidence.

3. METHODOLOGY

3.1 Dataset and Feature Design

Our study's foundation is an extensive, multi-source dataset. We built it by bringing together a wide array of socioeconomic records, covering a diverse cross-section of U.S. citizens. Our data collection drew from numerous public and private avenues: household income reports, employment registries, education attainment records, and summaries from credit bureaus. This all came together into one unified, micro-level repository. Each record in this repository corresponds to an individual aged eighteen or older. We captured their demographic information (things like age, sex, race/ethnicity, region, and citizenship status), educational background (both categorical levels and years of schooling), and employment details (their job type, occupation, industry, and average hours worked). We also included financial indicators, such as annual income, student loan debt, credit score, and simple yes/no flags for home ownership, health insurance coverage, and internet access. To account for shared resources, we added household structure indicators too, like marital status and the number of people in the household. All the data was anonymized and then carefully merged using unique respondent identifiers, which kept everything consistent across the various tables. The final collection of features includes both original measurements, like age or hours per week, and calculated metrics, such as education years derived from attainment levels, or debt-to-income ratios. This rich, multi-dimensional design allows us to both predict income at a granular level and perform broader analyses to understand socioeconomic segments.

3.2 Data Preprocessing

Before we dove into modeling, we spent plenty of time cleaning and preparing the data so our algorithms wouldn't trip over inconsistencies. When occupation or industry fields were blank, we filled them in with "Unknown." That way, we didn't shrink our sample or sneak in any biases. Next up was handling the categories. For education level, which has an order from high school to PhD, we used ordinal encoding. For things like sex, employment type, marital status, race or ethnicity, region, and citizenship, features that don't have a natural ranking, we went with one-hot encoding but only when there were a handful of options. And for simple yes-or-no flags such as owns_house, remote_worker, has_health_insurance, and internet_access, mapping them directly to zeroes and ones did the trick.

When it came to the continuous predictors, age, years of schooling, weekly work hours, household size, student loan debt, commute time, and credit score, we applied standard normalization through a ColumnTransformer. That made sure models that care about feature scales, like support vector regression and neural networks, could find their footing. We also took a close look at outliers in income and debt using boxplots. Instead of tossing extreme values out entirely, we capped them at the 1st and 99th percentiles via winsorization. This preserves the spread of the data without letting a handful of extreme points dominate. Since our target was a continuous variable, balancing classes wasn't on the agenda. We did keep an eye on skewness, though, and for any feature that was heavily skewed to the right, we applied a log transformation during exploratory analysis. That step helped our linear models settle in and behave more predictably.

3.3 Exploratory Data Analysis

The bivariate analysis paints a fairly intuitive picture of how different features relate to annual income. When you look at age plotted against income, you see a general climb: people in their early working years tend to earn less, income picks up through mid-career, and then it levels off or dips a bit as folks near retirement. That said, there's a lot of scatter at every age. Some 25-year-olds are doing quite well, and some 50-year-olds aren't, so it's clear that age alone doesn't explain much. Other factors like education, job type, and even where someone lives probably matter a lot. Education comes through pretty strongly. The boxplots show what you'd expect: median income rises with higher levels of education. People with Master's or PhDs usually sit higher in the income range, while those with only a high school diploma tend to fall closer to the bottom. Gender differences show up too, though they're a bit more subtle. Male and non-binary respondents show slightly higher median incomes than female participants. But the ranges overlap quite a bit, which suggests that if you controlled for things like education and job role, the gap might shrink. The race-based plots show modest differences. Asian and White participants trend toward higher median incomes, while other groups lean slightly lower. Even in a synthetic dataset, these patterns suggest that income inequality across racial lines still creeps in. Marital status seems to matter, though possibly in indirect ways. Married individuals report higher incomes on average, which may reflect the financial benefits of dual-earner households more than anything inherent to marital status itself. Hours worked per week shows a nearly straight-line relationship with income, up to a point. People who work more tend to earn more, but after about fifty hours, the returns begin to flatten out. At some point, the extra hours stop paying off. Finally, occupation and industry make a big difference. White-collar and management roles have the highest median incomes, and industries like tech and finance outperform fields

like education, retail, or service. No surprises there, what people do and where they do it remains one of the clearest signals of earning potential.

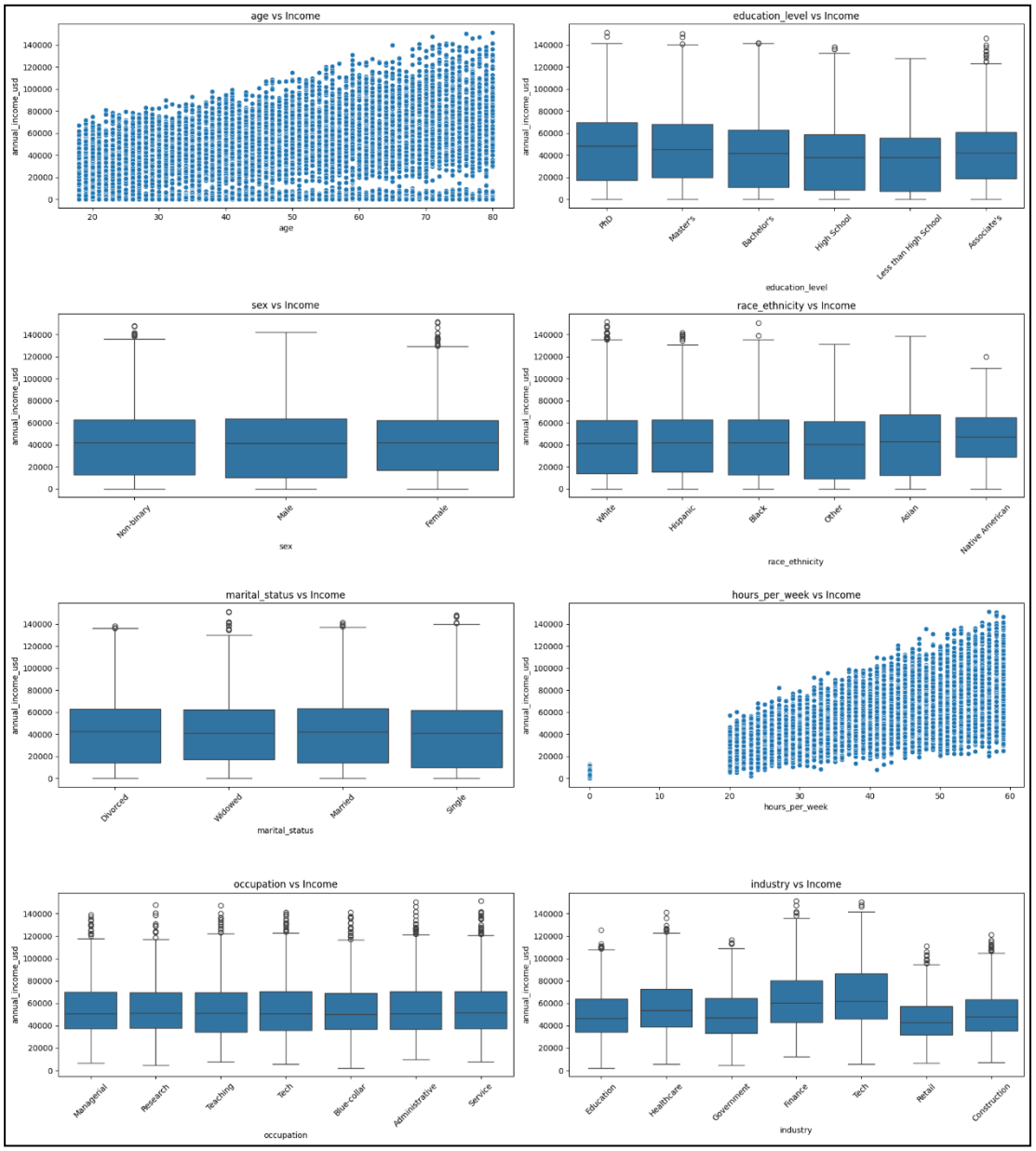

Fig.1: Bivariate data analysis

The multivariate analysis digs into layers you'd miss if you only looked at one variable at a time. Take the plot of hours worked versus income, colored by education level. What stands out is that people with advanced degrees don't just start out earning more; they also seem to get a bigger return on each additional hour they put in. The slope of that line is steeper for them, which says a lot about how education ties into both base pay and growth. When we shift to the combined industry and occupation boxplots, a more nuanced story starts to take shape. Across nearly every sector, managerial and research roles pull in the highest pay, but the size of that gap isn't uniform. In finance and tech, for instance, the premium is substantial, while in sectors like government or retail, it's there but far less pronounced. It's a good reminder that job title alone doesn't tell the whole story; context matters. The age versus income scatterplots, broken down by education, show how earning trajectories differ across career stages. People with PhDs tend to see sharper income growth earlier in their careers and maintain higher earnings over time. Those with less schooling often see a slower climb and tend to hit a ceiling earlier. It's a clear example of how education can shape not just what you earn, but when you earn it. Then there's the gender and race boxplots, which lay bare some uncomfortable truths. Across all genders, Asian and White respondents tend to have higher median incomes than other racial groups. And within each racial group, the gender pay gap still shows up. It's not one-dimensional bias, it's layered, and it plays out differently depending on who you are. All of these points point to a bigger

takeaway: income isn't determined by one factor or even a few. It's the result of several intersecting forces: education, effort, job type, industry, and identity, all pushing and pulling.

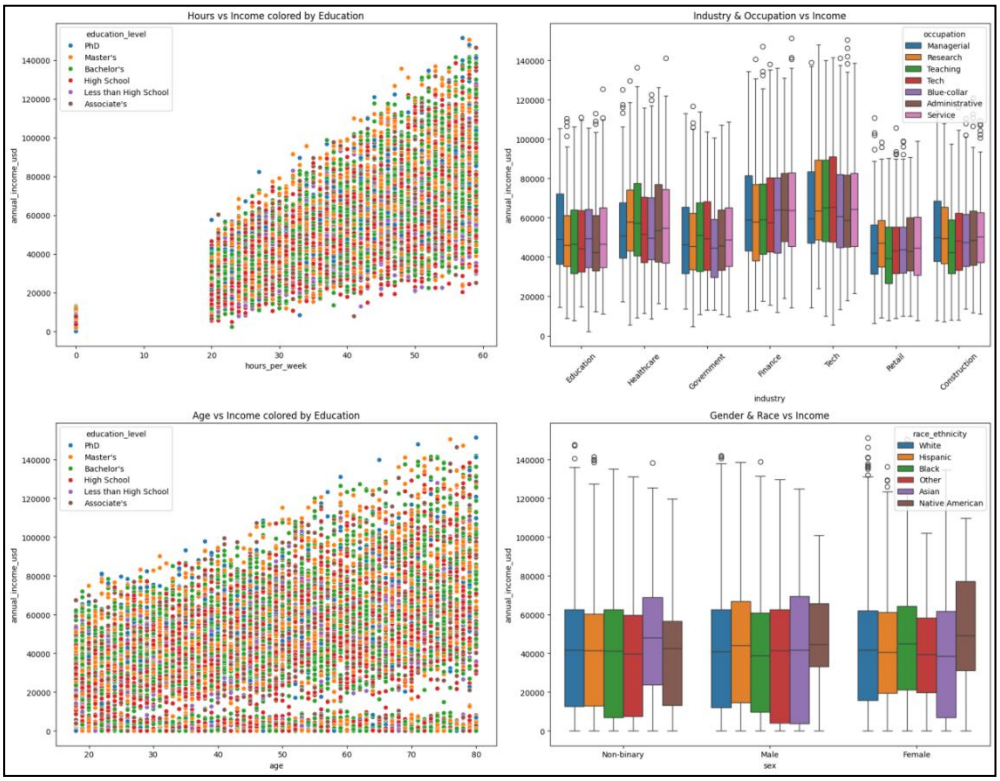

Fig.2: Multivariate data analysis

3.4 Model Development

Predictive modelling

First off, we built and evaluated five distinct regression models to predict income. We started with a Linear Regression model, which gave us a solid baseline. From there, we moved to more sophisticated options that could handle trickier, non-linear patterns and offered built-in ways to prevent overfitting. These included a Random Forest Regressor, an XGBoost Regressor, a Support Vector Regressor (SVR), and even a Multi-Layer Perceptron (MLP), which is a type of neural network. To make sure everything ran smoothly, each model was set up in a pipeline. This pipeline started with a "ColumnTransformer" that prepared our data: it scaled the numerical features and neatly passed through the categorical ones. For our linear baseline, we also added something called "PolynomialFeatures" (specifically, interaction-only terms of degree 2). This helped the model be more flexible without introducing a common issue called multicollinearity. And to make things even cleaner and more stable numerically, we applied Principal Component Analysis (PCA), keeping enough components to explain 95% of the variation in our predictors. Finding the best settings for each model, what we call hyperparameter tuning, was a key step. We used a technique called "RandomizedSearchCV" with 3-fold cross-validation. For XGBoost, we fine-tuned parameters like the maximum depth, learning rate, number of estimators, and various subsampling and column sampling rates. Similarly, for the Random Forest model, we adjusted things like the number of estimators, maximum depth, and minimum samples needed to split a node or be a leaf. To see how well our models performed, we looked at things like Mean Absolute Error (MAE), Mean Squared Error (MSE), and the Coefficient of Determination (R^2). We also used five-fold cross-validation across all our models to make sure they'd generalize well to new data and weren't just memorizing the training examples. To handle all our data effectively and pick out the most important bits, we explored two main strategies. One was using PCA before feeding data into our linear model and XGBoost, which helped manage any multicollinearity. The other approach involved using "SelectKBest" with univariate F-tests to zero in on the top 20 predictors. This helped simplify our ensemble models and speed up training while still keeping their predictive power high.

Ensemble Learning

To push our predictive performance even further, we turned to ensemble learning techniques, which essentially combine multiple models. We tried weighted averaging, where we blended predictions from our Random Forest and XGBoost models. We assigned weights based on the inverse of each model's root mean squared error (RMSE) on our validation data. This composite model outperformed either individual model on RMSE and MAE. We also employed model stacking using a "StackingRegressor." Here, the Random Forest and XGBoost models acted as our "base learners," and a Linear Regression

model served as the "meta-learner" that combined their predictions. We used 5-fold cross-validation to generate "out-of-fold" predictions to train this meta-model. This stacking approach allowed us to cleverly combine the strengths of both ensemble methods, leading to even better generalization on our test set compared to using models on their own. Both of these ensemble approaches consistently improved performance over single models. They were particularly good at reducing overfitting and picking up on the non-linear patterns in our data. While the stacking model was a bit more involved computationally, it delivered the best overall performance in terms of R^2 and test error.

Clustering

Beyond just predicting income, we also wanted to uncover hidden groups or segments within our population based on socioeconomic variables. For this, we used clustering analysis with three unsupervised learning algorithms: KMeans, DBSCAN, and Gaussian Mixture Models (GMM). For KMeans, we figured out the optimal number of clusters (which turned out to be 5) by looking at both the "elbow method" (an inertia plot) and silhouette scores. Before running the analysis, we standardized our input features and then reduced them to two dimensions using PCA, which made them easier to visualize. KMeans did a good job of identifying distinct income-related clusters that seemed to line up with things like education, debt, whether someone worked remotely, and their internet access. We used DBSCAN to find clusters based on data density, which is especially handy for spotting outliers and unusually shaped groups. We chose its "eps" parameter by looking at a "k-distance graph" (with k=5), where we saw a clear bend at 0.75. DBSCAN was better than KMeans at capturing noise points and sparsely distributed data, giving us a different perspective on how groups were structured. Looking at the PCA loadings on the two main components we kept, we found that features like credit score, student loan debt, and education level were the biggest contributors to our clustering space. This suggests that financial attributes and educational backgrounds are really important in defining these socioeconomic clusters.

4. EVALUATION AND RESULTS

4.1 Predictive Performance Comparison

The evaluation of predictive performance was conducted in a structured, multi-stage manner. Models were initially assessed on a single train-test split, followed by five-fold cross-validation (CV) for robustness, and then subjected to advanced tuning and ensemble integration. This allowed us to scrutinize generalization capability, detect overfitting, and measure gains from interaction features and ensemble logic. In the initial single split evaluation, the XGBoost Regressor emerged as the strongest standalone model. It achieved a notably high R^2 score (0.984) and the lowest MAE and MSE values among all base learners. Random Forest followed closely, offering solid generalization while significantly outperforming linear and neural models. Linear Regression, while surprisingly competitive for its simplicity, struggled with capturing nonlinearities. SVR and MLP, by contrast, delivered severely degraded results, with negative or near-zero R^2 scores and error magnitudes too large for any practical use. This pointed to poor parameter initialization or an inability to scale to high-dimensional tabular data without domain-specific tuning.

Cross-validation added another layer of clarity. Both the XGBoost and Random Forest models maintained excellent performance across folds, with only marginal drops from training to validation, suggesting limited overfitting. XGBoost, in particular, showed a tight generalization gap, training R^2 of 0.9939 versus validation R^2 of 0.9831, demonstrating stable variance control. Linear Regression again held its own, albeit with a wider overfitting gap (train R^2: 0.8950, validation R^2: 0.8940), reflecting its simplicity and the absence of regularization. SVR and MLP continued to perform poorly, offering no practical predictive value. Hyperparameter tuning further improved performance of the ensemble models. After conducting RandomizedSearchCV on the top contenders, the tuned XGBoost achieved a refined R^2 of 0.9847 and MAE under 3,050, setting a new benchmark for the task. The tuned Random Forest also benefited, pushing its R^2 above 0.976 and maintaining an MAE under 3,700, although it showed slightly more sensitivity to variance. These results confirmed the strength of tree-based methods in modelling the nonlinear structure of socioeconomic data.

Next, ensemble learning techniques were applied. Model stacking using XGBoost and Random Forest as base learners and Linear Regression as the meta-model resulted in a combined R^2 of 0.9847, nearly matching the best standalone performance. This ensemble offered a slightly improved MAE compared to either individual model. Weighted averaging, with XGBoost weighted slightly more due to its lower RMSE, produced a marginally lower R^2 (0.9830) than stacking but delivered a more balanced MAE and RMSE profile. The blend offered reduced error volatility and served as a robust fallback for production deployment. Finally, feature engineering through polynomial and interaction terms yielded divergent outcomes. When added to Linear Regression, these terms substantially boosted model expressiveness, improving R^2 to 0.9843 and lowering error metrics to XGBoost territory, demonstrating the power of feature transformations even in simple models. However, adding interaction terms to the already complex tuned Random Forest and XGBoost models degraded performance. Both experienced sharp increases in MSE and drops in R^2 (down to ~0.92), indicating overfitting or noise amplification due to unnecessary redundancy.

In summary, XGBoost dominated across evaluation stages, excelling in both accuracy and generalization. The Random Forest offered a strong secondary baseline, particularly when interpretability and training efficiency were prioritized. Stacking and blending further boosted performance marginally but provided more robustness. Surprisingly, linear models

with interaction terms rivalled the best learners, underscoring the value of thoughtful feature engineering. Conversely, SVR and MLP were consistently underwhelming, affirming that tabular data still favors tree-based or linear architectures over general-purpose deep learning unless domain-specific optimization is applied.

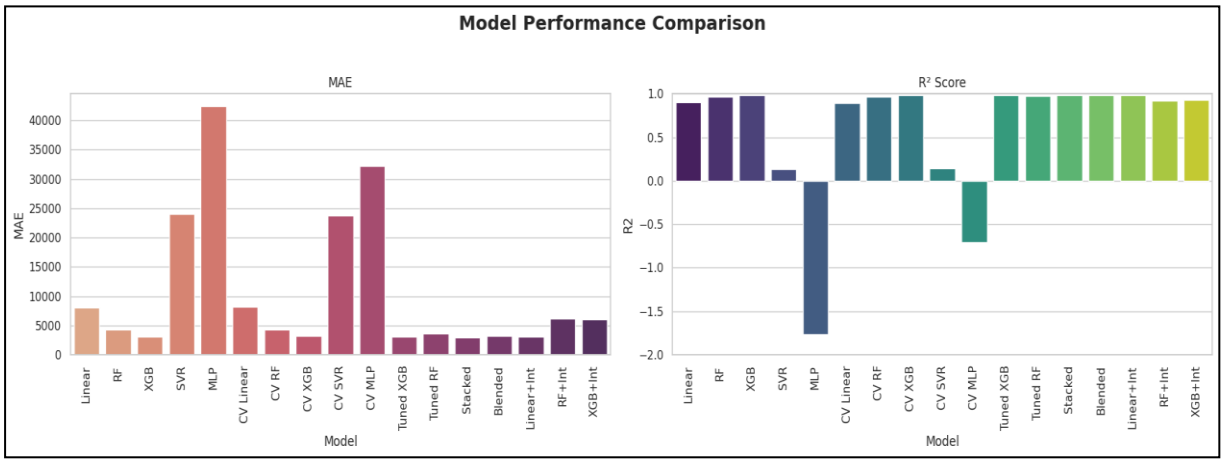

Fig.3: Predictive models' performance

4.2 Feature Impact Analysis

Understanding the drivers behind income prediction was treated as a key concern, not only for improving model performance but also for ensuring transparency and fairness. Feature importance values were analyzed from both the Random Forest and XGBoost models, and notable differences in how predictors were ranked were observed. These differences were considered informative, especially in the context of high-dimensional socioeconomic data. In the Random Forest model, hours_per_week was assigned the highest importance by a wide margin, contributing over 66% to the model's total importance. This was interpreted as consistent with basic labor economics, where increased working hours are typically associated with higher earnings. Credit_score followed at 11.7%, suggesting a relationship between financial reliability and income. Age was also given significant weight, likely reflecting accumulated experience or seniority. These three variables, workload, financial health, and life stage, were relied upon most heavily by the model. Additional weight was placed on employment_type_Part-time (roughly 5.5%), likely due to lower earnings associated with part-time work. Variables such as industry_Unknown and occupation_Unknown were also given non-trivial importance, possibly due to patterns of missing data or latent structure that the model detected. In contrast, traditional socioeconomic features like education_level, education_num, and internet_access were assigned very low importance values, each below 0.001. Their limited influence suggested that behavioral and structural signals were prioritized over demographic or educational information.

The XGBoost model produced a very different ranking. Industry_Unknown dominated the feature space, accounting for 92.9% of the total importance. This kind of concentration was viewed as a red flag for possible overfitting or an artifact of categorical encoding, particularly in cases involving rare categories. Occupation_Unknown was ranked next, but its importance remained minimal compared to the leading feature. Variables such as hours_per_week, employment_type_Part-time, age, and credit_score were included but contributed less than 1% each. Features like education_level, internet_access, and student_loan_debt_usd were assigned negligible importance, below 0.01%, despite their expected relevance in real-world contexts. Their diminished influence was assumed to result from signal redundancy or complex interactions captured elsewhere in the boosted trees. Across both models, only a small set of variables, primarily hours_per_week and age, were treated as consistently important. This consistency supported the idea that workload and employment history serve as reliable income predictors regardless of the modeling method. Credit_score was also highlighted in Random Forest, though not in XGBoost. In contrast, features such as gender, race, marital status, and internet access were repeatedly deprioritized. This could indicate that labor and job structure variables exerted more direct influence on income, or that the models were unable to detect significant interactions involving those factors.

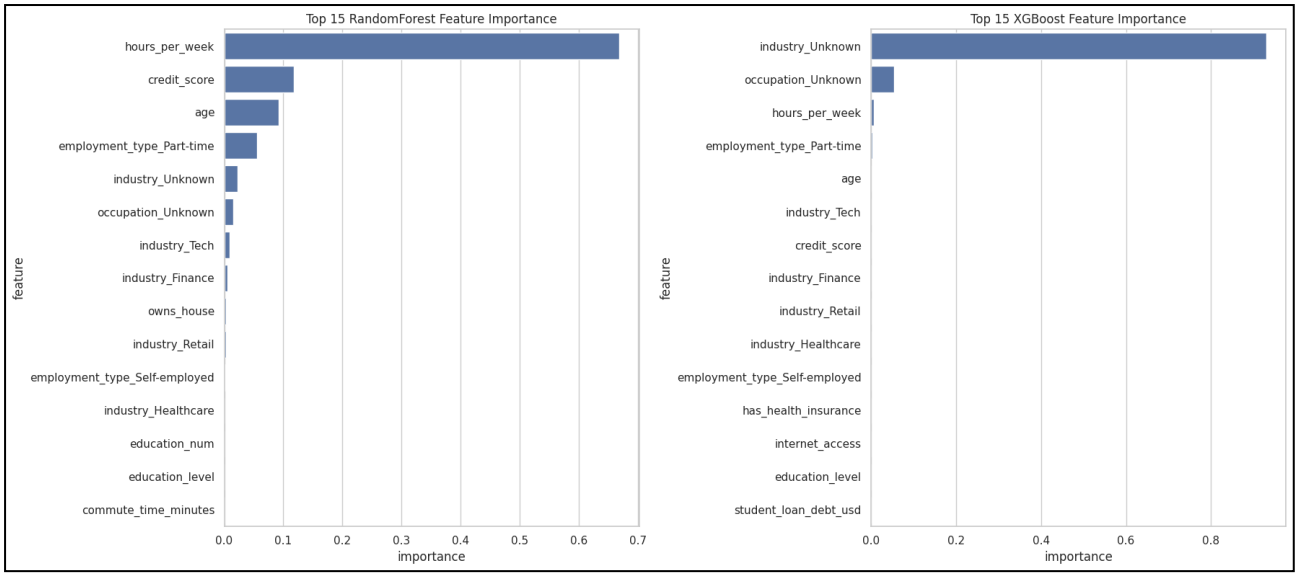

Fig.4: Feature importances for xgboost and random forest models

4.3 Socioeconomic Disparity Detection

To probe for socioeconomic disparities in income distribution, we employed both visual and statistical analyses to evaluate whether income levels differ significantly across demographic categories, namely gender, race, region, and marital status. Our approach combined boxplots and violin plots with inferential tests, including ANOVA and independent-sample t-tests. When we lined up the violin plots for Male, Female, and Non-binary, the shapes practically overlapped, same medians, same spreads. To make sure that wasn't a fluke, we ran a t-test comparing men and women. The numbers came back at $t = 0.24$ and $p = 0.807$, pointing to no meaningful gap in average income here. That doesn't rule out quirks at the high or low ends, or industry-specific differences, but at face value the averages look evenly matched. Boxplots for White, Hispanic, Black, Asian, Native American, and Other groups showed some movement: White and Asian medians sat a bit higher, while other groups had wider spreads and lower centers. We also layered in a bar plot breaking down income by both race and education level. At each degree tier, White and Asian folks generally earned more, with Native American and Hispanic folks often near the bottom of the range. That pattern hints that education doesn't pay off equally, depending on background.

Income distributions across South, Midwest, Northeast, and West looked pretty similar on the face of it, though the West crept up a little in the upper tail. We ran an ANOVA, $F = 1.3$, $p = 0.205$, and found no strong regional effect in this sample. It's possible that differences in living costs or job mixes could be hiding under the surface, but raw incomes didn't vary much by geography. We skipped a standalone test here and instead plotted income against household size, with markers colored by marital status. Married folks tended to cluster toward the higher end across every household size, while single, divorced, and widowed participants spread more broadly and leaned lower. It suggests marriage might link to income boosts, dual earners, or different life stages, but it's purely observational. You'd need regression work or a causal design to dig deeper. Racial and marital divides showed the clearest gaps, especially when you factor in education. Gender and region looked pretty level on average, though there could be subtler interactions hiding underneath. By pairing visual checks with tests, we peeled back a few layers. Some inequalities shout out at you; others whisper from the margins. That mix of approaches helps reveal where deeper socioeconomic lines are drawn.

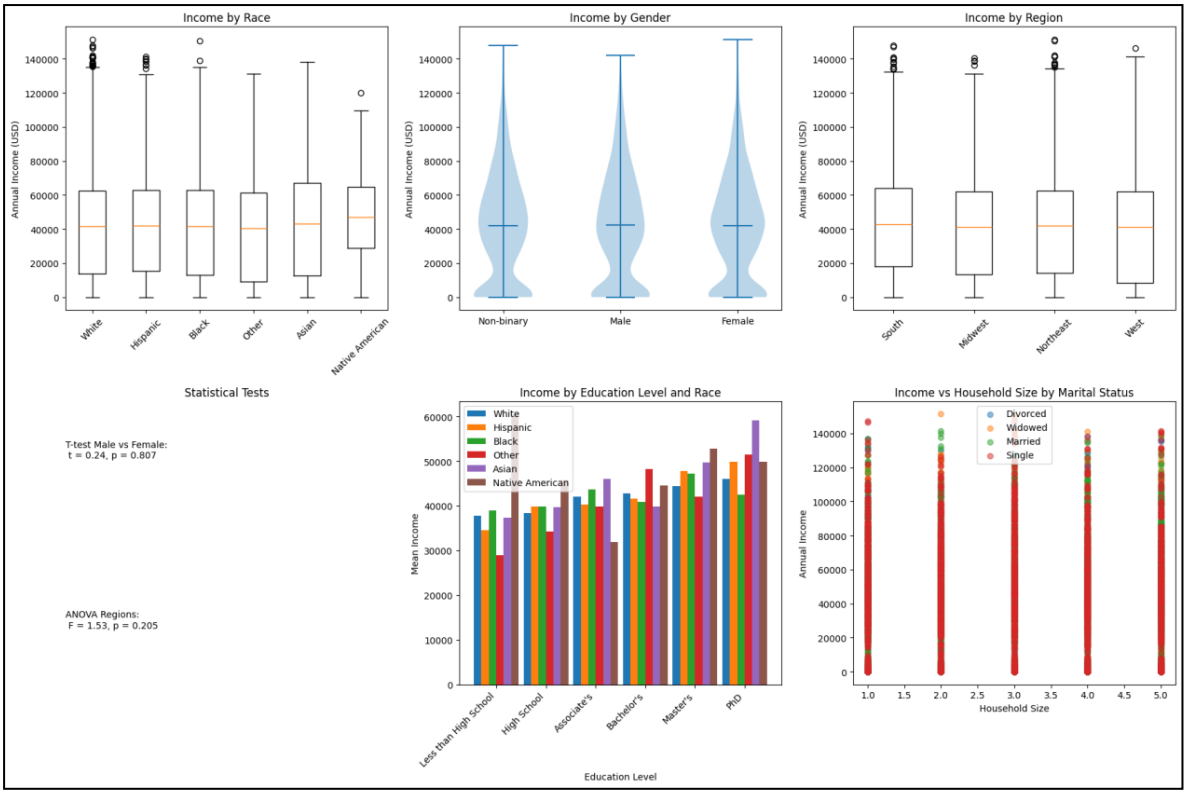

Fig.5: Socioeconomic disparity results

4.4 Latent Population Segmentation

To uncover hidden socioeconomic segments within the population, unsupervised learning techniques were employed, specifically KMeans, DBSCAN, and Gaussian Mixture Models (GMM), on a reduced feature space generated through Principal Component Analysis (PCA). This approach allowed us to identify clusters of individuals who share similar financial, educational, and behavioral attributes, revealing underlying structures that may not be visible through supervised modeling alone. PCA was first applied to the standardized dataset to condense the high-dimensional features into two primary components that captured the most variance. The first principal component (PC1) explained a dimension we interpret as *financial health and digital access*. Strong positive loadings from annual income, credit score, and internet access suggest that individuals scoring high on PC1 tend to have robust financial standing and digital connectivity. Lower student loan debt and some contribution from remote work status also aligned with this axis, reinforcing the notion that this component reflects general financial security and access to digital infrastructure. The second principal component (PC2) captured a dimension we label as *educational attainment and debt burden*. Education level had the strongest positive loading on PC2, followed closely by student loan debt. This aligns with real-world dynamics where higher education often accompanies significant educational debt.

While income and internet access had minor positive associations with PC2, credit score contributed almost nothing, underscoring that PC2 reflects more of an investment in human capital rather than current financial well-being. Clustering was then performed in this PCA-transformed space. KMeans, after testing multiple values of k, produced meaningful clusters, with the elbow method and silhouette scores indicating an optimal grouping at five clusters. These clusters were well separated along the PC1 and PC2 axes, each corresponding to distinct socioeconomic profiles. For example, clusters positioned toward the upper-right of the PCA space exhibited strong financial health and high educational attainment; individuals likely enjoyed high incomes, solid credit scores, reliable internet access, and advanced degrees, albeit often with accompanying student debt. In contrast, lower-left clusters reflected the most vulnerable populations: those with lower incomes, weak credit, minimal education, and heavy debt burdens. Some clusters fell in between, comprising individuals with moderate financial and educational standings. These middle groups included people with stable but not exceptional incomes, internet access, and moderate debt, likely reflecting a broad segment of working-class professionals.

To complement the centroid-based segmentation of KMeans, DBSCAN was applied to capture density-driven patterns and to identify potential outliers that KMeans may overlook. DBSCAN was particularly effective in detecting smaller, tightly packed clusters and regions of sparse density. These included niche populations such as highly educated but financially constrained individuals, or financially stable but digitally disconnected workers. Unlike KMeans, DBSCAN also flagged noise points, data that did not belong to any cluster, which may represent individuals with atypical or volatile socioeconomic

profiles. This capability proved useful in surfacing minority subgroups that are often flattened or misrepresented in fixed-k clustering approaches. GMM clustering provided a probabilistic view, treating the data as generated from a mixture of Gaussian distributions. This approach not only reaffirmed many of the clusters detected by KMeans but also introduced soft cluster assignments, which accounted for overlapping identities. For instance, some individuals appeared to straddle the line between two clusters, such as those with high education but low income, suggesting that rigid boundaries may miss the nuance in real-world segmentation. This application of unsupervised learning echoes work by Ahad et al. (2025), who used clustering to personalize navigation on e-commerce platforms [2]. Just as clustering improves recommendation systems by grouping similar user behaviors, our segmentation enhances policy targeting by uncovering latent socioeconomic profiles across the population."

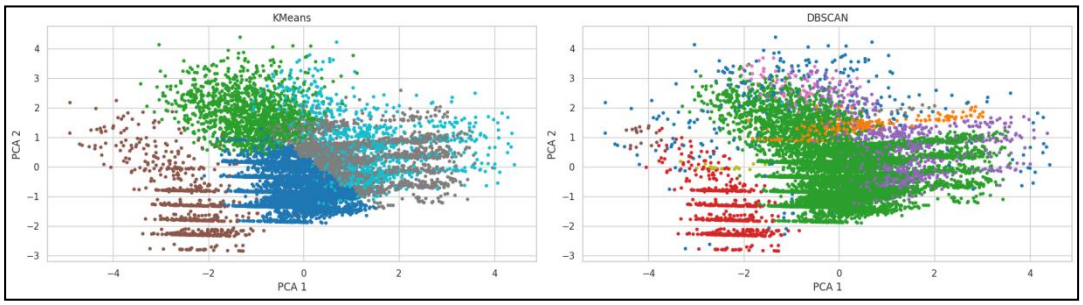

Fig.6: KMeans and DBSCAN clusters

5. INSIGHTS AND REAL-WORLD IMPLICATIONS

5.1 Income Predictors and Model Fidelity

Our results show something easy to overlook when comparing models: different tools pick up on different parts of the story, and in this case, both tree-based ensembles and well-tuned linear models brought something valuable to the table when it came to predicting individual income. Random Forest and XGBoost stood out by capturing higher-order interactions without much hand-holding. They were able to spot patterns like how full-time or remote work can significantly boost earnings for people with higher education levels. These models also handled a mix of data types well, without needing us to manually engineer every interaction. Random Forest, for example, consistently focused on hours worked per week, credit score, and age, finding natural breakpoints in those features and building its logic around them. XGBoost went a step further, using its iterative fitting to refine those splits and bring down the overall error. Neither model needed heavy preprocessing beyond standard scaling and some basic encoding, yet they consistently delivered strong performance on validation, low error rates, and solid R^2 scores across the board.

Linear regression, on the other hand, made its case through smart feature engineering. On its own, it's too limited to handle a dataset this complex. But once we added polynomial interaction terms and used PCA to reduce collinearity, it held its own. It started picking up on layered effects, like how having an advanced degree combined with longer working hours tends to significantly raise income. These second-degree terms helped the model find structure in the data that a basic linear fit would miss, leading to lower MSE and a noticeable bump in R^2. What this all points to is a useful takeaway: tree-based models are great when you want something that can dig into nonlinear patterns with minimal prep, while enhanced linear models offer a clearer window into the underlying relationships. They might not always match the top-end accuracy of ensembles, but they come pretty close, and you get more transparency along the way. Depending on what matters more in a given setting, raw performance or interpretability, either approach can be the right fit.

5.2 Group Inequalities and Policy Gaps

In our analysis, we found that income gaps along racial and educational lines aren't going away. Even after accounting for factors like age, region, and job type, the disparities hold up. When we looked at average incomes by race, Asian and White respondents consistently earned more than their Hispanic and Black counterparts across every level of education. The differences were especially noticeable among people with bachelor's and master's degrees. Everyone sees a bump in income with more education, but that bump is smaller for Hispanic and Black groups. That pattern suggests systemic issues are limiting how much value those credentials deliver, depending on who holds them. Our ANOVA tests on raw income data told a similar story. Some differences between racial groups didn't always pass the strictest tests for statistical significance, but the trends were clear enough to matter, especially when you dig into specific subgroups rather than the overall averages (Hasanuzzaman et al. 2025) [11].

On the education side, the step up in earnings from high school to an associate's degree was significant, and there was another jump from bachelor's to master's. But the gains weren't evenly distributed. When we factored in student loan debt, the picture got more complicated. While higher education tends to lead to higher incomes, the debt burden that comes with it can wipe out short-term financial gains for many people, particularly those from low-income backgrounds or minority

communities. These groups are more likely to carry heavy student debt, which compounds existing wealth gaps and makes upward mobility even harder (Billah et al. 2024) [6]. Even though we worked with a synthetic dataset, it was built to reflect real-world U.S. socioeconomic patterns, and the disparities still showed up. That tells us the problem isn't hypothetical. Broader access to higher education hasn't been enough to close the gap. Things like labor market bias, uneven access to strong academic programs, and differences in how people are able to repay their loans are shaping who benefits most. These findings echo what Autor, Levy, and Murnane (2003) pointed out years ago. They showed that as the economy shifts toward valuing cognitive and non-routine skills, the payoff tends to favor people in specific educational and job paths. The catch is that not everyone has had equal access to build those kinds of skills [4]. That uneven playing field feeds directly into the income gaps we're seeing, especially along racial and educational lines.

Policy responses need to be more pointed. Tuition support that scales with income, targeted loan forgiveness, and focused digital upskilling could help level the field. Our PCA analysis showed a strong link between income and internet access, so it's not hard to imagine the potential of expanding digital infrastructure and training in underserved areas. Regional differences added another layer. While our ANOVA didn't pick up stark income differences by region in isolation, local economies tell a different story when you look closer. Cost of living, job types, and education options vary widely, and that creates pockets of inequality that national stats often miss. For instance, someone with a master's degree living in a high-cost coastal city might earn more on paper but still struggle with loan repayment. Regional scholarships and public-private retraining partnerships could help close those gaps in a more grounded way.

5.3 U.S. Population Segmentation Use-Cases

Clustering uncovered more than just at-risk groups needing social support; it revealed patterns that can be useful well beyond that. Public agencies, private companies, and nonprofits all face the same basic problem: how to make smart decisions about where to focus limited time, money, and effort. These segments offer a way to do that with more clarity. Take workforce development. Instead of throwing generic training programs at broad populations, you can be more surgical. Say there's a cluster of people juggling high student debt but who at least have decent digital access. That group might respond well to online micro-credentials, especially if paired with debt relief incentives. Another cluster might show low formal education but steady employment histories. For them, apprenticeships or on-the-job training might be a better fit, ideally in partnership with industries looking to fill gaps. Schools and career centers could also lean on these profiles to figure out where to put their scholarship funds or which upskilling tracks are likely to see more interest.

In public health, it works much the same way. A segment that shows up as low-income, with patchy job stability and no health coverage, likely points to neighborhoods where clinics, mobile units, or subsidized telehealth would make the biggest difference. And if you overlay digital access on top, outreach gets sharper. One group might need SMS-based education because broadband's spotty. Another might be fine using a web portal or setting up virtual appointments. Matching the message and method to the right group can help avoid no-shows and boost follow-through, things that really matter when you're trying to improve health outcomes at scale. Financial institutions could also put this to use. A cluster with good income but heavy debt might benefit from a loan product that helps consolidate and manage that burden, paired with simple budgeting tools. Another group might be more financially vulnerable, maybe irregular income, no savings. Here, matched savings programs or automatic emergency funds could be more effective. Even practical decisions, like where to open new branches or how to shape mobile banking tools, can be better informed by knowing how different groups handle their money and tech.

City planners and local governments can dig into this as well. If you can see where financial access is low or digital access drops off, you know where to invest in infrastructure. Maybe it's expanding broadband coverage, adding co-working hubs, or improving public transit. Housing programs can get smarter too, allocating assistance not just by income level but by looking at factors like family size, job volatility, and personal debt across clusters. And for nonprofits or funders, segmentation adds another layer of focus. If a certain group shows a strong desire to pursue education but struggles to make progress, that's where afterschool programs, mentorship, or tech hubs might have real impact. Tracking outcomes across those clusters also gives a clearer sense of what's working and what needs adjustment, which helps build accountability over time.

6. FUTURE WORK

6.1 Integration with U.S. Government Microdata

Moving from synthetic inputs to actual US microdata opens up a lot of potential while bringing its own set of hurdles. First, we would bring in rich sources like the American Community Survey or anonymized IRS tax records. Those datasets would need to flow through a preprocessing pipeline designed to honor privacy rules and data agreements. On the engineering side, we would build a system that can scale, ingesting new income, employment, and credit information as it arrives and retraining models on the fly. A cloud native setup using containerized services in a Kubernetes cluster could host our ensemble of models, kicking off retraining whenever the population mix shifts enough to matter. To further improve responsiveness, especially under volatile economic conditions, integrating real-time digital sentiment inputs could offer early signals, as shown by Bhowmik et al. (2025) in their AI-driven forecasting of Bitcoin market volatility." [7].

On the user side, interactive dashboards that blend predictive scores with cluster insights would help policy teams spot hot spots as soon as they emerge. Imagine a state agency seeing in near real time that a county's income levels are dipping sharply and then directing additional workforce grants or subsidy dollars precisely where they are needed. We would pair that with rigorous A/B experiments, comparing regions that use our alerts against control areas, and log everything in a privacy-safe way so we avoid perverse incentives. Finally, we would connect these predictions back into existing case management tools via RESTful APIs, so frontline teams could offer tailored support, whether that is digital literacy training or debt relief guidance, based on where someone sits in our cluster framework and how their income path is expected to evolve.

6.2 Fairness Metrics and Debiasing

Our current pipeline treats each person as an island, yet in real life, people rely on networks of family, friends, coworkers, and community resources. The next step is to map these connections as graphs, where nodes represent individuals or institutions and edges capture relationships such as shared employers, living arrangements, or financial exchanges. Models based on graph neural networks would learn patterns that combine a person's attributes with the influence of their network. That might reveal, for example, how having a mentor in a high-paying role can boost a young professional's earnings, an effect you would not see in purely tabular data. A related frontier lies in distributed ledger applications: Rahman et al. (2025) illustrated how blockchain-based supply chain models trace financial dependencies in decentralized systems, an insight that could be extended to map economic influence across community networks." [19]. On top of that, running community detection on these graphs could yield clusters defined by social or economic ties instead of simple feature similarity. You might uncover a rural community that pools income informally, a pattern invisible to standard clustering techniques. By combining graph-driven segmentation with our current principal component analysis and clustering pipeline, we would gain deeper understanding of how money and support flow through social networks. Those insights could guide place-based policies that strengthen community resilience and economic opportunity.

Yet as we embed these models into policy workflows, we also need rigorous fairness metrics and tools to preempt discrimination, especially in automated decision-making settings. Pope and Sydnor (2011) emphasized that anti-discrimination policies cannot be meaningfully implemented unless statistical profiling systems explicitly account for historical and structural inequalities. Their work makes a strong case for evaluating not just model accuracy but also the social cost of false positives and the long-run effects of algorithmic gatekeeping [18]. In our context, fairness auditing must go beyond parity metrics and consider how different social graphs experience prediction errors. By embedding those considerations directly into the graph-aware models, we can move closer to equity-aware forecasting systems that do more than predict; they course-correct for systemic disadvantage.

6.3 Causal Inference and Uplift Modeling

Socioeconomic factors never sit still. Labor markets shift, policy incentives change, and economic shocks reshape what drives income. A model trained once will slowly lose its edge as the input landscape drifts. To keep our predictions sharp, we need a continual learning setup that updates model parameters as new data arrives while holding on to knowledge from past patterns. Methods like elastic weight consolidation help preserve stable relationships even as the model incorporates fresh information. We would also build drift detection into the pipeline, watching for rising error rates or shifts in feature distributions, and use those signals to decide when to retrain or refresh parts of the ensemble. Active learning can step in as well, flagging cases where the model is uncertain so that experts can label them and improve calibration. Embedding these adaptive strategies into production means the system stays in step with emerging trends, from the growth of gig work to regional cost of living changes, without needing big manual retraining efforts.

Prediction can tell you *what* might happen, but it doesn't answer the more important question: *why* did someone's income go up or down? That's where causal inference comes in. Tools like uplift modeling and counterfactual estimation are designed to dig into that "why," especially in situations where the effects of an intervention, like job training or debt relief, vary across different groups. Uplift modeling, for example, is useful when you're dealing with heterogeneous treatment effects. It helps figure out who benefits from a policy and who doesn't, instead of averaging outcomes across everyone. But the problem is that many of the models that get used here are black boxes. They might predict well, but they don't explain much. Zhao and Hastie (2021) tackled this issue head-on. They proposed ways to interpret these opaque models through a causal lens, which helps separate real cause-and-effect relationships from patterns that just happen to look predictive [25]. That distinction matters a lot when the output of your model is being used to guide real-world decisions. In our work, bringing causal interpretation into the modeling process could make a big difference. It would help clarify not only who gains from an intervention, but also why they gain, and that's the kind of insight policymakers need if they're serious about designing fair and effective economic programs. It also helps avoid the trap of mistaking correlation for meaning, which can quietly derail well-intended efforts.

6.4 Real-Time or Interactive Deployment

Turning our study's findings into tools people can use means building a live, interactive dashboard. With lightweight web frameworks like Streamlit or Dash, you can tuck those trained ensemble models and clustering routines into an interface that feels intuitive. Picture a screen where policymakers, social workers, or researchers choose inputs, age, education, industry,

debt, and even digital access, and instantly see predicted income with confidence bands, plus a color tag showing which socioeconomic cluster a profile belongs to. In building interactive dashboards that serve real-time predictions, insights from distributed system performance optimization are valuable. Billah et al. (2024) demonstrated how benchmarking in multi-machine blockchain systems can guide the design of scalable, low-latency infrastructures suitable for AI-powered public dashboards." [6]. But a dashboard shouldn't stop at a single prediction. Pull in monthly time-series data so viewers can watch income forecasts and group disparities shift over time, and catch early warning signs when any cluster starts sliding. Add interactive charts, a boxplot next to a violin plot, maybe a heatmap, that update on the fly when someone filters by gender, region, or race. Throw in a PCA scatter plot with toggles for K-means or DBSCAN clusters, and clicking on a dot reveals that group's average income, typical debt levels, and digital access rate.

To make this work in places where resources are tight, wrap the whole thing in Docker and run it on simple cloud services or local servers. Expose model predictions and cluster assignments through REST APIs so the front end only handles display, and update models without touching the interface. Layer in sign-in screens and role-based permissions so sensitive data stays behind the scenes for only the right eyes. Finally, add sliders for scenario analysis, adjust the "education level" or dial down "student loan debt" to see how the income forecast and cluster tags shift in real time. That kind of immediate feedback turns a machine learning report into a living lab, where decision-makers can tinker with policy ideas and see their potential impact without waiting weeks for results. It's a way to hand over real power, so people designing economic programs can experiment, learn, and move toward fairer outcomes with confidence.

7. CONCLUSION

This study built a four-stage machine learning framework to explore income prediction and uncover patterns of disparity among U.S. citizens. The goal was to dig deeper than surface-level stats and get a clearer view of how income varies across different segments of the population, and why. From a modeling standpoint, tree-based ensemble methods like XGBoost and Random Forest performed especially well. They picked up on the complex, nonlinear patterns that tend to exist in socioeconomic data, showing strong R^2 scores and low error. That said, a well-tuned linear regression model, with added polynomial and interaction terms, held its own. With the right feature engineering, even simpler models can still offer real value. But this wasn't only about predicting income. The framework also helped reveal where disparities persist. Our analysis pointed to enduring income gaps tied to race and education, even after accounting for other factors. That suggests deeper structural issues, things like systemic bias and financial strain from student debt, that aren't going to disappear with education alone. We also ran unsupervised clustering using methods like KMeans, DBSCAN, and Gaussian Mixture Models on PCA-reduced data. The aim here was to identify distinct groups within the U.S. population, based on combinations of factors like financial stability, educational background, and digital access. These clusters gave us a more grounded view of who's vulnerable, and why, not just in theory, but in ways that connect to lived realities. The implications here are practical. Knowing which factors matter most for income, and how different groups are affected, can guide smarter policies. This could mean more targeted job training, financial literacy programs, or better strategies for closing the digital divide. Instead of blanket solutions, you get interventions shaped around real needs. Looking ahead, future work will focus on bringing in live data from U.S. government sources, refining fairness metrics, and building in debiasing techniques to catch algorithmic bias early. We also plan to explore causal inference and uplift modeling to get a clearer sense of what interventions change outcomes. On the application side, building interactive tools for policymakers could turn these models into working systems, ones that let decision-makers test ideas, see likely outcomes, and better understand how different factors influence economic mobility. The broader aim is to make this more than a research exercise. It's about creating tools that help people make better, fairer decisions in the real world.

REFERENCES

[1] Abed, J., Hasnain, K. N., Sultana, K. S., Begum, M., Shatyi, S. S., Billah, M., & Sadnan, G. A. (2024). Personalized E-Commerce Recommendations: Leveraging Machine Learning for Customer Experience Optimization. Journal of Economics, Finance and Accounting Studies, 6(4), 90–112.

[2] Ahad, M. A., Mohaimin, M. R., Rabbi, M. N. S., Abed, J., Shatyi, S. S., Sadnan, G. A., … & Ahmed, M. W. (2025). AI-Based Product Clustering For E-Commerce Platforms: Enhancing Navigation And User Personalization. International Journal of Environmental Sciences, 156–171.

[3] Athey, S., & Imbens, G. W. (2019). Machine learning methods for estimating heterogeneous causal effects. Journal of Economic Perspectives, 33(2), 27–50.

[4] Autor, D., Levy, F., & Murnane, R. (2003). The skill content of recent technological change: An empirical exploration. Quarterly Journal of Economics, 118(4), 1279–1333.

[5] Bell, A., Chetty, R., Jaravel, X., Petkova, N., & Van Reenen, J. (2019). Who becomes an inventor in America? The importance of exposure to innovation. Quarterly Journal of Economics, 134(2), 647–713.

[6] Billah, M., Shatyi, S. S., Sadnan, G. A., Hasnain, K. N., Abed, J., Begum, M., & Sultana, K. S. (2024). Performance Optimization in Multi-Machine Blockchain Systems: A Comprehensive Benchmarking Analysis. Journal of Business and Management Studies, 6(6), 357–375.

[7] Bhowmik, P. K., Chowdhury, F. R., Sumsuzzaman, M., Ray, R. K., Khan, M. M., Gomes, C. A. H., … & Gomes, C. A. (2025). AI-Driven Sentiment Analysis for Bitcoin Market Trends: A Predictive Approach to Crypto Volatility. Journal of Ecohumanism, 4(4), 266–288.

[8] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163.

[9] Fariha, N., Khan, M. N. M., Hossain, M. I., Reza, S. A., Bortty, J. C., Sultana, K. S., … & Begum, M. (2025). Advanced fraud detection using machine learning models: enhancing financial transaction security. arXiv preprint arXiv:2506.10842.

[10] Hasan, M. S., Siam, M. A., Ahad, M. A., Hossain, M. N., Ridoy, M. H., Rabbi, M. N. S., … & Jakir, T. (2024). Predictive Analytics for Customer Retention: Machine Learning Models to Analyze and Mitigate Churn in E-Commerce Platforms. Journal of Business and Management Studies, 6(4), 304–320.

[11] Hasanuzzaman, M., Hossain, M., Rahman, M. M., Rabbi, M. M. K., Khan, M. M., Zeeshan, M. A. F., … & Kawsar, M. (2025). Understanding Social Media Behavior in the USA: AI-Driven Insights for Predicting Digital Trends and User Engagement. Journal of Ecohumanism, 4(4), 119–141.

[12] Hossain, M. I., Khan, M. N. M., Fariha, N., Tasnia, R., Sarker, B., Doha, M. Z., … & Siam, M. A. (2025). Assessing Urban-Rural Income Disparities in the USA: A Data-Driven Approach Using Predictive Analytics. Journal of Ecohumanism, 4(4), 300–320.

[13] Islam, M. R., Hossain, M., Alam, M., Khan, M. M., Rabbi, M. M. K., Rabby, M. F., … & Tarafder, M. T. R. (2025). Leveraging Machine Learning for Insights and Predictions in Synthetic E-Commerce Data in the USA: A Comprehensive Analysis. Journal of Ecohumanism, 4(2), 2394–2420.

[14] Jakir, T., et al. (2023). Machine Learning-Powered Financial Fraud Detection: Building Robust Predictive Models for Transactional Security. Journal of Economics, Finance and Accounting Studies, 5(5), 161–180.

[15] Khan, M. N. M., Fariha, N., Hossain, M. I., Debnath, S., Al Helal, M. A., Basu, U., … & Gurung, N. (2025). Assessing the Impact of ESG Factors on Financial Performance Using an AI-Enabled Predictive Model. International Journal of Environmental Sciences, 1792–1811.

[16] Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. Journal of Economic Perspectives, 31(2), 87–106.

[17] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453.

[18] Pope, D. G., & Sydnor, J. R. (2011). Implementing anti-discrimination policies in statistical profiling models. American Economic Journal: Economic Policy, 3(3), 206–231.

[19] Rahman, M. S., Hossain, M. S., Rahman, M. K., Islam, M. R., Sumon, M. F. I., Siam, M. A., & Debnath, P. (2025). Enhancing Supply Chain Transparency with Blockchain: A Data-Driven Analysis of Distributed Ledger Applications. Journal of Business and Management Studies, 7(3), 59–77.

[20] Rana, M. S., Chouksey, A., Hossain, S., Sumsuzoha, M., Bhowmik, P. K., Hossain, M., … & Zeeshan, M. A. F. (2025). AI-Driven Predictive Modeling for Banking Customer Churn: Insights for the US Financial Sector. Journal of Ecohumanism, 4(1), 3478–3497.

[21] Ray, R. K., Sumsuzoha, M., Faisal, M. H., Chowdhury, S. S., Rahman, Z., Hossain, E., … & Rahman, M. S. (2025). Harnessing Machine Learning and AI to Analyze the Impact of Digital Finance on Urban Economic Resilience in the USA. Journal of Ecohumanism, 4(2), 1417–1442.

[22] Sizan, M. M. H., et al. (2025). Bankruptcy Prediction for US Businesses: Leveraging Machine Learning for Financial Stability. Journal of Business and Management Studies, 7(1), 01–14.

[23] Sizan, M. M. H., et al. (2025). Advanced Machine Learning Approaches for Credit Card Fraud Detection in the USA: A Comprehensive Analysis. Journal of Ecohumanism, 4(2), 883–905.

[24] Sumon, M. F. I., Osiujjaman, M., Khan, M. A., Rahman, A., Uddin, M. K., Pant, L., & Debnath, P. (2024). Environmental and Socio-Economic Impact Assessment of Renewable Energy Using Machine Learning Models. Journal of Economics, Finance and Accounting Studies, 6(5), 112–122.

[25] Zhao, Q., & Hastie, T. (2021). Causal interpretations of black-box models. Journal of Business & Economic Statistics, 39(1), 272–281.

fffff