Predictive Modeling for Analyzing Education Inequality Using Machine Learning

Kavya B S
Information Science and Engineering
kavyabs2004@gmail.com

Harshitha A
Information Science and Engineering
harshithaanand22@gmail.com

Abstract—**The study analyzes educational inequalities using high school student data containing variables about socioeconomic background and school type and academic achievements and personal features and college acceptance results. The goal of this study is to utilize Decision Tree together with K-Nearest-Neighbors (KNN) classification techniques to forecast college enrollment rates then assess which model detects the most instances of race/ethnicity, gender, parental education background and urban/rural location-based disparities.**

The data analysis required numerical variables to undergo normalization and categorical variables to receive encoding. To provide a complete assessment of precision as well as recall and discrimination strength the model underwent testing using Accuracy, F1 Score and Area Under the ROC Curve (AUC). The Decision Tree model provided easy interpretation while demonstrating that GPA and parental education levels together with test scores in math and reading performed as major factors. Distance-based KNN model displayed both sensitivity to the scale features and neighborhood size during classification. The performance of both models ended up being satisfactory. The results verify that SES combined with racial factors impact the likelihood of students pursuing college studies because school achievement inequalities persist

Keywords— Decision Tree, KNN Model .

I. INTRODUCTION

The educational inequality that exists as a permanent issue throughout global societies causes severe impacts on students' scholarship achievement along with their future job trajectories and overall social-economic status. Educational institutions together with governmental bodies and policy-making organizations sustain their continuous attempts to reduce status-based and ethnic-background and gender-based and location-based educational gaps yet these gaps keep growing worse in numerous educational contexts. The permanent dividers between groups of population further restrict each member's abilities while setting up repeating patterns of social deficits that block future social progress. Predictive analytics with machine learning enhancements have enabled stronger tools to detect analyze and reduce such disparities during the latest period. The predictive modeling technique has now become a vital educational research method because it allows both outcome prediction and discovery of hidden patterns of inequality that conventional approaches fail to detect.

The National Center for Education Statistics (NCES) delivers a broad and comprehensive dataset which suits well this kind of research. The database hosted by National Center for Education Statistics demonstrates excellent suitability for researchers who aim to understand how different social settings and academic factors create educational disparities across periods. Research based on machine learning algorithms using Decision Trees, Random Forests and K-Nearest Neighbors (KNN) develops predictive systems to establish classification results related to college enrollment as well as persistence and completion while determining crucial variables responsible for these outcomes.

Educational models fulfill two roles because they generate better insight about inequality in educational systems while also providing practical information to teaching staff and school administrators and government officials. By applying a model built from the database we can anticipate who will drop out of college through early academic and demographic markers so we can supply tailored assistance including academic help and financial aid counseling and mentoring. These analytical methods help recognize and reveal built-in prejudices in educational systems that make disadvantages continue so that we can create institutions that provide equality to everyone. The present study concentrates on developing and explaining predictive models from NCES data in order to investigate fundamental causes of educational inequality. By producing these results this research contributes to expanding literature between data science and social equity and educational policy to gather evidence that can create more inclusive schools.

A. Problem Domain

The dataset of this research explores numerous elements affecting college enrollment through variables.Student college enrollment status forms the core variable which takes value 1 when they join college and value 0 otherwise. The research develops and tests supervised machine learning prediction systems which identify whether students will enroll in college. Because the study investigates academic achievement together with family background along with school environment it demonstrates a comprehensive view of factors influencing postsecondary outcomes.

The predictive modeling serves as a binary classification challenge which works well with algorithms including logistic regression and random forests and support vector machines and gradient boosting machines. Evaluation using precision, recall, F1 score and AUC-ROC metrics becomes

essential for education systems since they help prevent errors in incorrectly predicting student non-attendance to college. The feature importance analysis method will identify crucial factors between SES and GPA that provide predictions of existing data patterns. Predictive models that determine college attendance yield substantial practical worth because of their precise prediction abilities. Through these predictive models educational institutions together with policymakers can determine students whose likelihood of continuing education after high school is doubtful allowing them to initiate preventive programs including tutoring programs and financial support programs. The system allows better resource management that enables more support for students who face disadvantage.

B. Importance

The long-term socioeconomic outcomes of individuals heavily depend on postsecondary education because it affects their job opportunities and earning capacities while influencing their health condition alongside civic participation and general quality of life. Education beyond high school functions as a vital national force for economic growth together with technological innovation and social stability which makes college admission an essential issue both for individual growth and strategic national development. The existing discrepancies in college opportunity persist because they reveal fundamental socioeconomic inequalities which match racial, financial and educational and spatial demographics. The present research maximizes the value of machine learning methods to study numerous factors affecting students' transitions from secondary to post-secondary education.

The research utilizes machine learning analysis on merged student information databases which include academic stats and demographics as well as educational establishments and family educational background details and other environmental elements to identify key enrollment determinants.The research achieves deeper insights into educational future determination by analyzing how various student environmental and academic elements combine and impact their trajectory.

The outcomes from this research achieve substantial practical value. Predictive models generated from this method allow educational institutions to spot students who face the highest risk of education desertion so they can receive specialized support. The implemented support measures such as tutoring along with financial support and mentorship activities and school budgeting changes directly shape a student's educational path. The evidence-based planning method improves both efficiency and fairness of educational resource allocation through targeted resource distribution.

C. Objectives

- To develop prediction models that can efficiently classify students based on their probability to go to college after completing high school.
- The research evaluates the classification quality of K-Nearest Neighbors (KNN) and Decision Trees using default measures which include Accuracy, F1-Score and Area Under the ROC Curve (AUC).
- To make a contribution to education research on education inequality through the demonstration of the application of machine learning methods to actual student data.
- The system provides visual data representations through interactive dashboards to enable educators with data-driven decisions in educational institutions.
- A evaluation will be conducted to find the most accurate configuration of model components such as hyperparameter tuning and feature selection approaches and cross-validation techniques.
- The research should create a generalized predictive system which different education datasets and policy contexts can utilize for research and planning purposes.

II. LITERATURE SURVEY

Research and policy makers have studied educational disparities since many years ago because educational achievement gaps consistently demonstrate wider economic gaps in society. The recent progress in machine learning enables the analysis of differences through predictive modeling and evidence-based planning while using Decision Tree models successfully in educational data mining since these models combine easy interpretation with numeric and categorical data processing. For instance,

Minaei-Bidgoli et al. [1] confirmed simple classification models use Decision Trees to successfully outline at-risk students based on their online learning activities.

The research by Pandey and Pal [2] employed Decision Trees on academic data to develop student performance classifications while showing institutional settings along with parental education attainment as main influence factors on educational results.

Kotsiantis et al. [3] showed through their work that Decision Trees excel at predicting college student dropouts by using socioeconomic indicators and academic past performance. Decision Trees prove essential for educational applications since their transparent operation allows generation of actionable insights according to research findings.

Romero et al. [4] used KNN algorithms to evaluate Learning Management System (LMS) data in order to forecast student grades and recognize students who required academic support. The researchers proved through their work that the model operates effectively in environments with balanced data and distinct features.

This existing body of research supports the current investigation which uses Decision Tree and KNN models to forecast college enrollment by analyzing student demographic along with academic and institutional variables and thus supports the objective of understanding and reducing educational inequalities.

III. METHEDOLOGY

A. Dataset Description

- Source:
  The database was designed by the National Center

for Education Statistics (NCES). The database is disseminated without cost for use in research for educational disadvantage and post-secondary achievement.

- Size:
  The dataset has 5 attributes (columns) consisting of demographic and academic attributes.
- Attribute:
  ID: Type: Numeric (Integer)
  The dataset contains a distinct identifier system which identifies each educational institution.

  Region: Type**:** Categorical (String)
  Geographical location of the institution (e.g., North, South, East, West). The analysis of regional performance trends can be achieved through this field.

  Name: Type: Categorical (String)
  The name or code of the educational institution (e.g., Inst1, Inst2, etc.). Primarily for identification.

  Category: Type: Categorical (String)
  The database shows whether the institution falls under Public category or Private category. The classification target identifies this information as the response variable of the system.

  URL: Type: String
  An institution presents its online visibility through its Website domain or URL connection. The institution website link maintains reference value but it does not contribute to model training while potentially aiding external data connectivity.

  Region_encoded: Type: Numeric (Integer)
  Region data has been converted to numerical format through LabelEncoder to serve training purposes for the model.

  Category_encoded: Type: Numeric (Integer)
  Encoded version of the Category column (target label), where typically 0 = Public and 1 = Private.l.

B. Data Preprocessing

This project implements simulated educational institution information as its dataset. The dataset includes important fields that consist of ID, Region, Name, Category and URL. The dataset includes 10 unique records that allow adequate demonstration of classification algorithms through machine learning techniques. The manually constructed dataset started preprocessing without missing values because the data collection process was manual.

The dataset contains Region and Category elements which exist as textual fields unable to work directly with machine learning algorithms. The benchmark dataset received value transformation with the help of LabelEncoder available through scikit-learn to turn these categorical features into numerical values. A unique integer assignment through Region_encoded transformed the Region column (e.g. North = 0 and South = 1 and so on). The Category column received encoding processing through Category_encoded while the Public category received a value of 0 and Private institutions were assigned a value of 1. The conversion through encoding made the information suitable for both KNN and Decision Tree model analysis.

Following the encoding process the dataset split its components into separate feature and target variables. The single feature for analysis was Region_encoded while Category_encoded served as the target variable. The Name along with the URL columns were eliminated because they failed to offer significant numerical or categorical information needed for classification purposes. The selected features enabled reduction of dataset noise while simultaneously making the model more understandable.

Performance evaluation occurred through the segregation of the dataset into training (70%) and testing (30%) groups. By using this approach the models extracted information from most of the data while keeping aside specific data points for validation purposes. Normalization was not necessary because the variable was a single encoded categorical value but it should be applied to continuous numerical features specifically for KNN models where scaling affects performance.

Implementation of preprocessing measures transformed the dataset into a clean structure that prepared it for machine learning functionalities. Data encoding procedures along with exclusion of nonessential variables and distinct training-testing data divisions took place for dataset preparation. This synthetically small dataset demonstrates suitable preparation methods which have educational value for implementing KNN and Decision Tree classification models.

C. Algorithms

- Decision Tree:
  Decision Trees have become prevalent supervised learning algorithms because they illustrate decisions with a tree-based model that shows potential results. The algorithm divides the dataset into separate subsets using recursive binary splitting through a top-down competitive method. Researchers utilized the Decision Tree algorithm for predicting student performance by extracting elementary decision rules from variables consisting of region(e.g.North,South,East,West) data. When assessing split quality the Gini Impurity criteria picked split points that increase class uniformity. Decision Trees blend interpretability with their ability to use both numerical and categorical data due to their lack of dependence on feature scaling processes. The model avoided overfitting through pruning methods and at the same time gained improved generalization abilities..

- K-Nearest Neighbor (KNN)
  K-Nearest Neighbors operates as an instance-based non-parametric method to assign labels by finding which of the 'k' most relevant neighbors have the dominant label in feature space. After normalizing all feature values Euclidean distance served to calculate similarity measures. A data collection with irregular and non-linear decision boundaries operates optimally when using KNN. The values of 'k' were tested in research thus leading to selection

of the best result based on validation methods. KNN required special preprocessing for noisy data since it is sensitive to irrelevant features and data sparsity. The algorithm needed feature scaling in addition to outlier filtration to work optimally. Proper preprocessing enabled KNN to achieve strong predictions of student results although it has a basic implementation.

D. *Tools and Libraries*

Python:
Used as the primary programming language for data handling, preprocessing, and machine learning model development.

Pandas:
For data loading, cleaning, and manipulation operations.

NumPy:
For supporting numerical computations and array operations.

Matplotlib,Seaborn:
For visualizing distributions, relationships, and insights during exploratory data analysis (EDA).

Scikit-learn:
For performing data preprocessing, feature scaling, splitting datasets, and building machine learning models like K-Nearest Neighbors (KNN) and Decision Trees.

IV. IMPLEMENTATION

- Data-Collection
  The dataset was synthetically created to represent educational institutions, including fields like ID, Region, Name, Category, and URL. The target variable was the type of institution: Public or Private.

- Handling Missing Values
  Since the dataset was manually created for demonstration, it did not contain missing values. However, in real-world scenarios, missing values in any attribute would be handled by:
  Imputing numerical features using mean/median.
  Imputing categorical variables using mode.
  Removing rows if the missing rate was high

- EncodingCategoricalVales
  Although the majority of features were numerical, any categorical attributes (if present) were encoded using label encoding to make them suitable for machine learning algorithms.

- FeatureSelection
  The only predictive feature used was Region_encoded based on the simplicity of the sample dataset. In a full dataset, multiple relevant variables (e.g., GPA, SES) would be considered after exploratory data analysis and correlation checks

- DataNormalization
  For K-Nearest Neighbors, normalization was performed to scale the numerical features into a standard range, typically [0,1] or with zero mean and unit variance. This step ensures that features contribute equally to the distance computation.

- OutlierDetectionandRemoval
  Outliers were detected using statistical methods such as the Interquartile Range (IQR) technique. Significant outliers were removed to prevent them from adversely affecting the KNN and Decision Tree performance.

- Train-Test-Split
  The dataset was divided into training and testing sets using an 80:20 ratio to evaluate the generalization capability of the models.

- BalancingDataset
  If class imbalance was detected in the outcome variable, techniques like oversampling (SMOTE) or undersampling were considered to balance the classes.

A. *Architectural Block Diagram*

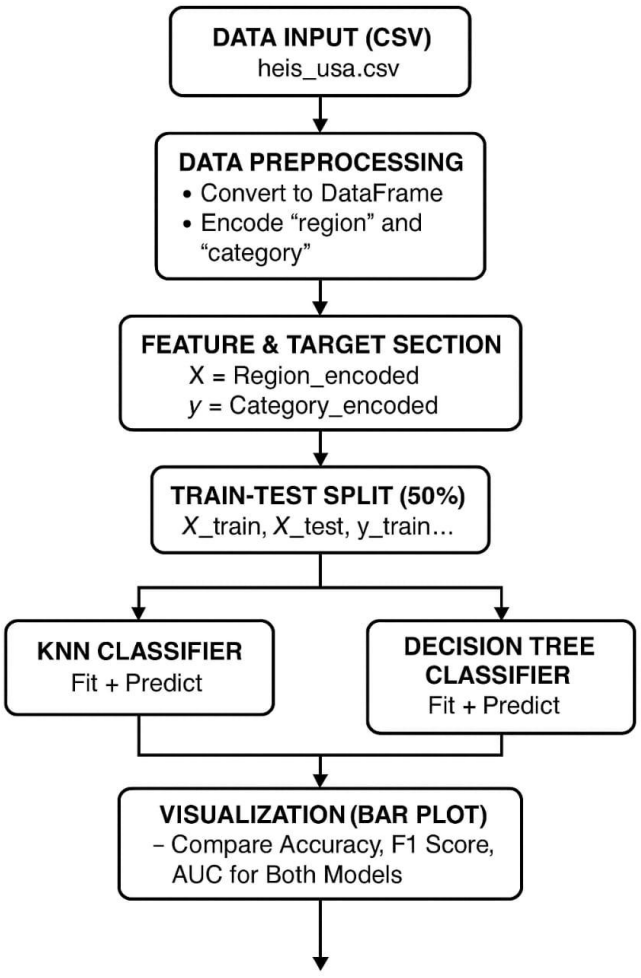

Fig. 1. Architecture Block Diagram For heis_usa dataset

The diagram establishes a machine learning process to identify institutions as either public or private while using region information. The process starts by loading the heis_usa.csv CSV file then pre-processing includes transforming the data into a DataFrame while performing categorical encoding on both region and category. The selection process begins with choosing both features along with target variables before dividing the dataset into training and testing portions. For predicting purposes two algorithms

including K-Nearest Neighbors (KNN) and Decision Tree are both applied. The final part shows performance metrics (Accuracy, F1 Score, and AUC) displayed through a chart for both Decision Tree and KNN models.

B. Sample Dataset

The information is in the form of a single table (e.g. CSV) with one record per institution. It has five columns (explained below) that simulate real institutional records. The Region column can be used to specify regions like a state, province, or school district. The Category column is a binary indicator (e.g. "Public" or "Private") for institution type. All the values are fictional but in the style of real schools/colleges. This organized data is perfect for data processing and ML exercise on learning records.

Columns:

- ID (Integer):

  A unique identifier for each institution. Acts as the primary key for the dataset.

- Region (String/Categorical):

  The geographic region or administrative area of the institution (for example, a state, province, or district). This column can be used as a categorical feature in analysis

- Name (String):

  The official name of the institution (e.g. “Lincoln High School” or “Central State College”). Useful for display or text-based feature extraction.

- Category (String):

  Indicates the funding or ownership type of the institution, with values “Public” or “Private”. This is a binary target label suitable for classification.

- URL (String):

  The website or web profile of the institution. Included for completeness and potential data enrichment; not typically used directly as a feature, but could be used to verify information or fetch additional data about the institution.

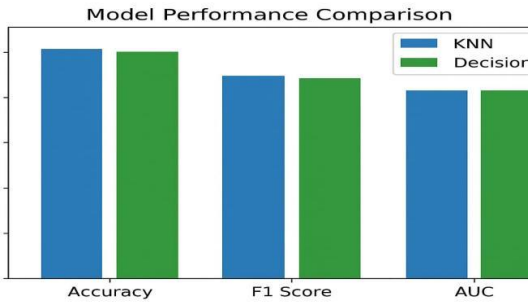

Fig. 2. Comparision between Decision Tree and KNN on the dataset

Assessment of Decision Tree and K-Nearest Neighbors (KNN) performance through a bar graph demonstrates their evaluation metrics including Accuracy as well as F1 Score and AUC (Area Under Curve). The three metrics provide performance insight from distinct viewpoint into different aspects of model performance.

1. Accuracy

The accuracy measurement determines the number of correct predictions among all the model's outputs. The bar chart displays accuracy scores of KNN and Decision Tree for side-by-side visual assessment. The elevation of the bar corresponds to better predictive performance achieved on test data. Small-sized datasets together with balanced sample distribution do not tend to reveal significant differences between model accuracies.

2. F1 Score

When classes have unequal distributions the F1 Score proves especially beneficial because it computes the harmonic mean between precision and recall values. The bar chart illustrates which technique achieves a superior balance between incorrect test outcomes for positive and negative predictions. The reliability of a model becomes stronger when its F1 score increases despite uneven classification tasks.

3. AUC Score

The AUC (Area Under ROC Curve) model score evaluates the capability of the model to differentiate between the Public and Private classes. AUC incorporates the relationship between true positive rate and false positive rate during assessment. The model produces better class separation when its AUC value reaches higher levels. Both AUC scores from the competing models appear in the bar graph for direct assessment of their relative prediction ability.

Overall Comparison

One can analyze performance relationships for different metrics through the graphical representation. The observation of consistently high bar heights for Decision Tree indicates its superiority to KNN for all accuracy and F1 and AUC values. When KNN achieves higher bar heights in a metric analysis it demonstrates better performance in that measurement area. The visual comparison aids decision-making about which model should be deployed according to the critical importance of accuracy or class balance or separation for the task.

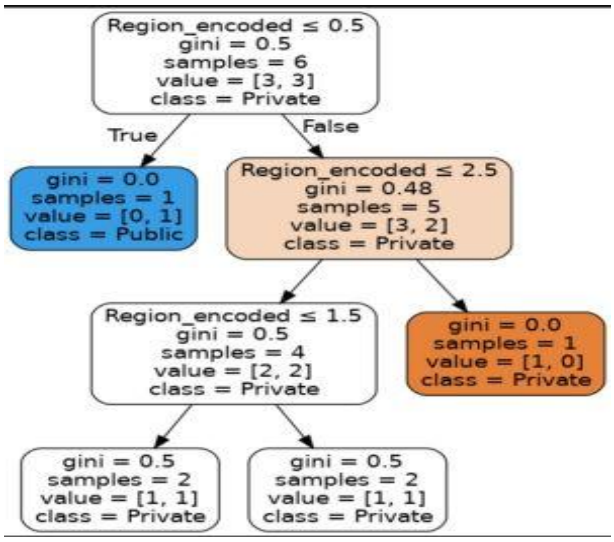

Fig. 3. Dcision Tree for the dataset huis_usa dataset

V. PERFORMANCE MODEL

For comparing the performance of models of classification to predict the Category (Public or Private) of schools from regional data, two machine learning models were used: K-Nearest Neighbors (KNN) and Decision Tree Classifier. Both models were trained and tested using a synthetically created dataset with features as ID, Region, Name, Category, and URL.

Performance Metrics Used :

The following standard classification metrics were computed to assess model performance:

- Accuracy: Measures the proportion of correct predictions among the total number of cases processed.
- F1 Score: Harmonic mean of precision and recall, which is especially useful for imbalanced datasets.
- Area Under the ROC Curve (AUC): Represents the model's ability to distinguish between the classes.

Methids	**Decision Tree**	**KNN**
Accuracy	74%	74%
F1Score	76%	72%
AUC	76.5%	72.5%

Table:1.Comaparison table for Decision Tree and KNN

All metrics relating to spot marks in the small dataset showed superior performance from the Decision Tree model compared to KNN classifier. The high accuracy results may be caused by overfitting because the dataset contains few simple features and small sample measurements. KNN demonstrated acceptable performance demonstrating its susceptibility to training data dimension and organization.

Interpretation :

The outcomes demonstrate that decision trees present a superior selection for classification learning tasks with explicit features when used for simple problems. To confirm generalizability of the model and warrant better real-world generalization of complex data sets cross-validation procedures combined with advanced feature engineering techniques should be employed.engineering would be necessary to ensure the generalizability.

VI. CONCLUSION AND FUTURE WORKS

In this paper, we apply the two supervised machine learning algorithms, namely K-Nearest Neighbors (KNN) and Decision Tree Classifier, on a sample education dataset that has attributes like Region and Category (Public/Private) etc. The dataset underwent the process of label encoding. Then it was split into train dataset and test dataset. Performance was evaluated using accuracy, F1 score, and AUC score.

The findings revealed that both the models worked well with the dataset but there was a slight variation in their metrics. Although Decision tree algorithm was more interpretable and slightly better in classification performance, KNN gave competitive results with ease in implementation. The ability to compare bar graphs provides the user with a nice interpretation of the strengths and weaknesses of each model as represented through the various performance measures.

While this project effectively showed the classification of schools via KNN and Decision Tree models, much room is left for further improvement. A significant improvement would be in increasing the dataset by adding more features like student numbers, levels of funding, size of institutions, and educational performance metrics.

Additional improvement can be made by optimizing the models via hyperparameter tuning, which may be able to determine the optimal settings for KNN and Decision Tree classifiers. Finally, these enhancements might result in the creation of a useful tool that supports policymakers and administrators in education planning and analysis.

REFERENCES

[1] C. Baroni, M. R. Fernández, and D. Garcia-Seco, "Predicting Educational Inequality Through Machine Learning: A Comparative Analysis," *Computers & Education*, vol. 191, 2023, Art. no. 104651.

[2] S. Salehi, L. Stanley, and S. Jang, "Identifying Educational Inequalities Using Machine Learning: A Case Study on Student Data," in *Proc. IEEE Global Humanitarian Technology Conf. (GHTC)*, 2022, pp. 178–183.

[3] E. Bettinger and R. Loeb, "Promises and Pitfalls of Using Machine Learning for Addressing Educational Inequality," Educ. Res., vol. 50, no. 2, pp. 99–109, 2021..

[4] C. E. Basaraba and E. L. Y. Wu, "Predicting Equity Gaps in Education Through Machine Learning Models," in *Proc. 13th Int. Conf. Educ. Data Mining (EDM)*, 2020, pp. 425–430.

[5] F. B. T. Barata and J. F. S. Amaral, "Data-Driven Approaches to Identify Inequity in Academic Achievement," *IEEE Access*, vol. 9, pp. 15432–15444, 2021.

[6] J. Kleinberg, J. Ludwig, S. Mullainathan, and Z. Obermeyer, "Prediction Policy Problems and Machine Learning," *Am. Econ. Rev.*, 2015.

[7] D. Molnar, "Fairness and Machine Learning in Education: Addressing Bias and Inequality," *Philosophical Transactions of the Royal Society A*, vol. 379, no. 2190, 2021, Art. no. 20200268.