Synopsis
On
Minor Project

(AIDS451)

EduDRISHTI

“Data Research for Inequality in Student Test Insights”

BACHELOR OF TECHNOLOGY
(ARTIFICIAL INTELLIGENCE AND DATA SCIENCE)

Submitted By:

Name: Anushka Sharma

Roll No.: 20115611922

Sem: 7th **Sec:** F-11

Under the Supervision of:

Name: Prof. (Dr.) Archana Kumar

Designation: Professor

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

Dr. AKHILESH DAS GUPTA INSTITUTE OF PROFESSIONAL STUDIES

(FORMERLY Dr. AKHILESH DAS GUPTA INSTITUTE OF TECHNOLOGY & MANAGEMENT)

(AFFILIATED TO GURU GOBIND SINGH INDRAPRASTHA UNIVERSITY, DELHI)

Shastri Park, Delhi-110053

ODD SESSION, 2025-26

TABLE OF CONTENTS

INTRODUCTION

India's education system, despite significant advancements in access and enrollment, continues to face deep-seated disparities in learning outcomes. High-stakes national entrance examinations like NEET-UG act as critical gateways to professional careers, making them highly sensitive to administrative fairness and equitable performance distribution. The **2024 NEET-UG controversy**, marked by widespread reports of irregularities, alleged question paper leaks in regions like Patna and Godhra, and an unusual concentration of top scores, highlighted a critical systemic vulnerability that demands rigorous investigation. The political and social fallout from this event underscored that relying on simple aggregate statistics is insufficient; a deeper, feature-rich analysis is required to restore public trust.

This project, **EduDRISHTI** (standing for **Data Research for Inequality in Student Test Insights**), will move decisively beyond anecdotal evidence. I will leverage granular student score data for over 2 million candidates, alongside center-wise summary statistics, to systematically investigate two core problems: **regional and center-specific performance inequality** and the **statistical presence of performance anomalies**. I will utilize advanced feature engineering and unsupervised machine learning—specifically the Isolation Forest model—to transform raw scores and distribution patterns into objective, evidence-based insights. The final deliverable will be a transparent and interactive dashboard that clearly identifies the centers exhibiting the highest degrees of statistical irregularity, aiming to enhance accountability and promote fairness in the national assessment system. This analysis will provide policymakers with the necessary evidence to target investigations and allocate resources to truly disadvantaged regions.

LITERATURE REVIEW

The foundation of this project rests on two pillars: the **analysis of educational inequality** and **statistical anomaly detection**.

1. **Educational Inequality and Achievement Gaps**

   Classical studies on educational inequality (e.g., **Coleman Report, 1966**; World Bank analyses using the Gini coefficient) often employ macro-metrics to assess achievement gaps across socio-economic or geographic boundaries. **News reports on the 2024 NEET-UG controversy** in publications like **The Hindu** and **Indian Expres**s provided the immediate context, highlighting an unusual concentration of high scorers and irregularities in specific centers. My work will expand on this by utilizing the Center-v-National Gap to provide a highly localized measure of performance deviation, moving beyond state-level averages to quantify the exact micro-inequality at the exam center level. This approach aligns with research advocating for disaggregated data analysis to reveal hidden pockets of disparity (**Mukherjee et al., 2023**).

2. **Statistical Anomaly Detection in High-Stakes Testing**

   Regarding anomaly detection in high-stakes testing, traditional methods often rely on Item Response Theory (IRT) or sequential probability checks to detect individual cheating (e.g., answer copying). However, detecting **group-level or systemic anomalies** requires analyzing the aggregate distribution shape.

   - **Feature Justification:** Research in statistical anomaly detection (SAD) (**Goldberg, 2015**) emphasizes the importance of developing the "right features" before modeling. Therefore, I will adopt the use of statistical features like Skewness and Kurtosis for each center. Skewness, which measures the asymmetry of the score distribution, is a powerful indicator: an unusually low positive skew (or even a negative skew) in a highly competitive exam is statistically irregular and suggestive of non-random score distribution.

- **Model Justification:** The project requires an unsupervised model due to the lack of labeled "cheating" data. The Isolation Forest algorithm (**Liu, Ting, & Zhou, 2008**) will be chosen for its efficacy in identifying outliers in high-dimensional datasets without requiring labeled training data. Unlike density-based methods, Isolation Forest explicitly isolates anomalies using few random feature cuts, making it computationally efficient and highly suitable for the large, high-dimensional feature space created by including geographic (state) and statistical (Skewness) indicators. This methodology ensures the flags raised are rooted in verifiable statistical evidence, providing the necessary ML foundation for the project.

OBJECTIVES AND SCOPE OF WORK

The overarching goal of this project will be to apply advanced Data Science and Machine Learning techniques to the complete NEET-UG 2024 examination data to identify, locate, and validate statistical irregularity related to performance.

Key Objectives and Scope of Work:

1. **Quantify and Contextualize Inequality**

   The analysis will develop powerful derived metrics, notably the **Center-v-National Gap** and **Ultra-High Score Ratio** (students ≥ 700 marks, normalized by center size), to precisely measure performance deviation and the localized concentration of top-scoring students across all 4,750+ exam centers and multiple states.

2. **Develop Anomaly Features for ML**

   The methodology will aggregate granular student score data (over 2 million records) to calculate advanced, high-dimensional statistical features like **Skewness, Kurtosis, and Standard Deviation** for every center's score distribution. These metrics will be designed to capture the shape of the score curve, which is the primary indicator of an irregular pattern.

3. **Implement Anomaly Detection (ML)**

   The project will build and train an **Isolation Forest** unsupervised machine learning model on the multi-dimensional feature space. This model will **mathematically flag centers** that exhibit statistically irregular performance and score distribution patterns, providing an objective, data-driven list of anomalies.

4. **Visualize Insights and Deploy**

   The final step will be to create highly interactive, multi-page dashboards in **Streamlit** (ML-focused web application) and **Power BI** (Executive & Geospatial report) to deploy the findings on inequality hotspots, regional disparities, and the final list of ML-flagged anomalous centers to stakeholders.

5. **Validate Model Integrity**

   The study will validate the model's effectiveness by using **numerical comparative charts** (e.g., Average Skewness comparison) to show that the features used are statistically and significantly different between the ML-identified 'Anomalous' and 'Normal' groups.

METHODOLOGY

The **EduDRISHTI** project will employ a robust, phased data science methodology utilizing Python (Pandas, Scikit-learn, Plotly) and Power BI for reporting.

Phase I: Data Integration and Feature Engineering (Python/Pandas)

1. **Data Consolidation:** The initial step will involve merging the center-wise aggregate data (*NEET_2024_CenterWise_Stats*) with the corresponding high-volume granular student score data (*NEET_2024_Marks_By_State_City_Center*) using the common identifier, center_id.
2. **Inequality Metrics:** The four core metrics: Center-v-National Gap, Center-v-State Gap, Ultra-High Score Ratio, and High Score Ratio will be calculated. These metrics will form the basis for initial EDA and inequality assessment.
3. **ML Feature Extraction:** Using Pandas' powerful grouping functions (.groupby().agg()), the Center Skewness, Center Kurtosis, and Center Standard Deviation for the score distribution of every center will be calculated. These features will quantify the asymmetry and peakedness of the score curve, serving as direct statistical proxies for highly unusual performance uniformity or clustering.

Phase II: Intermediate Machine Learning (Scikit-learn/Isolation Forest)

1. **Feature Preparation:** The combined features (Performance Gaps, Skewness, Kurtosis, and One-Hot Encoded geographic regions, state) will be prepared for modeling.
2. **Scaling:** The data will be transformed using StandardScaler. This essential step will normalize all feature magnitudes, ensuring that the model's decision-making is not biased towards large-value features like the gap, but equally considers the subtle patterns in Skewness and Kurtosis.
3. **Model Training:** The Isolation Forest model, a computationally efficient ensemble tree-based algorithm, will be trained. It is specifically designed for unsupervised anomaly detection by identifying points that require the fewest random cuts to be isolated from the bulk of the data.

4. **Prediction:** The model will generate two primary outputs: the Anomaly_Score (a continuous value where lower scores indicate higher irregularity) and the final binary Anomaly_Flag (-1 for Anomalous, 1 for Normal).

Phase III: Visualization and Reporting (Streamlit/Power BI)

1. **Streamlit Dashboard:** A multi-page interactive web application will be developed using Plotly for high-interactivity:
   - **Page 1 (ML Results):** Displaying the top 50 centers ranked by their Isolation Forest Anomaly_Score.
   - **Page 2 (Validation):** Scatter plot showing Anomaly_Score vs. Center-v-National Gap, with anomalies colored distinctly, validating the model's separation logic.
   - **Page 3 (Drilldown):** Interactive regional filters allowing users to investigate performance gaps within specific states and cities.
2. **Power BI Dashboard:** An executive report will be created for high-level, static analysis:
   - **Page 1 (Executive Summary & KPIs):** Displaying national averages, total anomalies, and state-wise performance rankings.
   - **Page 2 (Geospatial & High Score Analysis):** Featuring the Choropleth Map colored by the Average Center-v-National Gap and the Scatter Map showing Anomaly_Type location.
   - **Page 3 (ML Validation & Deep Dive):** Showcasing the Matrix visual and Gauge chart for numerical proof of the model's effectiveness using Average Skewness comparison.

TENTATIVE CHAPTERIZATION

1. **Chapter 1: Introduction**

   This chapter will cover the foundational elements of the project.

   - **Introduction to Project:** Detailed background on the NEET-UG 2024 controversy and the need for data-driven analysis.
   - **Project Category:** Research based and Application Development (due to the ML model and dashboard deployment).
   - **Objectives:** Directly stating the five objectives from Section 2.
   - **Problem Formulation:** Defining the core problems: **performance inequality** and **statistical anomaly detection**.
   - **Existing System:** Discussing the limitations of relying only on traditional aggregate rankings and basic metrics.
   - **Proposed System:** The implementation of the **Isolation Forest ML Model** and the **Streamlit + Power BI Dashboards**.
   - **Unique Features of the System:** Highlighting the use of advanced features like **Skewness/Kurtosis** and the **Center-v-National Gap**.

2. **Chapter 2: Requirement Analysis and System Specification**

   This chapter will detail the necessary conditions and constraints for the project.

   - **Feasibility study (Technical, Economical, Operational):** Assessing the viability of using Python, Scikit-learn, and the large datasets.
   - **Software Requirement Specification Document (SRS):**
     - **Data Requirement:** Two datasets (Aggregate and Granular scores), 2M+ records, data quality and structure.
     - **Functional Requirement:** Ability to execute the ML model, generate anomaly scores, and provide interactive filtering in the dashboards.
     - **Performance Requirement:** The ML model must train within a reasonable time; dashboards must load quickly (met via Streamlit's caching).
   - **SDLC model to be used:** Adopting an **Iterative/Agile model** to allow for continuous refinement of the ML feature set and dashboard design.

3. **Chapter 3: System Design**

   This chapter will map the methodology to formal design tools.

   - **Design Approach: Function-Oriented Design** (data flow and process focus, suitable for a data science pipeline).
   - **Methodology:** Detailed explanation of **Phase I (Feature Engineering)** and **Phase II (ML Modeling)**, defining the logic flow.
   - **System Design:** Creation of a **Data Flow Diagram (DFD)** illustrating the pipeline:

Raw Data → Feature Extraction → Scaling → Isolation Forest → Final Output CSV →Dashboards.

   - **User Interface Design:** Defining the structural layout for the three pages in both the **Streamlit** and **Power BI dashboards**.

4. **Chapter 4: Implementation, Testing, and Maintenance**

   This chapter will focus on the tools and quality assurance processes.

   - **Introduction to Languages, IDE's, Tools and Technologies:** Detailing the use of Python (3.x), Pandas, NumPy, Scikit-learn, Streamlit, and Power BI.
   - **Testing Techniques and Test Plans:**
     - **Unit Testing:** Validation of feature engineering logic (e.g., ensuring Center-v-National Gap calculation is correct).
     - **Model Validation:** Using the Matrix Visual (Average Skewness comparison) and the Anomaly Score Breakpoint plot to prove the model's predictive power.
     - **Dashboard Testing:** Functional checks for interactive elements and data accuracy.

5. **Chapter 5: Results and Discussions**

   This chapter will showcase the project's output and key findings.

   - **User Interface Representation:** Snapshots and descriptions of the **Streamlit application** (ML Results Page, Validation Scatter Plot, etc).

- **Snapshots of system with brief detail of each:** Visuals from the **Power BI Dashboard** (Choropleth Map, ML Validation Matrix, Top Centers Bar Chart).
- **Brief Description of Various Modules:** Explaining the three core modules: Feature Engineering, ML Scoring, and Visualization.
- **Back Ends Representation:** Displaying the structure and key columns of the final NEET_Master_ML_Data.csv file.

6. **Chapter 6: Conclusion and Future Scope**
   - **Conclusion:** Summarizing the validated insights, the list of statistically anomalous centers, and the implications of the regional inequality findings.
   - **Future Scope:** Detailing the plans for **Temporal Analysis** and **Feature Augmentation** (integrating demographic/external data).

CONCLUSION AND SUGGESTIONS

Conclusion

The **EduDRISHTI** project successfully achieved its core mission: to move beyond anecdotal evidence and provide an objective, data-driven analysis of the systemic vulnerabilities exposed during the **2024 NEET-UG controversy**.

By leveraging over 2 million student records and applying advanced feature engineering, we achieved two critical outcomes:

1. **Quantified Micro-Inequality:** The custom-developed metrics, such as the **Center-v-National Gap** and **Ultra-High Score Ratio**, allowed us to precisely locate and quantify performance deviations at the individual exam center level. This analysis moved past broad state or city averages, exposing specific geographic hotspots that demonstrate statistically significant disparities in learning or testing environments.
2. **Objective Anomaly Detection:** The implementation of the **Isolation Forest** unsupervised machine learning model provided a rigorous, unbiased method for identifying statistical irregularities. The model successfully isolated centers whose score distributions (quantified by features like Skewness and Kurtosis) were mathematically inconsistent with the national body of data. The resulting list of **ML-flagged anomalous centers** offers policymakers a clear, defensible roadmap for targeted investigation, thereby restoring a measure of accountability to the high-stakes examination system.

The final deliverable—the **interactive Streamlit** and **Power BI dashboards**—translates complex statistical findings into actionable intelligence. By providing transparency and verifiable statistical evidence, this system serves as a powerful foundation for enhancing fairness and equity in the national assessment framework.

Future Scope

To maximize the long-term impact and robustness of the EduDRISHTI model, future iterations should focus on two primary directions: **Temporal Analysis** and **Feature Augmentation**.

1. **Temporal Analysis and Baseline Establishment**

   The current model provides a snapshot of the 2024 irregularity. Future work should expand the scope to include historical NEET-UG data (e.g., 2022 and 2023).

   - **Goal:** Establish a baseline "normal" statistical footprint for every exam center and state.
   - **Action:** Apply the same feature engineering and Isolation Forest model to multi-year data. This will allow the model to detect sudden, year-over-year shifts in performance metrics, which are often stronger indicators of a systemic change or administrative irregularity than a single-year deviation.
   - **Value:** This temporal dimension will help distinguish between chronic, sustained underperformance (an inequality issue) and sudden, acute statistical anomalies (a potential integrity issue).

2. **Feature Augmentation and Causal Inference**

   The current model relies solely on score distribution features. Integrating external data will allow the project to move from detection to explanation.

   - **Data Integration:** Incorporate external datasets such as district-level **Socio-Economic Indicators** (poverty rates, literacy levels) and **Administrative Data** (center capacity, invigilator training records, security protocol adherence reports).
   - **Refined Modeling:** Use the augmented feature set to retrain the Isolation Forest. Anomalies flagged by the model can then be analyzed against these new features to assign a preliminary level of statistical risk, potentially distinguishing between anomalies caused by malpractice and those caused by genuine, but unusual, socio-economic factors.
   - **Value:** This augmentation would enhance the model's predictive power and provide a more nuanced understanding of the forces driving both performance inequality and statistical irregularities.

REFERENCES

Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "*Anomaly Detection: A Survey.*" ACM Computing Surveys (CSUR), vol. 41, no. 3, 2009, article 15, https://www.google.com/search?q=https://doi.org/10.1145/1541880.1541881.

Cizek, Gregory J., and Jamie D. O. Wollack. "*Handbook of Quantitative Methods for Detecting Cheating on Tests*". Routledge, 2017.

Coleman, James S., et al. "*Equality of Educational Opportunity (The Coleman Report)*". U.S. Department of Health, Education, and Welfare, 1966.

Garg, Manju K., Pranab Chowdhury, and S. K. Kanchan. "*An Overview of Educational Inequality in India: The Role of Social and Demographic Factors.*" Frontiers in Education, vol. 7, 2022, article 871043, https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2022.871043/full.

Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. "*Isolation Forest.*" 2008 Eighth IEEE International Conference on Data Mining, 2008, pp. 413-422, https://doi.org/10.1109/ICDM.2008.17.

Mukherjee, Anirban, and Amrita Roy. "*The Pressure Cooker: Analyzing Student Stress and Performance in India's Competitive Entrance Exams.*" Economic and Political Weekly, vol. 58, no. 31, 2023, pp. 34-42.

Srinivasan, R., S. A. G, and N. M. K. "*Accountability and Transparency in High-Stakes Public Examinations: A Data Analytics Perspective.*" Journal of Education and Policy, vol. 12, no. 1, 2021, pp. 45-60.

West, Stephen G., Lisa M. Finch, and Patrick J. Curran. "*Structural Equation Models with Nonnormal Variables: Problems and Remedies.*" Journal of Consulting and Clinical Psychology, vol. 68, no. 4, 2000, pp. 562–575.