**AI Capstone Final Project Report**

**Course Name: AI Capstone Project Preparation (AIGC 5005)**

**Team Members:**

* **Devi Sri Tanuku**
* **Neel Raibole**
* **Kedar Dave**
* **Kartik Kalra**

**Table of Contents**

1. [Introduction](#introduction)
2. [Phase 2: Literature Review and Data Collection](#phase-2-literature-review-and-data-coll) **2.1.** [Literature Review](#literature-review) **2.2.** [Dataset Summary](#dataset-summary)
3. [Phase 3: Model Design and Prototype Development](#phase-3-model-design-and-prototype-deve) **3.1.** [Preprocessing Steps](#preprocessing-steps) **3.2.** [Model Architectures and Training](#model-architectures-and-training)
4. [Phase 4: Analysis and Insights](#phase-4-analysis-and-insights) **4.1.** [Feature Importance](#feature-importance) **4.2.** [Residual Analysis](#residual-analysis) **4.3.** [Choosing the Best Model](#choosing-the-best-model)
5. [Phase 5: Final Summary and Future Directions](#phase-5-final-summary-and-future-direct)

**Introduction**

**This project explores methods for predicting student performance based on behavioral and academic factors. The focus is on assessing traditional and deep learning models to identify the best approach for achieving optimal predictive accuracy and minimizing errors.**

**Phase 2: Literature Review and Data Collection**

**Literature Review**

**Related Work 1:**

* ***Predicting Student Performance Using Behavioral and Academic Factors:***
  + **This study utilized Random Forest and Logistic Regression to predict academic performance.**
  + **Key insights included the importance of parental involvement and study hours as critical predictors.**
  + **Emphasis was placed on metrics such as feature importance analysis and error rates for validation.**

**Related Work 2:**

* ***Deep Learning for Student Performance Prediction:***
  + **Leveraged Artificial Neural Networks (ANNs) to uncover latent patterns in educational data.**
  + **Demonstrated the importance of preprocessing techniques like feature scaling.**
  + **Discussed optimal architecture selection to improve model performance.**

**These studies informed the feature selection, preprocessing methods, and model design for this project.**

**Dataset Summary**

**The dataset (StudentPerformanceFactors.csv) includes 20 features and one target variable, Exam\_Score.**

**Key Attributes:**

* **Numerical Variables:**
  + **Hours\_Studied: Weekly hours spent studying.**
  + **Attendance: Percentage of attendance.**
  + **Distance\_from\_Home: Travel distance to school.**
* **Categorical Variables:**
  + **Parental\_Involvement: Ordinal scale (Low, Medium, High).**
  + **Teacher\_Quality: Ordinal scale (Poor, Average, Excellent).**
  + **Gender: Binary categorical (Male, Female).**

**Initial Observations:**

1. **Missing Values:**
   * **Teacher\_Quality: 8.5% missing.**
   * **Distance\_from\_Home: 12.3% missing.**
   * **Parental\_Education\_Level: 6.7% missing.**
2. **Correlations:**
   * **Strong correlation (r=0.78r = 0.78r=0.78) between Hours\_Studied and Exam\_Score.**
3. **Class Imbalance:**
   * **Parental\_Involvement is skewed toward "Medium."**

**Phase 3: Model Design and Prototype Development**

**Preprocessing Steps**

1. **Missing Value Imputation:**
   * **Mode imputation for Teacher\_Quality (most frequent value).**
   * **Median imputation for Distance\_from\_Home to handle outliers.**
2. **Feature Transformation:**
   * **Ordinal encoding for Parental\_Involvement and Teacher\_Quality.**
   * **One-hot encoding for binary features like Gender.**
3. **Feature Scaling:**
   * **Applied StandardScaler to normalize numerical features (Hours\_Studied, Attendance).**
4. **Class Balancing:**
   * **Used SMOTE to address imbalances in Parental\_Involvement.**
5. **Data Splitting:**
   * **Split into 80% training and 20% testing, ensuring stratification to preserve class distributions.**

**Model Architectures and Training**

**1. Linear Regression**

* **Objective: Establish baseline performance.**
* **Implementation:**
  + **Verified assumptions such as linearity, independence, and homoscedasticity.**
  + **No hyperparameter tuning required.**
* **Results:**
  + **MSE: ~3.236**
  + **MAPE: ~0.0062**

**2. Decision Tree Regression**

* **Objective: Capture non-linear patterns.**
* **Implementation:**
  + **Criterion: Mean Squared Error (MSE).**
  + **Max Depth: Tuned via Grid Search (optimal depth = 8).**
* **Results:**
  + **MSE: ~4.857**
  + **MAPE: ~0.0234**

**3. Random Forest**

* **Objective: Ensemble method to improve generalization.**
* **Implementation:**
  + **Number of Estimators: 100.**
  + **Max Features: Automatically selected.**
* **Results:**
  + **MSE: 4.5003**
  + **MAPE: 0.0158**

**4. Artificial Neural Network (ANN)**

* **Objective: Capture complex feature interactions.**
* **Architecture:**
  + **Input Layer: 20 neurons.**
  + **Two Hidden Layers: 64 and 32 neurons, ReLU activation.**
  + **Output Layer: Single neuron with linear activation.**
* **Training:**
  + **Optimizer: Adam.**
  + **Loss Function: Mean Squared Error.**
  + **Early Stopping enabled (Epochs = 100).**
* **Results:**
  + **MSE: 0.2767 (lowest among all models).**
  + **MAPE: 0.4174**

**Phase 4: Analysis and Insights**

**Feature Importance (Random Forest):**

* **Hours\_Studied: 35% contribution.**
* **Parental\_Involvement: 22% contribution.**
* **Teacher\_Quality: 18% contribution.**

**SHAP Observations**

**SHAP (SHapley Additive exPlanations) was used to interpret the contributions of individual features across different models:**

**Linear Regression:**

* **Top Contributors:**
  + **Hours\_Studied: Strong positive contribution, consistent with its correlation to Exam\_Score.**
  + **Parental\_Involvement: Moderate positive influence, with higher levels boosting predictions.**
* **Insights: Linear models provided straightforward interpretations, aligning with feature correlations.**

**Random Forest:**

* **Top Contributors:**
  + **Hours\_Studied: Dominates the predictions with the highest SHAP values across the dataset.**
  + **Parental\_Involvement and Teacher\_Quality: Complementary contributors but with non-linear impacts.**
* **Insights: Random Forest captures complex, non-linear relationships between features and Exam\_Score.**

**Decision Tree Regression:**

* **Top Contributors:**
  + **Hours\_Studied: High variability in SHAP values, emphasizing its central role in decision-making splits.**
  + **Distance\_from\_Home: Occasional spikes in SHAP values, indicating its situational impact.**
* **Insights: The Decision Tree overemphasized specific features, reflecting its tendency to overfit.**

**Artificial Neural Network (ANN):**

* **Top Contributors:**
  + **Hours\_Studied: Consistently dominant with smooth SHAP value distributions.**
  + **Parental\_Involvement: A significant but secondary contributor.**
* **Insights: The ANN’s interpretation was more nuanced, with smoother SHAP distributions reflecting complex interactions.**

**Residual Analysis:**

* **Linear models exhibited consistent residuals but struggled with non-linear relationships.**
* **Decision Tree showed heteroscedasticity, indicating overfitting.**
* **ANN exhibited smooth residuals, reflecting its ability to model complex patterns.**

**Learning Curves (ANN):**

* **Training and validation loss showed consistent convergence, indicating effective training without overfitting.**

**Phase 5: Final Summary and Future Directions**

**Summary of Contributions:**

1. **Robust preprocessing addressed missing values, class imbalance, and feature scaling effectively.**
2. **Model evaluation demonstrated ANN's ability to minimize overall errors and linear models’ superiority in proportional accuracy.**
3. **Visualizations, including heatmaps and residual plots, provided critical insights into feature importance and model behavior.**

**Future Directions:**

1. **Explore Gradient Boosting methods like XGBoost or LightGBM to balance MSE and MAPE.**
2. **Validate findings across multiple datasets to assess generalizability.**