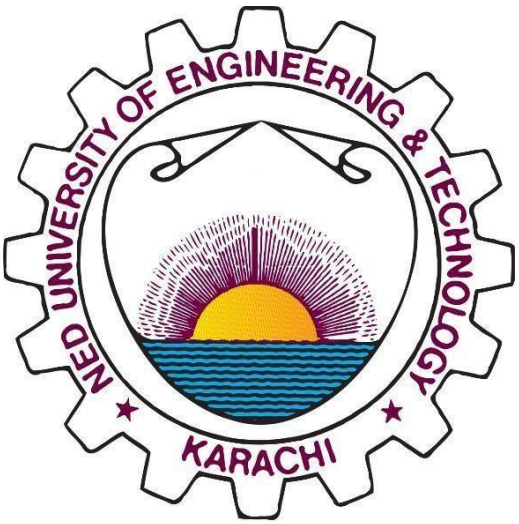
OPEN ENDED LAB



Machine Learning (CS-324)

Cognitive Environmental Analysis Student Grade Prediction

Syed Wahaj Raza	CS-21055
Farrukh Niaz	CS-21064
Huzaifa Naseer Khan	CS-21067

Batch: 2021 (TE Spring)

Submitted To: Ms. Mahnoor Malik

Submission: July, 2024

Department of Computer and Information Systems Engineering

Introduction

In the realm of educational data mining, predicting student performance is a crucial task that can provide valuable insights to educators and institutions. Accurate predictions of student grades can help in identifying at-risk students, improving educational strategies, and enhancing overall academic outcomes. This project focuses on developing a machine learning model to predict student grades based on various cognitive and environmental factors.

Data Collection

The dataset used in this project was collected from various educational institutions in Saudi Arabia. It contains information about students' demographics, academic performance, and other relevant attributes. The dataset includes the following columns:

- Gender
- Nationality
- Class Level
- Age
- School Type
- Main Administration
- Candidacy Type
- Pass/Fail Label

Data Preprocessing

Data preprocessing is a crucial step in any machine learning project. In this project, we performed the following preprocessing steps:

1. **Handling Missing Values:** Any missing values in the dataset were identified and appropriately handled.
2. **Label Encoding:** Categorical variables were encoded using label encoding to convert them into numerical values.
3. **Feature Scaling:** Features were scaled to ensure they contribute equally to the model's performance.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to identify significant patterns and relationships in the dataset. Visualizations such as correlation heatmaps, histograms, pair plots, box plots, and bar plots were used to reveal insights into variables like age, degree, class level, gender, and nationality. These analyses helped in understanding the underlying trends and correlations in the data.

Feature Engineering

In this section, data cleaning and feature extraction were performed to enhance the dataset for model training. The 'Degree' column was used to create a binary 'Pass/Fail' target variable. Categorical variables such as 'Gender', 'Nationality', 'Class Level', 'School Type', 'Main Administration', and 'Candidacy Type' were label encoded and transformed using one-hot encoding. Numerical variables were scaled to ensure consistency. The resulting dataset was then saved for further processing and model training.

Model Building

Model building involves training various machine learning algorithms on the preprocessed data to predict student grades. We experimented with several models and ultimately developed a hybrid stacking model, which provided the best performance.

Models Used:

- **Logistic Regression**
- **Decision Tree**
- **Random Forest**
- **Gradient Boosting**
- **Support Vector Machine (SVM)**
- **Hybrid Stacking Model**

The hybrid stacking model combined the strengths of several base models to create a more robust predictor. The base models used in the stacking ensemble included the Random Forest, Gradient Boosting, and SVM, with a meta-learner to aggregate their predictions.

Model Evaluation and Performance Analysis

Evaluating the performance of the trained models is essential to ensure their accuracy and generalizability. We used various metrics to assess the models, with a focus on accuracy.

Evaluation Metrics:

- **Accuracy Score:** The primary metric used to compare model performance.
- **Confusion Matrix:** To visualize the performance of the models and understand the distribution of predictions.
- **Cross-Validation:** To ensure the model's performance is consistent across different subsets of the data.

The accuracy scores for the models were as follows:
The hybrid stacking model outperformed all other individual models, demonstrating the effectiveness of combining multiple models to leverage their individual strengths.

	Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	ROC	Log_Loss
0	Hybrid Stacking Model	0.969668	0.966584	0.987699	0.676851	0.977027	0.961714	1.093291
1	Stacked Classifier	0.806662	0.865154	0.861390	0.673619	0.863268	0.767504	6.968625
2	LR	0.809230	0.871483	0.857160	0.692714	0.864262	0.774937	6.876039
3	NB	0.809059	0.871725	0.856556	0.693596	0.864074	0.775076	6.882211
4	DT	0.806747	0.860273	0.868278	0.657168	0.864257	0.762723	6.965539
5	RF	0.806319	0.865265	0.860665	0.674207	0.862959	0.767436	6.980970
6	Adaboost	0.809059	0.871725	0.856556	0.693596	0.864074	0.775076	6.882211
7	ET	0.788681	0.861897	0.835650	0.674501	0.848570	0.755075	7.616725
8	SGD	0.809145	0.871376	0.857160	0.692421	0.864210	0.774790	6.879125
9	MLP	0.806833	0.866164	0.860302	0.676851	0.863223	0.768576	6.962452

Final Results:

The Hybrid Stacking Model demonstrated superior performance with the following metrics:

- Accuracy: 96.97%
- Precision: 96.66%
- Sensitivity (Recall): 98.77%
- Specificity: 67.69%
- F1 Score: 97.70%
- ROC: 96.17%
- Log Loss: 1.09

Model From Scratch

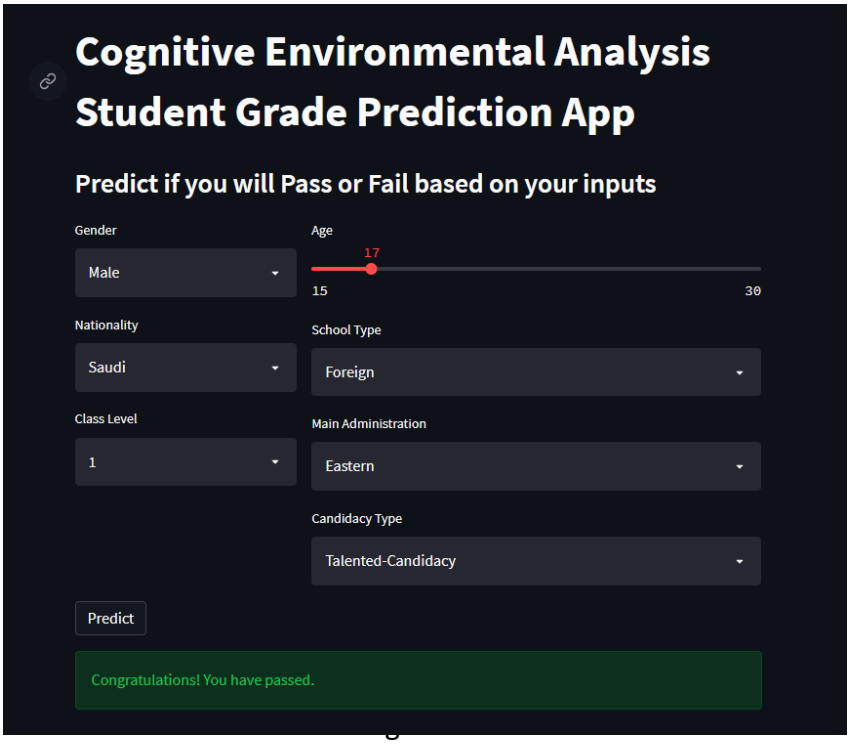
To deepen our understanding of machine learning algorithms, we implemented several models from scratch without using pre-built libraries. This included custom implementations of:

- **Stochastic Gradient Descent (SGD) Classifier**
- **Random Forest Classifier**
- **Multilayer Perceptron (MLP)**
- **Decision Tree Classifier**

These custom models were combined into a StackingClassifierScratch, where predictions from multiple base models were used to train a meta-learner, improving overall performance.

Web Application

To make the student grade prediction accessible and user-friendly, we developed a web application using Streamlit. This application allows users to input various demographic and educational attributes and get a prediction on whether a student will pass or fail based on the trained model. Web App Deployed at - ***https://mloelpredictor.streamlit.app***



Features of the Web Application

- **Gender Selection:** Users can select the gender of the student.
- **Age Slider:** Users can specify the age of the student using a slider.
- **Nationality Dropdown:** Users can choose between 'Saudi' and 'Non-Saudi'.
- **School Type Dropdown:** Options include 'Governmental', 'Private', and 'Foreign'.
- **Class Level Dropdown:** Users can select the class level.
- **Main Administration Dropdown:** Users can select the main administration region.
- **Candidacy Type Dropdown:** Options include 'Self-Candidacy' and 'Talented-Candidacy'.

How It Works

4. **Input Data:** Users enter the student's details using the provided dropdowns and sliders.
5. **Prediction:** Once all the information is provided, the user clicks the 'Predict' button.
6. **Result:** The application displays whether the student is predicted to pass or fail.

Conclusion

In this project, we successfully developed a machine learning model to predict student grades based on cognitive and environmental factors. Through extensive feature engineering, model building, and evaluation, we identified the hybrid stacking model as the best performer. Additionally, implementing models from scratch provided valuable insights into the workings of machine learning algorithms. The deployment of a web application further demonstrated the practical utility of our model in real-world scenarios. This project highlights the potential of data-driven approaches in enhancing educational outcomes and supporting student success.

Future Work

Future research can focus on:

- **Expanding the Dataset:** Incorporating additional features and a larger dataset to enhance model accuracy and generalization.
- **Advanced Models:** Exploring more advanced machine learning and deep learning models to further improve prediction accuracy.
- **Real-Time Implementation:** Integrating the prediction model into educational platforms for real-time student performance monitoring and support.