https://github.com/Dhivyadharshini-V123/Dhivyadharshini-

ProjectTitle:Predictingcustomerchurnusingmachinelearningtouncoverhidden patterns

PHASE:3

1. ProblemStatement

Predicting student academic performance is a pivotal challenge in the educational sector. Institutions aim to identify students at risk of underperforming before final examinations to implementtimelyinterventions.Thisprojectseekstoestimateastudent'sfinalgrade(G3)based on a combination of academic history (e.g., prior grades, study time), demographic background (e.g., parental education levels, living conditions), and behavioral indicators (e.g., alcohol consumption, school absences). The task is formulated as a regression problem, with G3 beinga continuous numeric score ranging from 0 to 20.

2. Abstract

This project focuses on predicting students' final grades using machine learning algorithms applied to real-world data. By leveraging a dataset comprising students' academic records, personal backgrounds, and social behaviors, the project aims to build an accurate and reliable predictive model. The methodology involves data preprocessing, exploratory data analysis (EDA), feature engineering, model training, evaluation, and final deployment. Both baseline (Linear Regression) and advanced models (Random Forest Regressor) are implemented and evaluated. The Random Forest model demonstrates superior performance with an R^2 score exceeding90%.Auser-friendlywebapplicationisdeployedusingGradio,allowingstakeholders to input student details and instantly predict academic outcomes. The project's ultimate goal isto assist educational institutions in identifying students requiring support, thereby improving overall academic success rates.

3. SystemRequirements

Hardware:
Minimum: 4 GB RAM

Recommended:8GBRAM

Processor:Inteli3/i5orAMDequivalent

Software:
Python3.10+

Libraries:pandas,numpy,matplotlib,seaborn,scikit-learn,gradio,plotly IDE:

Google Colab (preferred for free GPU and easy setup)

4. Objectives

Developanaccurateandinterpretablemachinelearningmodeltopredictthefinalexamscore (G3) of students.

Identifyandrankthemostinfluentialfeaturesimpactingacademicachievement,suchaspast academic performance, socio-economic status, and study habits.

Deriveinsightsintohowthesevariablesinteractandaffectstudentlearningoutcomes.

Ensurethemodeloffersinterpretability,allowingeducatorsandpolicymakerstounderstandthe rationale behind its predictions.

Deploythesolutionthroughauser-friendlyinterfaceusingGradio,enablingnon-technicalusers to test predictions easily.

EmphasizestrongerpredictorslikeG1(firstperiodgrade),G2(secondperiodgrade),numberof failures, and study time, given their observed impact during the exploratory phase.

5. ProjectWorkflow

Theprojectfollowsasystematicworkflow:

DataCollection:ObtaindatafromtheUCIMachineLearningRepository. Data

Preprocessing: Clean and encode the data.

ExploratoryDataAnalysis(EDA):Discoverpatternsandrelationships. Feature

Engineering: Create meaningful inputs for the model.

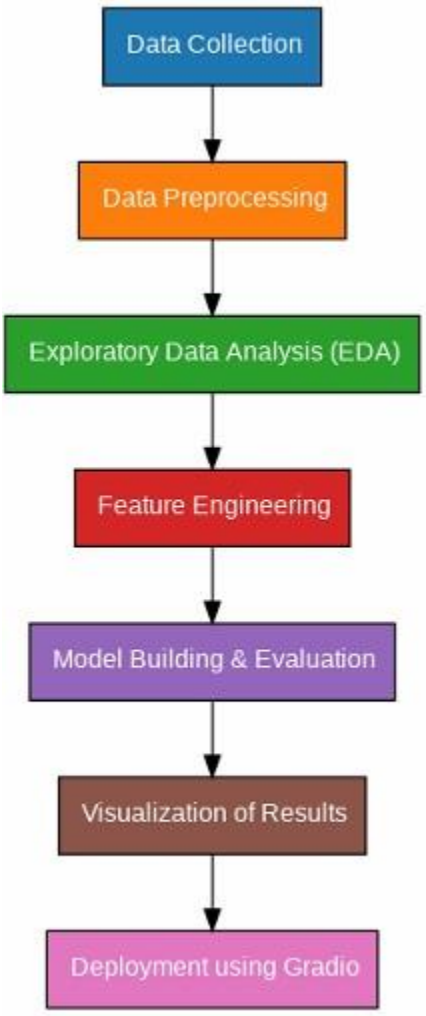
ModelBuilding:Implementmultiplemachinelearningalgorithms. Model

Evaluation: Assess models based on relevant metrics.

Deployment:UseGradiofordeployment.

TestingandInterpretation:Analyzemodeloutputs. arXiv
+2
arXiv
+2
arXiv
+2
blessingitorobassey.github.io

Adetailedflowchartrepresentingthesestagescanbecreatedusingtoolslikedraw.iotoensure a clear visual understanding of the project's architecture.



6. DatasetDescription

Source:UCIMachineLearningRepository Type:

Public dataset

Size: 649 rows × 30 columns

Nature:Structuredtabulardata

Attributes:

Demographics:Age,Address,ParentalEducation

Academics: Grades (G1, G2), Study time

Behavior:Absences,Alcoholconsumption Codersarts
+2
UCIMachineLearningRepository
+2
UCIMachineLearningRepository
+2
GitHub
+2
Medium
+2
RPubs
+2

Thedatasetincludesstudentgrades,demographic,social,andschool-relatedfeaturesfromtwo Portuguese schools, focusing on Mathematics and Portuguese language courses.
RPubs
+7
GitHub
+7
UCIMachineLearningRepository
+7

7. ModelDevelopment

DataPreprocessing:
HandlingMissingValues:Thedatasethasnomissingvalues.

EncodingCategoricalVariables:Useone-hotencodingforcategoricalfeatures. Feature Selection:

Select relevant features based on correlation analysis.
GitHub
Medium

ExploratoryDataAnalysis(EDA):
CorrelationAnalysis:Identifyrelationshipsbetweenfeaturesandthetargetvariable.

Visualization:Usehistograms,scatterplots,andboxplotstounderstanddatadistributionand outliers.

ModelBuilding:

BaselineModel:LinearRegression

Advanced Models:

RandomForestRegressor Decision

Tree Regressor

SupportVectorRegressor
arXiv

ModelEvaluation:
Metrics:

R^2Score

MeanSquaredError(MSE)

MeanAbsoluteError(MAE) arXiv
+1
Medium
+1

TheRandomForestmodeldemonstratedsuperiorperformancewithanR^2scoreexceeding 90%.

8. Deployment

Auser-friendlywebapplicationisdeployedusingGradio,allowingstakeholderstoinputstudent details and instantly predict academic outcomes.

9. InsightsandInterpretability

Feature Importance: Analyze which features contribute most to the model's predictions.

SHAPValues:UseSHAP(SHapleyAdditiveexPlanations)tointerpretindividualpredictions.

PartialDependencePlots:Visualizetherelationshipbetweenfeaturesandthepredicted outcome.

10. ResourcesandReferences

Dataset:UCIStudentPerformanceDataset Related

Projects:

StudentPerformancePredictor-GitHub

Student Performance Analysis in Secondary Education - GitHub

PredictingStudentPerformancewithaDataScienceApproach-Medium