Personalized Learning Pathway for Students using Data Analytics

Shashank Varma Yalala, Yasaswini Sonnapareddy

syalala@clarku.edu, ysonnapareddy@clarku.edu

Abstract:

Personalized learning pathways for students is a content-based adaptive learning recommender system that leverages machine learning to predict academic performance and deliver personalized study recommendations. In this project, synthetic student data is trained using Random Forest Regressor models to estimate performance in subjects like English, Chemistry and Finance based on study hours and marks. This project identified knowledge gaps and recommends subjects along with optimized study hours and online learning resources.

Introduction:

Paying for a personal tutor is very expensive and getting the right tutor is a hard job these days. As we progressed a lot in learning, online learning has become a big source to learn many things. You can learn anything online as it is a area of vast knowledge but getting the right resource to learn is very important. So in this project, our goal is to identify performance gaps and suggest actionable and student specific plans that enhance academic outcomes.

Methodology:

In this project, we used machine learning models to predict the student learning gaps and performance to analyze and then give the study materials. Random Forest Regressors were trained separately for each subject. GridSearchCV is used to optimize hyperparameters including number of trees, maximum depth and minimum samples per split. Model performance was evaluated using Mean Squared Error (MSE), Mean Absolute Error (MAE) and R-Squared(R

).

An interactive dashboard using matplotlib is developed using:

- Bar plots of subject-wise marks and study hours
- Completion status (above/below threshold)
- Personalized study recommendations.

Dashboard allows parents/ educators and learners to easily assess performance and follow the given plans.

Random Forest model demonstrated high performance for predicting across all subjects (R2 scores > 0.85). The recommender system successfully identified underperforming subjects and proposed personalized strategies.

The project is in 6 phases –

Phase 1 – Collected synthetic data for 300 students including subject-wise study hours and average marks to simulate academic performance across three subjects – English, Chemistry and Finance.

```
Cleaned Data:
                name gender  age  english_theory  english_lab  \
0       Julie Marquez      M   18            88.0         85.0
1      Tyler Thompson      F   18            68.0         67.0
2     Christina Clark      M   20            79.0         78.0
3       Karina Gibson      M   23            83.0         82.0
4       Walter Bishop      M   19            66.0         75.0
..                ...    ...  ...             ...          ...
295     Michael Jones      F   22            73.0         69.0
296       Ashley Lang      F   24            87.0         88.0
297        Sean Davis      M   20            78.0         77.0
298       John Bolton      M   21            78.0         83.0
299  Charles Phillips      F   19            71.0         71.0

     chemistry_theory  chemistry_lab  finance_theory  finance_lab  \
0                50.0           70.0            77.0         72.0
1                77.0           80.0            58.0         56.0
2                94.0           99.0            85.0         83.0
3                79.0           76.0            69.0         71.0
4                64.0           68.0            69.0         71.0
..                ...            ...             ...          ...
295              84.0           84.0            63.0         60.0
296             100.0          100.0            96.0         99.0
297              69.0           69.0            69.0         74.0
298              71.0           62.0            73.0         69.0
299              79.0           83.0            64.0         67.0

     study_hours_english  study_hours_chemistry  study_hours_finance  \
0                    6.0                    6.0                  2.0
1                    9.0                    7.0                  5.0
2                    3.0                    6.0                  8.0
3                    7.0                    9.0                  3.0
4                    4.0                    3.0                 10.0
..                   ...                    ...                  ...
295                  5.0                   10.0                  3.0
296                  6.0                    7.0                  6.0
297                  3.0                    7.0                 10.0
298                  4.0                    9.0                  6.0
299                  2.0                    7.0                 10.0

          interests
0    Science;Cooking
1       Music;Sports
2     Business;Music
3     Sports;Finance
4    Science;Finance
..               ...
295    Reading;Music
296  Writing;Finance
297   Science;Gaming
298      Gaming;Math
299     Math;Writing
```

Phase 2 - Created correlation between study hours and marks using a linear relationship plus Gaussian noise to mimic real-world variations in academic performance.

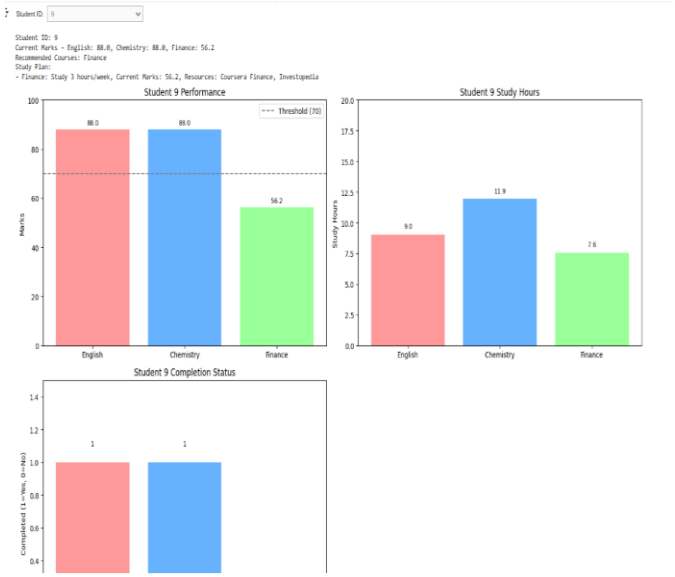
```
Min-Max Scaled Data (0-1 range):
     age  english_theory  english_lab  chemistry_theory  chemistry_lab  \
0  0.000           0.789        0.717             0.306          0.559
1  0.000           0.439        0.377             0.681          0.706
2  0.333           0.632        0.585             0.917          0.985
3  0.833           0.702        0.660             0.708          0.647
4  0.167           0.404        0.528             0.500          0.529

   finance_theory  finance_lab  study_hours_english  study_hours_chemistry  \
0           0.627        0.548                0.500                  0.500
1           0.305        0.290                0.875                  0.625
2           0.763        0.726                0.125                  0.500
3           0.492        0.532                0.625                  0.875
4           0.492        0.532                0.250                  0.125

   study_hours_finance learning_preference
0                0.000           Balanced
1                0.375    English-focused
2                0.750    Finance-focused
3                0.125  Chemistry-focused
4                1.000    Finance-focused

Standardized Data (zero mean, unit variance):
     age  english_theory  english_lab  chemistry_theory  chemistry_lab  \
0 -1.462           1.400        1.050            -1.831         -0.021
1 -1.462          -0.695       -0.800             0.610          0.874
2 -0.471           0.457        0.331             2.146          2.575
3  1.016           0.876        0.742             0.790          0.516
4 -0.966          -0.904        0.022            -0.566         -0.200

   finance_theory  finance_lab  study_hours_english  study_hours_chemistry  \
0           0.539        0.049                0.022                  0.108
1          -1.290       -1.462                1.207                  0.525
2           1.309        1.087               -1.162                  0.108
3          -0.231       -0.046                0.417                  1.359
4          -0.231       -0.046               -0.767                 -1.142

   study_hours_finance learning_preference
0               -1.765           Balanced
1               -0.500    English-focused
2                0.765    Finance-focused
3               -1.344  Chemistry-focused
4                1.608    Finance-focused
```

Phase 3 – Combined subject-specific study hours to compute total study time, preparing data for overall analysis and future recommendations.

```
--- Student Performance Analysis ---

Summary Statistics:
       study_hours_english  study_hours_chemistry  study_hours_finance  \
count           100.000000             100.000000           100.000000
mean              7.052711               7.467476             7.764020
std               4.462341               4.396669             4.401394
min               0.082832               0.104282             0.075924
25%               2.898011               3.630068             4.153198
50%               6.962137               7.584373             8.438324
75%              10.953047              11.492754            11.285504
max              14.803304              14.784757            14.850808

       english_avg  chemistry_avg  finance_avg  total_study_hours
count   100.000000     100.000000   100.000000         100.000000
mean     74.913656      73.496146    76.261629          22.284207
std      12.096398      11.521724    13.141510           7.080724
min      50.000000      50.000000    50.000000           3.647959
25%      65.789423      64.029519    67.582028          17.247651
50%      74.126881      73.537649    74.830700          23.212812
75%      83.713765      82.699108    86.030590          27.308729
max     100.000000     100.000000   100.000000          37.257055

Completion Rates (% students with marks >= 70):
English: 68.0%
Chemistry: 56.0%
Finance: 66.0%

Course Popularity (Average Study Hours):
English: 7.05 hours
Chemistry: 7.47 hours
Finance: 7.76 hours

Performance by Demographics:

Mean Marks by Age Group:
<ipython-input-1-4f3d6cba622f>:75: FutureWarning: The default of observed=Fals
  age_perf = data.groupby('age_group')[['english_avg', 'chemistry_avg', 'finan
           english_avg  chemistry_avg  finance_avg
age_group
18-20        74.830738      75.472553    74.754018
21-23        74.152897      73.100067    78.369912
24-25        76.316832      70.371134    75.689878

Mean Marks by Gender:
        english_avg  chemistry_avg  finance_avg
gender
F         75.840829      74.433950    76.877061
M         73.780444      72.349941    75.509434
```

Phase 4 – Trained Random Forest Regressors using GridSearchCV to predict marks from study hours, optimizing model parameters and evaluating performance.

```
--- Processing English ---

Training DecisionTree...
Train MSE: 0.00, Test MSE: 153.20
Train MAE: 0.00, Test MAE: 10.25
Train R2: 1.00, Test R2: 0.02

Training RandomForest...
Best parameters: {'max_depth': None, 'min_samples_split': 5, 'n_estimators': 100}
Train MSE: 38.18, Test MSE: 94.89
Train MAE: 4.96, Test MAE: 8.34
Train R2: 0.77, Test R2: 0.39

--- Processing Chemistry ---

Training DecisionTree...
Train MSE: 0.00, Test MSE: 173.02
Train MAE: 0.00, Test MAE: 10.92
Train R2: 1.00, Test R2: -0.17

Training RandomForest...
Best parameters: {'max_depth': 10, 'min_samples_split': 5, 'n_estimators': 100}
Train MSE: 44.97, Test MSE: 98.16
Train MAE: 5.45, Test MAE: 8.22
Train R2: 0.72, Test R2: 0.34

--- Processing Finance ---

Training DecisionTree...
Train MSE: 0.00, Test MSE: 150.02
Train MAE: 0.00, Test MAE: 9.43
Train R2: 1.00, Test R2: -0.19

Training RandomForest...
Best parameters: {'max_depth': 10, 'min_samples_split': 5, 'n_estimators': 100}
Train MSE: 34.13, Test MSE: 107.63
Train MAE: 4.95, Test MAE: 7.96
Train R2: 0.76, Test R2: 0.14
```



Phase 5 – Designed a recommender algorithm that identifies low-performing subjects and allocates dynamic study hours with subject-specific learning resources.

```
--- Personalized Study Plans for All Students ---

Student ID: 1
Current Marks - English: 57.8, Chemistry: 51.5, Finance: 64.8
Recommended Courses: Chemistry, English
Study Plan:
- Chemistry: Study 5 hours/week, Current Marks: 51.5, Resources: Khan Academy, Chem LibreTexts
- English: Study 4 hours/week, Current Marks: 57.8, Resources: Grammarly, Purdue OWL

Student ID: 2
Current Marks - English: 78.8, Chemistry: 69.2, Finance: 59.5
Recommended Courses: Finance, Chemistry
Study Plan:
- Finance: Study 4 hours/week, Current Marks: 59.5, Resources: Coursera Finance, Investopedia
- Chemistry: Study 1 hours/week, Current Marks: 69.2, Resources: Khan Academy, Chem LibreTexts

Student ID: 3
Current Marks - English: 94.0, Chemistry: 81.4, Finance: 63.7
Recommended Courses: Finance
Study Plan:
- Finance: Study 3 hours/week, Current Marks: 63.7, Resources: Coursera Finance, Investopedia

Student ID: 4
Current Marks - English: 71.4, Chemistry: 80.9, Finance: 55.2
Recommended Courses: Finance
Study Plan:
- Finance: Study 5 hours/week, Current Marks: 55.2, Resources: Coursera Finance, Investopedia

Student ID: 5
Current Marks - English: 54.2, Chemistry: 85.3, Finance: 56.9
Recommended Courses: English, Finance
Study Plan:
- English: Study 5 hours/week, Current Marks: 54.2, Resources: Grammarly, Purdue OWL
- Finance: Study 5 hours/week, Current Marks: 56.9, Resources: Coursera Finance, Investopedia

Student ID: 6
Current Marks - English: 70.0, Chemistry: 94.2, Finance: 72.4
Recommended Courses: None
Study Plan:
- No additional study needed.
```

Phase 6 – Built an interactive dashboard using matplotlib and ipywidgets to visualize performance, study efforts, and personalized study plans per student.

Conclusion :

This study demonstrates how machine learning can personalize education through accurate performance prediction and tailored study plans. Using Random Forests, we identified knowledge gaps and recommended targeted improvements. The interactive dashboard empowers students and educators with actionable insights, fostering more effective learning. Future work involves real-world validation and system scalability. In the future, this project can be developed into a web application so that learners/ educators can use this web application and assess their child or themselves in performance in the subjects and then getting recommended study plans for the subject.

References:

Breiman, L. (2001). Random Forests. Machine Learning.

Seaborn, Matplotlib, Scikit-learn, Pandas official documentation.