Agnel Charities'

Fr. C. Rodrigues Institute of Technology, Vashi, Navi-Mumbai

(An Autonomous Institute & Permanently Affiliated to University of Mumbai)

Department of Computer Engineering

Big Data Analytics Mini Project-2024-25

Title: Personalized Learning

Abstract

This project leverages PySpark to develop a personalized learning system based on student performance data. The dataset includes variables such as student ID, gender, study time, and grades (G1, G2, G3). Using Spark's capabilities, the dataset is first cleaned and processed, with missing values handled through imputation. Feature engineering is applied to calculate an average grade and encode categorical features like gender and study time.

KMeans clustering is utilized to group students into clusters based on their performance. These clusters are then used to provide personalized learning resource recommendations. The project concludes with a visual analysis of the clustering results, demonstrating patterns in student performance and offering tailored learning materials for each group. This approach showcases the potential of data-driven, personalized educational experiences aimed at improving student outcomes.

Tool/Technology

1) PySpark:
Used for distributed data processing, loading the student performance dataset, data cleaning, and feature engineering. SparkSession was used to initialize and manage the Spark application. Functions from pyspark.sql like StringIndexer, OneHotEncoder, and VectorAssembler were applied for data transformation and preparation.
2)KMeans Clustering:
Implemented using PySpark's machine learning library (pyspark.ml.clustering.KMeans) to group students based on their grades and study time.
3)Seaborn and Matplotlib:
Used for data visualization, particularly to create scatter plots that highlight the clustering results and provide insights into student performance distribution.
4)Pandas:
Used for converting Spark DataFrames into Pandas DataFrames for easier manipulation and plotting with visualization libraries.

5)Jupyter Notebook/Colab:

The project was executed in a notebook environment (likely Google Colab), providing an interactive platform for running PySpark code, data exploration, and visualizing results in real-time.

Data set used

The dataset used in this project is a student performance dataset containing information about students' demographics, study habits, and grades. The key columns in the dataset include:

1. student_id: A unique identifier for each student.
2. gender: The gender of the student (categorical feature).
3. study_time: The amount of time the student spends studying, categorized into ranges (e.g., "<2 hours", "2-5 hours", etc.).
4. G1, G2, G3: The grades obtained by students in three different assessments or grading periods.

Additional transformations were applied, such as creating an `average_grade` feature based on the three grading periods (G1, G2, and G3) and converting categorical features like `gender` and `study_time` into numeric format using encoding techniques. The dataset was cleaned, with null values in the grades being filled with 0.

Screen Shots

```
root
 |-- student_id: string (nullable = true)
 |-- gender: string (nullable = true)
 |-- study_time: string (nullable = true)
 |-- G1: integer (nullable = true)
 |-- G2: integer (nullable = true)
 |-- G3: integer (nullable = true)
```

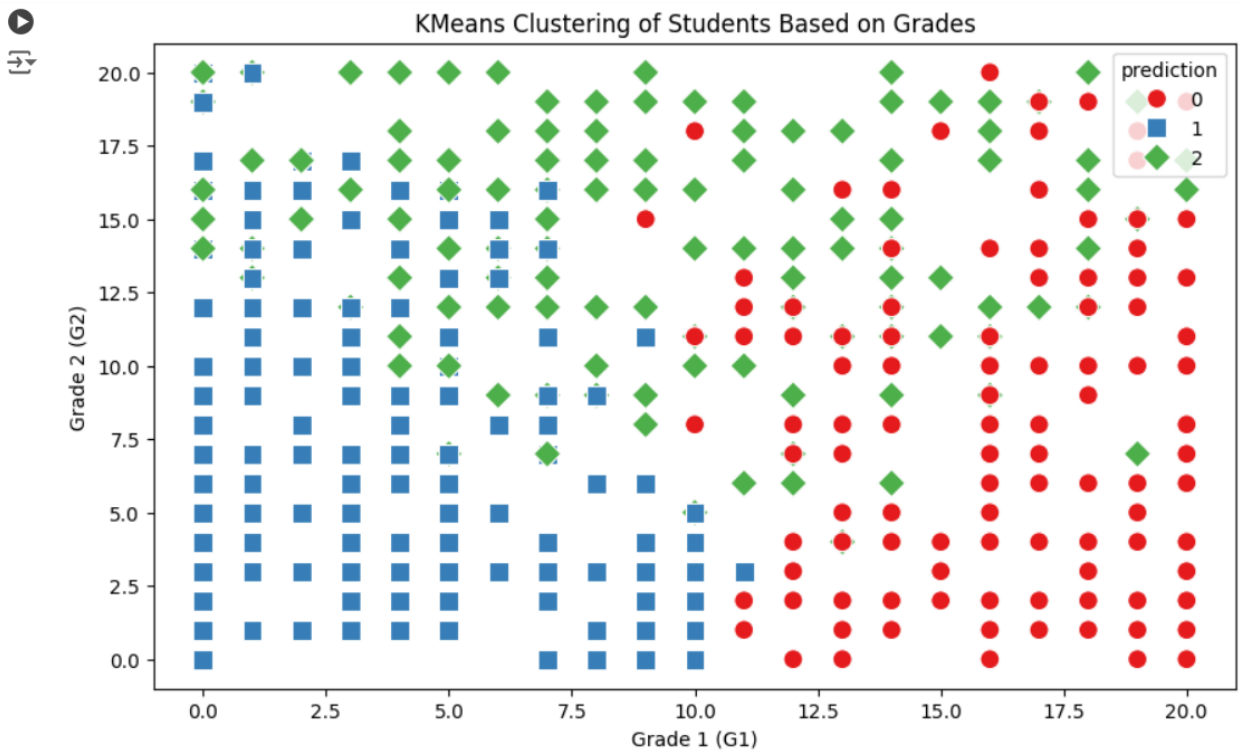
summary	student_id	gender	study_time	G1	G2	G3
count	500	500	500	500	500	500
mean	NULL	NULL	NULL	9.65	10.064	9.66
stddev	NULL	NULL	NULL	6.199060954162883	5.976566128252137	5.941295986229936
min	STUDENT_1	F	2-5 hours	0	0	0
max	STUDENT_99	M	>10 hours	20	20	20

student_id	gender	study_time	G1	G2	G3
0	0	0	0	0	0

```
root
 |-- student_id: string (nullable = true)
 |-- gender: string (nullable = true)
 |-- study_time: string (nullable = true)
 |-- G1: integer (nullable = true)
 |-- G2: integer (nullable = true)
 |-- G3: integer (nullable = true)
```

```
+----------+----------+--------------------+----------+
|student_id|study_time|            features|prediction|
+----------+----------+--------------------+----------+
| STUDENT_1|5-10 hours|[3.0,16.0,18.0,10.0]|         2|
| STUDENT_2|  <2 hours|  [0.0,8.0,0.0,12.0]|         1|
| STUDENT_3|  <2 hours|      (4,[2],[14.0])|         1|
| STUDENT_4| >10 hours|  [1.0,20.0,7.0,5.0]|         0|
| STUDENT_5|5-10 hours|  [3.0,19.0,1.0,8.0]|         0|
| STUDENT_6|  <2 hours|  [0.0,12.0,2.0,7.0]|         0|
| STUDENT_7| >10 hours|[1.0,15.0,13.0,10.0]|         0|
| STUDENT_8|5-10 hours|[3.0,12.0,13.0,18.0]|         2|
| STUDENT_9| >10 hours| [1.0,13.0,10.0,8.0]|         0|
|STUDENT_10| 2-5 hours| [2.0,2.0,15.0,16.0]|         2|
|STUDENT_11|  <2 hours|  [0.0,5.0,3.0,11.0]|         1|
|STUDENT_12| 2-5 hours|[2.0,17.0,19.0,13.0]|         2|
|STUDENT_13|5-10 hours| [3.0,18.0,4.0,14.0]|         0|
|STUDENT_14| 2-5 hours| [2.0,4.0,20.0,15.0]|         2|
|STUDENT_15| 2-5 hours|[2.0,14.0,13.0,19.0]|         2|
|STUDENT_16|5-10 hours|  [3.0,1.0,15.0,0.0]|         1|
|STUDENT_17| >10 hours|  [1.0,9.0,17.0,9.0]|         2|
|STUDENT_18| 2-5 hours|  [2.0,17.0,6.0,7.0]|         0|
|STUDENT_19| >10 hours| [1.0,12.0,9.0,17.0]|         2|
|STUDENT_20| 2-5 hours|   [2.0,4.0,6.0,0.0]|         1|
+----------+----------+--------------------+----------+
```

```
+----------+--------------------------------------------------+
|student_id|recommended_resources                             |
+----------+--------------------------------------------------+
|STUDENT_1 |[Statistics Course, Data Analysis Video]          |
|STUDENT_2 |[Advanced Calculus Book, Calculus Quiz]           |
|STUDENT_3 |[Advanced Calculus Book, Calculus Quiz]           |
|STUDENT_4 |[Basic Algebra Book, Introductory Algebra Video]  |
|STUDENT_5 |[Basic Algebra Book, Introductory Algebra Video]  |
|STUDENT_6 |[Basic Algebra Book, Introductory Algebra Video]  |
|STUDENT_7 |[Basic Algebra Book, Introductory Algebra Video]  |
|STUDENT_8 |[Statistics Course, Data Analysis Video]          |
|STUDENT_9 |[Basic Algebra Book, Introductory Algebra Video]  |
|STUDENT_10|[Statistics Course, Data Analysis Video]          |
|STUDENT_11|[Advanced Calculus Book, Calculus Quiz]           |
|STUDENT_12|[Statistics Course, Data Analysis Video]          |
|STUDENT_13|[Basic Algebra Book, Introductory Algebra Video]  |
|STUDENT_14|[Statistics Course, Data Analysis Video]          |
|STUDENT_15|[Statistics Course, Data Analysis Video]          |
|STUDENT_16|[Advanced Calculus Book, Calculus Quiz]           |
|STUDENT_17|[Statistics Course, Data Analysis Video]          |
|STUDENT_18|[Basic Algebra Book, Introductory Algebra Video]  |
|STUDENT_19|[Statistics Course, Data Analysis Video]          |
|STUDENT_20|[Advanced Calculus Book, Calculus Quiz]           |
+----------+--------------------------------------------------+
only showing top 20 rows
```



Conclusion

This project demonstrates the application of PySpark for analyzing and clustering student performance data. By utilizing KMeans clustering, students are grouped based on their study habits and grades, allowing for personalized learning recommendations. The results offer valuable insights into patterns of student performance, with recommendations tailored to each cluster (e.g., resources for algebra, calculus, or statistics). The use of distributed computing with PySpark ensures scalability for larger datasets, while the visualizations provide a clear understanding of the clustering results. This approach highlights the potential of data-driven educational tools for enhancing personalized learning experiences.