Project Documentation

Advanced Knowledge & IS Applications

Analysis Topic:

Student Academic Performance Prediction

Contents

1 Power BI Implementation

1.1 Dataset Description

The project utilizes the **"Exam Score Prediction"** dataset to analyze factors influencing student academic performance. The dataset includes granular records for individual students, organized into the following categories:

- **Identifier:** `student_id` (Unique identifier).
- **Demographics:** `age`, `gender`.
- **Academic Info:** `course`, `class_attendance` (%), `exam_score`, `passed`.
- **Habits:** `study_hours`, `sleep_hours`, `internet_access`, `study_method`.

1.2 ETL Steps (Power Query)

The Extract, Transform, and Load (ETL) process was executed using Power Query Editor to ensure data integrity:

1. **Data Extraction:** Loaded raw data from `Exam_Score_Prediction.xlsx`.
2. **Data Transformation:**
    - **Type Conversion:** Converted `student_id` to Text to prevent aggregation. Ensured `exam_score` and `study_hours` were set to Decimal Number for precision.
    - **Cleaning:** Checked for duplicates and null values.
    - **Formatting:** Capitalized column headers for professional presentation.

1.3 DAX Implementation

Data Analysis Expressions (DAX) were created to derive actionable metrics:

Key Measures:

- **Total Students:** `COUNTROWS(Exam_Score_Prediction)`
- **Average Score:** `AVERAGE(Exam_Score_Prediction[exam_score])`
- **Success Rate:** Percentage of students marked as "Passed".

Calculated Columns:

- **Performance Level:** Classifies scores into Low (< 60), Medium ($60 - 79$), High ($80 - 89$), and Excellent (≥ 90).
- **Attendance Status:** Segments attendance into Low, Moderate, and High groups.

1.4 Dashboard Visualizations

The report consists of three main pages:

- **Overview:** Features KPI cards (Total Students, Success Rate) and a Pie Chart for Pass/Fail distribution.
- **Analysis:** Contains a Scatter Chart correlating Study Hours vs. Exam Score and a Slicer for Internet Access.
- **Details:** A tabular view of student records.

Interactive Features: Includes Bookmarks for resetting filters and Report Page Tooltips showing average scores by gender on hover.

2 KNIME Machine Learning Workflow

2.1 Workflow Overview

The machine learning phase was executed in KNIME to predict student success (*Passed*). To ensure a realistic model, the **Exam Score** was explicitly excluded from the training data to prevent data leakage.

Preprocessing Pipeline:

- **Missing Values:** Imputed using the Mean method.
- **Normalization:** Applied Min-Max Normalization (scale 0-1) for *Study Hours* and *Attendance*.
- **Partitioning:** Data split into **80% Training** and **20% Testing**.

2.2 Visualizations (EDA)

Exploratory Data Analysis was conducted to understand feature distributions.

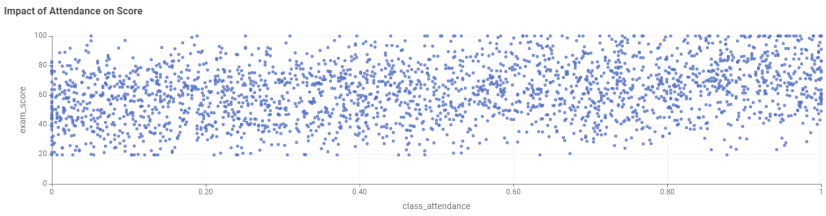

Figure 1: *Correlation between Attendance and Exam Score.*

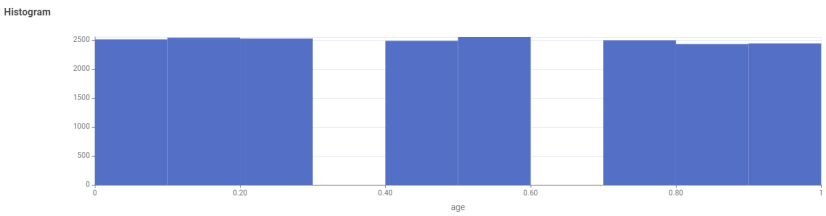

Figure 2: *Distribution of Student Ages.*

Figure 3: *Gender Distribution.*

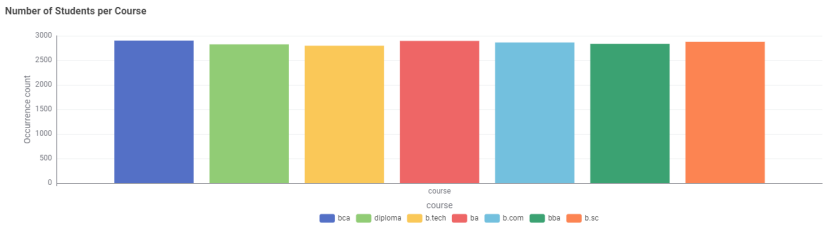

Figure 4: *Student enrollment per course.*

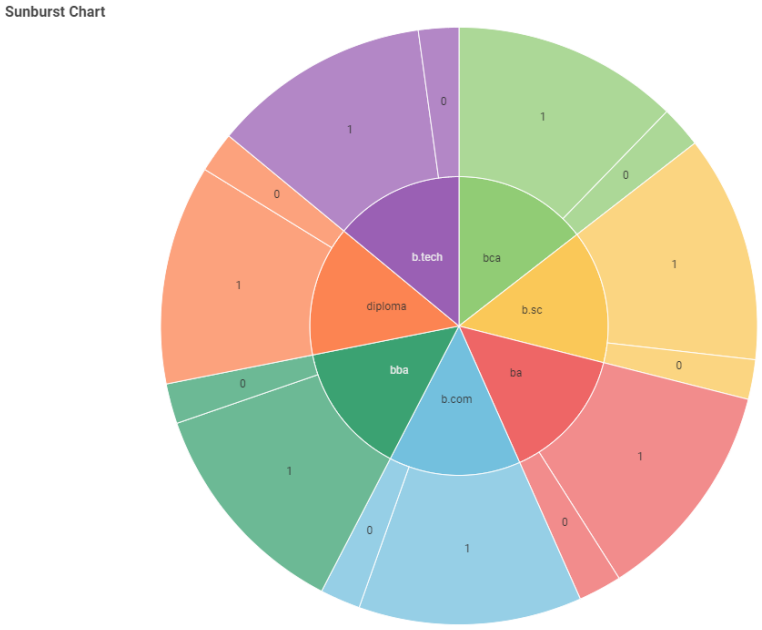

Figure 5: *Academic Success Rate by Course Hierarchy*

2.3 Machine Learning Models

Two classification models were trained:

1. **Decision Tree:** A baseline model providing interpretability via decision rules.
2. **Random Forest:** An ensemble model using multiple trees to capture complex patterns (e.g., Study Method impact).

2.4 Evaluation & Comparison

Both models were evaluated on the test dataset. The **Random Forest** outperformed the Decision Tree.

Table 1: *Model Performance Metrics*

Metric	Decision Tree	Random Forest
Accuracy	84.0%	**86.2%**
True Positives (Passed)	2,705	**2,783**
False Positives	346	**216**
False Negatives	294	336
Cohen's Kappa	0.566	**0.617**

Comparison Analysis: While both models performed well, the Random Forest achieved higher accuracy (**86.2%**) and significantly better precision (fewer False Positives). This indicates it is more reliable in identifying truly successful students without misclassifying failing students.

2.5 Conclusion

The **Random Forest Classifier** is selected as the final model for this project. Its ensemble nature allowed it to capture subtle non-linear relationships between *Sleep Quality*, *Study Habits*, and *Success* that the single Decision Tree missed.