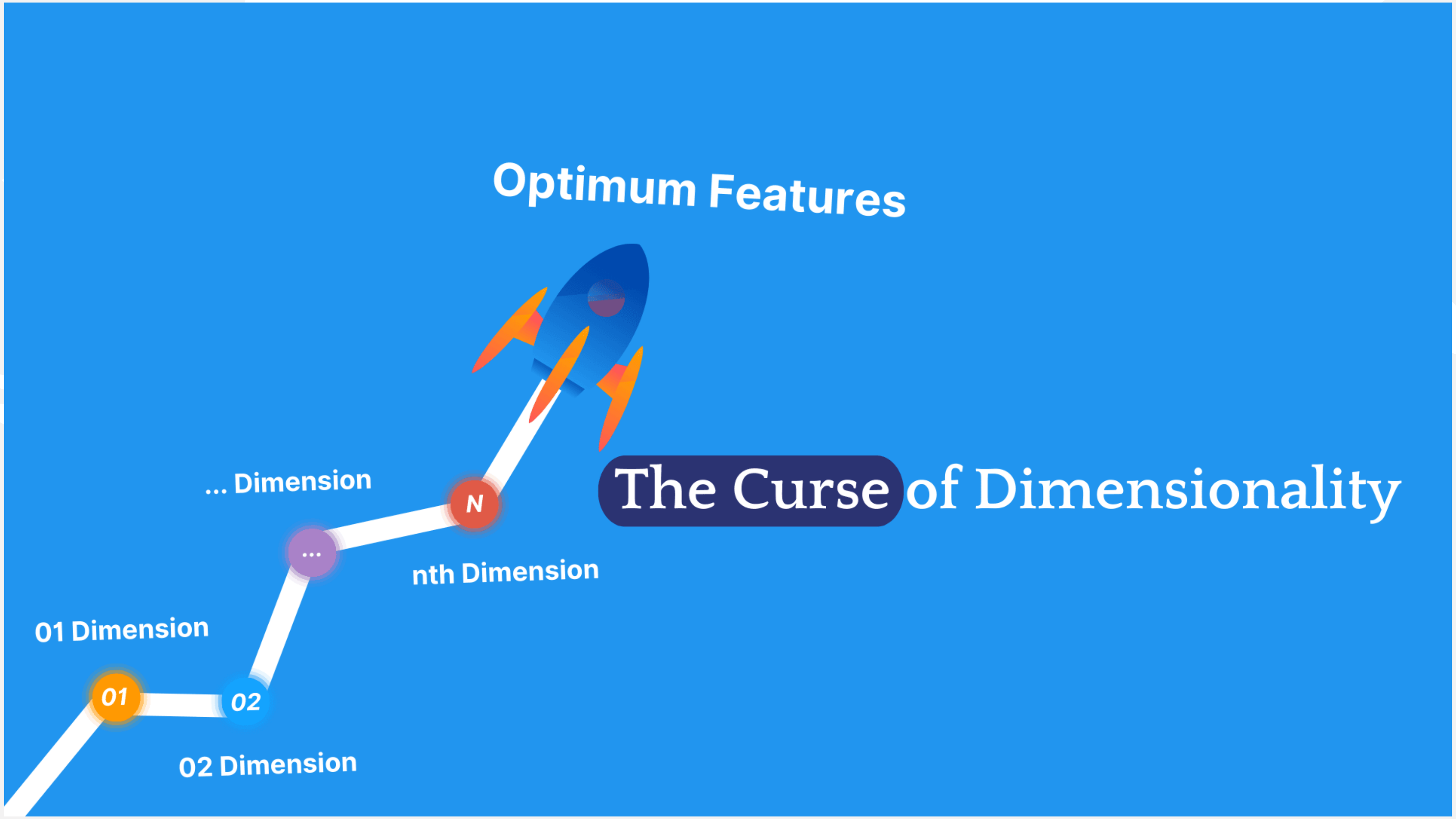
Project on Dimentionality Reduction



Problem Statement

Artifact Submission

Recommendation

Data Description

Challenge Evaluation Criteria

Vibelines & Bounties

Problem Statement - Student Performance Clustering

Goal

The primary goal of this project is to use **dimensionality reduction techniques** to analyze high-dimensional student performance data, aiming to uncover natural groupings or segments among the students. These segments can reveal hidden patterns of performance (e.g., students who excel in STEM but struggle in humanities, or vice-versa) that are not immediately obvious from raw grade averages.

Dataset

Student Grades Data: The dataset contains individual student grades across multiple subjects, including: **Math, Science, Computer Science (CS), Band, English, History, Spanish, and Physical Education (PhysEd)**, along with a unique student_id.

Objectives

Dimensionality Reduction: Apply **Principal Component Analysis (PCA)** to reduce the feature space (grades across all subjects) to 2 or 3 principal components.

Visualization Comparison:

- Plot the student data using the reduced PCA components.
- Apply a **non-linear technique** like **t-SNE** or **UMAP** to the data and plot the results.

Analysis & Interpretation: Visually compare the outputs of PCA and t-SNE/UMAP to **identify clusters** of students. Interpret the meaning of these clusters (e.g., "Cluster A" represents students who score consistently high across all subjects).

Deliverable

A report and visualization showing the effectiveness of PCA versus t-SNE/UMAP in separating and visualizing student performance clusters, leading to actionable insights for academic advising or curriculum adjustment.

Data Description

Student Grades Data: The dataset contains individual student grades across multiple subjects, including: **Math, Science, Computer Science (CS), Band, English, History, Spanish, and Physical Education (PhysEd)**, along with a unique student_id.

Column Name	Description
student_id	Unique identifier for each student.
math	Grade in Mathematics.
science	Grade in Science.
cs	Grade in Computer Science.
band	Grade in Band (or a similar elective).
english	Grade in English.
history	Grade in History.
spanish	Grade in Spanish (or a foreign language).
physed	Grade in Physical Education.

Artifact Submission

Your submission must include the following five artifacts, all packaged within a single GitHub repository.

1. Jupyter Notebook (.ipynb) (Add separate ipynb files for PCA and t-SNE)

This is the core of your submission. Your Jupyter Notebook should be a complete, well-documented narrative of your data analysis journey. It must include:

- **Detailed Explanations:** Use Markdown cells to explain your thought process, the questions you are trying to answer, and the insights you've uncovered.
- **Clean Code:** The code should be well-structured, easy to read, and free of unnecessary clutter.
- **Comprehensive Comments:** Use comments to explain complex logic and the purpose of different code blocks.
- **Key Visualizations:** All visualizations should be clear, properly labeled, and directly support your findings.

2. Presentation (.pptx or .pdf)

Create a compelling presentation that summarizes your team's analysis and key findings. This presentation should serve as your final pitch. It must include:

- **Executive Summary:** A concise overview of your findings.
- **Key Insights:** The most important takeaways from your analysis.
- **Data-Driven Recommendations:** Actionable steps that can be taken based on your insights.
- **Supporting Visualizations:** A selection of your best visualizations to illustrate your points.

3. README File (.md)

The README file is the first thing we'll look at. It should serve as a quick guide to your project and provide essential details. It must include:

- **Project Title :**
- **Brief Problem Statement:** A summary of the project and your approach.
- **Summary of Findings:** A bullet-point summary of your most significant insights.

4. Attached Dataset

Please include the original dataset (.csv or other format) within your repository. This ensures the judges can reproduce your analysis without any issues.

5. GitHub Repository

Your final submission will be your GitHub repository. The repository name **must follow this exact format:**
Dimentionality_Reduction_ProjectName_TMP

Challenge Evaluation Criteria

Criteria Name	Criteria weight
Data Understanding and Exploratory Data Analysis	20%
Data preprocessing and feature engineering	25%
Model building and evaluation	30%
Business Recommendation	15%
Coding guidelines and standards	10%

Recommendation for PCA Implementation

1. Data Preprocessing:

Selecting only the numerical columns representing grades (math, science, cs, band, english, history, spanish, physed).

Crucially, performing feature scaling on these grade columns. Since grades are typically on the same scale (e.g., 0-100), standardizing them (mean 0, variance 1) ensures that each subject contributes equally to the principal components, preventing subjects with slightly larger score ranges from dominating the analysis.

2. PCA Implementation:

Applying the PCA algorithm to the scaled student grade data.

Calculating the **explained variance ratio** for each principal component to understand how much of the total variation in student performance each new component captures.

3. Determining the Optimal Number of Components:

Using a **scree plot** and the cumulative explained variance ratio to decide how many principal components are sufficient to represent most of the information in the original eight subject grades (e.g., aiming to capture 85-95% of the total variance with fewer components).

4. Data Transformation & Visualization:

Transforming the original high-dimensional grade data into a lower-dimensional space (e.g., 2 or 3 principal components).

Creating a scatter plot of the transformed data, where each point represents a student. This visualization can reveal natural groupings of students based on their overall academic strengths or weaknesses, which might not be obvious from looking at individual grades.

5. Interpretation:

Analyzing the principal components to understand what underlying "latent factors" they represent. For example, PC1 might represent overall academic aptitude, while PC2 might differentiate between "STEM-focused" and "humanities-focused" students.

Observing student clusters in the reduced-dimensional space to identify different types of learners or performance groups.

Recommendation for t-SNE Implementation

1. Data Preprocessing:

- Selecting only the numerical columns representing grades (math, science, cs, band, english, history, spanish, physed).
- **Crucially, performing feature scaling** on these grade columns. While grades might be on a similar scale (e.g., 0-100), standardizing them (mean 0, variance 1) is a best practice for t-SNE, ensuring that all subjects contribute equally to the similarity calculations.

2. t-SNE Implementation:

- Applying the t-SNE algorithm to the scaled student grade data to reduce its dimensionality, typically to 2 components for direct visualization.
- Careful consideration and tuning of the perplexity hyperparameter will be necessary to achieve a meaningful and stable visualization.

3. Visualization:

- Creating a scatter plot of the 2-dimensional t-SNE output. Each point on the plot will represent a student. The relative proximity of points will indicate similarity in their overall academic performance profiles.
- The points can potentially be colored or labeled based on other available student attributes (if any, beyond grades) to add further context to the clusters.

4. Interpretation:

- Analyzing the resulting visualization to identify if distinct clusters of students emerge. For example, one cluster might represent students strong in STEM subjects, another in humanities, and yet another representing students with balanced performance or those struggling across the board.
- Understanding what common characteristics define students within each visually identified cluster.
- This visual insight can help educators to:
  - Identify student groups who might benefit from differentiated teaching strategies.
  - Spot outliers (students with unusual performance profiles).
  - Gain a holistic understanding of academic performance patterns across the student body.

The outcome of this project will be a powerful visual representation that simplifies the complex, multi-dimensional student grade data, making it easier to identify and understand underlying academic groupings and patterns.

VIBELINES

Kickoff Vibes - T
Wrapoff Vibes - T + 7
Spotlight Vibes - T + 14

Lets Go