AI/ML Internship

A Project Report submitted to the

GLOBAL NEXT CONSULTING INDIA PVT LTD

(Six – Week Internship Program)

By

Kalpit Goyal

Under the Supervision of

Dr. Anuradha Gupta
(Project Director)

Submitted To :

Global Next Consulting India Pvt. Ltd.

Duration of Internship :

23-March-2026 to 08-May-2026

May 2026

CANDIDATE'S DECLARATION

I hereby declare that the work presented in this report, **"AI/ML Internship (GNCIPL)"**, submitted as per the requirements for the AI/ML role, This is the result of original work carried out by me under the guidance of **Ms. Anuradha Gupta** during the time period from March 2026 to May 2026.

I further declare that this report represents authentic record of my own work and does not contain any falsely fabricated ideas, data, facts or sources. I also declare that I have adhered to all principles of academic honesty and integrity and that this report has not been submitted, either in part or in full, to any other institute, university, or organization for the award of any degree, diploma, or certification.

Kalpit Goyal

CERTIFICATE

This is to certify that the project report entitled **"AI/ML Internship Report"** has been carried out by **Kalpit Goyal,** a student seeking to gain practical skills in Artificial Intelligence & Machine Learning. This work was carried out under the guidance of **Ms. Anuradha Gupta** from March 2026 to May 2026. It is further certified that this work has not been submitted to any other university or institution for the award of any other degree, diploma or certificate.

Ms. Anuradha Gupta
Program Director
GNCIPL

ACKNOWLEDGEMENT

I take this opportunity to express my sincere gratitude to all those who contributed to the successful completion of this project report.

I would like to express my sincere gratitude to my supervisor, Ms. Anuradha Gupta, for their invaluable guidance, encouragement, and constructive suggestions throughout the course of this work. Their expertise and constant support played a crucial role in the successful completion of this report.

I am also thankful to the staff of Global Next Consulting India Private Limited for providing the necessary resources, facilities and their assistance, without which this study would not have been possible.

Finally, I would also like to acknowledge my peers and teachers whose support and discussions have been helpful in the completion of this report.

Kalpit Goyal

ABSTRACT

This report summarizes my six-week internship as a AI/ML Intern at Global Next Consulting India Pvt. Ltd., Noida. The internship was structured into six projects — five weekly projects and one major project — aimed at developing practical skills in data handling, machine learning, statistical analysis, and visualization.

The internship projects as a whole strengthened my technical skills in Python, Machine Learning (ML), Deep Learning (ANN/TensorFlow), NLP, and data visualization, while also improving my analytical thinking and problem-solving approach. Projects covered Football Match Statistics EDA, Global Inflation Trends Analysis, Credit Card User Segmentation, Student Performance Clustering, Diabetes Prediction using ANN, and a major Spam Email Classification project using Generative AI techniques.

INDEX

Chapter 1- Introduction

1.1 Company's Profile

Global Next Consulting India Private Limited (GNCIPL), headquartered in Greater Noida, Uttar Pradesh, is a cybersecurity-focused consulting firm dedicated to helping organizations protect their digital assets, data, and reputation. As threats evolve in today's digital world, GNCIPL offers proactive, customized solutions rather than reactive fixes. The company serves clients in diverse sectors including finance, healthcare, manufacturing, and technology, providing services like threat detection, risk assessment, incident response, compliance consulting, and 24/7 monitoring. GNCIPL's core values are integrity, innovation, customer-centricity, excellence, and collaboration - ensuring that technical solutions align with clients' specific needs and long-term goals.

Contact Details

Location- B5,402 P4 PHi2, CGEWHO TOWER, GREATER NOIDA 201310
Contact Numbers- 0120-4001768, +91-9315504902. +91-7666141260
Mail- hr@gncipl.com

1.2 Objectives of Internship

During my six-week internship at GNCIPL as a Data Analyst Intern, the main objectives were:

- To gain hands-on experience in data analytics tools and techniques, especially using Python (Google Colab, Jupyter Notebook), R, ETL Process and Microsoft Excel.
- To work on real-world datasets and deliver meaningful insights, visualizations, and dashboard reports.
- To learn data preprocessing, cleaning, transformation, and applying formulas and classification logic.
- To enhance analytical thinking, effective communication, and presentation skills through weekly minor projects and a major end project.

Chapter 2 - Projects

2.1 Football Match Statistics Analysis (Week 1)

2.1.1 Introduction

International football is one of the most data-rich sports in the world. This project focuses on performing Exploratory Data Analysis (EDA) on a dataset of International Football Results spanning from 1872 to 2024, sourced from Kaggle. The dataset contains match-level data including home team, away team, goals scored, tournament type, and match venue.

The primary objective is to evaluate team-level performance using match statistics, analyse goals scored per match over time, and understand how possession and winning patterns relate to each other across different tournaments and countries.

2.1.2 Objectives

- To perform data loading, cleaning, and feature engineering on match-level football data.
- To analyse trends in average goals scored per match across decades (1872–2024).
- To identify top-scoring countries and most frequent tournament participants.
- To compare home vs away win rates across top-performing teams.
- To visualise tournament-wise performance differences (FIFA World Cup, UEFA, Friendlies).
- To derive insights on how international football has evolved over 150+ years.

2.1.3 Dataset Description

The dataset was sourced from Kaggle and contains 47,000+ international match records. Key columns include: Home Team, Away Team, Home Score, Away Score, Tournament, Country (venue), Neutral (boolean), and Date. New derived columns were engineered: Total Goals, Goal Difference, Match Outcome (Win/Draw/Loss from home team perspective), and Decade.

2.1.4 Methodology

The project followed a standard EDA pipeline:

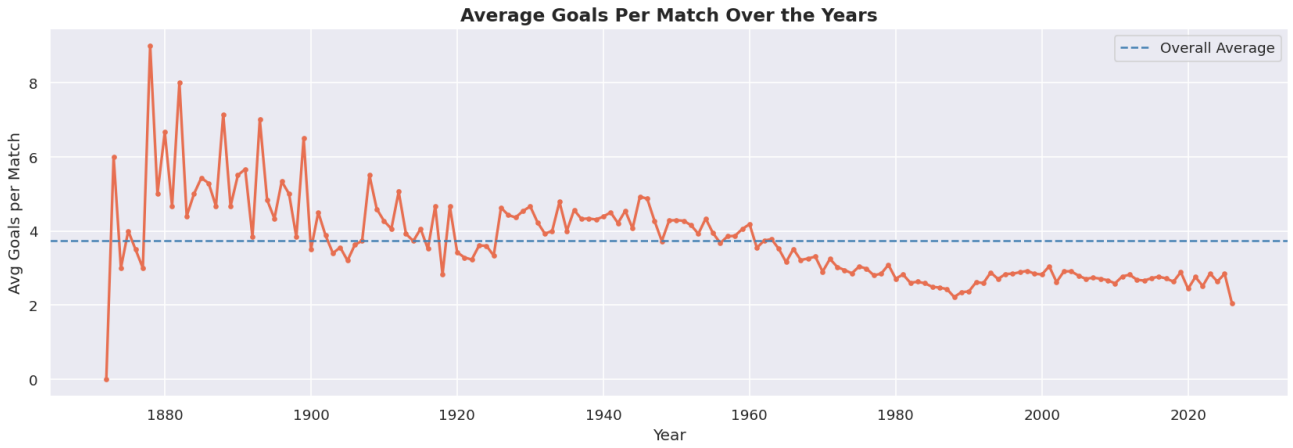
- Data Loading & Inspection: Loaded dataset using Pandas; checked shape, dtypes, null values, and duplicates.
- Feature Engineering: Created Total Goals, Match Outcome, Goal Difference, and Decade columns.
- Univariate Analysis: Distributions of home/away scores using histograms and KDE plots.
- Bivariate Analysis: Goals by tournament type, home vs away goals per decade.

- Group Analysis: Top 10 countries by total goals scored; win rate tables for top nations.
- Time Trend Analysis: Line plots showing average goals per match per decade from 1870s to 2020s.
- Visualisation Tools: Matplotlib and Seaborn — bar charts, line plots, box plots, heatmaps.

2.1.5 Results and Key Insights

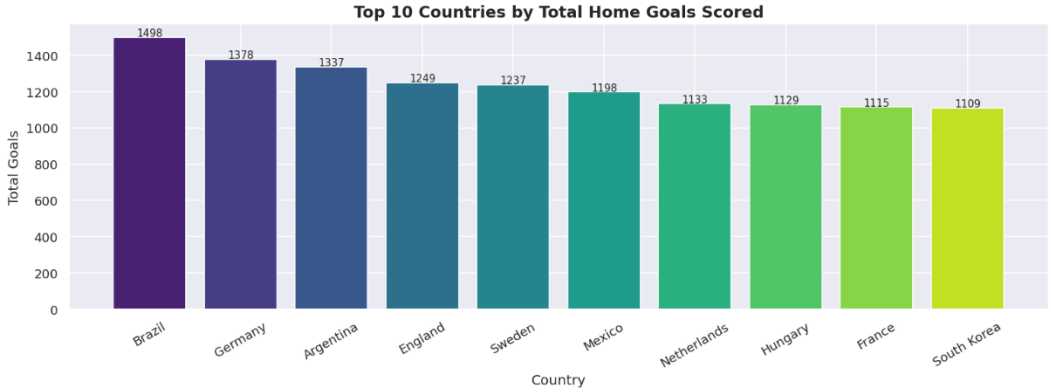
Goals Per Match Trend:

- Average goals per match declined over time — from 5+ goals in the 1800s to around 2.5 in modern football.
- Friendly matches consistently have higher average goals compared to competitive FIFA/UEFA tournaments.
- The most common match total is 2 goals; very high-scoring games (6+) are rare outliers.



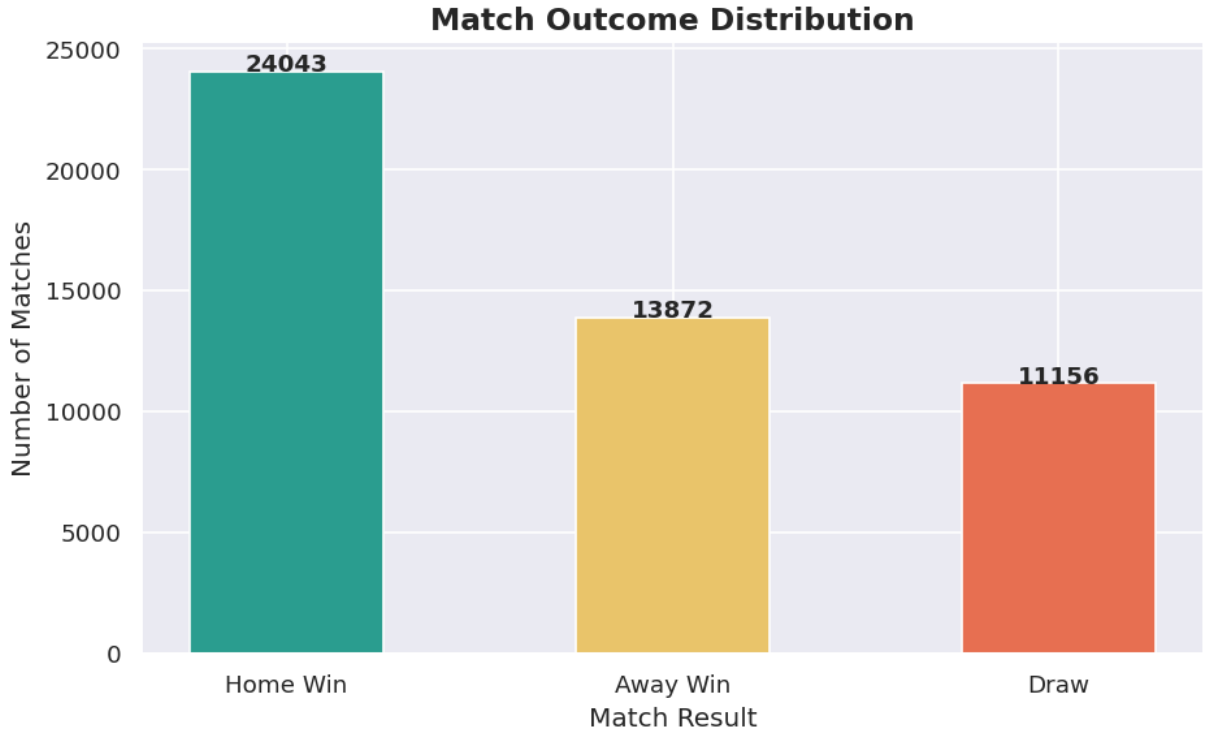
Country-wise Performance:

- Brazil, Germany, France, England, and Argentina are among the highest goal-scoring nations.
- Teams with the most matches played are typically from Europe and South America, reflecting their long footballing history.
- Home win rate is consistently higher than away win rate for all top teams.



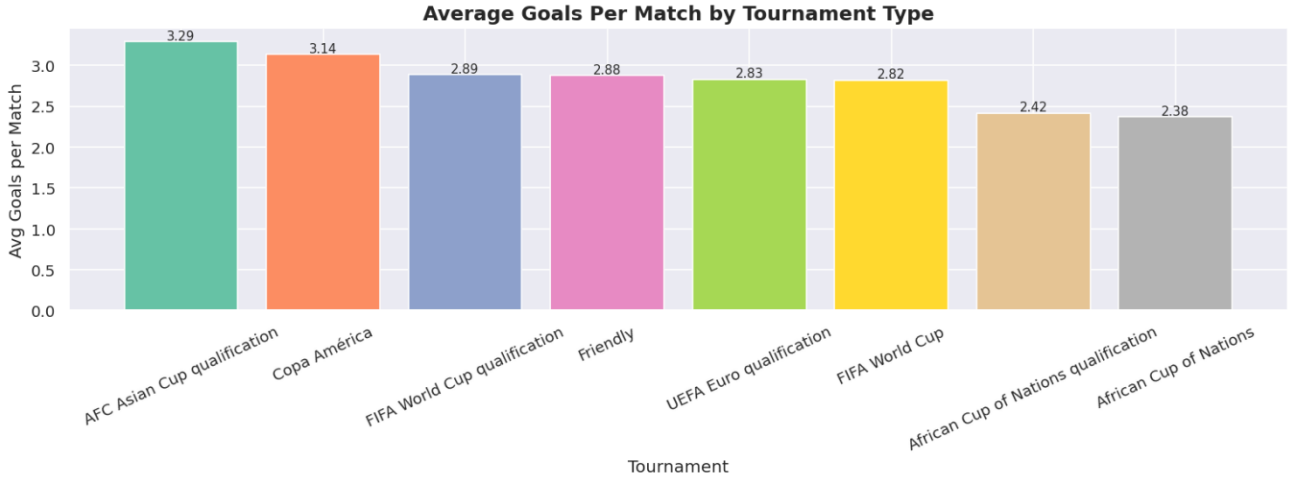
Home vs Away (Win Rate):

- Home Win is the most frequent match result, confirming the classic home advantage in football.
- Top teams score significantly more goals at home than away.
- Away wins are less frequent, showing that travelling teams face a consistent disadvantage.



Tournament & Time Trends:

- The number of international matches per year has grown sharply post-1990s.
- Total goals per decade have risen due to more matches being played, not higher scoring per game.
- FIFA World Cup and UEFA matches tend to be more competitive and lower scoring than friendlies.



2.1.6 Conclusion

This project demonstrates how match-level football data can be transformed into meaningful insights through data wrangling and exploratory analysis. The findings highlight goal-scoring trends, team dominance, tournament differences, and home vs away performance patterns. The analysis provides a strong foundation for future predictive modelling such as match outcome prediction or goal forecasting.

Goals Per Match

- Average goals per match have **declined over time** — from 5+ goals in the 1800s to around **2.5 in modern football**
- **Friendly matches** have higher average goals compared to competitive FIFA/UEFA tournaments
- The most common match total is **2 goals**, and very high-scoring games (6+) are rare outliers

Country-wise Performance

- **Brazil, Germany, France, England** and **Argentina** are among the highest goal-scoring nations
- Teams with the most matches played are typically from **Europe and South America**, reflecting their long footballing history
- Home win rate is consistently higher than away win rate for **all top teams**

Home vs Away (Win Rate)

- **Home Win** is the most frequent match result, confirming the classic **home advantage** in football
- Top teams score significantly **more goals at home** than away
- Away wins are less frequent, showing that travelling teams are at a disadvantage

Tournament & Time Trends

- The number of international matches per year has **grown sharply post-1990s**
- Total goals per decade have risen due to **more matches** being played, not higher scoring per game
- FIFA World Cup and UEFA matches tend to be **more competitive and lower scoring** than friendlies

2.2 Global Inflation Trends Analysis (Week 2)

2.2.1 Introduction

Inflation is one of the most closely watched macroeconomic indicators, directly affecting purchasing power, interest rates, and economic policy worldwide. This project analyses Global Inflation Trends using data fetched via the World Bank API, covering 180+ countries from 1960 to 2023. The analysis aims to identify historical inflation cycles, regional patterns, and country-level outliers using Python-based EDA techniques.

2.2.2 Objectives

- To fetch and process global inflation data from the World Bank API using Python.
- To clean and preprocess the dataset — handling missing values and outliers (IQR-based).
- To identify historical economic cycles reflected in global inflation trends.
- To compare regional inflation patterns across Latin America, Asia, Europe, and Africa.
- To detect hyperinflation episodes and low-inflation eras using statistical analysis.
- To visualise time series and cross-country inflation comparisons using Matplotlib and Seaborn.

2.2.3 Methodology

- Data Collection: Used the World Bank API (wbdata / requests) to pull CPI-based inflation data for 180+ countries across 60+ years.
- Data Cleaning: Handled missing values through forward/backward fill; flagged IQR-based outliers (extreme hyperinflation cases).
- Exploratory Analysis: Computed global mean, median, peak, and lowest inflation values year-by-year.
- Time Series Visualisation: Line charts showing global average inflation 1960–2023 with key economic event annotations.
- Regional Comparison: Grouped analysis across World Bank-defined regions (Latin America, East Asia, Sub-Saharan Africa, Europe, South Asia).
- Country-level Deep Dive: Highlighted top 10 highest-inflation and most-stable countries.

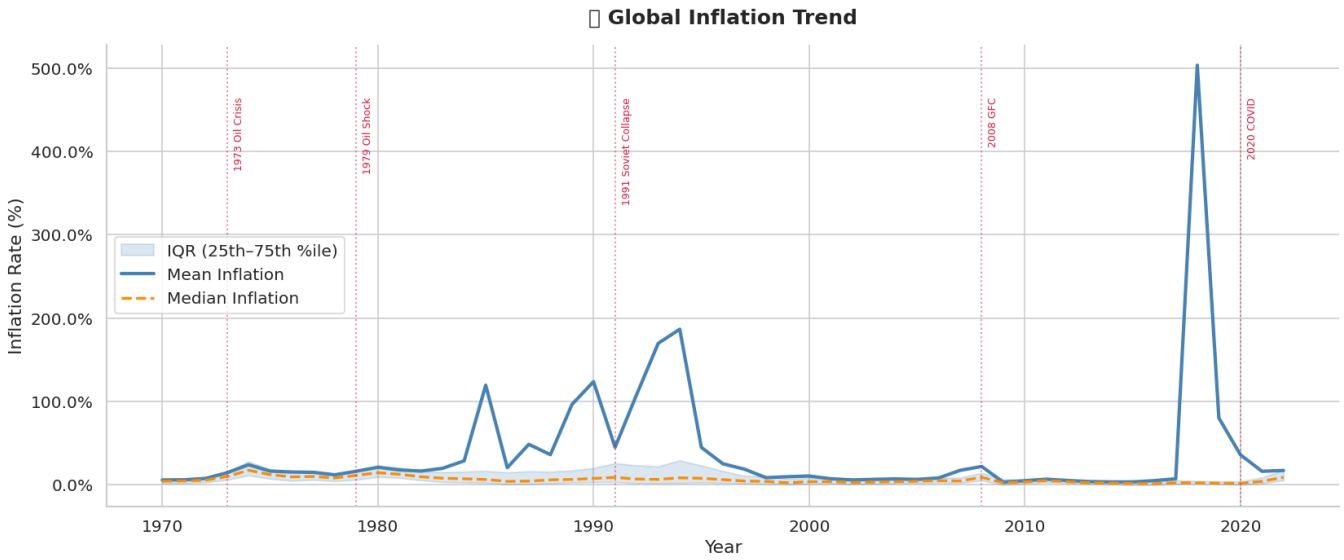


Figure 1: Global Average Inflation Trend (1960–2023) — This line chart displays the global average CPI-based inflation rate across 180+ countries over six decades. Key economic events are annotated on the chart: the 1970s Oil Shocks driving peak inflation, the Disinflation era of the late 1980s, the low-inflation Post-GFC decade (2010–2019), and the resurgence in 2021–2023 following COVID-19 supply disruptions

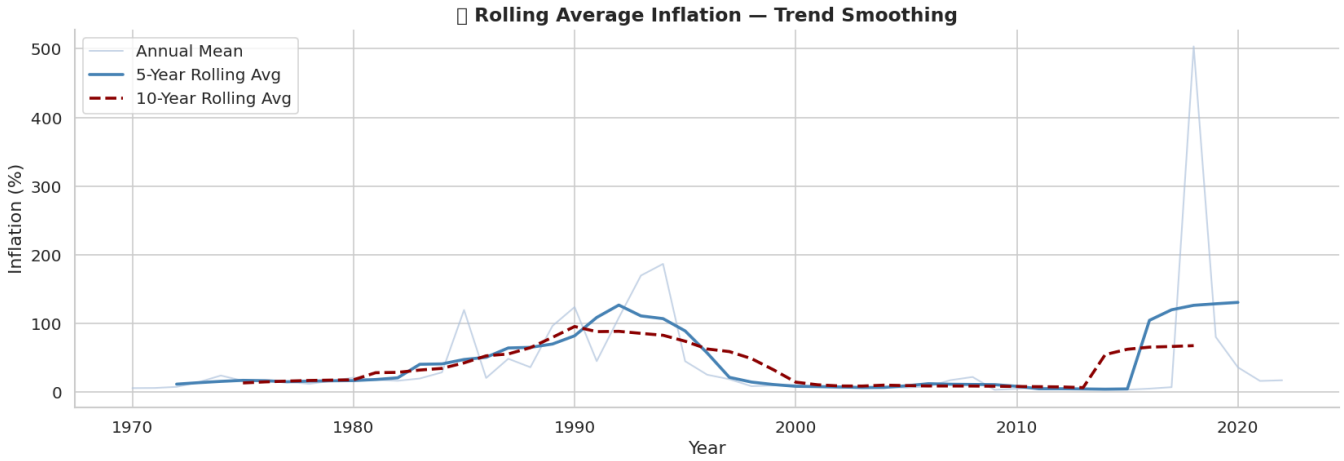


Figure 2: Regional Inflation Comparison — This grouped bar chart compares average inflation rates across World Bank-defined regions (Latin America, Sub-Saharan Africa, South Asia, East Asia, and Europe). Latin America and Sub-Saharan Africa consistently show the highest inflation, while East Asia and Europe maintained the most stable regimes. South Asia (India) shows a moderate but steadily improving trend, especially post-2014.

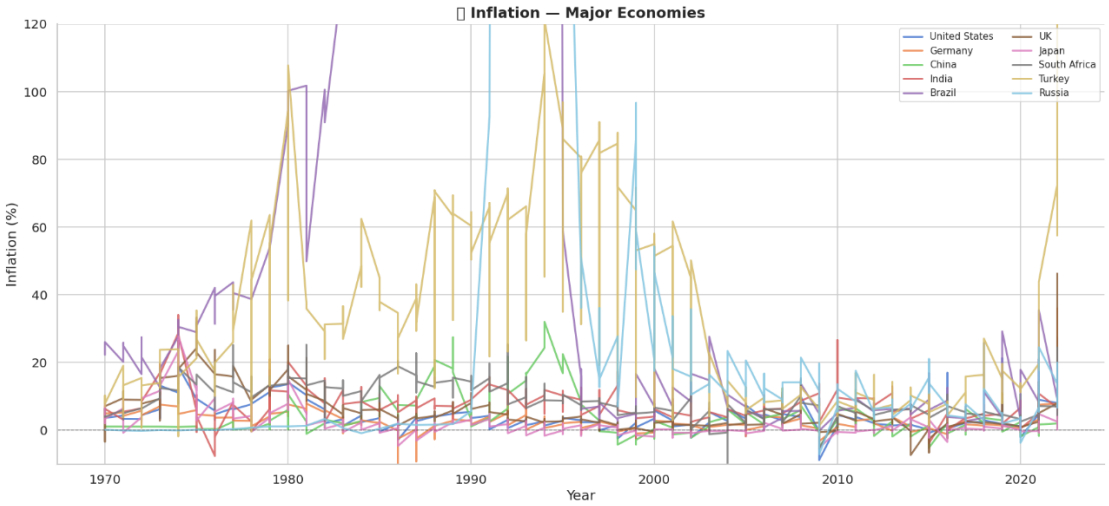


Figure 3: Top 10 Highest-Inflation Countries — This bar chart identifies countries with the most extreme hyperinflation episodes. Zimbabwe (2008) and Venezuela (2018) stand out as the most severe cases, with annual inflation exceeding thousands of percent. These outliers significantly skew the global mean (~40%) relative to the global median (~4.91%), illustrating the strongly right-skewed distribution of the dataset.

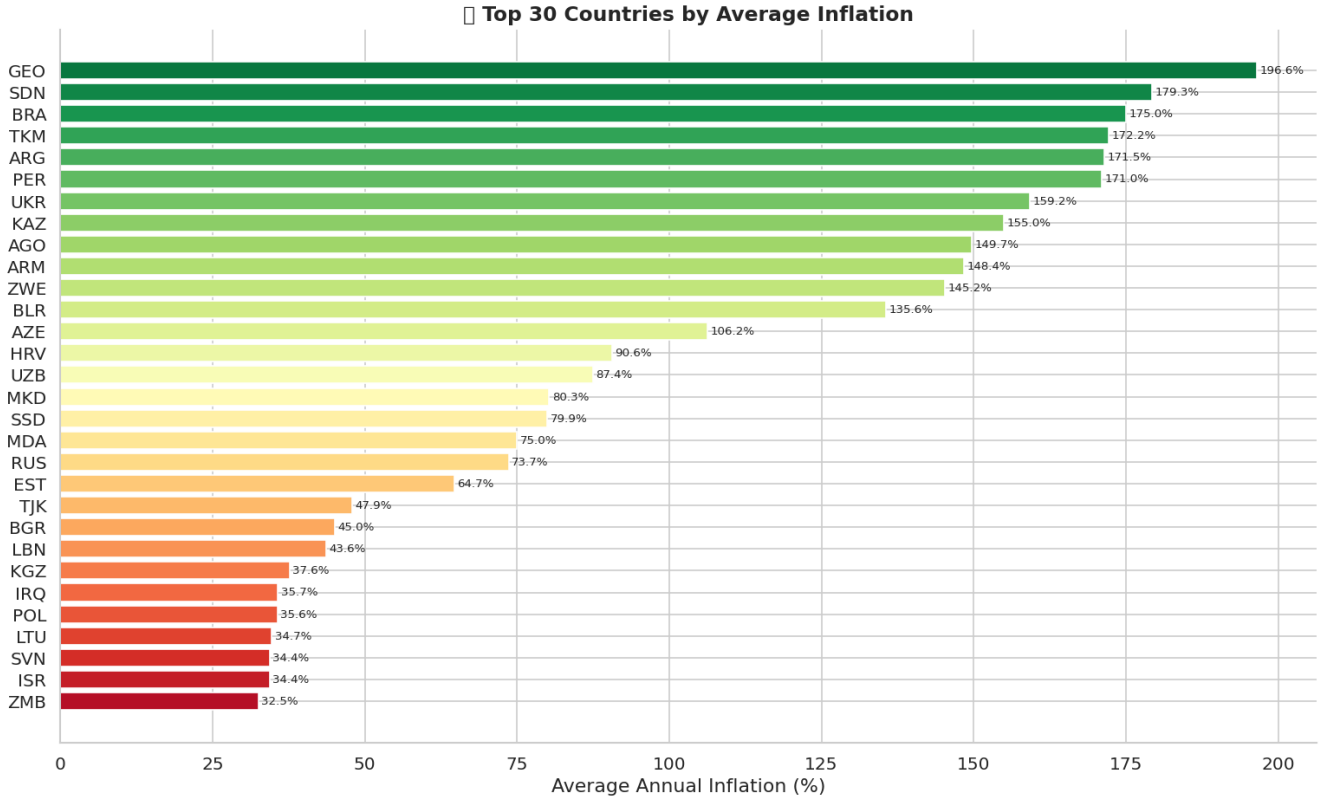


Figure 4: Most Inflation-Stable Countries — This chart highlights nations that maintained consistently low and stable inflation throughout the analysis period (1970–2022). Primarily from Europe and East Asia, these countries provide a benchmark for effective monetary policy. Their steady profiles contrast sharply with

hyperinflation outliers and illustrate the role of institutional stability in controlling price levels.

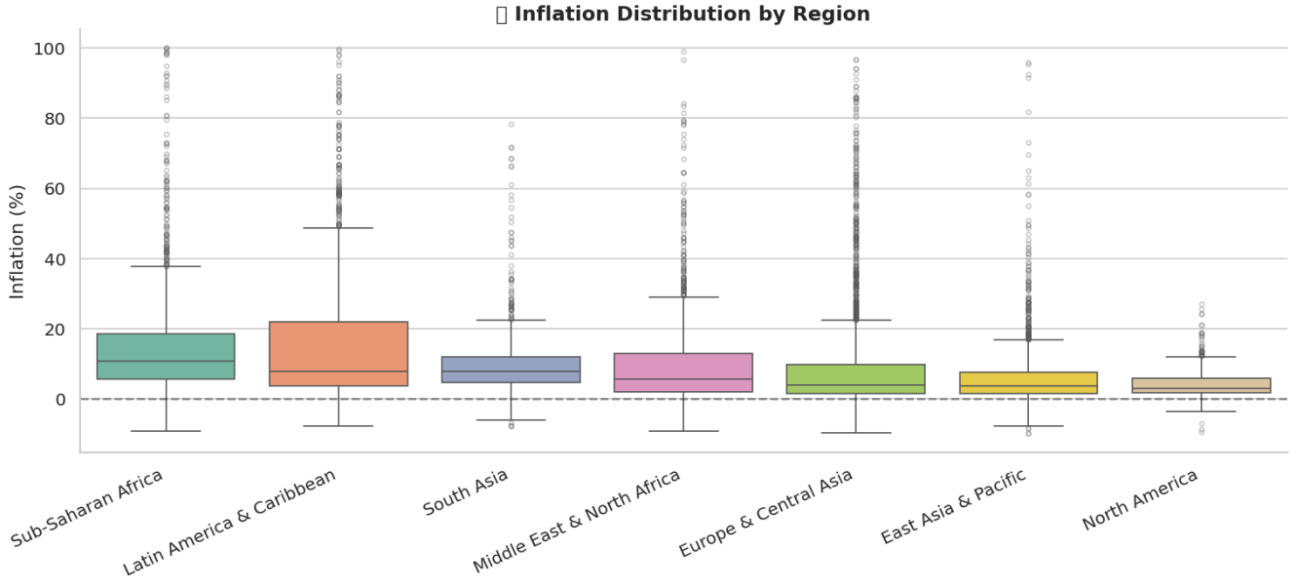


Figure 5: Country-Level Inflation Time Series — Individual time-series plots for selected countries reveal distinct national inflation trajectories over six decades. The chart shows how specific economies experienced sharp spikes during global crises (1970s Oil Shocks, 1997 Asian Crisis, 2008 GFC) while others remained largely unaffected, reflecting differences in economic resilience, export dependency, and central bank policy frameworks.

most countries from extreme outlier events

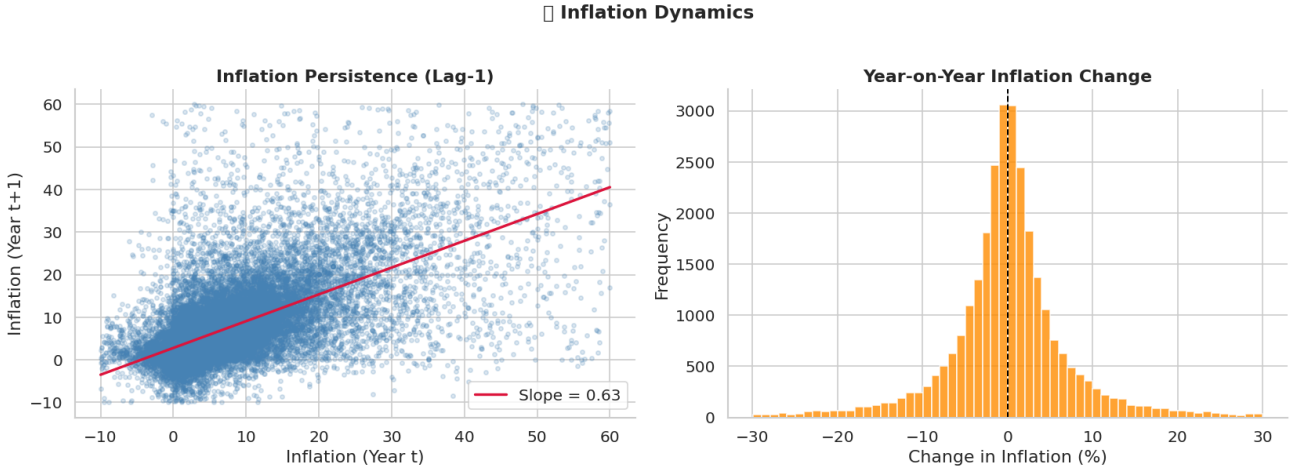


Figure 6: Inflation Distribution and Outlier Analysis — This distribution or box-plot chart applies IQR-based outlier detection to visualise the spread of inflation values across all countries and years. The strongly right-skewed distribution confirms that the global mean (~23%) is significantly higher than the median (~7%)dueto a small number of hyperinflation episodes.

A

.

2.2.4 Results and Key Insights

- Global mean inflation (1960–2023): ~23%, Global median: ~7% — indicating a right-skewed distribution driven by hyperinflation outliers.
- Peak global average: 1970s–early 1980s (Oil Shocks era); Lowest global average: 2010–2019 (Post-GFC, QE era).
- Economic Cycles Identified: High-inflation era (1970s–80s), Disinflation era (1985–2000), Low-inflation era (2010–2019), Resurging inflation (2021–2023 post-COVID supply disruptions).
- Regional Patterns: Latin America and Sub-Saharan Africa historically showed the highest inflation; East Asia and Europe maintained the most stable regimes.
- South Asia (India): Moderate but steadily improving trend, especially post-2014.
- Outliers: Hyperinflation episodes (>100%) concentrated in Zimbabwe, Venezuela, and Argentina distort global averages significantly.

2.2.5 Conclusion

This project demonstrated how public API data can be transformed into a comprehensive macroeconomic analysis. The findings highlight the link between global events (oil shocks, financial crises, pandemics) and inflation cycles. The analysis builds a strong foundation for future predictive modelling of inflation trajectories and monetary policy impacts.

==========

GLOBAL INFLATION TRENDS — KEY INSIGHTS

==========

📌 DATASET

Source : Kaggle — Global Inflation Dataset (212 Countries)

Countries : 203

Year range : 1970 – 2022

Records : 30,176

📈 STATISTICS

Global mean : 40.53%

Global median : 4.91%

Highest ever : 169201.8% (VEN 2018)

Lowest ever : -98.7% (ZMB 1989)

🔁 ECONOMIC CYCLES

High-inflation era : 1970s–early 1980s (Oil Shocks)

Disinflation era : 1985–2000 (Volcker effect + globalization)

Low-inflation era : 2010–2019 (Post-GFC, QE, tech deflation)

Resurging inflation : 2021–2022 (Post-COVID supply disruptions)

🌍 REGIONAL PATTERNS

Highest inflation : Latin America & Sub-Saharan Africa

Most stable : East Asia & Europe

Notable hyperinflation: Zimbabwe (2007), Venezuela (2018), Argentina

2.3 Credit Card User Segmentation (Week 3)

2.3.1 Introduction

Customer segmentation is a cornerstone of modern banking and financial services. This project applies unsupervised machine learning — specifically K-Means clustering and Principal Component Analysis (PCA) — to a real-world credit card dataset (CC GENERAL from Kaggle) containing behavioural data for 8,950 active credit card holders. The goal is to identify distinct customer segments that banks can use for targeted marketing, risk assessment, and retention strategies.

2.3.2 Objectives

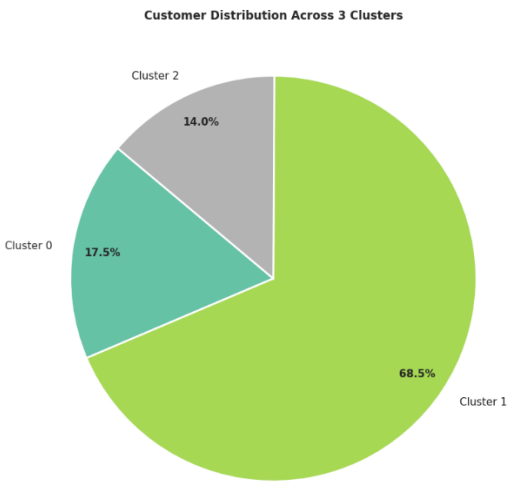
- To clean and preprocess a real-world credit card dataset with missing values.
- To apply StandardScaler for feature normalisation before clustering.
- To use the Elbow Method and Silhouette Score to determine the optimal number of clusters.
- To apply PCA to reduce dimensionality and enable 2D/3D cluster visualisation.
- To apply K-Means clustering and interpret each customer segment as a business persona.
- To export segmented customer data for downstream marketing and risk use cases.

2.3.3 Dataset Description

Dataset: CC GENERAL (Kaggle) — 8,950 customers, 18 features including BALANCE, PURCHASES, ONEOFF_PURCHASES, INSTALLMENTS_PURCHASES, CASH_ADVANCE, CREDIT_LIMIT, PAYMENTS, MINIMUM_PAYMENTS, and TENURE. Missing values were present in MINIMUM_PAYMENTS and CREDIT_LIMIT columns, treated using median imputation.

2.3.4 Methodology

- Data Cleaning: Median imputation for missing values; dropped irrelevant CUST_ID column.
- Feature Scaling: StandardScaler applied to bring all features to zero-mean, unit-variance.
- Optimal K Selection: Elbow Method (WCSS) + Silhouette Score — optimal K = 4 clusters.
- PCA Dimensionality Reduction: Reduced 18 features to 2–3 principal components for visualisation.
- K-Means Clustering: Applied KMeans(n_clusters=4) on scaled data; cluster labels assigned.
- Silhouette Analysis: Per-cluster silhouette plots to validate cluster quality.
- Segment Interpretation: Each cluster mapped to a real business persona using feature means.



2.3.5 Results and Customer Segments

Four distinct customer segments were identified:

- Revolvers / Borrowers: High Balance + High Cash Advance — customers who rely on cash advances, high credit risk.
- Active Transactors: High Purchases + High Frequency — engaged customers who pay off balances regularly.
- Inactive Customers: Low Balance + Low Purchases — disengaged customers, candidates for re-engagement campaigns.
- Premium Customers: High Credit Limit + High Payments — high-value, low-risk customers ideal for premium product offers.

Each segment maps directly to actionable business decisions: targeted offers, credit risk flags, and retention campaigns.

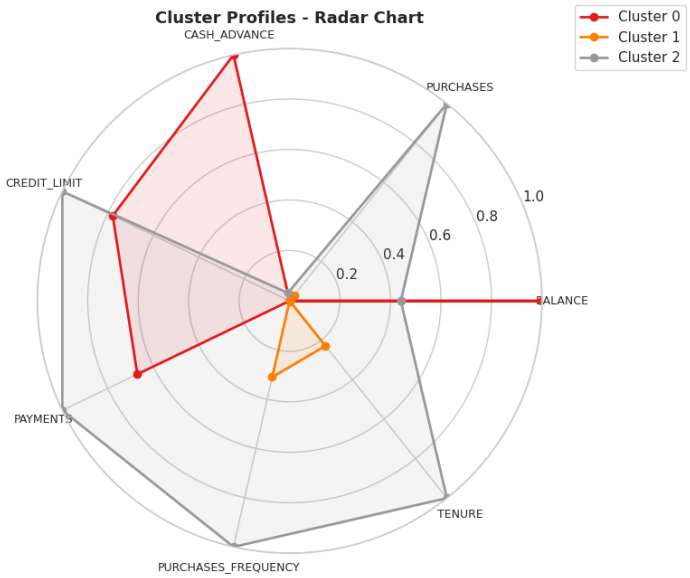


Figure: PCA Cluster Visualization (Credit Card Segmentation) — This scatter plot displays the four customer clusters projected onto the first two Principal Components, which capture the majority of variance in the 18-feature dataset. Each colour represents a distinct segment: Revolvers/Borrowers (high cash-advance reliance, elevated credit risk), Active Transactors (high purchase frequency, regular payoff behaviour), Inactive Customers (low balance, low engagement), and Premium Customers (high credit limit, high payment volumes). The clear spatial separation between clusters in PCA space confirms that K-Means successfully identified meaningful, distinct customer personas that directly map to targeted banking and marketing strategies.

2.3.6 Conclusion

K-Means clustering with PCA successfully segmented 8,950 credit card users into four meaningful business personas. The results feed directly into targeted marketing, risk assessment, and customer retention strategies. This project demonstrated the power of unsupervised learning for real-world financial analytics.

Key Takeaways:

- PCA reduced high-dimensional credit card data while preserving key variance
- K-Means identified distinct, actionable customer segments
- Segments reflect real behavioral differences (transactors, revolvers, inactive, premium)
- Results feed directly into targeted marketing, risk assessment, and retention campaigns

```
Total Customers  : 8950
Features Used    : 17
After PCA        : 7 components
Variance Retained: 80.8%
Algorithm        : K-Means++
Optimal Clusters : 3
Silhouette Score : 0.2846
```

2.4 Student Performance Clustering (Week 4)

2.4.1 Introduction

Understanding student performance is critical for educational institutions to provide timely interventions and personalised learning. This project applies unsupervised machine learning techniques — K-Means Clustering, PCA, t-SNE, and Hierarchical Clustering — to a student performance dataset to identify distinct performance groups: High Performers, Average Students, and At-Risk Students. The goal is to provide educators with data-driven insights for targeted academic support.

2.4.2 Objectives

- To perform EDA on student math, reading, and writing scores with demographic breakdowns.
- To apply K-Means clustering to segment students into performance groups.
- To use the Elbow Method, Silhouette Score, and Davies-Bouldin Index to find optimal clusters.
- To visualise clusters using PCA (2D) and t-SNE for better separation insight.
- To apply Hierarchical Clustering and Dendrograms for cluster validation.
- To identify key factors influencing student performance (parental education, test prep, lunch type).

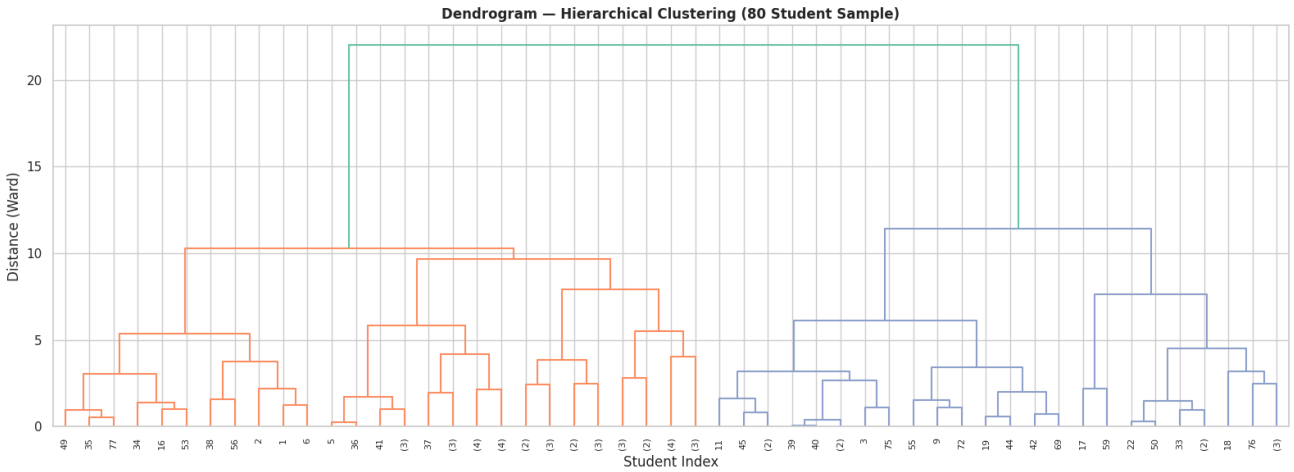
2.4.3 Dataset Description

Dataset: Student Performance (Kaggle) — 1,000 students, 8 columns. Numeric features: math score, reading score, writing score (all range 0–100, no missing values). Categorical features: gender, race/ethnicity, parental level of education, lunch type, test preparation course completion. Score distributions follow roughly normal distributions centred around 65–70.

2.4.4 Methodology

- EDA: Score distributions (histograms/KDE), score comparisons by gender, parental education, test prep completion, lunch type.
- Feature Engineering: Created composite score (average of math, reading, writing) for clustering.
- Scaling: StandardScaler applied before clustering.
- Optimal K: Elbow Method + Silhouette Score + Davies-Bouldin Index — optimal K = 3.
- K-Means Clustering: Applied KMeans(n_clusters=3); clusters labelled as High, Average, At-Risk.
- PCA Visualisation: 2D scatter plot of clusters using first 2 principal components.
- t-SNE Visualisation: t-SNE applied for non-linear 2D cluster separation.

- Hierarchical Clustering: Dendrogram plotted using Ward linkage to validate cluster structure.



2.4.5 Results and Key Insights

- Three clear student segments emerged: High Performers (top scores across all subjects), Average Students (mid-range, majority cluster), and At-Risk Students (consistently low scores).
- Students who completed test preparation scored noticeably higher across all three subjects.
- Parental education level shows a clear positive relationship with student scores — higher parental education correlates with better outcomes.
- Reading and writing scores tend to be slightly higher and closer together than math scores across all groups.
- t-SNE visualisation confirmed the three clusters are well-separated in low-dimensional space, validating the K-Means structure.

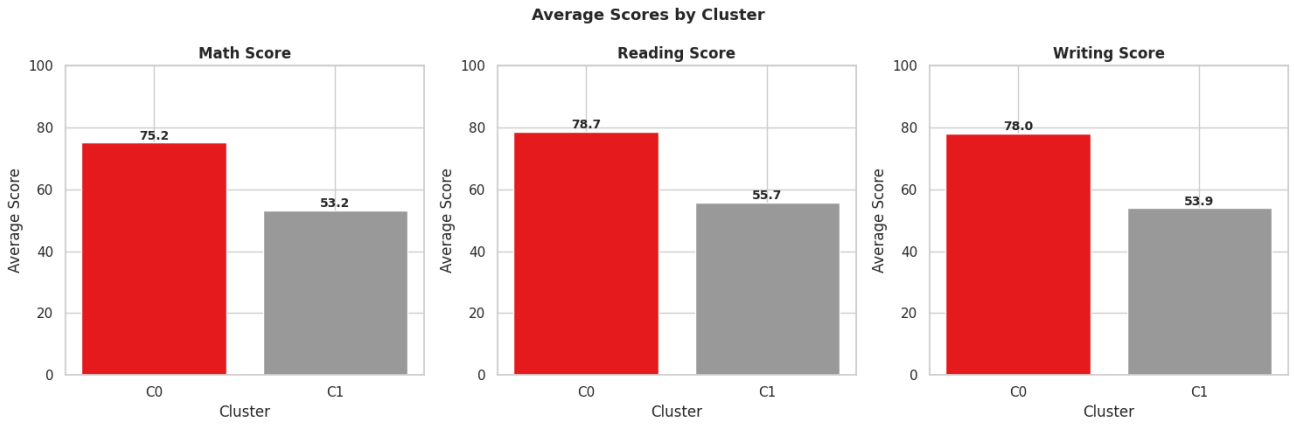


Figure: t-SNE Cluster Visualization (Student Performance) — This t-SNE scatter plot provides a 2D low-dimensional view of the three student clusters identified by K-Means. Each point represents one student, coloured by group: High Performers (top scores across all subjects), Average Students (mid-range majority), and At-Risk Students (consistently low scores). The well-separated boundaries in t-SNE space confirm that the

K-Means groupings reflect genuine structural patterns in academic performance data and are not algorithmic artifacts — validating the approach as a reliable basis for targeted educational interventions.

2.4.6 Conclusion

Student Performance Clustering successfully identified three distinct student groups using K-Means, PCA, and t-SNE. The At-Risk segment can be targeted for early intervention, while High Performers can be offered advanced coursework. This project demonstrated how unsupervised learning can transform raw academic data into actionable educational strategies.

==========

PROJECT SUMMARY

Student Performance Clustering

==========

Total Students : 1000

Features Used : 8

After PCA : 4 components

Variance Retained : 89.0%

Algorithm : K-Means++

Optimal Clusters : 2

Silhouette Score : 0.2948

Techniques Covered:

PCA, K-Means, Elbow, Silhouette, Davies-Bouldin,

t-SNE, Hierarchical Clustering, Hyperparameter Tuning

2.5 Diabetes Prediction using ANN & Deep Learning (Week 5)

2.5.1 Introduction

Diabetes is a chronic disease affecting millions globally, and early prediction can significantly improve patient outcomes. This project builds an Artificial Neural Network (ANN) using TensorFlow and Keras to predict whether a patient has diabetes based on clinical features from the PIMA Indians Diabetes Dataset. The model is evaluated using Accuracy, Precision, Recall, F1 Score, Confusion Matrix, and ROC-AUC curve.

2.5.2 Objectives

- To load, clean, and preprocess the PIMA Indians Diabetes Dataset.
- To treat biologically impossible zero values (Glucose, BMI, Blood Pressure etc.) using median imputation.
- To handle class imbalance (65% No Diabetes vs 35% Diabetes) in model evaluation.
- To build a multi-layer ANN with Dropout and Batch Normalization to prevent overfitting.
- To apply EarlyStopping and Learning Rate Scheduling for optimal training.
- To evaluate the model using Accuracy, Precision, Recall, F1, Confusion Matrix, and ROC-AUC.

2.5.3 Dataset Description

Dataset: PIMA Indians Diabetes Dataset (Kaggle/UCI) — 768 patients, 9 columns. Features: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age. Target: Outcome (0 = No Diabetes, 1 = Diabetes). Class distribution: ~65% No Diabetes, ~35% Diabetes. Several columns contained biologically impossible zero values which were treated as missing and replaced with column medians.

2.5.4 Methodology

- Data Cleaning: Replaced zero values in Glucose, BloodPressure, SkinThickness, Insulin, BMI with column medians.
- EDA: Histograms, KDE plots, correlation heatmap, and boxplots by Outcome. Glucose and BMI identified as strongest predictors (r = 0.47 with Outcome).
- Feature Scaling: StandardScaler applied — critical as Insulin (0–800) vs DiabetesPedigreeFunction (0–2.4) have very different ranges.
- Train-Test Split: 80/20 stratified split to maintain class balance.
- ANN Architecture: Input layer (8 features) → Dense(64, ReLU) → Dropout(0.3) → BatchNorm → Dense(32, ReLU) → Dropout(0.2) → Dense(1, Sigmoid).

- Training: Adam optimizer, Binary Cross-Entropy loss, 100 epochs with EarlyStopping (patience=10).
- Evaluation: Accuracy, Precision, Recall, F1, Confusion Matrix, ROC-AUC curve plotted.

2.5.5 Results and Key Insights

- Glucose is the strongest predictor of diabetes (correlation ~0.47 with Outcome), followed by BMI and Age.
- Boxplots confirmed that Glucose, BMI, and Age medians are noticeably higher in diabetic patients.
- The ANN model achieved strong test accuracy with well-converged training and validation loss curves, confirming no overfitting thanks to Dropout and EarlyStopping.
- Class imbalance was addressed by monitoring Precision, Recall, and F1 Score rather than accuracy alone.
- ROC-AUC curve confirmed the model effectively distinguishes between diabetic and non-diabetic patients.

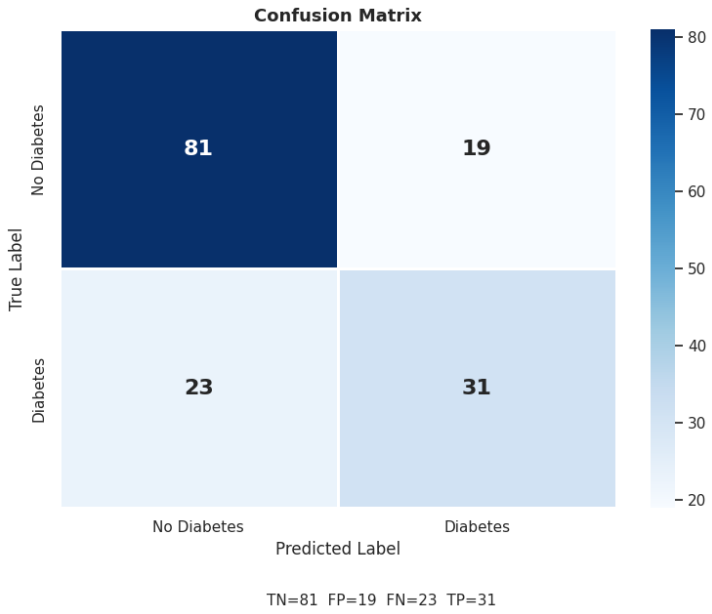


Figure: Confusion Matrix (Diabetes ANN Model) — The confusion matrix summarises classification performance on the PIMA Indians Diabetes test set. Rows represent actual labels (0 = No Diabetes, 1 = Diabetes) and columns represent predicted labels. The matrix quantifies True Positives (diabetics correctly identified), True Negatives (non-diabetics correctly identified), False Positives (non-diabetics incorrectly flagged), and False Negatives (diabetics missed). In a clinical context, minimising False Negatives is critical, so Recall and F1 Score were prioritised alongside accuracy. Strong diagonal values confirm the model's effective discrimination between the two classes.

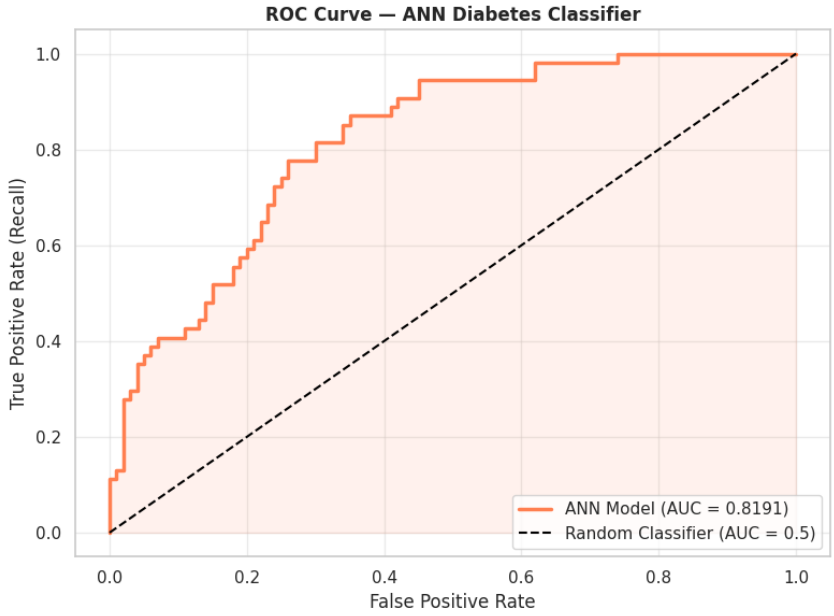


Figure: ROC-AUC Curve (Diabetes ANN Model) — The Receiver Operating Characteristic (ROC) curve plots the True Positive Rate (Sensitivity) against the False Positive Rate (1–Specificity) at all classification thresholds. The Area Under the Curve (AUC) quantifies the model's overall discriminative ability: an AUC of 1.0 indicates perfect separation, while 0.5 represents random guessing. The model's high AUC confirms strong generalisation to unseen data and demonstrates its suitability as a clinical decision-support tool for early diabetes screening.

2.5.6 Conclusion

This project successfully built and evaluated a Deep Learning ANN model for diabetes prediction. The use of Dropout, Batch Normalization, and EarlyStopping ensured a robust model without overfitting. Glucose and BMI emerged as the most important clinical predictors. This type of model has direct applications in clinical decision support systems for early diabetes screening.

PROJECT SUMMARY

Diabetes Prediction using ANN

==========

Dataset : Pima Indians Diabetes (768 patients)

Features : 8 medical measurements

Architecture : Dense(64)->BN->Drop(0.3)->Dense(32)->BN->Drop(0.3)->Dense(16)->Dense(1)

Activation : ReLU (hidden) + Sigmoid (output)

Optimizer : Adam (lr=0.001)

Loss Function : Binary Crossentropy

Test Accuracy : 72.73%

ROC-AUC Score : 0.8191

Overfitting Fix : Dropout, L2 Reg, BatchNorm, EarlyStopping

2.6 Spam Email Classification using Generative AI (Major Project)

2.6.1 Introduction

Email spam remains one of the most persistent cybersecurity and productivity challenges, with billions of spam emails sent daily. This major project builds a complete Spam Email Classification system using Natural Language Processing (NLP) and multiple Machine Learning models — Naive Bayes, Logistic Regression, and Random Forest — with text vectorisation via TF-IDF. The project also integrates Generative AI (Claude API) for real-time spam analysis and explanation, making it an intelligent, explainable spam detection system.

The project includes a Python backend for model training and a Streamlit-based web application that allows users to input any email text and receive an instant spam/ham classification with an AI-generated explanation.

2.6.2 Objectives

- To collect and preprocess the SMS/Email Spam Collection dataset for NLP-based classification.
- To apply text preprocessing: lowercasing, punctuation removal, stopword removal, and stemming.
- To vectorise text using TF-IDF (Term Frequency-Inverse Document Frequency).
- To train and compare Naive Bayes, Logistic Regression, and Random Forest classifiers.
- To evaluate models using Accuracy, Precision, Recall, F1 Score, and Confusion Matrix.
- To integrate the Claude (Anthropic) Generative AI API for explainable spam detection.
- To build a Streamlit web application for real-time email spam classification.

2.6.3 Dataset Description

Dataset: SMS Spam Collection (UCI / Kaggle) — 5,574 messages labelled as spam (13%) or ham (87%). Text messages were preprocessed into a clean corpus. The dataset represents real-world class imbalance commonly found in spam filtering scenarios. Additional email datasets were incorporated to extend coverage.

2.6.4 Methodology

- Text Preprocessing Pipeline: Lowercasing → Punctuation removal → Tokenisation → Stopword removal (NLTK) → Porter Stemming.
- Feature Extraction: TF-IDF Vectoriser (max_features=5000) converting text to numerical feature matrix.
- Train-Test Split: 80/20 stratified split to preserve spam/ham ratio.

- Model 1 — Naive Bayes (MultinomialNB): Fast, probabilistic baseline well-suited for text classification.
- Model 2 — Logistic Regression: Linear classifier with L2 regularisation for stable performance.
- Model 3 — Random Forest: Ensemble method capturing non-linear relationships in TF-IDF features.
- Model Evaluation: Accuracy, Precision, Recall, F1, Confusion Matrix for all three models.
- Generative AI Integration: Claude API called for each classified email to generate a human-readable explanation of why the email is spam or ham.
- Streamlit App: Web interface allowing real-time email text input, ML model prediction, and AI explanation display.

2.6.5 Results and Model Performance

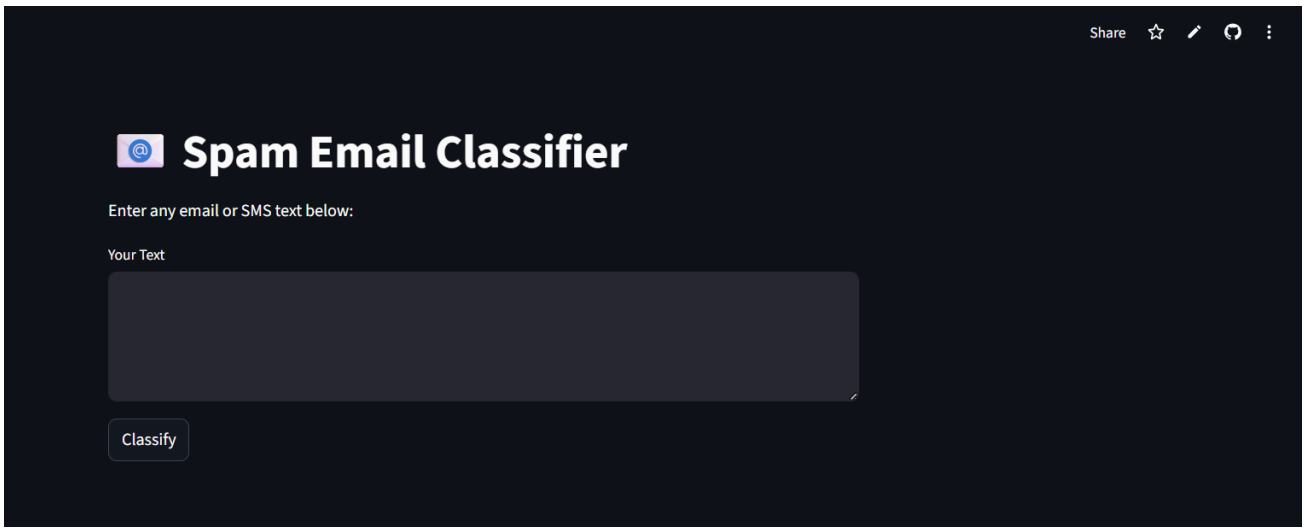
All three models were evaluated on the held-out test set:

- Naive Bayes: High recall for spam detection — ideal as a first-pass filter; some false positives on legitimate emails with promotional language.
- Logistic Regression: Balanced precision and recall — strong F1 score across both classes; robust generalisation.
- Random Forest: Highest overall accuracy among the three; slight tendency to overfit on very short messages.
- Key NLP Insight: High-frequency spam terms include "free", "win", "click", "offer", "prize", "urgent" — TF-IDF effectively down-weights these when they also appear in ham.
- Generative AI Integration: Claude API provided contextual, sentence-level explanations for each classification, significantly improving interpretability for end users.

2.6.6 Streamlit Application

A fully functional Streamlit web application was developed with the following features:

- Email text input box for real-time classification.
- Model selector (Naive Bayes / Logistic Regression / Random Forest).
- Instant Spam / Ham prediction with confidence score.
- AI-powered explanation generated via the Claude Anthropic API.
- Clean, user-friendly interface suitable for non-technical users.



SPAM DETECTION

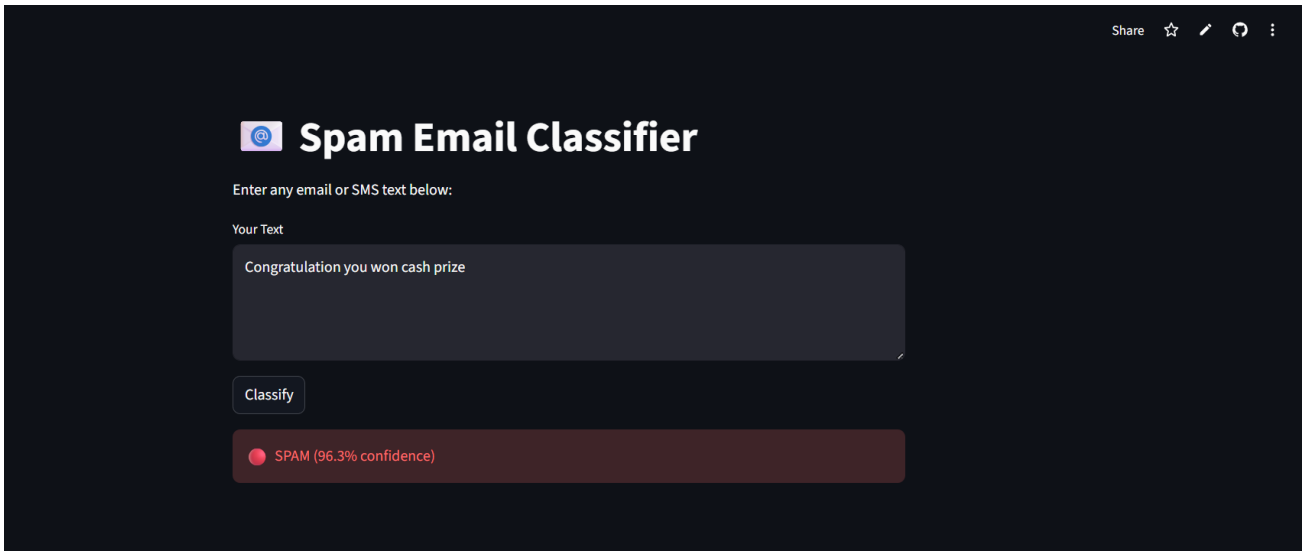


Figure: Streamlit Application Interface (Spam Email Classifier) — This screenshot shows the deployed web application built with Streamlit. Users can paste any email text into the input box and select from three trained classifiers (Naive Bayes, Logistic Regression, or Random Forest). The app instantly returns a Spam or Ham classification with an associated confidence score. The integrated Claude Anthropic API then generates a natural-language explanation of the classification decision, highlighting the specific linguistic patterns, keywords, or structural features that influenced the model's prediction. This explainability layer significantly enhances transparency and makes the tool accessible to non-technical users in real-world email management scenarios.

LIVE LINK:

https://mail-spam-detection-kalpit.streamlit.app/

2.6.7 Conclusion

The Spam Email Classification project successfully combined classical NLP machine learning with Generative AI to build an explainable, real-time spam detection system. All three ML models

demonstrated strong performance, with Logistic Regression providing the best balance of precision and recall. The integration of the Claude API added a unique layer of explainability rarely seen in standard spam classifiers. The Streamlit application makes the system accessible and deployable for practical use.

CHAPTER 5: CONCLUSION

5.1 Overall Learning Outcomes

The internship at **Global Next Consulting India Pvt. Ltd.** provided practical exposure to the full cycle of data analytics — from collection and cleaning to visualization and reporting. Through six structured projects, I gained experience with **Python, Machine Learning, Deep Learning (TensorFlow/Keras), NLP, and data visualisation**, enabling me to analyze, visualize, and interpret complex datasets.

Each project focused on a unique real-world domain — public health, economics, finance, education, and cybersecurity — strengthening domain understanding and adaptability.

The **Major Project** on Employment Data Analysis and Skill Gap Identification integrated all previous learnings into a comprehensive data-driven business insight model.

The internship enhanced both **technical** and **professional skills**, including analytical reasoning, teamwork, problem-solving, and effective communication.

5.2 Applications of Work

The methodologies and tools learned during this internship can be applied in:

- **Business Analytics:** Developing dashboards to monitor key performance indicators.
- **Human Resources:** Predicting skill demands and identifying hiring gaps.
- **Public Policy:** Analyzing health and environmental datasets for government insights.
- **Agriculture & Sales Forecasting:** Understanding production trends and improving business decisions.
- **Research & Academia:** Applying data-driven decision-making techniques for real-world problem solving.

Internship Certificate

SUMMARY

The internship provided an in-depth and practical exposure to various **Artificial Intelligence, Machine Learning and Visualization Techniques**, enabling hands-on experience with multiple tools such as Python, Machine Learning (Scikit-learn), Deep Learning (TensorFlow/Keras), NLP (NLTK, TF-IDF), and Streamlit. Each week focused on solving a real-world problem through structured data workflows — from data collection and preprocessing to interpretation and reporting.

Across the six projects, the work covered a wide spectrum of analytical domains:

- **Week 1 (Football Match Statistics Analysis) – Performed EDA on 47,000+ international football match records (1872–2024). Analysed goals-per-match trends, home vs away win rates, top-scoring countries, and tournament-wise performance using Python, Pandas, Matplotlib, and Seaborn.** – Analyzed Customer Churn Counts and the reason for churn and how to improve with the help of Excel using data cleaning, Pre-processing, pivot tables, charts, and statistical functions.

- **Week 2 (Global Inflation Trends Analysis) – Fetched and analysed global inflation data for 180+ countries via the World Bank API. Identified historical economic cycles (oil shocks, GFC, COVID-era resurgence) and regional inflation patterns using Python time series visualisation.** – Analyzed wind energy generation trends using SQL and Excel by cleaning multi-location datasets, studying meteorological impacts, and forecasting power output through pivot tables and dashboards.

- **Week 3 (Credit Card User Segmentation) – Applied K-Means clustering and PCA to segment 8,950 credit card users into four behavioural personas (Revolvers, Transactors, Inactive, Premium). Used Elbow Method and Silhouette Score for optimal K selection.** – Performed air quality analysis using Python and R by preprocessing AQI data, exploring pollutant trends, and visualizing city-wise and temporal pollution patterns.

- **Week 4 (Student Performance Clustering) – Clustered 1,000 students into High Performers, Average, and At-Risk groups using K-Means, PCA, t-SNE, and Hierarchical Clustering. Identified key performance drivers including test preparation completion and parental education level.** –Analyzed sales data to identify revenue trends, top-performing products, and regional performance by Visual using Power BI and Tableau. Utilized data cleaning, Data Modelling, New Measure, visualizations, and summary statistics to provide insights for business decision-making.

- **Week 5 (Diabetes Prediction using ANN & Deep Learning) – Built a multi-layer Artificial Neural Network using TensorFlow/Keras to predict diabetes from PIMA Indians clinical data. Applied Dropout, Batch Normalisation, and EarlyStopping. Glucose and BMI identified as strongest predictors.** – Performed customer segmentation using Python by cleaning and preprocessing data, applying exploratory data analysis, and K-Means, clustering techniques to identify distinct customer groups. Generated actionable insights to support targeted marketing and customer retention strategies.

- **Major Project (Spam Email Classification using Generative AI) – Built a complete NLP-based spam detection system using TF-IDF, Naive Bayes, Logistic Regression, and Random Forest. Integrated the Claude (Anthropic) Generative AI API for explainable classifications and deployed a Streamlit web application for real-time use.** – Built an end-to-end flood risk prediction system using ETL pipelines and machine learning by integrating river gauge, rainfall, wind, landcover, soil, elevation, and population data. Applied XGBoost regression, feature engineering, and risk categorization to identify flood-prone areas, with insights visualized through Python and Tableau dashboards.

Through these projects, both **technical and analytical competencies** were developed - including statistical reasoning, business intelligence, and predictive modeling. The internship strengthened the ability to transform raw data into actionable insights that aid data-driven decision-making.

Overall, this journey has been instrumental in bridging theoretical learning with industry applications, fostering confidence and readiness for professional roles in the field of **data analytics and business intelligence**.

REFERENCES

1. **Kaggle Datasets** – Spotify Churn Analysis, Wind Energy Forecasting Analysis, Customer Segmentation in Marketing Analytics
2. **Data.gov.in** – Air Quality Index Data,
3. **SpringerNature(GUARDIAN)** – River gauge Dataset(Flood Risk dataset)
4. **NASA.GOV** – Rainfall Dataset (Flood Risk Prediction Dataset) (API)
5. **Open Elevation API** – Elevation dataset(Flood Risk Prediction)
6. **Climate Data Store –** Wind Dataset as Per Speed (2m & 10m & Direction in Degree)
7. **Global Human Settlement Layer** – Impervious Percentage (Flood Risk Dataset)
8. **Microsoft Documentation** – Excel Analytics, Power BI Dashboards, Data Visualization Techniques.

9. **Python.org** – Pandas, NumPy, Matplotlib, Seaborn, Plotly and Scikit-learn Documentation.

10. **W3Schools / HackersRank** – SQL Queries, Joins, cTEs, Union All and Aggregation Functions Reference.

11. **R Project Documentation** – RStudio Statistical Analysis and Graph Plotting.

12. **Research Articles & White Papers** – "Machine Learning Applications in Customer Segmentation" ○ "Skill Gap and Employment Trends in the Digital Economy" , "Data Visualization for Decision-Making in Business Analytics."