COVID-19 Data Analysis Capstone Project

Author: Prajwal M
Date: 28/11/2025
Project Type: Data Science Capstone
Institution: Internship Studio

Executive Summary

This capstone project encompasses a comprehensive data analysis and machine learning application across two distinct datasets. Part 1 focuses on exploratory data analysis (EDA) of COVID-19 global pandemic data, while Part 2 applies predictive machine learning models to loan default prediction. The project demonstrates proficiency in data manipulation, statistical analysis, data visualization, and machine learning model development using Python.

Project Overview

Part 1: COVID-19 Data Analysis using Python

Objective

To perform comprehensive exploratory data analysis on global COVID-19 pandemic data, extracting meaningful insights about disease spread, economic indicators, and human development across different continents and countries.

Dataset Information

- **Format:** CSV
- **Key Variables:** continent, location, date, totalcases, totaldeaths, gdppercapita, humandevelopmentindex

Project Scope

1. Data Import and High-Level Understanding

- Import dataset using Pandas from the provided URL
- Identify dataset dimensions (rows, columns)
- Analyze data types of all columns
- Generate descriptive statistics and data frame information

2. Low-Level Data Understanding

- Count unique values in location column
- Identify continent with maximum frequency
- Calculate maximum and mean values in totalcases column
- Determine 25th, 50th, and 75th percentiles for totaldeaths
- Identify continent with maximum human development index
- Identify continent with minimum GDP per capita

3. Data Filtering and Selection

- Filter dataframe to retain only relevant columns: continent, location, date, totalcases, totaldeaths, gdppercapita, humandevelopmentindex
- Update dataframe with filtered columns

4. Data Cleaning

- Remove duplicate observations
- Identify and document missing values across all columns
- Remove observations where continent column is missing (using subset parameter)
- Fill remaining missing values with 0

5. Date-Time Conversion

- Convert date column to datetime format using pandas.to_datetime()
- Extract month information and create new month column

6. Data Aggregation

- Apply groupby operation on continent column to find maximum values
- Create new dataframe (df_groupby) with aggregated results
- Use reset_index() for proper indexing

7. Feature Engineering

- Create new feature: death_to_case_ratio = totaldeaths / totalcases
- This ratio helps identify severity of pandemic across regions

8. Data Visualization

- **Univariate Analysis:** Histogram of GDP per capita using seaborn distplot
- **Bivariate Analysis:** Scatter plot showing relationship between totalcases and gdppercapita
- **Multivariate Analysis:** Pairplot of df_groupby dataset to identify correlations
- **Categorical Analysis:** Bar plot of continent vs totalcases using seaborn catplot

9. Data Export

- Save processed df_groupby dataframe to CSV format for future analysis

Key Insights Expected

- **Geographic Distribution:** Understanding which continents were most affected by COVID-19
- **Socioeconomic Correlation:** Analyzing relationship between GDP per capita, human development index, and pandemic severity
- **Temporal Patterns:** Month-wise trends in COVID-19 cases and deaths
- **Regional Disparities:** Identifying differences in pandemic impact across economic and development indicators

---

Part 2: Machine Learning Model Application on Loan Dataset

Objective

To build and evaluate a predictive machine learning model for loan default prediction, demonstrating the application of classification algorithms to real-world financial data.

Dataset Information

- **Dataset:** loan_dataset.csv
- **Problem Type:** Binary Classification
- **Target Variable:** Loan default status (Yes/No)
- **Model Type:** Logistic Regression

Project Scope

1. Data Loading and Cleaning

- Read loan_dataset.csv file
- Perform exploratory analysis on dataset structure
- Handle missing values and outliers
- Encode categorical variables as needed
- Prepare features and target variable for modeling

2. Model Selection and Justification

- **Selected Model:** Logistic Regression
- **Rationale:** Logistic Regression is optimal for binary classification tasks. It provides:
  - Interpretable coefficients showing feature importance
  - Probabilistic predictions (0-1 range)

    - Efficient computation with good performance on loan data
    - Clear decision boundary for binary outcomes

3. Model Development

- Split data into training (70%) and testing (30%) sets
- Train Logistic Regression model on training dataset
- Perform hyperparameter tuning if necessary
- Generate predictions on test dataset

4. Model Evaluation

- **Accuracy Score:** Overall proportion of correct predictions
- **Confusion Matrix:** True Positives, True Negatives, False Positives, False Negatives
- **Precision and Recall:** Analyze model's ability to identify defaulters
- **ROC-AUC Curve:** Evaluate model discriminative ability across thresholds
- **Classification Report:** Detailed metrics for each class

5. Model Interpretation

- Analyze feature coefficients to identify key predictors
- Determine which factors most influence loan default
- Provide business insights for risk management

Expected Outcomes

- A trained Logistic Regression model with documented accuracy metrics
- Confusion matrix visualization showing prediction performance
- Feature importance analysis for business decision-making
- Model ready for deployment in loan underwriting process

Technical Implementation

Technologies and Libraries Used

Python 3.x

- pandas: Data manipulation and analysis
- numpy: Numerical computations
- matplotlib: Static data visualization
- seaborn: Statistical data visualization
- scikit-learn: Machine learning model development

- jupyter notebook / Python script: Code execution environment

Development Environment

- **IDE:** Jupyter Notebook / VS Code
- **Submission Format:** .ipynb (Jupyter) and .py (Python script) files

---

Methodology

Part 1 Workflow

Dataset Import
↓
High-Level EDA (shape, dtypes, info)
↓
Low-Level EDA (unique values, statistics)
↓
Data Filtering (column selection)
↓
Data Cleaning (duplicates, missing values)
↓
DateTime Conversion & Feature Extraction
↓
Data Aggregation (groupby continent)
↓
Feature Engineering (ratio creation)
↓
Data Visualization (multiple plot types)
↓
Results Export (CSV format)

Part 2 Workflow

Dataset Loading & Cleaning
↓
Exploratory Data Analysis
↓
Data Preprocessing & Encoding
↓
Train-Test Split (70:30)
↓
Model Training (Logistic Regression)
↓
Model Prediction
↓
Performance Evaluation (Metrics & Confusion Matrix)
↓
Results Interpretation & Documentation

Expected Deliverables

Part 1: COVID-19 Data Analysis

- **File Format:** .ipynb (Jupyter Notebook) and .py (Python Script)
- **Contents:**
  - Data loading and exploration code
  - Data cleaning and preprocessing steps
  - Aggregation and feature engineering code
  - Visualization plots with interpretations
  - Summary statistics and insights

Part 2: Loan Default Prediction

- **File Format:** .ipynb (Jupyter Notebook) and .py (Python Script)
- **Contents:**
  - Data loading and cleaning code
  - EDA and preprocessing steps
  - Logistic Regression model implementation
  - Model evaluation metrics and confusion matrix
  - Feature importance analysis
  - Business recommendations

Project Report

- Comprehensive documentation of methodology
- Data insights and findings
- Model performance summary
- Code files with comments and documentation

Key Learnings and Skills Demonstrated

Data Science Skills

- Pandas data manipulation and cleaning
- Statistical analysis and EDA techniques
- Data aggregation using groupby operations

- Feature engineering for enhanced analysis
- Time series date-time handling

Visualization Skills

- Histogram and distribution analysis
- Scatter plots for relationship analysis
- Pairplots for multivariate exploration
- Categorical bar charts and comparisons

Machine Learning Skills

- Classification model selection and application
- Train-test data splitting
- Model evaluation using confusion matrix
- Accuracy measurement and performance interpretation
- Feature importance analysis

Professional Development

- Code documentation and comments
- Project organization and file management
- Version control with Git/GitHub
- Report writing and professional communication

Conclusion

This capstone project provides practical experience in end-to-end data science workflows. Part 1 demonstrates competency in exploratory data analysis and visualization of complex pandemic data, while Part 2 showcases machine learning model development and evaluation. Together, these projects validate technical skills in Python programming, data manipulation, statistical analysis, and predictive modeling—essential competencies for a data science professional.

The project successfully bridges theoretical knowledge with practical application, preparing for real-world data science challenges in various domains including public health analytics and financial risk assessment.
