SmartBridge -Forecasting Economic Prosperity: Leveraging Machine Learning For GDP Per Capita Prediction

Milestone 1: Project Initialization and Planning Phase

The "Project Initialization and Planning Phase" forecast GDP per capita using machine learning for improved economic planning. It involves gathering diverse economic, demographic, and environmental data, followed by rigorous analysis to inform feature engineering. Model selection and validation strategies ensure robust predictions, addressing risks like data biases and volatility, while adhering to ethical considerations and maintaining clear documentation for effective communication and transparency throughout the process.

Activity 1: Define Problem Statement

The project aims to develop and deploy machine learning models to accurately predict GDP per capita, leveraging diverse economic, demographic, and environmental data sources. The goal is to enhance economic decision-making by providing reliable forecasts that account for complex interactions and potential risks, ensuring sustainable development and informed policy formulation**.**

A **Problem Statement Report: Click Here**

Activity 2: Project Proposal (Proposed Solution)

The proposed solution involves applying advanced machine learning techniques to analyze comprehensive economic, demographic, and environmental data. By selecting and training robust models, the project aims to improve the accuracy of GDP per capita predictions, facilitating better-informed policy decisions and fostering sustainable economic growth.

Project Proposal Report: Click Here

Activity 3: Initial Project Planning

The objective of predicting GDP per capita, acquiring diverse datasets, forming a multidisciplinary team, establishing a structured timeline with milestones, allocating resources, assessing risks, and implementing a communication plan for effective collaboration and progress tracking**.**

Project Planning Report: Click Here

Milestone 2: Data Collection and Preprocessing Phase

The Data Collection and Preprocessing Phase involves executing a plan to gather relevant loan

application data from Kaggle, ensuring data quality through verification and addressing missing values. Preprocessing tasks include cleaning, encoding, and organizing the dataset for subsequent exploratory analysis and machine learning model development.

Activity 1: Data Collection Plan, Raw Data Sources Identified, Data QualityReport

This project gathers historical economic indicators, demographic data, and socio-economic factors from reliable sources such as World Bank, IMF, and national statistical agencies. The dataset includes variables like GDP growth rates, inflation rates, population demographics, education levels, healthcare spending, and infrastructure development indices.

Data Collection Report: Click Here

Activity 2: Data Quality Report

The collected data from sources like World Bank and national statistical agencies is comprehensive and reliable, spanning multiple years and countries. Variables are consistently formatted and include minimal missing values, ensuring robustness in analysis. Data preprocessing techniques such as normalization and outlier detection further enhance dataset integrity for accurate machine learning model training.

Data Quality Report: Click Here

Activity 3: Data Exploration and Preprocessing

Initial exploratory data analysis revealed correlations between GDP per capita and variables like education levels and healthcare spending. Missing values were handled through imputation methods, and categorical variables were encoded appropriately. Feature scaling was applied to ensure uniformity in data distribution for machine learning model training.

Data Exploration and Preprocessing Report: Click Here

Milestone 3: Model Development Phase

The Model Development Phase entails crafting a predictive model . It encompasses strategic feature selection, evaluating and selecting models (Random Forest, Decision Tree, KNN, XGB), initiating training with code, and rigorously validating and assessing model performance for informed decision-making in the lending process.

Activity 1: Feature Selection Report

Feature selection was performed using techniques such as correlation analysis, recursive feature elimination, and feature importance from ensemble models. Key variables influencing GDP per capita include education levels, healthcare expenditure, infrastructure development, and inflation rates, which were identified as significant predictors for model development.

Feature Selection Report: Click Here

Activity 2: Model Selection Report

After evaluating various machine learning algorithms including linear regression, decision trees, random forest, and gradient boosting, the random forest model demonstrated superior performance based on metrics such as R-squared and Mean Absolute Error (MAE). Hyperparameter tuning further optimized model accuracy and robustness for predicting GDP per capita across different countries.

Model Selection Report: Click Here

Activity 3: Initial Model Training Code, Model Validation and Evaluation Report

The RandomForestRegressor model achieved an R-squared score of 0.85 and a Mean Absolute Error (MAE) of $1200 when evaluated on the test set, demonstrating robust performance in predicting GDP per capita based on selected features like education levels, healthcare spending, and infrastructure development indices.

Model Development Phase Template: Click Here

Milestone 4: Model Optimization and Tuning Phase

Hyperparameters of the RandomForestRegressor were optimized using techniques like GridSearchCV, focusing on parameters such as max_depth, min_samples_split, and max_features. Cross-validation helped fine-tune the model, achieving an improved R-squared of 0.87 and reducing MAE to $1100, enhancing predictive accuracy for GDP per capita prediction.

Activity 1: Hyperparameter Tuning Documentation

The logistic regression model was selected for its superior performance, exhibiting high accuracyduring hyperparameter tuning. Its ability to handle complex relationships, minimize overfitting, and optimize predictive accuracy aligns with project objectives, justifying its selection as the final model.

Activity 2: Performance Metrics Comparison Report

The Performance Metrics Comparison Report contrasts the baseline and optimized metrics for various models, specifically highlighting the enhanced performance of the linear regression model. This assessment provides a clear understanding of the refined predictive capabilities achieved through hyperparameter tuning.

Activity 3: Final Model Selection Justification

The Final Model Selection Justification articulates the rationale for choosing logistic regression as the ultimate model. Its exceptional accuracy, ability to handle complexity, and successful hyperparameter tuning align with project objectives.

Model Optimization and Tuning Phase Report: Click Here

Milestone 5: Project Files Submission and Documentation

For project file submission in Github, Kindly click the link and refer to the flow. **Click Here**

For the documentation, Kindly refer to the link. **Click Here**

Milestone 6: Project Demonstration

In the upcoming module called Project Demonstration, individuals will be required to record a video by sharing their screens. They will need to explain their project and demonstrate its execution during the presentation.