Name:Papa Tekyi Ward

ID:10201100131

**Documentation**

1. Instructions for Each Module:

-Regression: Load numeric data as a CSV file. Choose the prediction target column. Optionally scale the features, and the model fits a Linear Regression model for predictions. You also have the facility to input own data for making predictions.

-Neural Network: Upload a classification data, select a target column, and tune the hyperparameters (learning rate and number of epochs) of the neural network. Train the model and you can even see the training. You can even upload test examples to obtain predictions.

-Clustering: Upload a data, select clustering features, and tune the number of clusters (k). You can plot clusters as a 2D plot and save the clustered data.

LLM Q&A (Budget Q&A): Upload a PDF of the Ghana 2025 Budget. The system splits the document into chunks, saves them in a vector database (FAISS), and creates answers from the document content using a HuggingFace model.

2. -Regression Module:

Dataset: A CSV file uploaded by the user that includes the features (independent variables) and the target variable (dependent variable) for regression analysis.

Model: Linear Regression, a basic statistical model utilized to model the relationship between the features and the target variable.

-Neural Network Module:

Dataset: A CSV file for classification problems, with features and a target column of categorical labels.

Model: A neural network built with TensorFlow and Keras made up of layers such as Dense with ReLU activation for hidden layers and softmax for the output layer. Categorical cross-entropy as the loss function and Adam optimizer.

-Clustering Module:

Dataset: A numeric feature CSV file to cluster. To-be-clustered features will be used to cluster similar data points.

Model: K-Means clustering algorithm, a simple unsupervised learning algorithm that divides data into k clusters by minimizing the within-cluster variance.

-LLM Q&A Module:

Dataset: The 2025 Ghana Budget PDF file, uploaded by the user.

Model: Falcon-RW-1B, a causal language model for question-answering. The document is embedded using sentence-transformer embeddings and indexed in a FAISS index for retrieval-based question answering.

3. -Data Loading & Preprocessing:

The PDF document (Ghana 2025 Budget) is loaded by the user and stored temporarily.

The document is loaded through PyPDFLoader, and it's broken down into smaller segments through RecursiveCharacterTextSplitter.

-Vectorization:

The segments are embedded through the HuggingFaceEmbeddings model (sentence-transformers/all-MiniLM-L6-v2), converting the text to numerical vectors that can be efficiently searched.

-Storage & Retrieval:

The document embeddings are stored in a FAISS vector store so that it can retrieve appropriate document chunks in a timely manner when a user asks a question.

-Language Model:

The Falcon-RW-1B language model is used for generating the answer. It's initialized using the HuggingFace AutoModelForCausalLM API and wrapped in a text generation pipeline.

Question-Answering Chain (RAG - Retrieval-Augmented Generation):

A RetrievalQA chain is utilized where the retrieved relevant document fragments are input into the language model to generate answers.

-Novelty: The approach leverages a combination of a FAISS vector store for rapid document retrieval and a powerful language model (Falcon-RW-1B) for generating human-like answers, which can be termed novel in real-time PDF-based Q&A systems.

4. Document Upload & Preprocessing:

The 2025 Ghana Budget PDF is uploaded by the user, and it is read and broken into text chunks. In this manner, long documents are handled efficiently in chunks.

Text Embedding:

The text chunks are mapped to high-dimensional embeddings through the HuggingFaceEmbeddings model. This allows the system to grasp the semantic content of the document and facilitates similarity search efficiently.

Document Retrieval:

FAISS is used for indexing the embeddings and allows for retrieval of the most relevant document chunks efficiently whenever a query is made. Cosine similarity between the query and document embeddings powers the retrieval operation.

Answer Generation:

The text chunks obtained are input to Falcon-RW-1B for generating answers. The model takes the input and outputs a natural language response to the question.

The system receives the question of the user, pipes the question through, and renders the machine's generated response in real-time in such a manner that interactive as well as dynamic questioning of the document is facilitated.

5. Strengths: The system's retrieval-augmented generation (RAG) allows for more domain-specific and accurate answers compared to ChatGPT, which may give more general and broader answers.

Limitation: The responses by ChatGPT could be general, while the LLM in this system can be more specific when answering questions about the specific document on which it has been trained.