**Assignment task: AIQOD**

**Analysis Projects**

**Domain Introduction**: The project focuses on analyzing text data to classify it into multiple categories using machine learning techniques.

**Project Introduction**: The project aims to develop a text classification model using logistic regression and TF-IDF vectorization to predict the categories of text data.

**Objective**: The objective is to accurately classify text data into multiple categories using machine learning algorithms.

**ELT Approach**: The ELT (Extract, Load, Transform) approach involves extracting text data, transforming it into TF-IDF vectors, and loading it into a machine learning model for classification.

**EDA Findings**: Exploratory Data Analysis (EDA) reveals that the text data is well-distributed among different categories, with no significant class imbalances.

**Feature Engineering**: No specific feature engineering techniques are applied, as the TF-IDF vectorization process handles feature creation.

**Statistical Technique**: Statistical significance tests are not applicable in this context, as the focus is on predictive modeling rather than statistical inference.

**Conclusion**: The key findings of the project include the successful classification of text data into multiple categories using logistic regression and TF-IDF vectorization.

**Business Suggestion/Solution**: The developed model can be used to automatically classify text data, streamlining the process of categorizing large volumes of textual information.

**Text-Extraction Projects**

**Project Introduction:** The project involves analyzing text data using the TF-IDF (Term Frequency-Inverse Document Frequency) vectorization technique for further classification and analysis.

**Objective:** The objective is to extract features from text data using TF-IDF vectorization and build a classification model to predict the categories of the text data accurately.

**Text Extraction:** The text extraction process involves converting text data into numerical features using TF-IDF vectorization, which assigns weights to words based on their frequency in a document and their rarity across documents.

**Text Cleaning:** Text cleaning techniques such as removing special characters, converting to lowercase, and removing stop words are applied to the text data before TF-IDF vectorization to improve the quality of the features.

**Data Storage:** The extracted features are stored in a structured format for easy access and analysis, enabling efficient processing of large volumes of text data.

**Conclusion:** The project successfully utilizes TF-IDF vectorization to extract features from text data and build a classification model for predicting the categories of text data.

**Business Suggestion/Solution:** The TF-IDF approach can be used to analyze and categorize large volumes of text data, enabling businesses to extract valuable insights and make informed decisions based on textual information.

**Model Building Projects**

**Domain Introduction:** The project focuses on building a text classification model using machine learning algorithms to predict the categories of text data.

**Problem Statement:** The project aims to build a text classification model that can accurately predict the categories of text data based on its content.

**Data Cleaning and Preprocessing:** Data cleaning and preprocessing involve handling null values, outliers, and other data quality issues to ensure the quality of the dataset. In this project, the data is preprocessed by combining text columns, removing missing values, and converting labels to numeric type for model compatibility.

**EDA Findings:** Exploratory Data Analysis (EDA) reveals insights into the distribution of text data across different categories and helps identify patterns in the data. The analysis shows a balanced distribution of text data among different categories, with no significant class imbalances.

**Feature Engineering:** Feature engineering techniques such as TF-IDF vectorization are used to convert text data into numerical features for model building. TF-IDF vectorization is applied to the combined text data to transform it into a format suitable for machine learning models.

**Statistical Significance:** Statistical significance tests are not applicable in this context, as the focus is on predictive modeling rather than statistical inference.

**Class Imbalance Technique:** Class imbalance techniques are not required in this project, as the dataset does not exhibit significant class imbalances.

**Model Building:** A logistic regression model is chosen as the base model for its simplicity and effectiveness in handling multi-class classification tasks. The model is trained using the TF-IDF transformed text data and the multi-output logistic regression classifier.

**Model Selection:** The model is selected based on its performance metrics such as accuracy, precision, recall, and F1-score on the validation dataset. The model is evaluated using the F1-score as it provides a balance between precision and recall for multi-class classification tasks.

**Final Model:** The final model is selected based on its performance on the validation dataset and its ability to generalize to unseen data.

**Conclusion:** The key findings of the project include the successful development of a text classification model using logistic regression and TF-IDF vectorization.

**Business Suggestion/Solution:** The developed model can be used to automatically classify text data, enabling faster and more accurate categorization of textual information.