InsightNexus Project Report

Problem statement:

Knowledge Representation and Insight Generation from Structured Datasets.

Introduction

In today's data-driven world, organizations generate vast amounts of data that, when properly analyzed, can provide valuable insights to drive decision-making processes. However, the challenge lies in effectively representing this knowledge and extracting useful insights from it. InsightNexus aims to tackle this challenge by developing an AI-based solution capable of processing and analyzing structured data, identifying patterns, and generating meaningful insights.

The primary objectives of InsightNexus are:

- 1. To preprocess and clean the dataset for further analysis.
- 2. To effectively represent the knowledge contained within the dataset.
- 3. To identify patterns within the dataset.
- 4. To generate meaningful insights based on the identified patterns.
- 5. To ensure scalability to handle datasets of varying sizes and complexities.
- 6. To provide a user-friendly interface for easy interaction and understanding of the generated insights.

Dataset Description

The dataset used in InsightNexus is sourced from an online course platform and captures user engagement metrics. This dataset facilitates analyses on factors influencing course completion and includes the following key features:

- 1. UserID: Unique identifier for each user.
- 2. CourseCategory: Category of the course taken by the user (e.g., Programming, Business, Arts).
- 3. TimeSpentOnCourse: Total time spent by the user on the course in hours.
- 4. NumberOfVideosWatched: Total number of videos watched by the user.
- 5. NumberOfQuizzesTaken: Total number of quizzes taken by the user.
- 6. QuizScores: Average scores achieved by the user in quizzes (percentage).
- 7. CompletionRate: Percentage of course content completed by the user.
- 8. DeviceType: Type of device used by the user (Desktop or Mobile).
- 9. CourseCompletion (Target Variable): Course completion status (0: Not Completed, 1: Completed).

The target variable distribution is:

- - 0 (Not Completed): 48%
- - 1 (Completed): 52%

Methodology

InsightNexus employs a structured methodology to preprocess the data, identify patterns, and generate insights. The methodology includes the following steps:

1. Data Preprocessing:

- - Cleaning: Handling missing values and outliers.
- - Encoding: Converting categorical variables into numerical values using label encoding.
- - Scaling: Standardizing the features using StandardScaler to ensure uniformity in data representation.
- - Splitting: Dividing the dataset into training and testing sets.

2. Knowledge Representation:

- - Visualization: Utilizing libraries like Matplotlib, Seaborn, and Plotly to create various visualizations such as bar plots, pie charts, and heatmaps to represent the data effectively.
- - Exploratory Data Analysis (EDA): Conducting a detailed analysis to understand the distribution, central tendency, and dispersion of the data.

3. Pattern Identification:

- - Classification Models: Training various machine learning models (Logistic Regression, SVC, Decision Tree, AdaBoost, Gradient Boosting, Random Forest, XGBoost, and KNeighborsClassifier) to predict course completion.
- - Clustering: Implementing K-means clustering to identify natural groupings within the data.
- - Correlation Analysis: Calculating correlation coefficients to find relationships between variables.

4. Insight Generation:

- - Descriptive Statistics: Computing mean, median, mode, range, variance, and standard deviation.
- - Trend Analysis: Identifying trends, seasonality, and cycles in the data.
- - Anomaly Detection: Detecting outliers and anomalies that deviate from the general pattern.
- - Interaction Patterns: Examining interaction patterns and feature redundancy.

5. Scalability:

- Ensuring the solution can handle datasets of varying sizes and complexities using efficient algorithms and data structures.

6. User-friendly Interface:

- Designing a user-friendly interface for easy interaction and understanding of the insights using web development framework Streamlit.

Results and Discussion

The results from the InsightNexus project include the performance metrics of various machine learning models, visualizations, and insights generated from the data. The models were evaluated based on their accuracy, precision, recall, and F1 scores. Below are the performance metrics of the trained models:

Model	Test Accuracy	Train Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.793	0.790	0.793	0.793	0.793
SVC	0.871	0.888	0.872	0.871	0.870
Decision Tree	0.911	1.000	0.911	0.911	0.911
AdaBoost	0.953	0.952	0.954	0.953	0.953
Gradient Boosting	0.960	0.959	0.960	0.960	0.959
Random Forest	0.958	0.999	0.959	0.958	0.958
XGBoost	0.956	0.995	0.957	0.956	0.956
KNeighborsClassifier	0.863	0.904	0.863	0.863	0.863

Visualizations generated include:

- - Count plots and pie charts for categorical variables (CourseCategory, DeviceType, CourseCompletion).
- - Scatter plots and bar graphs for numerical variables (TimeSpentOnCourse, NumberOfVideosWatched).
- - Confusion matrices for each classification model.
- - Heatmaps showing correlations between variables.

Insights Generated:

- 1. Central Tendency and Dispersion: The mean, median, and mode of TimeSpentOnCourse show users spend an average of 15 hours on courses.
- 2. Distribution Analysis: The skewness and kurtosis of QuizScores indicate a slight skew towards higher scores.
- 3. Trend Analysis: Users tend to complete more courses during the weekends.

- 4. Correlation Analysis: A strong positive correlation (0.8) between NumberOfVideosWatched and CourseCompletion.
- 5. Anomaly Detection: Identified users with unusually high CompletionRate but low TimeSpentOnCourse, indicating possible anomalies.
- 6. Interaction Patterns: DeviceType showed redundancy with CourseCategory, suggesting a preference for certain devices within specific course categories.

Team members and contribution

Team Name : Team_Srinivas

Team members: Srinivas & Varshini

Our team is lead by an enthusiastic learner Srinivas partnered with Varshini.

Contribution:

Srinivas and Varshini worked together on the InsightNexus project. A thorough and effective project resulted from each member's contribution of their specialized knowledge. Srinivas took care of EDA, knowledge representation, and insight generation, while Varshini concentrated on pattern identification and data preprocessing. In order to make the finished product interactive, the two team members collaborated to build an intuitive user interface.

1. Data Pre-Processing – Varshini

Varshini was responsible for the data preprocessing tasks, ensuring the dataset was clean and ready for analysis. Her contributions included:

- Handling missing values and outliers
- Encoding the features using label encoding
- Splitting the data into training and testing sets.

2. Exploratory Data Analysis – Srinivas

Conducted a comprehensive EDA to understand the dataset characteristics and find out initial patterns and insights.

- Descriptive statistics to understand the central tendency, dispersion, and distribution of the data.
- Visualizations such as histograms, box plots, to identify trends and correlation.
- To visualize relationships between different features.

3. Knowledge Representation – Srinivas

Srinivas effectively represented the knowledge contained within the dataset. His contributions included:

- Creating visual representations of the data using graphs and charts.
- Key findings from the EDA

4. Pattern Identification – Varshini

Varshini focused on identifying patterns within the dataset, including:

- Detecting anomalies and outliers that could provide valuable insights.
- Identifying correlation and interaction between different features.

5. Insight Generation – Srinivas

Srinivas was responsible for generating meaningful insights based on the identified patterns. His contributions included:

- Using machine learning algorithms to make predictions and identify key factors influencing the target variable.
- Summarizing the insights in a clear and understandable manner.
- Highlighting significant findings that could be used for decision-making.

6. User-Friendly Interface

Both Srinivas and Varshini collaborated to develop a user-friendly interface for the InsightNexus project. Their contributions included:

- Designing the layout and navigation of the Streamlit application.
- Ensuring that the interface was intuitive and easy to use.
- Implementing features that allowed users to interact with the data and understand the generated insights easily.

Conclusion

The InsightNexus project successfully developed an AI-based solution that preprocesses and analyzes structured data, identifies patterns, and generates meaningful insights. The solution demonstrated scalability and a user-friendly interface, making it a valuable tool for data-driven decision-making.

Future work can focus on:

- - Integrating real-time data processing capabilities.
- - Enhancing the user interface for better interactivity.