**ADTA-5340 – Mid Term (Theory)**

Q1)

Artificial Intelligence (A.I. or AI) is everywhere in the news now-a-days. AI is the most trending topic in this decade. We are seeing many applications of AI in our day-to-day life for example, YouTube recommendations, Photo editing, self-driving cars and ChatGPT. AI is one of the technologies which is growing rapidly, and many businesses started using it in improving sales and profits.

Did you know who is Father of Computer Science? “Alan Turing”, is referred as Father of Computer Science. In the year of 1950, he was denoted birth of Artificial Intelligence in his seminal work on “Computer Machinery and Intelligence”. Alan referred a test called “Can Machines Think?” which often called “Turing Test”. This test briefly distinguishes human and computer text response. This test remains a significant part in the history of AI. The term "artificial intelligence" was introduced by John McCarthy in 1956 during a conference at Dartmouth College.

In the 1980s and 1990s, AI research focused on building machine learning algorithms, which helped machines to learn from data and improve their performance. This led to the development of neural networks using AI, which were modeled after the structure of the human brain. There is a project started by IBM named “Blue Brain Project”, which is about developing artificial brain using super computers.

In the 2000s, advances in computing power and data storage enabled the development of deep learning algorithms, which allowed machines to process vast amounts of data and make more accurate predictions. In 2011, IBM's Watson defeated two former champions on Jeopardy! a popular quiz show. Watson was a question-answering AI system that could process and understand natural language. In 2017, Google's AlphaGo Zero defeated the original AlphaGo by learning the game from scratch without any human data. This breakthrough showed the potential of unsupervised learning in AI. In 2020, OpenAI had launched a new tool called ChatGPT, which can help with your queries more than google can does. At present, few days ago, a new version of ChatGPT launched, GPT4 which has outstanding capabilities. It can create an image, a movie, or a complete book for you. There is an exponential growth in developments in the area of AI. There are many AI tools coming which can help in writing, copywriting, building nicely looking UIs of websites etc.

Q2)

Big data and machine learning have had a crucial role on various sectors of the U.S. economy, leading to new opportunities and innovation. I would like to explain this question with examples of three sectors:

1. **Healthcare Industry:**

Big data and machine learning have revolutionized the healthcare industry by providing new tools to diagnose and treat diseases, improve patient outcomes, and lower costs. Some of the ways big data and machine learning have impacted healthcare include:

1. Personalized medicine: Machine learning algorithms can analyze large datasets to identify personalized treatment options based on a patient's genetic makeup and medical history.
2. Medical imaging: Machine learning algorithms can analyze medical images to detect diseases such as cancer, reducing the need for invasive procedures.
3. Predictive analytics: Big data analytics can help healthcare providers identify patients at high risk of developing chronic diseases, allowing for early intervention and prevention.
4. **Financial Industry:**

They have transformed the financial industry by providing new tools for fraud detection, risk management, and investment decision-making. Some of the ways big data and machine learning have impacted finance include:

1. Fraud detection: Machine learning algorithms can analyze large datasets to identify patterns of fraudulent activity, helping to prevent financial crime.
2. Risk management: Big data analytics can help financial institutions identify and manage risks associated with investments and loans, improving portfolio management.
3. Trading and investment decisions: Machine learning algorithms can analyze financial data to identify patterns and trends, providing insights for trading and investment decisions.
4. **Retail industry:**

Big data and machine learning have disrupted the retail industry by providing new tools to personalize marketing, optimize inventory management, and improve customer experience. Some of the ways big data and machine learning have impacted retail include:

* 1. Personalized marketing: big data analytics can help retailers analyze customer data to personalize marketing and advertising campaigns, improving customer engagement and loyalty.
  2. Inventory management: Machine learning algorithms can analyze sales data to optimize inventory management, reducing waste and improving profitability.
  3. Customer experience: Machine learning algorithms can analyze customer behavior and preferences to provide personalized recommendations and improve the customer experience.

Overall, big data and machine learning have had a transformative impact on various industreis of the U.S. economy, leading to new opportunities for innovation, efficiency, and growth. However, these technologies also raise concerns about privacy, security, and bias, highlighting the need for responsible and ethical use.

Q3)

**Supervised learning:**

It is a machine learning technique in which an algorithm is trained to learn a mapping between input and output data based on labeled data. In other words, the algorithm is provided with a dataset in which the inputs and corresponding outputs are already known, and it learns to generalize this mapping to make predictions on new, unseen data.

Supervised learning can be used for both classification and regression tasks. In classification tasks, the goal is to predict a categorical output (such as whether an email is spam or not), while in regression tasks, the goal is to predict a continuous output (such as the price of a house).

**Unsupervised Learning:**

It is a type of machine learning in which an algorithm is trained to identify patterns and relationships in data without any labeled data. Instead of providing the algorithm with input-output pairs, the algorithm is simply given a dataset and must find structure or patterns on its own.

Unsupervised learning can be used for a variety of tasks, such as clustering similar items, dimensionality reduction, and anomaly detection. The main difference between supervised and unsupervised learning is that in unsupervised learning, the algorithm is not told what the output should look like, and it is left to discover patterns and relationships in the data on its own.

**Semi-supervised Learning:**

It is a type of machine learning that combines aspects of both supervised and unsupervised learning. In this approach, the algorithm is trained on a combination of labeled and unlabeled data, where labeled data is a subset of the dataset that has known input-output pairs and unlabeled data refers to the rest of the data without labels.

Semi-supervised learning is useful in situations where labeled data is scarce or expensive to obtain. By leveraging both labeled and unlabeled data, semi-supervised learning can improve the accuracy of the model compared to using only labeled data.

Q4)

**Structured data:**

The data is said to structured when the values are organized such that they can be accessible or retrievable. In the traditional relational systems, data is structured in the form of tables. For example, Oracle database, MySQL, MS-SQL etc. In Oracle DB, data is represented in rows and columns. There are also other tools where you can see structured data like Microsoft Excel. Structured data is easy to manage, read and migrate during any kind of operations on data. It would be easy to do analysis on structured data when compared to other types of data.

**Unstructured Data:**

The data without any defined structure is said to be unstructured data. 80-90% of the data collected by businesses is unstructured. Doing analysis on real-time data will give you more accurate results. Hence the problem for Bigdata is 80% of data available for analysis is of unstructured type. As this data cannot be stored in relational databases, it has other ways to store and manage like schema-free and schema-on-read kind of databases. NoSQL is best for flexible storage. Apache Hadoop can store both structured and unstructured data.

**Semi-structured Data:**

The data in semi-structured format do not have proper format however it is flexible to format in the way we need. Examples of semi-structured data are HTML, JSON. We can also take Emails which we use on daily basis as an example which contains ‘To’ address, subject line and body however you can add text, attachment or image in the body, which is flexible to us. Since semi-structured data has some format to it, hence it is bit easy to analyze.

Q5)

1. **Volume:** In Bigdata world, the data huge and critical for analysis to get best results. According to a report, it is estimated to be 463 Exa Bytes of data generated each day. So, storing, managing and analyze of huge data is crucial. Microsoft and Amazon together can store 1200 Peta Bytes.
2. **Velocity:** The speed at which data receives is called velocity. Velocity can be achieved by establishing high speed media of communication, low latency source systems and higher bandwidths. Now-a-days, optical fiber cables are used to transfer data in seconds. It is estimated that data transferred using optical fiber cables are 200,000 Kilometers per second.
3. **Variety:** Data is available in wide varieties like text, audio and video. When it comes to data analysis, each source of data is important. Hence bigdata need to store, manage all kinds of data.
4. **Veracity:** Veracity is defined as how accurate or truth the data is. These days data is huge, we cannot distinguish true or false data. Hence it is veracity of data is crucial in Bigdata world.
5. **Value:** Even other 4 V’s are in place but nothing can be achieved without value for the business. This is the most important characteristic out of all. Data must be effective to generate accurate results.

Q6)

The CRISP-DM (Cross-Industry Standard Process for Data Mining) framework is a widely used analytics life cycle that outlines a structured approach to solving business problems using data analytics. The framework consists of six major steps, each of which is further broken down into sub-steps. In this answer, we will discuss each step below:

**Business Understanding:**

This is the first step of the CRISP-DM framework is understanding the business problem at hand and defining the project goals. This step involves collaboration between the analytics team and the business stakeholders to identify the specific business questions that need to be answered, the relevant metrics, and the success criteria for the project.

Steps:

a. Identify the business problem or opportunity.

b. Determine the project goals and objectives.

c. Define the scope of the project.

d. Develop a preliminary plan to achieve the project goals.

**Data Understanding:**

The second level of the CRISP-DM framework is understanding the data that will be used in the analysis. This step involves collecting, exploring, and describing the data to gain a comprehensive understanding of its characteristics and quality.

Steps:

a. Collect initial data.

b. Describe the data using statistical summaries and visualization techniques.

c. Explore the data for patterns, trends, and anomalies.

d. Verify data quality.

**Data Preparation:**

The third process of the CRISP-DM framework is for preparing the data for analysis. This step involves transforming the data into a format that is suitable for analysis by cleaning, integrating, and selecting relevant features.

Steps:

a. Clean the data by removing duplicates, correcting errors, and filling missing values.

b. Integrate data from multiple sources.

c. Transform the data by creating new variables or aggregating data.

d. Select relevant features for analysis.

**Modeling:**

The fourth step of the CRISP-DM framework is for selecting and building the models that will be used to solve the business problem. This step involves choosing appropriate algorithms and tuning their parameters to achieve the desired performance.

Steps:

a. Select appropriate modeling techniques.

b. Train the models using the prepared data.

c. Evaluate the performance of the models using validation techniques.

d. Refine the models by adjusting their parameters or selecting different algorithms.

**Evaluation:**

The step 5 of the CRISP-DM framework is evaluating the performance of the models and determining whether they meet the project goals. This step involves assessing the accuracy, reliability, and stability of the models.

Sub-steps:

a. Evaluate the models using the success criteria defined in the first step.

b. Assess the accuracy, reliability, and stability of the models.

c. Determine whether the models are suitable for deployment in a production environment.

**Deployment:**

The final step of the CRISP-DM framework involves deploying the models into a production environment and monitoring their performance over time. This step involves implementing the models, testing them in real-world scenarios, and maintaining them to ensure that they continue to provide accurate and reliable results.

Steps:

a. Plan and prepare for deployment.

b. Implement the models in a production environment.

c. Monitor and maintain the models to ensure their ongoing performance.

d. Document the entire analytics life cycle for future reference.

Overall, the CRISP-DM framework provides a structured approach to solving business problems using data analytics. By following this framework, organizations can ensure that their analytics projects are well-planned, executed, and documented, which can lead to better decision-making and improved business outcomes.

Q7)

1. **Scikit-learn:**

It is an open-source machine learning library for Python that provides a range of supervised and unsupervised learning algorithms for data analysis and data mining tasks. It is built on top of popular scientific Python libraries such as NumPy, SciPy, and Matplotlib, and is designed to be user-friendly, modular, and efficient.

Scikit-learn provides a comprehensive suite of tools for various tasks such as classification, regression, clustering, dimensionality reduction, and model selection. It includes a wide range of popular algorithms such as decision trees, support vector machines, random forests, gradient boosting, k-nearest neighbors etc.

1. **Scipy:**

It is an open-source scientific computing library for Python that provides a collection of numerical algorithms and tools for scientific research, engineering, and data analysis. It is built on top of the NumPy library and extends its functionality to provide a broader range of numerical methods.

Scipy provides a wide range of modules and functions for various tasks such as numerical integration, optimization, interpolation, linear algebra, signal processing, statistics, etc.

1. **Re:**

The re (regular expression) library is a powerful and versatile Python library that provides support for regular expressions, which are a powerful and flexible way to search, manipulate and validate text data. Regular expressions are a sequence of characters that define a search pattern.

The re library provides a wide range of functions that allow users to perform various operations with regular expressions, including searching for patterns, replacing patterns, splitting strings, and more. The library provides a compact, efficient, and easy-to-use syntax for working with regular expressions.

1. **Requests:**

The requests library is a popular Python library that allows users to send HTTP requests using Python. It provides a simple and elegant way to interact with web services and web APIs.

With the requests library, users can easily send GET, POST, PUT, DELETE, and other HTTP requests to web servers and receive responses in a format that can be easily parsed and processed in Python. The library supports various authentication mechanisms, including basic authentication, OAuth, and token-based authentication.