

**NAME: PREM RAJESH PANCHAL**

**SECTION: ’B’**

**SUBJECT: INTRO TO DATA SCIENCE**

**GITHUD LINK:**

**REGISTRATION NO:** **2411021240011**

**SUBMITTED TO: MR.AMAN SHARMA**

1.**What is Data Science?**

Data Science is a field that uses various techniques and tools to extract insights and knowledge from data. It combines statistics, computer science, and domain expertise to analyze large datasets and solve complex problems. Think of it as a way to turn raw data into valuable information that can help make decisions.

**Key Components of Data Science:**

1. **Data Collection**: Gathering data from various sources such as databases, online sources, or sensors.
2. **Data Cleaning**: Preprocessing the data to remove any errors, inconsistencies, or missing values.
3. **Data Analysis**: Using statistical and computational methods to examine the data and find patterns or trends.
4. **Data Visualization**: Creating visual representations like charts and graphs to make the data easy to understand.
5. **Machine Learning**: Applying algorithms that can learn from data and make predictions or decisions without being explicitly programmed.

**CRISP-DM Process:**

The CRISP-DM (Cross-Industry Standard Process for Data Mining) is a widely used framework for tackling data science projects. It consists of six phases:

1. **Business Understanding**: Defining the goals and objectives of the project from a business perspective.
2. **Data Understanding**: Collecting and exploring the data to understand its characteristics and potential issues.
3. **Data Preparation**: Cleaning and transforming the data to make it suitable for analysis.
4. **Modeling**: Applying various algorithms and techniques to build predictive models.
5. **Evaluation**: Assessing the performance of the models and ensuring they meet the project's objectives.
6. **Deployment**: Implementing the models in real-world applications and monitoring their performance.

**Applying CRISP-DM to Real-World Problems:**

Let's look at how the CRISP-DM framework can be used to solve real-world problems, like predicting customer churn or recommending movies.

**Predicting Customer Churn:**

1. **Business Understanding**: The goal is to identify which customers are likely to leave the service.
2. **Data Understanding**: Gather data on customer interactions, purchase history, and feedback.
3. **Data Preparation**: Clean the data by removing duplicates and handling missing values.
4. **Modeling**: Use machine learning algorithms to build a model that predicts the likelihood of churn.
5. **Evaluation**: Test the model's accuracy and make adjustments if necessary.
6. **Deployment**: Implement the model to identify at-risk customers and take actions to retain them.

**Recommending Movies:**

1. **Business Understanding**: The goal is to enhance user experience by suggesting movies they might enjoy.
2. **Data Understanding**: Collect data on user preferences, watch history, and ratings.
3. **Data Preparation**: Clean and organize the data for analysis.
4. **Modeling**: Use recommendation algorithms to build a model that suggests movies.
5. **Evaluation**: Assess the model's effectiveness in providing relevant recommendations.
6. **Deployment**: Integrate the model into the platform to provide personalized movie recommendations.

* the system scales to millions of users while maintaining high accuracy.

**Netflix Recommendation System Case Study (CRISP-DM Model)**

**1. Business Understanding**

**Objective:** Enhance user experience by providing personalized recommendations to increase engagement and retention. **Key Business Questions:**

* How to personalize recommendations for each user?
* How to improve retention and engagement?
* What metrics measure system performance?

**2. Data Understanding**

Netflix collects vast amounts of structured and unstructured data: **Key Datasets:**

* **User Data:** Demographics, account details.
* **Viewing History:** Watched movies/shows.
* **User Ratings:** Explicit feedback.
* **Search Queries:** Keywords searched.
* **Content Metadata:** Genre, cast, director, release year.

**Data Characteristics:**

* **Volume:** Millions of users and content items.
* **Variety:** Structured (ratings) & unstructured (reviews, queries).
* **Velocity:** Continuously updated.
* **Veracity:** Ensuring data accuracy and consistency.

**3. Data Preparation**

**Steps:**

1. **Data Cleaning:** Remove duplicates, handle missing values.
2. **Data Transformation:** Convert categorical data to numerical (e.g., one-hot encoding for genres, normalize ratings).
3. **Feature Engineering:**
   * Calculate user engagement scores.
   * Create content similarity scores.
   * Extract text features using NLP.

**4. Modeling**

Netflix employs machine learning techniques for recommendations: **Techniques Used:**

1. **Content-Based Filtering:** Recommends similar content based on metadata.
2. **Collaborative Filtering:** Recommends based on user-item interactions (Matrix Factorization - SVD, ALS).
3. **Hybrid Model:** Combines content-based & collaborative filtering for robust recommendations.

**Advanced Techniques:**

* **Deep Learning Models:**
  + **Autoencoders:** Capture latent features.
  + **RNNs:** Model sequential viewing behavior.
  + **Transformers:** Personalization based on history.

**5. Evaluation**

**Metrics Used:**

* **Precision & Recall:** Measure recommendation relevance.
* **F1-Score:** Balances precision and recall.
* **RMSE & MAE:** Evaluate prediction accuracy.

**A/B Testing:**

* Compare different models via user engagement metrics.

**Evaluation Types:**

* **Offline:** Simulate user behavior using historical data.
* **Online:** Real-time measurement through user interactions.

**6. Deployment**

**Steps:**

1. **Backend Integration:** Deploy model within Netflix’s infrastructure.
2. **Real-Time Updates:** Ensure dynamic recommendation generation.
3. **Monitoring:** Continuously track system performance.

**Tools Used:**

* **Apache Spark:** Large-scale data processing.
* **AWS/Azure:** Cloud deployment.
* **Kafka:** Real-time streaming.
* **TensorFlow/PyTorch:** Deep learning models.

**7. Maintenance**

**Continuous Improvement:**

* **Periodic Retraining:** Keep model updated with new data.
* **User Feedback Loop:** Improve recommendations.
* **Feature Enhancements:** Add new features & algorithms.
* **Scalability Improvements:** Handle growing data efficiently.
* **Exploration of New Algorithms:** Graph-based recommendations, reinforcement learning.