Data Research Analyst

305 שעות לימוד אקדמיות

כולנו חיים בעידן המידע ועובדים או שואפים לעבוד בארגונים שמגדירים את עצמם כ- Data Driven Company. ארגונים אלה בוחרים אסטרטגית קבלת החלטות על בסיס ניתוח דאטה שנאסף.

תהליך ניתוח נתונים הוא תהליך מורכב ומאתגר. ככל שארגון אוסף יותר ויותר נתונים, נדרשת הבנה מעמיקה יותר של תהליכים עסקיים של הארגון ושליטה בכל הדאטה שנאגר. בסופו של דבר החיבור בין השניים מוביל להצלחה בהפקת תובנות עסקיות קריטיות חיוניות עבור הארגון.

נאיה קולג', חטיבת הדרכה בחברת נאיה טכנולוגיות, מתמחה בעולמות הדאטה מעל עשור ומציעה מסלול הכשרה וקורסים בתחום הדאטה במטרה להכשיר את הדור החדש של מומחי דאטה מקצועיים ובעלי ידע העדכני והנדרש ביותר בתעשיה.

מסלול הכשרה זה צמח מהשטח לנוכח הביקוש הגובר לאנליסטים בעלי יכולות טכניות גבוהות ובעלי ידע מעשי ופרקטי בעבודה עם טכנולוגיות מגוונות.

תיאור תפקיד:

כיום ,תפקידו של אנליסט הנתונים אינו תחום ייחודי למגזר העסקי בלבד, אלא הוא תפקיד הנדרש בכל ענף עתיר נתונים, כולל גם מגזר ציבורי, זרועות הממשל השונות וכמובן בענפי המדע השונים. האנליסטים עובדים עם כלים מגוונים, נדרשים לחשיבה אנליטית על מנת להפיק ידע מהנתונים, כישורי למידה עצמית, ולרוב הם בעלי תארים אקדמיים בתחומים מדויקים.

בשנים אחרונות תפקיד של אנליסט עבר אבולוציה מדהימה. הוא התחיל את דרכו הטכנולוגית בעבודה על קבצי אקסל, בעיקר, בהמשך נדרש לעבודה עם שפת SQL על מנת לבטל את התלות באנשי IT בארגון להשגת הנתונים הרלוונטיים לניתוח, רכש יכולת לעבוד עם כלי ויזואליזציה מתקדמים כגון Tableau ואחרים, ולבסוף כיום אנו עדים לדרישה בידע בכלים טכנולוגיים נוספים שמקפיצים את החשיבות ואת הערך שלו בארגון. ובפרט היכרות מעמיקה עם Python (כשפת תכנות שמהווה כלי נוח וגמיש יותר בעבודה עם דאטה לא טבלאי ותחקור אנליטי), אך גם יישום שיטות סטטיסטיות, עבודה עם נפחים גדולים של נתונים, היכרות עם עולם ה-Machine Learning ו-Big Data ויכולת לנתח נתוני Big Data.

בין המשימות היומיומיות של האנליסט אפשר למנות:

- עבודה עם כמויות גדולות של נתונים (גם מול בסיסי נתונים רלציוניים וגם NoSQL)
- Querying Data - כתיבת שאילתות SQL ו-Python
- Data Processing – אינטגרציה, עיבוד והכנת נתונים לניתוח
- Data Modeling – מידול נתונים - בניית מודל נתונים שיושב בבסיסו של כלי BI המתאר קשרים בין הנתונים
- Dashboards and Reports - בניה ותחזוקת דוחות, מדדים וערכים, ופיקוח על ביצועי המוצר
- Data Visualization - יישום טכניקות ויזואליזציה להצגת נתונים גם בכלי BI וגם באמצעות Python
- יישום שיטות סטטיסטיות, A/B Testing וניתוח היעילות שלהם
- Data Analysis Project – יכולת להוביל פרויקטים אנליטיים מקצה לקצה כולל איסוף ומניפולציה על נתונים, מיזוג ומידול, הגדרת מדדים והערכת ביצוע, פתרון בעיות
- עבודה עם בעלי תפקידים נוספים (מנהלי מוצר, מנהלי פרויקטים, מפתחים, אנליסטים, Data Scientists, Data Engineers ואחרים) כדי לבסס הבנה והיכרות עם הצרכים ומטרות ארגוניות

מבנה המסלול:

מסלול הכשרה זה כולל כלים רבים הנדרשים בשלבים השונים של העבודה עם נתונים בארגון.

בשלב הראשון נחזק ונעמיק את הידע **בשפת SQL** לכתיבת שאילתות מורכבות, יצירה של אובייקטים, שימוש בפונקציות אנליטיות לתחקור של הנתונים. הלימוד במודול זה יבוסס על Use Cases רלוונטיים במטרה לחבר בין החומר הנלמד לבעיות עסקיות אימיתיות איתן מתמודד אנליסט.

עולם ה-BI החדש שופע כיום בכלים מתקדמים המאפשרים **Self-Service BI** ותפקיד האנליסט תופס מקום יותר ויותר משמעותי שכן הוא אינו נדרש להסתמך על אנשי IT ו-BI. במודול השני נלמד לעבוד עם כלי מוביל בתחום BI – **Tableau**. הלימוד יכלול את כל התהליך מקצה לקצה, החל מיבוא נתונים, הכנתם לקראת הניתוח, נתייחס למידול נתונים, כולל חישובים בסיסיים ומתקדמים, וכמובן בניית Dashboard ועבודה עם טכניקות ויזואליזציה מתקדמות שהכלי מציע.

תחום נוסף שנלמד הקורס הינו אחד התחומים החמים בעולם האנליזה, אשר גרם להתפתחות וכניסה של כלים כמו Tableau – הינו תחום **Data Visualization**. במודול הזה נלמד עקרונות חשובים בוויזואליזציה של הנתונים, ובניית דשבורדים תוך כדי יישום של עקרונות אלה.

בחלקו הרביעי של המסלול נלמד לתכנת ב-**Python**, השפה המובילה כיום לתחקור הנתונים. נרכוש כלים לעבודה עם נתונים ממקורות שונים ולהצגתם. המודול כולל גם תכנות בסיסי וגם מתקדם עד לרמה של פיתוח מונחה-עצמים (Object-Oriented) כאשר גם בעולמות הדאטה תהליך הפיתוח מכיל עבודה עם אובייקטים.

בהמשך המודול נלמד את סט הכלים לעבודה עם נתונים בכלל ונתונים טבלאיים בפרט – חבילות יעודיות לעבודה עם נתונים, במרכזן חבילת ה-**pandas**. דרכה ניחשף לעקרונות שונים בהכנה ובוויזואליזציה של נתונים תוך כדי התייחסות לחבילות נוספות כגון **numpy**, **matplotlib** ו-**seaborn**. נלמד לעבוד עם קבצים ומקורות שונים כגון CSV, JSON ו-HTML, עבודה עם בסיסי נתונים ורשת האינטרנט. בעזרת הכלים הללו נכיר לעומק את מגוון השיטות של Exploratory Data Analysis (**EDA**). בסיומו של המודול סטודנטים יהיו מסוגלים לבנות Data Pipeline מקצה לקצה.

המודול הבא עוסק בסקירה של מונחים, **שיטות וטכניקות מעולם הסטטיסטיקה**. לאחר שנכיר את מושגי היסוד, נראה כיצד הטרמינולוגיה החדשה מסייעת לנו לתאר את הדאטה בצורה שלמה יותר. המונחים יעזרו לנו לנסח בצורה מתמטית יותר את ההתפלגות של הדאטה, ובכך יאפשרו להחיל על הדאטה הרלוונטי חוקים סטטיסטיים בכדי להחליט האם תופעה מסוימת צפויה ו/או מובהקת. המטרה העיקרית במודול הזה הינה לקשר בין העולם העסקי והדאטה הארגוני לתחום מדעי של סטטיסטיקה ולהבין איך בעזרת שיטות סטטיסטיות ניתן להפיק תובנות נוספות על הנתונים.

מודול הבא מציג סקירה מעמיקה של תחום **Machine Learning**. גם אם אנליסט נתונים לא מריץ אלגוריטם על דאטה, שכן זהו תפקידו של Data Scientist בארגון, עדיין ישנה חשיבות רבה להיכרות עם התחום. במודול זה נסקור את סוגי הבעיות הנפוצות תחום ML ונכיר מונחי יסוד. מטרת המודול לספק היכרות בסיסית עם עולמות ה-ML ולאפשר לאנליסט לתקשר ברמה מקצועית ובשפה משותפת עם חוקרים (Data Scientists) בארגון.

מודול אחרון במסלול הכשרה זה יחשוף את הסטודנטים לעולם ה-**Big Data**, ויעניק יכולת ניתוח נתונים בסביבת **Hadoop** - פרויקט-על מבוסס קוד פתוח של קרן התוכנה אפאצ'י, שמטרתו לעבד כמויות גדולות של נתונים (Big Data) בסביבת הפיתוח. נתחיל בהיכרות עם עולמות ה- Data המסורתיים וטכנולוגיות Big Data חדישות השולטות כיום בשוק המידע. ולאחר מכן נצלול ללמידה של כלים שונים המאפשרים ניתוח יעיל, מהיר ופשוט של נתונים בסביבת Hadoop, על מנת לייצר ערך עסקי לארגון מהנתונים שבבעלותו.

מעבר לתרגול רב בשיעורים, במהלך המסלול הסטודנטים עובדים על **תרגילים מסכמים לכל מודול** שמאפשרים לסטודנטים ליישם את הידע הנרכש במהלך הלימודים על בעיות עסקיות אמיתיות. בנוסף הסטודנטים יעבדו על פרויקט גמר.

כחלק מפרויקט גמר הסטונדטים יממשו תהליך עבודה הכולל את מגוון הכלים והטכנולוגיות שנלמדו במהלך הקורס. כחלק המפרויקט יבוצעו: עבודה אל מול בסיסי נתונים, חישוב והבנה של תובנות עסקיות (לרבות התפלגות) בעזרת שפות סקריפטים וקוד (SQL, Python) והנגשת המידע למשתמשי קצה בעזרת כלי BI (Tableau).

קהל יעד:

המסלול מיועד כל מי שמעניין להיכנס לעולם המרתק של ניתוח נתונים ברמה טכנולוגית מתקדמת ובעלי תארים בתחומים: סטטיסטיקה, מתמטיקה, מערכות מידע, תעשיה וניהול, מדעי המחשב ובעלי רקע ונסיון בסיסי בעבודה עם נתונים.

דרישות קדם:

- תואר אקדמי בתחומים בתחומים רלוונטיים
- שפת SQL ברמה בסיסית ומעבר מבחן כניסה ב-SQL
- נסיון בעבודה עם נתונים ומערכות מידע (כלי BI) - יתרון
- אנלגית ברמה גבוהה

תוכנית הלימוד:

Part 1: Basic SQL Baseline

- Getting to know your Data
- Basic Select
- Where Statement
- Order By
- Scalar Functions
    - Math functions
    - Text functions
    - Date & Time functions
    - Conversion Functions
- Isnull & Coalesce
- Case Statement
    - Simple Case
    - Advanced Case
- Group Functions
- Having
- Join
    - Inner Join
    - Outer Joins
- Union
- Excpt
- Intersect
- Sub Query
- Derived Tables
- DML

Part 2: Advanced SQL

Window Functions

- Ranking Functions

  - Row Number
  - Rank
  - Danse Rank
  - Ntile
- Aggregative Window Functions
- Offset Functions
  - Lag
  - Lead

- Previous Class Overview
- Window Functions: Controlling the Window
- Views
- CTE: Common Table Expression
- USE CASES
- FINAL PROJECT

Part 3: Tableau

- Connecting to Data
  - Setting Up Connections and Data Sources
  - Organizing Your Data
  - Relationship Data Modeling
- Desktop Basics
- Creating a Report
  - Displaying your data
  - Understanding Fields
  - Chart types
  - Filtering & Sorting
  - Organizing Data: Groups, Sets and Bins
  - Parameters – Dynamic and Statics
- Advanced Reporting
  - Creating and Editing Calculated Fields
  - Aggregations, String Functions, Date Calculations
  - Table Calculations
  - Level of Detail
  - Data Blending
  - Advanced Charts
  - Analytics and Forecasting
  - Mapping Data Geographically
- **Advanced Formatting**
- **Dashboards**
  - Combining views into a dashboard
  - Sizing, Layout and Formatting
  - Interactivity: Filters and Actions
- Final Project

Part 4: Data Visualization

עקרונות בהצגה ויזואלית של נתונים - Data Visualization

- מה זה דאטה ויז ולמה צריך את זה
- סוגי ויזואליזציות בעולם העסקי
- איך לספר את הסיפור של המספרים
- סוגי גרפים, מפות וטבלאות
- הטיות קוגניטיביות בהצגת נתונים + חוקי הגשטלש
- צבעוניות ומשמעותה
- ביצד להשתמש בטקסטים וטיפוגרפיה בצורה נכונה
- אינטראקטיביות ויתרונותיה בהצגת הדאטה

בניית דשבורדים שמתאימים למשתמשים

- מהו תהליך תהליך עבודה נכון על דשבורד
- צרכים של המשתמשים
- סוגי דשבורדים
- בניית היררכיה ולייאאוט, הגדרה מטרה ופעולה
- בחירת סוג הגרף, המפה או הטבלה
- שימוש בצבעים כדי להדגיש את הבעיות
- שימוש בטקסטים וטיפוגרפיה בצורה נכונה
- אינטראקטיביות בדשבורדים

Part 5: Python Programming

Basic Python

- Fundamentals
  - Intro
  - Python essentials
  - The working environment
- Data types
  - Numbers
  - Strings
  - Booleans
  - None
- Collections
  - Lists
  - Tuples
  - Dictionaries
  - Sets
- Control flow
  - if…else
  - for…in
  - list comprehension
  - while

- Textual interface
    - input
    - format

Intermediate Python

- Functions
    - User-defined functions
    - Built-in functions
    - Lambda expressions
- Text files
- The standard library
    - import
    - datetime

EDA and visualizations

- The *NumPy* library
    - Array
    - Broadcasting
- The matplotlib library
    - matplotlib objects
    - Plotting
- The *pandas* library
    - Series and Index
    - DataFrame
    - GroupBy
    - Visualizations
- Advanced visualizations
    - Seaborn
    - Plotly
- General tools
    - Intro to regular expressions (re)
    - API's and Connecting with Data Resources
        - JSON
        - Intro to working with DBs with SQLAlchemy package
- Final Project

Part 6: Introduction to Statistics and Data Analysis

- Descriptive Statistics
- Probability 1

- Probability 2
- Combinatorics
- Discrete Random Variables
- Continuous Random Variables
- Central Limit Theorem and Time Series
- Hypothesis testing and A/B testing
- Exercise

Part 7: Machine Learning Introduction

Introduction – short overview

- Background and motivation
- Programming fundamentals

Data Preparation

Regression

- Introduction and measures
- Linear regression

Classification

- Introduction and measures
- Decision tree
- Logistic regression
- k-nearest neighbors and the metric concept

Clustering

- Introduction and measures
- k-means
- Agglomerative clustering and the linkage concept

Miscellaneous

- Deep Learning
- Recommender systems
- Text mining and NLP

Part 8: Big Data Analytics

Data Technologies Introduction

- Introduction to Big Data
- Hadoop - a closer look
- NoSQL
- Technologies and trends in the world of Big Data

Hadoop Infrastructure and Ecosystem Services

- Hadoop core and ecosystem - The essentials (e.g. HDFS)
- Data ingestion technologies – Overview (Sqoop, Flume, Nifi)
- Using Hive to query HDFS data with SQL
- Using Impala to query HDFS data with SQL
- Deeper look into Hive and Impala
    - Understanding Hive tables
    - Working with partitions
    - Understanding file formats
    - Useful functions and tools for Hive (Regex, Ngrams, User-defined functions)
    - Working with unstructured data
    - Querying complex data types
    - Performance considerations – Statistics and execution plans
- Workflow managers and job schedulers - Introduction to Airflow and Oozie
- Final Project