



Hands-On DATA SCIENTIST

סילבוס קורס

מסמך זה מיועד למטרה שלשמה החמן ולא לעיונו של כל אדם אחר. אין להעתיק לשכפל ולגלות תוכן מסמך זה, חלקו או כולו, או לעשות בו שימוש כלשהו אלא על פי הסכמה מראש ובכתב של חברת GSTAT בע"מ

G-Academy

G-Academy, חברת הדרכות של קבוצת G-STAT, מתמחה בהעברת קורסים והכשרות לחברות וללקוחות פרטיים במגוון נושאים בעולם ניתוח הנתונים, Big Data, וה- (ML) Machine learning. מטרת G-Academy הינה: הכשרת הלקוחות ליישום פרקטי של הכלים והתוכן הנלמדים מתוך מטרה זו, אנו מקפידים על:

- ✓ מרצים מומחים בעלי רקע הדרכתי וניסיון פרקטי Hands-on מוכח בכלי ובתוכן הנלמד
- ✓ ליווי לקוח החל משלב תכנון צרכים וכתובת סילבוס ועד העברת כלל ההרצאות תוך מתן דגש ליחס אישי
- ✓ יעילות ואפקטיביות התכנים המועברים ומתן מחירים אטרקטיביים

G-Academy מכשירה ומדריכה סטודנטים בהתאם למקצועיות ולניסיון של חברת G-STAT שהינה החברה הגדולה והמובילה בארץ בניתוח נתונים, פיתוח ויישום מודלים סטטיסטיים ובעלת ותק של 20 שנות פעילות.

בנוסף, G-Academy הכשירה עד כה מעל ל-200 דאטה אנליסטים חדשים בישראל ורואה בכך גאווה וזכות גדולה. מכאן החזון שלנו התפתח לעבור לשלב הבא - להכשיר דאטה סיינטיסטים.

בברכה,

רו"ח אסף אלקן

מנכ"ל G-ACADEMY

Hands-On Data Scientist

משך הקורס: כ-5 חודשים, 320 שעות אקדמאיות*

תאריך התחלה: 3.5.2023

סיום משוער: 11.10.2023


ימי ושעות הקורס: ימי רביעי בשעות 17:30 – 21:30


כללי


*שעות מפגשי לימוד: 160


שעות תרגול ולמידה עצמית: כ-100


שעות עבודה עצמאית על הפרויקטים: כ-60

 הקורס מועבר באופן "היברידי" התואם את שוק העבודה של ימנו, רובו מועבר בזום וישנם מועדים ספציפיים המועברים פרונטלי (כ-7 מפגשי למידה פרונטליים מתוך כ-22 מפגשי למידה בקורס). לגבי המועדים הפרונטליים: אפשר להתחבר אליהם דרך הזום, מועברים במכון מופ"ת, רחוב שושנה פרסיץ 15, תל אביב ויש חניה חינם.

 בנוסף, מתקיימים מפגשי תרגול בימי ראשון בין השעות 19:00-21:00 בזום.

 כלל השיעורים מוקלטים וזמינים לצפייה לפחות שבועיים ממועד העברתם.

 הרישום מותנה במקום פנוי, בתשלום ובסיום תהליך מילוי טפסי הרישום המתקבלים לאחר מעבר ריאיון התאמה לקורס.

 נדרש מחשב אישי לקורס (מומלץ מחשב נייד עם מערכת הפעלה Windows, אם בכוונתכם להשתמש במערכת הפעלה אחרת אנא ידעו אותנו בבקשה מראש).

על תפקיד הדאטה-סיינטיסט

בעידן הנוכחי נוצר צורך לייצר ערך מכמויות הדאטה הנאספות מידי יום, הצורך בתובנות עיסקיות מנתונים הוא קריטי לארגונים, זאת כיוון שבאמצעות זאת ניתן לחקור להבין וללמוד את תנודות השוק, לשפר מכירות ולעשות עיבודי מידע, אלו יתנו לארגון יתרון אסטרטגי בשוק..
דאטה סיינטיסט הוא כל מי שמפיק תובנות עיסקיות מנתונים רבים, בעזרת מודלים מתמטיים מתוחכמים (או בשמם הנפוץ למידת מכונה Machine Learning). ביום יום משרה זו זורשת לדעת לתכנת/ת, להשתמש בשיטות סטטיסטיות ולרוב להבין את המתמטיקה של המודלים מעולם למידת המכונה על מנת ללמד אותם את המידע שבידיו/ה. מטרת הלמידה היא לפתור בעיות או לספק תובנות עיסקיות.

ארבעת התכונות הנדרשות עבור ה-DATA SCIENTIST:

- 📌 רקע מתמטי גבוה והבנה בכלים סטטיסטיים.
- 📌 יכולות תכנות כדי לכתוב קוד וכן להבין קוד של אדם אחר.
- 📌 הבנה עסקית עמוקה בתחום העיסוק, היכולת להבין ולנסח בעיה.
- 📌 אנגלית ברמה טובה, בקריאה וכתיבה.

קהל יעד

- 📌 בעלי רקע יישומי ב BI \ data analyst מסדי נתונים ומערכות מידע.
- 📌 בעלי ניסיון / רקע באחד או יותר תחומים המתוארים להלן:
 - רקע בתכנות בשפת ניתוח נתונים פופולארית בשוק (Python, R, SQL)
 - ניסיון בניתוח נתונים (אקסל מתקדם, SQL, כלי BI)
 - בעלי תארים אקדמאים: מדעי מחשב, הנדסת מערכות מידע, הנדסת תוכנה, הנדסה תעשייה וניהול, פיזיקה, סטטיסטיקה/מתמטיקה, כלכלה.
- 📌 יעוץ עם גורם מקצועי + מעבר מבדקת התאמה הבוחן יכולות אנליטיות של המועמד/ת
- 📌 שליטה טובה בקריאה וכתיבה בשפה האנגלית



סילבוס

INTRO TO DATA SCIENCE:

The first module presents a general description of the data scientist position, common business use-cases and how to solve them using Machine learning and Data science capabilities.

We will provide a detailed description of the course outline and describe the basic libraries which we will use during the course.

Course outline

- Business use-cases and solutions
- What is data science, and what does a data scientist do?
- The data science pipeline

STATISTICS

The second module is dedicated to the statistical knowledge needed from a Data-Scientist.

In order to process and define key features from Big-Data, there is a need to test and evaluate the full extent of the data, using statistical methods.

- mean, median, standard deviation
- Random selection VS population
- Data Distribution – Focusing on Gaussian
- statistical tests– Z, T,chi
- Type one and type two errors
- Variables type



EXPLORATORY DATA ANALYSIS (EDA) && VISUALIZATION

In this module we will focus on exploring the data, his behavior and shape, using a variation of graphs and visualizations. This process is the main part of the day-to-day work of a data scientist. We will present the science behind analyzing big amount of data and how to enable generalization capabilities on a population.

In this module:

- Data vitalizations
- Outliers detection and treatment
- Correlation and collinearity
- Graphing using Matplotlib and seaborn

FEATURE ENGINEERING

In this module, we will examine the steps and ways to select features for the future machine learning model. This section is devoted to data manipulation and selection.

We will present a few of the most effective strategies for organizing data and selecting the most appropriate features. This will ensure that the model can be taught in the most comprehensive manner.

In this module:

- Data cleaning
- Dealing with missing values
- Transformations and aggregations
- One-dimensional analysis
- Transformations in the time dimension



INTRO TO MACHINE LEARNING

In this module, we will explore the world of machine learning and the various existing models. We will understand which tasks can be solved using ML, what are the appropriate models for each task and what are their basic hypotheses.

In this module:

- Machine learning models
- Supervised vs. unsupervised
- Scikit-learn

SUPERVISED LEARNING

Supervised learning is a branch within machine learning where we train the model based on labeled information, "solved examples". The model learning itself is performed by searching for a hypothesis within the solution dimension and reducing the error.

In this module:

- Over-fit and Under-fit of the model
- Cross validation
- Train / Test – stratified
- Unbalanced data

SUPERVISED LEARNING – REGRESSION

Regression is a technique for investigating the relationship between independent variables and a dependent variable. It is used as a method for predictive modeling in machine learning, where an algorithm is used to predict a continuous number.

- Linear regression
- Decision tree and forests
- Time series
- Performance measure (MSE, R^2 , R^2 adj)



CLASSIFICATION

A labeling task classifies the records into selected classes (for example spam email versus non-spam - every email we receive will be classified into one of these labels according to the model's decision). This is a function from the space of examples (our data) to a space of labels (classes we want to categorize into). We will learn how to estimate the quality of the model and how to analyze its errors.

- Logistic Regression
- Decision trees
- Ensemble and bagging
- AdaBoost, XGBoost, Bagging
- Neural networks
- performance measures (AUC, Accuracy ,Confusion Matrix)
- Deep learning

UNSUPERVISED LEARNING AND CLUSTERING

Unsupervised learning uses machine learning algorithms to analyze and aggregate unlabeled data sets ("answer"/ label). Algorithms uncover patterns or groupings of data without human intervention by reducing dimensionality, unifying by analogy, or explaining variance.

- Unsupervised data
- Dimension reduction
- K-Means
- Hierarchical clustering
- PCA – principal component analysis
- Performance measures for Clustering



MODEL OPTIMIZATION

In this module we will learn how to optimize the ML model and how each of the parameters affects the learning of our model.

- Hyper parameter tuning
- Feature importance
- Model optimization

EXPLAINABILITY

In this module we will learn how to explain the results of the model and how each of the parameters affects the learning of our model. It is not enough to produce a model with high accuracy. You also need to explain the reasons and business benefits of the model's result. In addition, you need to demonstrate that it does not discriminate against certain populations and manage the model's accuracy over time.

- Model explainability – Shap
- feature importance
- Feature selection

TEAM WORK IN DATA SCIENCE

Tools for collaborative data science

In this module:

- Overview Jupyter notebooks
- Python best practices
- Git
- How to operate in a team



PRODUCTION ENVIRONMENT AND MODEL MONITORING

A production environment is a term primarily used to describe the environment in which software and models are actually put to use for their intended uses by end users. A production environment can be considered a real-time environment where programs are run for corporate or commercial activity. In this module we will discover how to upload an ML model to production, how to maintain it while it is there and how to allow access to others.

- Upload a model to production
- Model drift and data drift
- Bias over time
- API of a model

DEEP LEARNING

In this module, we will present the most significant innovation in machine learning- Deep Learning. We will explore the different uses of deep learning and the network architectures that enable each form of use. Text processing, image recognition, and extracting information from a video feed are some of the uses that we will present.

- Evolvement of deep learning
- Perceptron, NN
- PyTorch and TensorFlow
- Image recognition (CNN)
- Natural Language Process (RNNs)



CLOUD COMPUTING

Cloud computing is the delivery of computing services (including servers, storage, databases, networks, software, analytics, and machine learning) over the Internet ("the cloud") to offer faster innovation, flexible resources, and knowledge sharing.

In this module:

- Cloud computing Basics
- Cloud platforms
- Scaling

DEEP LEARNING ON THE “CLOUD”

Deep learning is one of the most advanced forms of machine learning, capable of processing and understanding both images and text. Deep learning requires a lot of computation capabilities (mainly GPUs). The most common solution is cloud computing, which involves processing data on computer power in the cloud. Another common use is a data science platform in the cloud. We will present the different uses of deep learning on one of the most known data science platforms in the world.

In this module:

- Machine learning on Cloud
- Machine learning in AWS
- Data science platform
- Implementation of Image recognition model



FINAL PROJECT

The concluding project of the course is a connecting factor, connecting all the parts learned during the course. As part of the project, the students will deal with an unfamiliar data set and will analyze it in-depth, beginning with forming analytical questions, performing the analysis, building a model, and explaining the results.

The students will carry out a practical and final project that will simulate a real work environment that includes:

1. Analysis of goals and business problems with the tools learned.
2. Work in a team accompanied and supported by an experienced data science team leader.
3. Work meetings and presentation of products in front of a business entity, and the decision maker