Data Science Bootcamp

560 שעות לימוד אקדמיות

תיאור התפקיד:

מטרת תפקידו של **מדען הנתונים – Data Scientist** הינה ביצוע מחקרי מידע מעמיקים בכדי להפיק תובנות עסקיות בהתבסס על חיזוי או זיהוי דפוסי התנהגות בנתונים עסקיים של הארגון. תהליך העבודה על הנתונים ארוך, מכיל הרבה שלבים וביניהם ניקוי וטיוב של הנתונים, סידור, השלמת פערים ותיקונים, הכנה של נתונים להרצה של מודל, הפעלת אלגוריתמים שונים של מידול, כריית מידע ו- **Machine Learning** על המידע וכמובן תקשור התוצאות למקבלי החלטות. מודל טוב יישום בתהליכים העסקיים של החברה ויניב תוצאות רצויות.

תיאור ההכשרה:

מסלול זה מקנה את הכלים הנדרשים לכל שלב ושלב בעבודתו של ה-Data Scientist עם דגש על פרקטיקה ויכולות תכנות מתקדמות. מעבר לתרגול השוטף שיתבצע כחלק מתהליך הלימוד של כל נושא, יינתנו במהלך המסלול פרויקטים "אמיתיים", כך שבסוף המסלול יהיה ברשות הסטודנט תיק עבודות מכובד שילווה אותו בהמשך דרכו. מטרת ההכשרה להכין את הבוגרים לראיונות עבודה לתפקידי Data Scientist Junior.

מסלול הכשרה זה בונה את הידע בהדרגתיות תוך כדי יישום של הטכניקות הרלוונטיות ביותר וחשיפה לטכנולוגיות השכיחות בתעשיה. התשתית של ההכשרה מאפשרת לסטודנטים לחוות אלמנטים שונים של יישום בשטח ודרכם, מעבר ללימוד התאורטי של התחום והנושא, להכיר איך מריצים פרויקט Data Science בחברות היום ועם אילו אתגרים מתמודדים הן בצד העסקי וכן בצד הטכנולוגי.

- ההכשרה מועברת בסביבת ענן Google Colab – סביבה שהופכת להיות יותר ויותר נפוצה, במיוחד בסטארט-אפים, ומאפשרת הרצה של הקוד והמודלים
- ההכשרה מבוססת על Use-cases רבים ומגוונים על מנת לחשוף את הסטודנטים לשאלות עסקיות שונות, סוגי דאטה ואתגרים שתמונים בה
- ההכשרה כוללת מספר פרויקטים מעשיים ורמתם עולה מאחד לשני. עבודה על הפרויקטים תתקיים בצוותים ותכלול הצגה – כהכנה לעבודה בשטח
- ההכשרה תכלול כלים וטכנולוגיות שיסייעו ל-Data Scientist מתחיל להתמצא בסביבה הטכנולוגית הארגונית.

תוך כדי הכשרה הסטודנטים ילמדו ויידעו לענות על שאלות מקצועיות מגוונות וביניהן:

- מה ההבדל בין מודלים שונים, ובאילו תרחישים נעדיף מודל מסוים
- איזה מדדים קיימים בסוגי בעיות Machine Learning וביצד נבחר את המדד המתאים לבעיה העסקית
- איך מתמודדים עם דאטה לא מאוזן
- הורדת כמות המשתנים בשיטות של Feature Selection ו-Dimensionality reduction
- הכרות בסיסית עם יסודות תיאורטיים בעולמות האלגברה והסטטיסטיקה
- ועוד שאלות רבות!

מסלול הכשרה זה מלווה בלפחות 50% תרגול מעשי במהלך השיעורים, ובנוסף הסטודנטים יקבלו משימות ופרויקטים לכל נושא רלוונטי, במסגרתם יוכלו לממש את הידע הנרכש במהלך השיעורים. תוצרים של הפרויקטים יוצגו במפגשי סיום לכל נושא, יועלו לחשבון GitHub של כל סטודנט ובכך ייצרו תיק עבודות עשיר ומקצועי להצגה בפני המעסיקים בהמשך.

קהל יעד ודרישות קדם:

בעלי נסיון ב-Data Analysis, BI, פיתוח, המעוניינים להעשיר את יכולותיהם בתחום תחקור הנתונים.

- בעלי נסיון באחד או יותר תחומים המתוארים להלן:
  - רקע בתכנות בשפה עילית כלשהי
  - נסיון בניתוח נתונים (SQL, כלי BI)
  - דרישות קדם בקרב מפתחים: נסיון בפיתוח תוכנה בסביבה של מוצרי דאטה, עם נסיון בממשק מול אנשי דאטה
  - תארים רלוונטיים: מדעי מחשב, מערכות מידע, הנדסה, מדעים מדויקים, מדעי החיים, סטטיסטיקה/מתמטיקה, תעשיה וניהול.
- יעוץ עם גורם מקצועי + מעבר מבדק התאמה הבוחן יכולות אנליטיות של המועמד/ת
- שליטה טובה בשפה האנגלית

תנאים לקבלת תעודת סיום קורס:

- 80% נוכחות מינימום
- הגשת כל הפרויקטים במהלך ההכשרה

תוכנית הלימודים:

Section 1 – Python

במודול זה נלמד לתכנת ב-**Python**, השפה המובילה כיום לתחקור הנתונים, ונרכוש כלים לעבודה עם נתונים ממקורות שונים ולהצגתם. נתכנן ונפתח קוד מאפס, נכיר את ספריית המודולים העשירה של השפה ונדע כיצד להיעזר בה. נתוודע אל סביבות העבודה שבהן נעבוד במהלך הקורס – Google Colab (עבודה עם Jupyter Notebooks), PyCharm ו-git. בסופו של הפרק כל סטודנט מתכנן ומיישם פרויקט ב-Python, ומנגיש את הגרסה המלאה שלו ב-GitHub.

- The working environment
- Data types
- Data structures (list, dictionary, etc.)
- Flow control (if-else, for-in, etc.)
- Textual interface
- Functions (inc. lambda)
- Working with files
- Object-Oriented Programming (OOP) basics
- Python API's
    - Python Standard Library
    - Modules and packages
    - datetime
    - Regular expressions

Section 2 – EDA - Exploratory Data Analysis

בפרק זה נסקור מושגים וכלים שימושיים בעבודתו היומיומית של כל data practitioner. נסקור את משפחת החבילות מ-**PyData**, המהוות את סט הכלים המושלם לעבודה עם נתונים כולל עבודה עם נתונים מובנים (structured) ולא מובנים (unstructured). נכיר לעומק את חבילת ה-**pandas**, דרכה ניחשף לעקרונות שונים בהכנה ובויזואליזציה של נתונים ולחבילות נוספות כגון **numpy**, **matplotlib** ו-**seaborn**. לאחר מכן נתוודע לפורמטים נפוצים (כגון JSON ו-HTML) ולמקורות נפוצים של נתונים, כגון בסיסי נתונים ורשת האינטרנט. נפנים עקרונות סטטיסטיים בסיסיים. בסופו של הפרק כל סטודנט יעשה פרויקט ובו אנליזה מלאה ומעמיקה לדאטה על פי בחירתו.

שלב EDA יכלול גם לימוד כלים טכנולוגיות הרלוונטיות לעבודה עם מקורות דאטה שונים, וביניהם ניחשף לטכנולוגיות Big Data המאפשרות לעבוד עם נתונים לא מובנים, נכיר טכנולוגיית Devops חשובה – Docker – שתסייע לסטודנטים להבין את מנגנוני העבודה בסביבות "אמיתיות" שכוללות Deployment.

- Tabular data
    - Mathematical packages (*scipy, numpy*)
    - Pre-processing with *pandas*
        - Basic concepts
        - Indexation and filtering
        - Aggregations and advanced manipulations
    - Visualization (matplotlib, seaborn)
    - Working with databases (SQLalchemy)
- Statistics
    - Combinatorics & Probability
    - Random variables and distributions
- Non-tabular data
    - Documents and JSON
    - Web scraping and HTML

- Big Data
    - Spark, Spark.sql (pyspark)

Section 3 – Machine Learning

בחלקו השלישי של המסלול נצלול לליבה של עבודת ה-Data Scientist, ובאמצעות use-case-ים שונים, המייצגים בעיות עסקיות מגוונות, ניחשף באופן שיטתי והדרגתי לעולם אינסופי של כלים, שיטות, אתגרים, עקרונות, וכמובן – מודלים סטטיסטיים. כל use-case יציב בפנינו אתגרים חדשים, שההתמודדות עימם תחשוף בפנינו עוד ועוד כלים ורעיונות.

נתוודע ל-**CRISP-DM**, המתודולוגיה המקובלת לפיתוח בעולם ה-Data Science, נבין את השלבים השונים שלה, וניישם אותם בפועל על אוסף רחב של בעיות עסקיות מעולמות תוכן שונים. נביר לעומק את החבילה הנפוצה ביותר בעולם ה-**Machine Learning**, הלא היא **Scikit-Learn**.

בסופו של דבר בסיום המודול סטודנטים יכירו לעומק את ההיבטים השונים של יצירת מודלים לחיזוי, כיצד הם באים לידי ביטוי ב-Scikit-learn וב-PySpark. במהלך הפרק הזה כל סטודנט יוצר פרויקט חיזוי ראשוני מקצה לקצה (עם בעיית רגרסיה).

פירוט הנושאים בחלק זה של הסילבוס אינו מייצג תהליך כרונולוגי, אלא מתמצת את הנושאים המרכזיים בהם נעסוק.

Concepts and models

- Supervised & unsupervised learning
- Variance-Bias trade-off
- Pipelines – Transformers & Estimators
- Feature engineering
- Dimensionality reduction
- Model selection – Cross-validation & grid search
- Overfitting & regularization
- Ensemble methods – Voting, bagging & boosting
- Imbalanced data
- Anomaly detection
- Clustering
- Metrics and similarities
- Scoring

Models

- Linear regression
- Logistic regression
- Decision trees (inc. random forest, XGBoost and CatBoost)
- K-nearest neighbors (k-NN)
- K-means
- Agglomerative clustering
- DBSCAN
- Naïve Bayes

Section 4 – Data Science in production

בחלק זה נעסוק בשיטות לפיתוח והנגשת מודלים בסביבות מציאותיות, ולשם כך נלמד קצת Linux וניצור אפליקציה ב-Flask. בסופו של הפרק הזה, ובהתבסס על הכלים שגובשו בפרק הקודם, כל סטודנט ייצור פרויקט חיזוי מקצה לקצה (עם בעיית קלסיפיקציה).

- Deployment
    - pickle
    - REST API
    - Flask
    - Argparse
- MLops
- Data Science in the cloud

Section 5 – Deep Learning

בפרק זה נעסוק ביתר פירוט בלמידה עמוקה – משפחת מודלים שעומדת בליבה של המהפכה הטכנולוגית של השנים האחרונות – ונבין כיצד לרתום את כוחם. נכיר את מושגי היסוד של רשתות נוירונים, נבין כיצד לתכנן אותם ב-keras, וניחשף לארכיטקטורות שונות וליישומים מגוונים שבהם הם התגלו כשימושיות מאוד.

- Neural networks & MLP
- Implementation with keras
- Important concepts (CNN, RNN, autoencoders)
- Advanced architectures and applications

Section 6 – Final Project

את חלקו אחרון של המסלול נקדיש לעבודה בפועל על פרויקט אישי מסכם, בו יוכל כל סטודנט להתנסות בטכניקות וברעיונות שנלמדו בקורס מול בעיה עסקית אמיתית. במהלך מפגשי הפרויקט הסטודנטים יסבירו את הבעיה העסקית שנבחרה ואת הנתונים המלווים אותה, ידגימו את תהליכי ה-pre-processing וה-feature engineering שלהם, ויציגו את המודלים שבהם בחרו להשתמש בסופו של דבר.