Practical Data Science

335 שעות לימוד אקדמיות

תיאור התפקיד:

הצורך להתמודד עם כמויות גדולות של מידע הוליד בשנים האחרונות תפקידים רבים והתמחויות שונות כגון ה-**Data Analyst**, ה-**Business Intelligence** וה-**Big Data**. עם זאת, היכולת לשלב בין כל אלו ולהוסיף עליהם נדבך ייחודי של חיזוי, נותרה נחלתם של מעטים, ובשנים האחרונות ביסס עצמו ה-**Data Scientist**/מדען הנתונים כ"מקצוע הנחשק ביותר של המאה ה-21".

תפקידו של **מדען הנתונים** הינו לבצע מחקרי מידע מעמיקים בכדי להפיק תובנות עסקיות לארגון, לנקות, לטייב ולסדר את המידע המשמש למחקרים השונים, להפעיל אלגוריתמים שונים של מידול, כריית מידע ו- **Machine Learning** על המידע, ולסייע בבניית תהליכי הכנת המידע ואופטימיזציה של האלגוריתמים השונים.

הכישורים הנדרשים מ-Data Scientist רבים ומגוונים ומתמקדים ב-4 שלבים עיקריים של עבודה עם המידע:

השגת המידע | חקירת המידע |ניתוח אנליטי של המידע | הצגת המידע

מסלול זה מקנה את הכלים הנדרשים לכל שלב ושלב בעבודתו של ה-Data Scientist עם דגש על פרקטיקה ויכולות תכנות מתקדמות.

וזה עוד לא הכל - בתום הלימודים מקבלים הבוגרים סדנת פיתוח קריירה, הכוללת כתיבת קורות חיים, סימולציית ראיון, וקישור למשרות ולארגונים מעסיקים!

לקראת ראיונות העבודה הם מקבלים ליווי והכנה לראיונות מקצועיים על ידי מומחי הדאטה הבכירים שלנו.

תיאור ההכשרה:

בחלקו הראשון של המסלול נלמד לתכנת ב-**Python** (מועבר בגרסת Python3), שהיא השפה המובילה כיום לתחקור הנתונים, ונרכוש כלים לעבודה עם נתונים ממקורות שונים ולהצגתם. נפתח מאפס קוד בסביבה מונחית-עצמים (Object-Oriented), שהיא המתודה הסטנדרטית כיום בפיתוח תוכנה ונבין לעומק את היתרונות הגלומים במתודה זו. בנוסף, נכיר את ספריית המודולים העשירה של השפה ונדע כיצד להיעזר בה.

בחלקו השני של המסלול נסקור את משפחת החבילות מ- **PyData**, המהוות את סט הכלים המושלם לעבודה עם נתונים בכלל ונתונים טבלאיים בפרט. ראשית 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. בעזרת הכלים הללו נתנסה במגוון שיטות של Exploratory Data Analysis (**EDA**).

בחלקו השלישי של המסלול נצלול לפרקטיקה היומיומית של ה-Data Scientist, ובאמצעות use-case-ים שונים ניחשף באופן שיטתי והדרגתי לעולם אינסופי של כלים, שיטות, אתגרים, עקרונות, וכמובן – מודלים סטטיסטיים. 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. נכיר לעומק את החבילה הנפוצה ביותר בעולם ה-**Machine Learning**, הלא היא **Scikit-Learn**.

את חלקו הרביעי 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.

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 (4 פרויקטי ביניים ופרויקט מסכם אחד), יועלו לחשבון GitHub של כל סטודנט ובכך ייצרו תיק עבודות עשיר ומקצועי להצגה בפני המעסיקים בהמשך.

קהל יעד ודרישות קדם:

בעלי רקע ב-Data Analysis, BI, פיתוח, מסדי נתונים ומערכות מידע, המעוניינים להעשיר את יכולותיהם בתחום תחקור הנתונים.

- בעלי נסיון מאחד או יותר תחומים המתוארים להלן:
  - רקע בתכנות בשפה כלשהי (שפות OOP, שפות סקריפטיות – SQL, POWERSHELL, BASH, R ועוד)
  - נסיון בניתוח נתונים (אקסל מתקדם, 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 מבדקת התאמה הבוחן יכולות אנליטיות של המועמד/ת
- שליטה טובה בשפה האנגלית

תנאים לקבלת תעודת סיום קורס:

- 80% נוכחות מינימום
- הגשת כל הפרויקטים במהלך הקורס

תוכנית הלימודים:

Section 1 – Python

בפרק זה נלמד לתכנת בשפת Python ונתוודע אל סביבת העבודה של הקורס – Google Colab (עבודה עם Jupyter Notebooks).

- 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

בפרק זה נסקור מושגים וכלים שימושיים בעבודתו היומיומית של ה-data scientist.

- Pre-processing with *pandas*
  - Basic concepts
  - Indexation and filtering
  - Aggregations and advanced manipulations
- Mathematical packages (*scipy, numpy*)
- Visualization packages (matplotlib, seaborn)
- Working with data resources (JSON files, databases & web)

Section 3 – Machine Learning

בחלק זה נראה use-case-ים, המייצגים בעיות עסקיות שונות ומגוונות. כל use-case יציב בפנינו אתגרים חדשים, שההתמודדות עימם תחשוף בפנינו עוד ועוד כלים ורעיונות. פירוט הנושאים בחלק זה של הסילבוס אינו מייצג תהליך כרונולוגי, אלא מתמצת את הנושאים המרכזיים בהם נעסוק. בפרק זה נבין לעומק את ההיבטים השונים של יצירת מודלים לחיזוי, ונראה כיצד הם באים לידי ביטוי ב-Scikit-learn

Concepts

- Supervised & unsupervised learning
- 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
- Deep Learning
    - Neural networks & MLP
    - Implementation with keras
    - Important layers (CNN, RNN, autoencoders)
    - Advanced architectures

Models

- Linear regression
- Logistic regression
- Decision trees (inc. random forest)
- K-nearest neighbors (k-NN)
- Neural networks
- K-means
- Agglomerative clustering

<u>Section 4 – Project</u>

בפרק זה נעבוד על פרויקט אישי מסכם (כל אחד על פרויקט משלו), כאשר עיקר ההתקדמות תתבצע בבית, ובכיתה ניפגש לקבל תמיכה וליווי, להתייעץ ולהחליף רעיונות. במהלך המפגשים נקיים שיעורי מבוא לנושאים NLP, Big Data (Spark) ועבודה בסביבת פיתוח (PyCharm).