פרויקט סיכום - תכנות מתקדם

מרצים: פרץ אור, גוטמן דוד

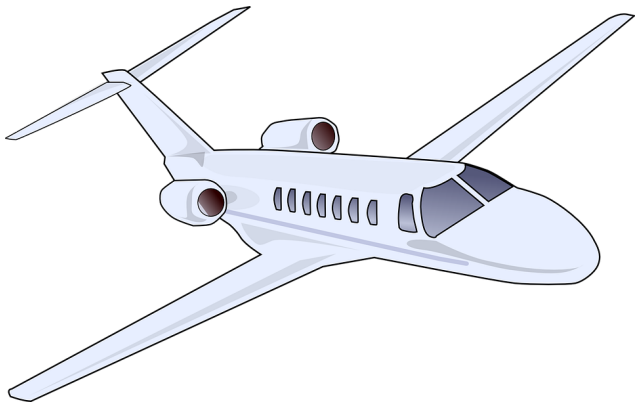
עוזרת הוראה: נטע רחמילביץ'

אדמיניסטרציה

1. פרויקט זה הינו פרויקט סיכום בקורס תכנות מתקדם. ציון הפרויקט יחולק עפ״י:
   a. 40% הפרויקט + דו״ח
   b. 30% הגנה בע״פ
2. בנוסף, בשקלול הסופי יתווספו 30% מטלות בית.
3. ניתן להגיש את הפרויקט בשפה העברית או בשפה האנגלית.
4. יש להגיש את הפרויקט (מחברת jupyter \ קובץ py) עם מסמך PDF אשר מסכם את עבודתכם (עפ״י הנדרש). תאריך ההגשה הינו **8.7.2021 בשעה 09:00** , וההגשה תתבצע דרך תיבת הגשה במודל. יש להוסיף ת״ז של כל אחד מחברי הקבוצה.
5. יש להשתמש בכללי תכנות נכונים אשר נלמדו לאורך הסמסטר: הימנעות מקוד כפול, שמות משתנים רלוונטים, ויתור על לולאות לא נחוצות ושימוש בפעולות וקטוריות ככל שניתן.

הקדמה

מחירי כרטיסי טיסה יכולים להשתנות בהתאם לביקוש, מסלול, שעות המראה \ נחיתה ואפילו תאריך. כלומר, מחירם הוא לא צפוי ויכול להשתנות בכל רגע נתון. בפרויקט זה, אוסף הנתונים המצורף מכיל מידע טיסות פנים בהודו בחודשים מרץ עד יוני 2019. המטרה היא לנתח את אוסף הנתונים המצורף כדי להבין מהם הגורמים המשפיעים ביותר על מחירי הטיסות בתוך הודו ולמצוא תכונות דומות עבור טיסות שונות



תכונות (Features)

Airline	שם חברת התעופה
Date	תאריך הטיסה
Source	מהיכן הטיסה המריאה
Destination	יעד הטיסה
Route	מסלול הטיסה, מופרד בנתונים ע"י פסיק
Dep_Time	שעת המראה
Arrival_Time	שעת נחיתה - אם הטיסה נחתה באותו יום שבו היא המריאה, מופיעה שעת נחיתה. אחרת, מופיעה גם תאריך הנחיתה (בנוסף לשעה)
Duration	משך זמן הטיסה
Total_Stops	עצירות
Additional_Info	מידע נוסף
Price	מחיר (הערך ב- Rupee, המטבע ההודי)

במידה ותרצו לנתח את הנתונים במטבע גלובלי, ערכו של 1 דולר אמריקאי הוא 72.5 רופי הודי.

דרישות

עליכם לבנות פרויקט מונחה נתונים עפ"י מודל CRISP-DM:

I. Business Understanding

קיימות שתי שאלות עסקיות עבור אוסף הנתונים הנ"ל:

שאלה 1: האם ניתן לחזות או להעריך את מחיר הטיסה של טיסת פנים בהודו על סמך נתיב, חברת תעופה או תאריך?

שאלה 2: לכמה קבוצות ניתן לחלק את הטיסות כך שבכל קבוצה תהיינה טיסות אשר דומות אחת לשניה (מבחינת חברה תעופה, נתיב וכו׳)?

עבור כל אחת מהשאלות, הסבירו מדוע היא SMART:

Specific ,**M**easurable, **A**ssignable, **R**ealistic, **T**ime-Related

II. Data Understanding

הסבירו את גישת ה"data-driven" עבור כל אחת מהשאלות. כלומר, הסבירו מדוע ניתן לבנות פרויקט זה ולנסות לענות על שאלות 1,2 בהתבסס על אוסף הנתונים המצורף. בתשובתכם, ניתן להוסיף מדדים פשוטים מתוך אוסף הנתונים (ממוצעים, דיאגרמות פשוטות, ועוד).

<u>טיפ:</u> כפי שלמדנו לאורך הסמסטר, בתור התחלה - להדפיס את ה- describe, info נותן אינטואיציה כללית ראשונית על האופן בו הנתונים בנויים.

III. EDA + Data Preparation

בחלק זה, יש לבצע תחקור נתונים (data analysis) עפ"י הכללים שלמדנו לאורך הסמסטר.

להלן רשימת רעיונות כיצד ניתן לחקור את הנתונים:

1. **תחקור סטטיסטי בסיסי** - השוואת ממוצעים של קבוצות שונות, ערך מקסימלי, ערך מינימלי וכו׳.
2. **התפלגויות** - כפי שראינו לאורך הסמסטר, גרפים ממחישים בצורה טובה כיצד ״נראית״ כל עמודה. השתמשו בפקודות הקיימות על מנת לחקור את העמודות השונות.

   חשוב: ניתן להציג גם התפלגויות של עמודות קטגוריאליות, כפי שראינו ע״י value_counts.
3. **הפרדת ערכים** - ניתן ליצור עמודות חדשות, המתבססות על העמודות הקיימות, בכדי ליצור נתונים שניתנים להמחשה. למשל, עמודת duration, ניתן להמיר את ערכה לערך נומרי ובכך להציג זמן ממוצע של משך טיסה עפ״י חברת תעופה, נתיב, וכו׳.

לאחר מכן, יש לבצע תהליך הכנת הנתונים לקראת מודל חישובי:

1. במידה וישנם ערכים חסרים, מחק אותם.
2. העתק את אוסף הנתונים הקיים ל-DataFrame נוסף. כעת, הראשון ישמש אותנו לclustering, והשני ישמש אותנו ל-classification.
3. עבור אוסף הנתונים ל-clustering:
   a. השאירו את עמודות Airline, Source, Destination, Price בלבד.
   b. המירו את הערכים הרלוונטים לקידוד one-hot-vector.
   c. המירו את הערכים הרלוונטים לנרמול MinMax.
4. עבור אוסף הנתונים ל-classification:
   a. המר את עמודת Airline באמצעות קידוד ״רגיל״, כלומר כל חברת תעופה תקבל מספר ייחודי: הראשונה תקבל 1, השניה 2, וכו׳.
   b. המירו את הערכים הרלוונטים לקידוד one-hot-vector. (חוץ מעמודת Airline).
   c. המירו את עמודת Price עפ״י הכללים הבאים:
      i. מחיר אשר נמוך מ7,000 רופי - המר ל״1״ (זול).
      ii. מחיר בין 7,000 ל-14,000 רופי - המר ל״2״ (ממוצע).
      iii. מחיר גבוה מ14,000 רופי - המר ל״3״ (יקר).

d. המר את עמודת התאריך כך שתכיל את החודש בלבד. כלומר, טיסות אשר נערכו במרץ יופיעו כ"3", אפריל כ"4" וכו'.
e. מחק את כל העמודות אשר מציינות זמנים (שעת המראה, שעת נחיתה וכו').
f. חלקו את הנתונים לאימון ומבחן, כאשר משקל קבוצת המבחן הוא 25%.

IV. Modeling

בחלק זה, עליכם להריץ שני מודלים:

1. מודל **Clustering** בעזרת אלגוריתם KMeans - הריצו ובדקו $K = 2, 3, \ldots, 10$. עבור כל ריצה, שמרו את סכום הטעויות ואת מדד הסילואט. השתמשו ב"שיטת המרפק", בחרו K אשר מתאים לפתרון הבעיה.
2. מודל **Classification** בעזרת עצי החלטה - הריצו מופע של Random Forest המשתמש ב25 עצים שונים.

V. Evaluation

יש להציג בצורה מפורטת את תוצאות האלגוריתמים מחלק 4. כלומר, יש להציג מדדי דיוק, מטריצת בלבול, ניתוח פשוט לאשכולות שיצאו בclustering וכו'.

VI. Final Report

יש להגיש את הפרויקט בליווי דו"ח מסכם. ניתן לעשות זאת במקביל בjupyter notebook, ולהשתמש בתאי markdown כדי לכתוב טקסט פשוט.

הדו"ח צריך להכיל 5 חלקים, עפ"י שלבי מודל CRISP-DM. בנוסף, יש להוסיף פרק "מסקנות" אשר מסכם את עבודתכם בפרויקט. יש לציין מה הן המסקנות הבולטות אשר נאספו לאורך הדרך.

בהצלחה !