


פרוייקט גמר – ניהול התפעול והייצור

קבוצה 7

עומר צינמן omerzin19@gmail.com

ערן שלזינגר eran4300@gmail.com

שימו  - לפרוייקט הזה קיימת repository ב git :

[omer7642/ManagementOperationsFinalProject \(github.com\)](https://github.com/omer7642/ManagementOperationsFinalProject)

לפנינו מידע על פציינטים, הלקוח מתוך מאגר המידע Kaggle. מאגר המידע מכיל 4,981 רשומות של פציינטים שונים כאשר כל רשומה מכילה את המידע הבא :

1. gender – מגדר הפציינט
2. age – גיל הפציינט (בשנים)
3. hypertension – האם הפציינט סובל מיתר לחץ דם
4. heart_disease – האם הפציינט קיימת מחלת לב
5. ever_married – האם הפציינט התחתן או לא
6. work_type – סוג העבודה של הפציינט – ילדים, ממשלתי, מובטל, מגזר פרטי או עצמאי.
7. residence_type – האם גר באזור כפרי או אורבני
8. avg_glucose_level – רמת גלוקוז ממוצעת בדם
9. BMI – מדד מסת הגוף של הפציינט
10. smoking_status – האם הפציינט מעשן, עישן בעבר, מעולם לא עישן או לא יודע.
11. stoke – האם הפציינט סבל משבץ או לא.

המאגר הורד למחשב כקובץ csv שמשקלו 279KB ואנו מבצעים עליו ניתוחים בעזרת ספריות שונות בפיתוח, על מנת לחזות מצבים שונים של יתר לחץ דם, שבץ וכו'.

כעת, נענה על התשובות הניתנות לנו בשאלון לפי כל מודל :

1. תיאור הבעיה :

א. מודל 1 :

במודל זה עלינו לחזות מגדר הפציינט בעזרת השדות גיל, מדד מסת גוף ורמת גלוקוז ממוצעת בדם. אני מצפה שנשים ייסבלו יותר מיתר לחץ דם, ושיהיה להן BMI נמוך יותר (מכיוון שנשים נמוכות יותר מגברים באופן ממוצע, מה שמשפיע על BMI). אני לא צופה שלגיל יהיה השפעה על מגדר. במודל זה עלינו להשתמש בעץ החלטה מסוג CART (עץ ההחלטה שלמדנו בכיתה).

ב. מודל 3 :

במודל זה עלינו לחזות האם הפציינט התחתן, וזאת בהתבסס על סטטוס העישון ורמת הגלוקוז הממוצעת בדם. אני איני רואה קשר בין סטטוס עישון לבין מצב נישואים ואין ספק שזה נבחר בצורה שרירותית, ולכן לא יכול להעריך שום דבר בצורה מבוססת. במודל זה עלינו להשתמש ברגרסיה לינארית על מנת לקבל תוצאה. איני חושב שרגרסיה לינארית מתאימה פה כי בסוף אני אמור לקבל תשובה של 0 או 1, ורגרסיה לינארית לא תתן תשובה כזו, אך לצורך זה ביצענו עיגול למספר הקרוב ביותר.

ג. מודל 4 :

במודל זה עלינו להשתמש ברגרסיה לוגיסטית על מנת לחזות הימצאות מחלת לב. עלינו לחזות על פי סטטוס עישון, מדד מסת גוף, רמת גלוקוז ממוצעת בדם וגיל. אני אשער שככל שהמדדים גבוהים יותר (מעשן, BMI גבוה יותר, גלוקוז גבוה יותר וגיל מבוגר יותר), כך הסיכוי להמצאות מחלת לב גדולה יותר.

ד. מודל 5 :

במודל זה עלינו להשתמש בשדות סוג העבודה, מדד מסת גוף ורמת גלוקוז ממוצעת בדם על מנת לחזות יתר לחץ דם. ההשערה שלי היא שככל שיש יותר סוכרים בדם בצורה ממוצעת, ורמת מסת הגוף גבוהה יותר, ככה הסיכוי גבוה יותר ליתר לחץ דם. עלינו להשתמש ב KNN כמודל.

2. **תיאור הדאטה :** כמו שציינתי בהתחלה, קיימים 4981 רשומות של מידע שנלקח מ Kaggle. בכל מודל קיימים שדות שונים מתוך קובץ ה CSV הכללי, וכל מודל בוחן אספקט שונה בהתבסס על הדאטה. לפירוט השדות וההשערות ראה סעיף 1.

סעיפים 3-7 נמצאים בקוד, כך ש :

- אנליזה תיאורית ואנליזה ראשית (סעיפים 3-4) ייכנסו תחת Variable Manipulation בכל מודל (אין הפרדה כי הן במילא קשורים קשר הדוק).
- הערכת מודל ומכשולים ייכנסו תחת Conclusion (מכשולים הם חלק מהערכת המודל – לא רואה טעם לפצל).
- עבודה עתידית תכנס תחת recommendations בכל מודל. (למעשה גם סעיף 1 נמצא בקוד בתחילת כל מודל).

8. סיכום קצר :

העבודה על הפרויקט הייתה מאוד מאתגרת. ראשית כל, עצם העובדה שאין שאלות אחידות ושהיינו צריכים לבצע הצלבות בשביל להבין מה אנחנו נשאלים. בהרבה מהמקרים השדות (features) היו רנדומליים לחלוטין ולא ראיתי בהם שום קשר עם השדה הנחקר (נניח, לחזות מגדר על בסיס גיל – לא נראה לי כל כך הגיוני אע"פ שעץ ההחלטה מצא קשר – על זה פירטתי במודל 1). צורת ההגשה הייתה מאוד בעייתי גם כן. מכיוון שיש לי 16 גרפים שונים וזה יצא לי כמעט 10 דפים (רק בשביל להציג את הגרפים). אני לא רואה דרך לבצע אנליזה כאשר הגרף עצמו לא נמצא מולי, ולכן היה לי מאוד טבעי לבצע את האנליזה כבר בתוך הקוד בעזרת markdown. אני מקווה שזה בסדר (השלבים בוצעו כמפורט למעלה). מכתיבת הקוד מאוד נהנינו. חשוב לי לתת קרדיטציה ל ChatGPT שהותר לשימוש בנוהל האוניברסיטה שעזר לשפוך insights לגבי המודלים. השתמשנו בו אבל כל מה שכתבנו הוא דברים שאנחנו רואים בעצמנו והוא רק עזר לשים את האצבע על אותן נקודות חלשות.

אנו מודים לצוות הקורס על קורס מאתגר,

עומר וערן.