בס"ד

**שאלות על מבנה הדאטה, תשובות ומסקנות**

1. לגבי העמודות employee\_department, employee\_job ושאר הנתונים הקבועים:

א. האם התוכן בעמודות אלו יהיה בעברית, או שכל הנתונים מלכתחילה ימולאו באנגלית? (2 נימוקים מדוע כדאי שזה יהיה באנגלית: א. כך נוכל לפרסם ולתרום את הדאטה לעולם. ב. נחסוך את ההמרה לאנגלית בשביל המודל ). אין בעיה שיהיה באנגלית

ב. האם משמעותיים שמות המחלקות/התפקידים? כמותם? אם כן, צריך לשנות בדאטהסט.

**נכון לעכשיו, זה מה שיש בקוד. אנחנו צריכות לקבל ממך את הערכים האמיתיים, עדיף באנגלית.**

        self.departments = ['טכנולוגיה', 'כספים', 'משאבי אנוש', 'מכירות', 'מחקר ופיתוח', 'ביטחון', 'משפטי', 'שיווק']

        self.jobs = ['מנהל', 'אנליסט', 'מפתח', 'רואה חשבון', 'יועץ', 'חוקר', 'מתכנת', 'מנהל פרויקט']

        self.countries = ['ארה"ב', 'גרמניה', 'צרפת', 'בריטניה', 'איטליה', 'יפן', 'איראן', 'סוריה', 'לבנון', 'קוריאה הצפונית']

        self.hostile\_countries = ['איראן', 'סוריה', 'לבנון', 'קוריאה הצפונית']

2. לגבי ה-employee\_id: נכון לעכשיו לכל עובד id 1, 2, 3 וכו' לפי הסדר - זה נקרא מספר עובד רץ? אם כן, זה תקין. תקין אבל מקובל לתת מספר אישי בגודל אחיד למשל כולם עם 3 ספרות: 001,002...350 (בכל אופן, לא כזה עקרוני).

3. לגבי ה-date: אני מבינה שאנחנו מגישות למודל דאטהסט שבו לכל עובד ולכל יום – יש שורה אחת בלבד, שבה מרוכזת פעולתו באותו יום. האם הבנתי נכון? כן בדיוק, חוץ מהמאפיינים האישיים שלא כ"כ משתנים לאורך הזמן אבל מתחקור שביצעתי אין בעיה להציג אותם באותה טבלה, המודל מבין שהנתונים הללולא תלויים בתאריך

4. האם יש בעיה בהדפסות/צריבות מקמפוס אחר? האם תמיד זה מחשיד? בהדפסות וצריבות זה די מחשיד כי הקמפוסים מפוזרים במקומות שונים בארץ ועל פניו אין סיבה שעובד מקמפוס A ידפיס/יצרוב נתונים מקמפוס B (בבקרת כניסה לעומת זאת, זה בסדר שעובד יבקר בקמפוס אחר לצורך פגישות, אבל אם יש שינוי פתאומי בכמות הביקורים כאשר התפקיד לא השתנה, זה גם מחשיד).

5. employee\_job – מה זאת אומרת "תפקידים עם ציפיות שונות מההתנהגות"? אשמח לדוגמה. יש תפקידים שבהם התנהגות מסוימת היא נורמלית ויש תפקידים שלא. למשל עובד ממחלקת רכש/שיווק רגילים לעבוד עם ספקים/לקוחות ולהדפיס/לצרוב כמויות גדולות של חומרים ועבורם זו לא התנהגות חריגה, לעומת מתכנת למשל שלא מתעסק עם ניירת ביום יום. לכן, אם יש אנומליה, צריך להשוות אותה לנורמה בקבוצת העובד ולא ביחס לכלל העובדים. צריך לשאול את אלה איך עושים זאת.

אפרת עושה את השאלה כללית יותר: יש התנהגות של אדם שגם בהשוואה לקבוצתו אינה נחשבת נורמלית, אבל בהשוואה לו – כן, היא נחשבת התנהגות אופיינית לו. דוגמה נוספת, מלבד הדוגמה של טיסה למדינה עוינת: עובד שקבוע מתחיל את יום העבודה שלו מוקדם, או קבוע משלים שעות במוצ"ש. אמנם יש לו המון שעות חריגות, אבל זה לא נכון להסיק מכך שהוא חשוד, היות שצריך לבדוק את החריגות גם ביחס לעובד עצמו.

תשובת אלה: מנרמלים את הנתונים ביחס לכל קבוצה, ואז מה נשאר בתור יוצא דופן? רק ההתנהגות שיש בה אנומליה. את האנומליה של האדם ביחס לעצמו ניתן למצוא באמצעות time-series.

6. employee\_seniority\_years - מהו טווח השנים? מ-0 ועד? בהנחה שניתן לעבוד מגיל 16 עד מקסימום 67 אז נגדיר שהטווח האפשרי הוא 0-50

7. is\_contractor עובד קבלן: האם אמור להיות יחס מסוים בין עובדי קבלן לעובדי חברה? כמובן הם נמצאים בקשר שוטף עם עובדי החברה. יש סוגים שונים של עובדי קבלן- חלקם יועצים ומגיעים לסייע לעובדים בנושאים מורכבים, חלקם עובדים בחברה במשרה חלקית (וביתר הימים עובדים בחברה אחרת), וחלקם עובדים באופן יומיומי בחברה.

8. employee\_classification רמת סיווג עובד: האם אמור להיות יחס מסוים בין הרמות השונות? מדובר ברמות הסיווג הבטחוני בישראל , סיווג שנקבע לפי השב"כ – מקובל לסמן במספרים:  
 1- בל"מס (בלתי מסווג), 2- שמור, 3- סודי, 4-סודי ביותר. אצלנו אין שום עובד שנכנס עם בלי סיווג, לכן 1 אצלנו זה שמור, אבל בהקשר של סיווג מסמכים 1 זה בלמ"ס (לא בטוחה שצריך להיכנס לרזולוציות האלו).

9. has\_foreign\_citizenship – א. האם בדאטהסט לא אמור להיות מידע איזו אזרחות זרה יש לעובד? זו שאלה טובה, האמת שאצלנו יש פירוט של מדינות והתלבטתי אם להכניס את זה למודל כי יש כבר המון שדות.

ב. אם יש אזרחות ממדינה עוינת זה יותר מחשיד. האם אתם מניחים שלא התקבלו לעבודה אנשים בעלי אזרחות ממדינה עוינת? לא תמיד אזרחות ממדינה עוינת זה יותר מחשיד, למשל רוסיה מוגדרת מדינה עוינת אבל עובדים רוסים לא בהכרח חשודים. לא מקבלים אצלנו ערבים באופן גורף. דרוזים יש ממש מעט. תלוי בנו לאיזו רזולוציה נרצה להיכנס, כרגע נראה לי שיש מספיק פיצ'רים.

ג. אם יש לאדם אזרחות ממדינה עוינת והוא טס לאותה מדינה, זה אולי פחות מחשיד מאשר אדם שאין לו אזרחות של אותה מדינה ובכל זאת הוא טס לשם. כן, זה הרבה פחות מחשיד כשהעובד הוא אזרח של אותה מדינה כי הוא יכול לטוס לבקר משפחה/ טיול שורשים (במקרה הזה צריך לבדוק שתדירות הביקורים שלו יחסית קבועה ולא עולה באופן פתאומי). לעומת זאת, עובד שטס למדינה עוינת בלי קשר למוצא שלו, זה מדליק נורה אדומה מיידית (גם אם טס אליה פעם אחת). צריך לשאול את אלה איך בודקים זאת.

תשובה: החלטנו להוסיף עמודה בדאטה - האם האדם טס לארץ מוצאו. בכך נחסוך את הצורך לשמור לכל אדם את כל האזרחויות השונות שלו.

10. לגבי הדפסות/הצריבות: כתבת: "במידה ועובד ביצע באותו יום הדפסות ממספר קמפוסים שונים, נבצע ONE HOT  או שנאלץ את הסימולטור שלא יהיה יותר מקמפוס אחד לעובד באותו תאריך".

א. אני חושבת שעדיף להיצמד למציאות שבה עובד יכול לעשות הדפסות וצריבות מכמה קמפוסים שונים באותו יום. מה החלטנו? (אני לא חושבת שיש בעיה לעשות ONE HOT או משהו אפילו פשוט יותר. אפשר לייצג את זה גם באמצעות ערך int). אפשרי

תשובה: בסוף החלטנו לא לעשות את זה. ראו להלן ב'.

ב. נניח שנצמדים למציאות שיכולים להיות כמה קמפוסי הדפסה/צריבה ליום - האם כמות פקודות ההדפסה/הצריבה תהיה רק סכימה של כולם או שנעשה גם כמות פר קמפוס? (זה מתבקש שיהיה פר קמפוס, כי אם היו הרבה פקודות הדפסה מהקמפוס של העובד ורק מיעוט מהקמפוס השני זה אמור להיות פחות מחשיד מאשר המציאות ההפוכה). תיאורטית את צודקת, פרקטית אני חוששת שנלך לאיבוד עם כ"כ הרבה שדות. מציעה שנציג לאלה את ה-DS ונבדוק מה היא חושבת על הסדר גודל

תשובה: החלטנו להוסיף עמודה בדאטה – האם הדפיס באותו יום גם מקמפוס שאינו הקמפוס שלו וכן האם צרב באותו יום גם מקמפוס שאינו שלו.

אם כך, אני חושבת שאין ערך לעמודה "קמפוס ביצוע ההדפסה" או "קמפוס ביצוע הצריבה". אלא אמור להיות במקומה "מספר קמפוסים שבהם הדפיס", "מספר קמפוסים שבהם צרב", בנוסף לעמודות "האם הדפיס מקמפוס שאינו שלו", "האם צרב מקמפוס שאינו שלו". האם אפשרי?

ג. האם הדפסות/צריבות מקמפוס אחר זה מחשיד? כן, עניתי על זה בסעיף 4

ד. מהו טווח הנפחים של הצריבה/כמות דפי ההדפסה?

האם חשוב לציין טווח? אם כן, צריך למקד את קלוד**. חושבת שחשוב לציין טווח, אבדוק אצלנו איך נראית ההתפלגות**

11. לגבי ההדפסות: א. האם יתכנו פקודות הדפסה יותר מכמות הדפים שהודפסו? לא, פקודת הדפסה הכוונה שאת לוחצת PRINT במחשב/ מעבירה שכפול במכונת צילום אז בכל פקודה חייב להיות לפחות עמוד אחד.

ב. האם המשוואה הבאה נכונה: כמות הדפסות צבע + כמות הדפסות שחור לבן = כמות עמודים שהודפסה? אם כן, האם אין פה איזו עמודה מיותרת? (לדוגמה סך העמודים שהודפסו - יכול להיות מחושב לפי הסכום, כעמודה מחושבת). זו אכן עמודה מחושבת, אני לא יודעת אם המודל "מבין" שהמדדים האלו יכולים להסתכם או שמסתכל עליהם כפיצרים שאין ביניהם קשר? צריך לשאול את אלה מה לעשות במקרה כזה והאם זה בסדר שיש מדדים שקשורים (למשל במודל ליניארי, הקשר ביו המשתנים צריך להיות בלתי תלוי).

ג. כנ"ל לגבי אחוז הדפסות בצבע מתוך כלל ההדפסות -האם לשמור את העמודה ממש כחלק מה-CSV או להוסיף אותה בשלב עיבוד הנתונים? נשאל את אלה

תשובת אלה: אין בעיה להוסיף את העמודות הנ"ל למודל ובכך "ללעוס" למודל את הנתונים עוד קצת. אלה לא זוכרת מודל מסוים שנתונים בלתי תלויים מפריעים לו.

צריך לברר: א. מהם המודלים שמסתמכים על כך שהפיצ'רים הם בלתי תלויים. ב. האם יש בכך בעיה מהותית? ג. האם המודלים האלו רלוונטיים לנו?

בנוסף, האם הנתונים הם מסוג float - שבר או int - כמספר האחוזים מתוך 100? לא קריטי, לצורך הקריאות עדיף להשאיר int ואם מתקבלת חלוקה עשרונית זה יתעגל למספר הקרוב ביותר

12. מה זאת אומרת שכתבת על רוב העמודותchanged over time, per employee+date ? שגם ה-employee וגם ה-date משתנים? הכוונה שהנתונים בין השורות משתנים כתלות בעובד+תאריך, למשל כמות הדפסות בשורה מסוימת היא כמות ההדפסות שעובד X ביצע בתאריך Y , לעומת זאת המאפיינים האישיים כמעט ולא משתנים (למשל ארץ מוצא לעולם לא תשתנה, תפקיד יכול להשתנות לעיתים רחוקות). זה בעיקר סימון עבורנו.

13. לגבי הצריבות:

בהדפסות חילקנו לכמות דפים במועד חריג וכמות פקודות במועד חריג.

מדוע בצריבות יש רק כמות פקודות במועד חריג, ואין נפח צריבה במועד חריג? (אם מחליטים להוסיף עמודה צריך להוסיף גם בדאטה). זה הפוך – בצריבות יש נתון של נפח ובהדפסות אין. בהדפסות כמות הדפים נותנת אינדיקציה מספקת על הנפח שהעובד הדפיס, לעומת צריבות שנצרבות על דיסק ואם לא נתעד את הנפח של הקובץ, לא תהיה דרך למדוד את גודל הצריבה באותה בקשה. כמות בקשות= מצביעה על תדירות, כמות דפים/ נפח צריבה= מצביעה על גודל .

אבל מה קורה במקרה שיש 3 פקודות במועד חריג ואחת לא במועד חריג, ה-3 במועד חריג היו בגודל 5KB והאחת שלא הייתה במועד חריג הייתה 20M? האם לא נציין את הנפח במועד חריג?

14. avg\_request\_classification – א. מה זאת אומרת סיווג ממוצע של ה+ U3:X4בקשות (שנשלחו באותו יום). הכוונה למצב שהעובד שלח באותו יום מספר בקשות – כאשר באחת מהן הקבצים הם ברמת סיווג נמוכה והשניה היא ברמת סיווג גבוהה. מכיוון שאנחנו מקבצות את הדאטה לרמה של עובד+תאריך צריך להחליט איך לייצג את רמת הסיווג של הקבצים.

ב. מדוע כתבת הערה להביא או את המקסימום או את הממוצע? Max כבר יש לנו, למה לא להביא גם ממוצע? זה תלוי במדיניות של המודל – האם להתריע על כל קובץ שיצא ברמת סיווג גבוהה(במקרה כזה נביא MAX) או שנרצה משהו חלק יותר כמו AVARAGE שלא יקפיץ לנו חריגות על כל בקשה. הסיכון הוא שעובדים יעלו על הלוגיקה, ויעבירו קובץ בסיווג 4 עם מלא קבצים בסיווג 1 ואז הממוצע יהיה קרוב ל-1 למרות שזה כלל קובץ ברמת סיווג סודי ביותר.

תשובת אלה: החלטנו לעשות את שתי העמודות ולא לוותר על אחת מהן.

ג. מהו טיפוס העמודה? אם זה ממוצע זה אמור להיות float, למה זה קטגורי 1-4? טעות שלי, התכוונתי שזה ממוצע בטווח של 1 עד 4, זה אכן float

15. trip\_day\_number - האם trip יכול להיות בארץ, או ש-trip\_day\_number הכוונה רק למספר יום שהותו בחו"ל? רק חו"ל, בארץ לא מעניין.

אם הכוונה רק למספר ימי שהותו בחו"ל, צריך לתקן בעיה בדאטה: יום אחד האדם נמצא בארץ, יום למחרתו הוא נמצא בחו"ל ביום השישי לטיולו. זה לא מספר ימים אלא מספר היום של הטיול- כלומר יום 1= יום יציאה מהארץ, יום 2= יום השני לטיול... יום N= יום חזרה לארץ. אני לא הצלחתי לצפות בדאטה אבל אם התכוונת שבתאריך X השדה 0 ובתאריך X+1 הוא שווה 6 זו באמת שגיאה כי זה חייב להיות רצף.

כן, זה מה שהתכוונתי.

16. num\_entries, num\_exits - האם מספר הכניסות זהה תמיד למספר היציאות? אם כן, האם אין פה עמודה מיותרת? זה המצב התקין, אלא אם כן העובד יצא עם תג של מישהו אחר ואז זה מאוד מחשיד אם לעובד מסוים יש כניסות בלי יציאות ולהיפך..

17. num\_entries - הכוונה לכמות הפעמים שהוא נכנס למערכות של הארגון גם מרחוק, או דווקא פיזית ממש? (אם פיזית ממש - צריך לוודא שבכל הזמן שהוא בחו"ל זה יהיה 0). הכוונה לכניסה פיזית, רעיון טוב להוסיף בדיקה על כניסה במהלך שהייה בחול.

18. total\_presence\_minute – אם האדם נכנס למערכות מרחוק - זה נחשב או לא? משמעותי כנ"ל. מערכת בקרת כניסה מתעדת כניסה פיזית בלבד, בכל פעם שעובד נכנס/יוצא פיזית לשער הקמפוס הוא צריך להעביר תג כדי להיכנס/ לצאת.

19. קלוד הוסיף לכל רשומה עמודה is\_malicious שאומרת האם מדובר בעובד זדוני.

איך הוא החליט מה למלא בכל שורה? קודם הוא החליט באופן רנדומלי מהו ה-id של העובדים החשודים, ואז הוא ג'ינרט להם התנהגות לא רגילה (תפקיד המודל המפענח הוא ההפך - לזהות על סמך ההתנהגות הלא רגילה מהו ה-ID של העובדים החשודים)

כשמדובר בעובד שנחשב זדוני - המודל תמיד שם 1 בעמודה is\_malicios שלו, כך בכל 180 הרשומות שלו. למעשה 180 יום ברצף.

האם עדיף לאפשר לעובד זדוני להיות בעל ימים "תמימים", כך שהלייבל יהיה per יום, האם היום היה יום זדוני או לא, או להשאיר כך – שסימון העמודה הוא ברמת העובד, ואם העובד הוא זדוני אוטומטית כל הימים שלו מסומנים כשייכים לעובד זדוני. באפשרות השניה אין התייחסות האם הפעולות באותו יום היו תמימות או זדוניות. נשאל את אלה.

מסקנה: החלטנו שהסימון יהיה ברמת העובד כך שכל 180 הימים יסומנו כשייכים לעובד זדוני, גם אם היו ימים שלא הייתה בהם שום פעילות זדונית.

19. כמות פקודות הדפסה/צריבות במועד חריג/כניסה במועד חריג – את ציינת את הקריטריונים (השעות המדויקות) על מנת שזה יחשב כחריג/ככניסה בלילה. אם הבנתי נכון, הקריטריונים הללו הם למראה עינינו בלבד – המודל אינו יודע מתי התבצעו הכניסות/ההדפסות, היות שמכניסים לו דאטה כבר מעובד שבו הוא רואה רק שהיו X כניסות חריגות. האם הבנתי נכון?

20. האם להקפיד על יחס מספרי מסוים בין רמות הסיווג השונות?

21. האם להקפיד על יחס מספרי מסוים בין עובדי הקבלן לעובדי החברה?

22. הדאטה דן במקום עבודה ללא היברידיות. הבנתי נכון?

23. avg\_request\_classification – כשהמודל מג'נרט את הדאטה, הוא יג'נרט כאן מספר כלשהו, על פי המצאה שלו. הוא לא יג'נרט על פי חישוב מדויק, היות שחסרים לו נתונים כדי לחשב (דרגת הסיווג של כל בקשה). האם הבנתי נכון?

הערה: שלא נשכח להכניס למודל את טווח ההדפסות/הצריבות.