פתרון שאלה 6 שאלות Data Integrity :

סעיף 1: יתרונות וחסרונות של תיקון הרשומה

יתרונות

- המידע המתוקן מייצג את ה־concept הנכון, כלומר הערך האמיתי וכך נמנעות טעויות עתידיות במודל.
- המודל יכול "ללמוד מהטעות" כי הרשומה הלא נכונה מוחלפת בגרסה המתוקנת והנכונה.
- איכות המדידות משתפרת – פחות רעש ויותר עקביות בין הנתונים לבין המציאות.

חסרונות

- עצם התיקון משנה את מערך הנתונים ממנו המודל לומד, ולכן יש סיכון שהמודל "יאבד" עקביות מול ביצועים שנמדדו בעבר.
- המדידה על נתונים מתוקנים משקפת תמונה אחרת מזו שהייתה בזמן אמת, ולכן קשה להשוות בין מודלים שנמדדו בתקופות שונות (לפני ואחרי תיקון).
- תהליך התיקון עשוי לגרום לכך שהנתונים יהיו "מנוקים מדי" ביחס לעולם האמיתי, וכך ההערכה של המודל תהיה פחות מייצגת את האתגרים בפועל.

סעיף 2: יתרונות וחסרונות של מחיקת הרשומה

יתרונות

- פשוט ומהיר: ברגע שזיהינו טעות, מוחקים.
- מונע כניסה של נתונים שגויים למודל.
- חוסך את הצורך בהכרעה מהו "התיקון הנכון".

חסרונות

- אובדן מידע – ייתכן שגם אם הרשומה לא נכונה במלואה, היא עדיין מכילה חלקים שימושיים.
- פגיעה בכיסוי הנתונים – ככל שמוחקים יותר רשומות, המודל נחשף לפחות דוגמאות.
- דוגמאות שבהן ההחלטה אינה חד־משמעית הן בעלות ערך אינפורמטיבי גבוה: הן מלמדות את המודל להתמודד עם מצבים אמיתיים של חוסר ודאות. מחיקתן מסירה מידע ייחודי זה.
- עלול להטות את המדידה, במיוחד אם נמחקות רשומות מסוג מסוים בתדירות גבוהה, וכך המדדים לא ישקפו את ביצועי המודל במציאות.

סעיף 3: השפעה של תיקון שגיאות לאורך זמן

כאשר מתקנים שגיאות באופן עקבי, אכן נמדוד על נתונים "נכונים" יותר, ולכן לכאורה ההערכה של המודל משתפרת. אבל יש כאן בעיות משמעותיות:

- **הטיה לפי המודל** - ההטיה נוצרת בהתאם למודל שבו משתמשים, כך שהתיקונים לא רק "מרימים" את הביצועים אלא גם מעוותים את בסיס הנתונים. לאורך זמן, נמדוד על נתונים יותר נכונים, אבל בפועל אנחנו מרחיקים את הנתונים שלנו מנתוני ה־production עליהם נמדד. ההערכה על ה־production תהיה נמוכה יותר (אפילו אם היא מוטעית ואמורה להיות יותר גבוהה). חמור יותר, ההטיה היא לפי המודל בו אנחנו משתמשים, כך שההשפעה היא לא רק העלאה של הביצועים אלא עיוות בסיס הנתונים בייצוגיות שלו.
- **ריחוק מנתוני ה־production** - הנתונים המתוקנים אינם זהים לנתונים שהמודל פוגש בפועל בסביבת ה־production. המשמעות היא שהערכת הביצועים שלנו מתרחקת מהתנאים האמיתיים, ולכן כאשר המודל ירוץ בפרודקשן, ייתכן שנראה תוצאות פחות טובות מהצפוי.
- **פער במדידה** - ההערכה על נתוני production (עם טעויות לא מתוקנות) תהיה נמוכה יותר, אפילו אם היא מוטעת, בעוד המדידה על הנתונים המתוקנים תציג ביצועים גבוהים יותר. הפער הזה פוגע באמינות ההשוואה.

לכן, חשוב מאוד להגדיר מדיניות אחידה וברורה:
או שנשמור גרסה "גולמית" של הנתונים לצורך מדידה מול המציאות, או שנבצע תיקונים אך נדאג לציין את ההבדל ולהבין את ההשלכות של ריחוק מה־production.