ניהול מידע מבוזר - פרויקט חלק ב'

מגישים:

ירין בן שטרית 206230021

רועי פאפו 316327451

תומר פרץ 318295029

חלק 3 - Learning:

במשימת החיזוי, בחרנו בנתוני מדידות המגיעים מקנדה בלבד, וזאת משתי סיבות:

1. כמות התחנות בקנדה היא 8910, המדינה השנייה בגודלה מבחינת כמות תחנות מדידה.
2. זוהי מדינה הומוגנית מבחינת אופי האקלים. האקלים במדינה קר ומושלג יחסית, לכן ציפינו לקבל דגימות דומות מבחינת ההתפלגות שלהן, דבר שיאפשר לבצע חיזוי אמין עבור דגימות חדשות.

בנוסף, הפיצ'רים שבחרנו הם הטמפרטורה המקסימלית (TMAX), הטמפרטורה המינימלית (TMIN) וכמות השלג (SNOW). בחרנו למצע את הפרמטרים האלו על פני חודש עבור כל תחנה על מנת לייצר יציבות בדאטה ולמתן מקרי קצה. כך גם כמות הגשם שאנו מנסים לחזות היא כמות גשם ממוצעת על פני חודש עבור תחנה ספציפית.

משימת החיזוי התחלקה למספר שלבים:

1. **עיבוד המידע:** שלב זה מתחלק לשני חלקים - Preprocessing & Postprocessing, כפי שתוארו בחלק השני של הפרויקט. חלק Preprocessing נמצא ב- mlib_data וחלק Postprocessing נמצא ב- mlib_data_post, אלו שתי טבלאות שנמצאות על השרת שיצרנו כדי לאחסן את המידע הרלוונטי לחלק זה. בנוסף, בחלק ה- Preprocessing יצרנו טבלה בשם StationsForClustering המכילה רק תחנות מקנדה, בנוסף לנתוני גובה וקווי אורך ורוחב שלהן.
2. **K-Means:** בהמשך להנחתנו הראשונית של ההומוגניות של הדגימות מקנדה, החלטנו לנסות לעדן ולשפר את ההומוגניות של הדגימות ע"י בחירת תחנות שהמרחק האוקלידי ביניהן קטן מספיק. חישבנו ידנית את סכום המרחקים של כל הנקודות ממרכז הקלסטר שלהן. עבור ערכי K שונים, הרצנו את אלגוריתם K-Means כך שיתקיימו שני תנאים:
   a. אחד מהקלסטרים שנוצרו מכיל כמות של אלפי דגימות.
   b. סכום הריבועים הכולל של כל דגימה ממרכז הקלסטר שלה נמוך ביחס לסכום המרחקים הכולל בשאר הקלסטרים, עבור כל קלסטר.

   לאחר מכן, וידאנו את אמינות המתודה שפיתחנו ע"י אימון הדאטה על מודל של רגרסיה לינארית ובדיקת מדד ה- RMSE עבור סט מבחן שנבחר באקראי בגודל של 20% מתוך כל הדגימות של הקלסטר מהתנאי הקודם. אכן ראינו שקיימת קורלציה גבוהה בין קלסטר שסכום ריבועי המרחקים ממרכזו נמוך, כמות הדגימות שבו גבוהה מספיק (אלפי דגימות, כפי שנדרש בתנאי a) לבין הכללה טובה של מודל הרגרסיה הלינארית שאומן על 80% מהדאטה בקלסטר ונבחן על 20% הנותרים.
   הפיצול לסט האימון והמבחן נבחר בצורה אקראית.
   מתוך כל תתי-הקבוצות שמקיימות את התנאים לעיל, בחרנו את תת-הקבוצה שלדעתנו מקיימת את האיזון הטוב ביותר בין כמות הדגימות לבין מדד RMSE. הדאטה שנבחר אוחסן בטבלת SQL בשם model_data, זוהי הטבלה הסופית בה אחסנו את הנתונים עבור אימון ובחינת המודל.
   הסיבה שבחרנו להשתמש במודל זה היא שזהו מודל פשוט, אינטואיטיבי להבנה וחשנו כי הוא ממדל נכונה את המציאות, היות ואנו מעריכים שמרחק אוקלידי בין אזורים גיאוגרפיים מעיד על התנהגות אקלימית דומה.

3. **אימון ושערוך המודל:** עד כה, בחרנו רק את הדאטה אשר נשתמש בה כדי לבצע אימון וחיזוי של המודל. המודל שבחרנו לבצע את החיזוי באמצעותו הוא מודל הרגרסיה הלינארית. הסיבה שבחרנו במודל זה היא הפשטות שלו והאפקטיביות שלו עבור הדאטה שבחרנו. בחרנו מסט הדגימות הכולל 20% בצורה רנדומלית שישמשו כסט המבחן ו- 80% כסט האימון. הסיבה לבחירת פיצול זה היא ש- 80% מהווה כמות מספקת לאימון המודל, ומצד שני, 20% מהווה כמות גדולה מספיק לבדיקת אמינות המודל על מספר רב של דגימות ותרחישים שונים. ה- RMSE שהתקבל עבור סט האימון (שנבחר רנדומלית בכל פעם) הוא 2.1.

המסקנות שקיבלנו מראות על קיום קשר לינארי בין הפיצ'רים שנבחרו.