בס"ד
מגיש : אליסף סינואני

קישור לגיט

קישור לdb

למידת מכונה שלב א'

חילוץ תכונות ובחירה בנתוני אירועי ירי

הקדמה

הפרויקט הזה שואף לחקור ולהחיל כלים שונים לחילוץ תכונות ולבחירת תכונות ממאגר נתונים המתעד אירועי ירי בעיר ניו יורק, כפי שנרשמו על ידי משטרת ניו יורק .(NYPD) המאגר כולל פרטים על תאריכי, שעות ומקומות האירועים, פרטי הקורבנות והחשודים ותוצאות האירועים. המטרה היא לזהות את התכונות המשמעותיות ביותר המשפיעות על אופי ותדירות אירועי הירי, באמצעות שימוש בטכניקות להפחתת מימדים כדי להבין את המבנה הבסיסי של מאגר הנתונים.

כלים שנמצאו בשימוש

1. **Pandas** : לניפוי נתונים ראשוני, כולל חילוץ עמודות חדשות.
2. **הקידוד הדמי של**(get_dummies) **Pandas** : לקידוד משתנים קטגוריאליים.
3. **SelectKBest עם f_classif** : לבחירת התכונות המשמעותיות ביותר.
4. **StandardScaler** : להכנת הנתונים ל.PCA-
5. **PCA (ניתוח רכיבים עיקריים) :** להפחתת מימדים.

סקירת השלבים :

בחינת הקוד לניתוח נתוני אירועי ירי של NYPD מספקת תובנות עמוקות לגבי התכונות המשפיעות על האירועים הללו. הבחינה הבאה מסכמת את השלבים העיקריים והכלים שהופעלו בתהליך:

טעינת ובחירת נתונים

הקוד מתחיל בטעינת מאגר הנתונים מקובץ CSV, תוך שימוש בספריית `pandas`. נבחרו עמודות רלוונטיות לניתוח, כולל מידע על בורו, תחנת משטרה, קוד גזרה משפטית, קבוצת גיל הקורבן, מין, גזע, ותאריך ושעה של האירוע. זהו שלב חשוב המסייע להתמקד במידע המשמעותי ביותר לניתוח.

קידוד משתנים קטגוריאליים

השימוש ב `get_dummies`- מאפשר קידוד של משתנים קטגוריאליים כגון מין וגזע של הקורבן. פעולה זו חיונית להמרת הנתונים לפורמט נומרי, תוך שמירה על מידע קטגוריאלי בצורה שניתן לנתח.

טיפול בערכים חסרים

ערכים חסרים בנתונים הנומריים מולאו באמצעות החציון, באמצעות `SimpleImputer`. טיפול זה מאפשר המשך עיבוד וניתוח של הנתונים ללא הפרעות הנובעות מערכים חסרים.

סקלינג של התכונות

הקוד מבצע נורמליזציה של הנתונים באמצעות `StandardScaler`, מהלך שנועד להבטיח שכל התכונות יתרמו באופן שווה לניתוח ה.PCA-

ניתוח רכיבים עיקריים(PCA)

הקוד מיישם PCA כדי להפחית את מימדיות הנתונים ולחשוף את המבנה הבסיסי של הנתונים. שני הרכיבים העיקריים הראשונים מוצגים בדיאגרמה, מהלך שמספק תובנות על הקורלציות והקבוצות בתוך מאגר הנתונים.

ויזואליזציה

באמצעות `matplotlib`, הקוד מציג את תוצאות ה PCA-בגרף פיזור. הוויזואליזציה מאפשרת זיהוי תבניות, קבוצות ובודדים במאגר הנתונים, מה שיכול להוביל לתובנות חשובות לגבי אופי האירועים.

במסגרת הפרויקט, כל שלב והכלי שנועד לו ממוקדים במטרה להפיק את המקסימום מהנתונים. התהליך משלב בין טכניקות עיבוד נתונים מתקדמות לטכניקות ויזואליזציה על מנת להבין טוב יותר את המידע הכמוס במאגר הנתונים.

דוח סופי :

במהלך פרויקט ניתוח נתוני אירועי ירי של NYPD, השתמשנו במגוון כלים לעיבוד נתונים, ניתוח סטטיסטי, ויזואליזציה, והפחתת מימדיות. הדיווח הסופי יסכם את נוחות ויעילות שימוש בכל אחד מהכלים, ויבצע השוואה ביניהם.

Pandas

Pandasהוא ספריית Python חזקה ונוחה לעיבוד וניתוח נתונים טבלאיים. השימוש ב Pandas-הוכיח גמישות רבה בטעינה, סינון, וקידוד הנתונים. יתרונותיו במיוחד ברורים בעבודה עם נתונים טקסטואליים וקטגוריאליים, הודות ליכולת להמיר בקלות עמודות לפורמט נומרי. עם זאת, עבור משימות ספציפיות כמו הפחתת מימדיות, הוא דורש שילוב עם ספריות נוספות.

SimpleImputer ו StandardScaler-מSklearn-

Sklearnמספקת כלים להכנת נתונים לניתוח מכונה, כולל טיפול בערכים חסרים ונורמליזציה . SimpleImputerהיה יעיל למילוי ערכים חסרים, אך הגביל אותנו לערכים נומריים בלבד, מה שדרש טיפול נפרד בנתונים לא נומריים StandardScaler. היה חיוני לסקלינג הנתונים, כשהוא מאזן בין תכונות שונות לפני ביצוע.PCA

PCA

PCA, ביצועו דרך Sklearn, מאפשר הפחתת מימדיות נתונים תוך שמירה על רוב המידע הרלוונטי. השימוש ב PCA-היה אינטואיטיבי ונוח, והוא מתאים במיוחד לנתונים נומריים גדולים שבהם קיימת רצון לחשוף את המבנה הבסיסי או לצמצם ממדים לפני ניתוח נוסף.

Matplotlib

ספריית הוויזואליזציה Matplotlib מספקת אפשרויות רחבות לייצוג גרפי של נתונים. השימוש בה להצגת תוצאות ה PCA-היה ישיר ופשוט, מה שמאפשר לחוקרים להבין טוב יותר את הנתונים באמצעות ויזואליזציה גרפית.

השוואה ומסקנות

בעוד ש Pandas-הוא כלי נהדר לניפוי ועיבוד ראשוני של נתונים Sklearn, היא הכרחית לטיפול בערכים חסרים ולהכנת הנתונים למודלים סטטיסטיים ולניתוח מכונה PCA. מספקת את היכולת לצמצם מימדים ולחשוף את המבנה העמוק של הנתונים, בעוד Matplotlib מאפשרת ויזואליזציה חזקה וברורה של התוצאות.

כל כלי מותאם למטרה אחרת ומשלים את האחר. למשל, עבור ניתוח ועיבוד ראשוני Pandas, הוא הבחירה הטובה ביותר, בעוד שלמשימות של הכנת נתונים למודלים סטטיסטיים Sklearn, מספקת כלים חיוניים PCA. מתאים לאנליזה של מערכות נתונים גדולות ומורכבות, ו Matplotlib-מוביל ביכולת להציג נתונים באופן ויזואלי המקל על ההבנה והניתוח שלהם.

Bibliographic List

1. **Pandas**:
   - McKinney, W. (2010). Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference* (pp. 51-56).
2. **Scikit-learn (SimpleImputer, StandardScaler, PCA)**:
   - Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Dubourg, V. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(Oct), 2825-2830.
3. **Matplotlib**:
   - Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90-95.