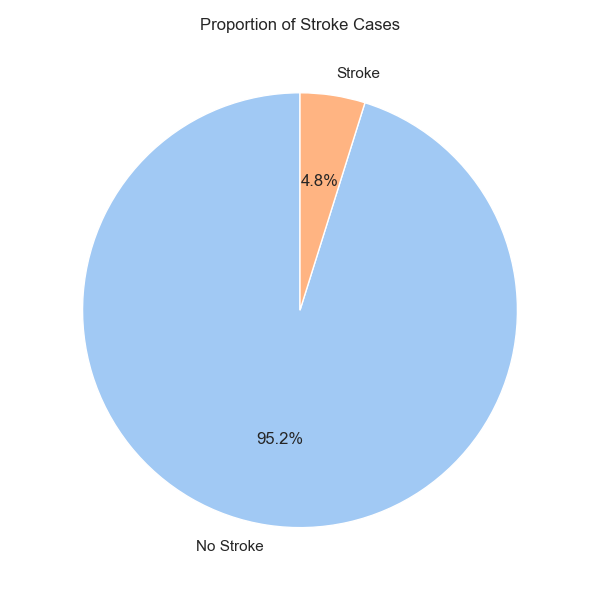
**ניתוח גורמי סיכון ומודל חיזוי לשבץ מוחי: פרויקט במסגרת קורס תכנות מתקדם בפייתון למדעי המוח:**

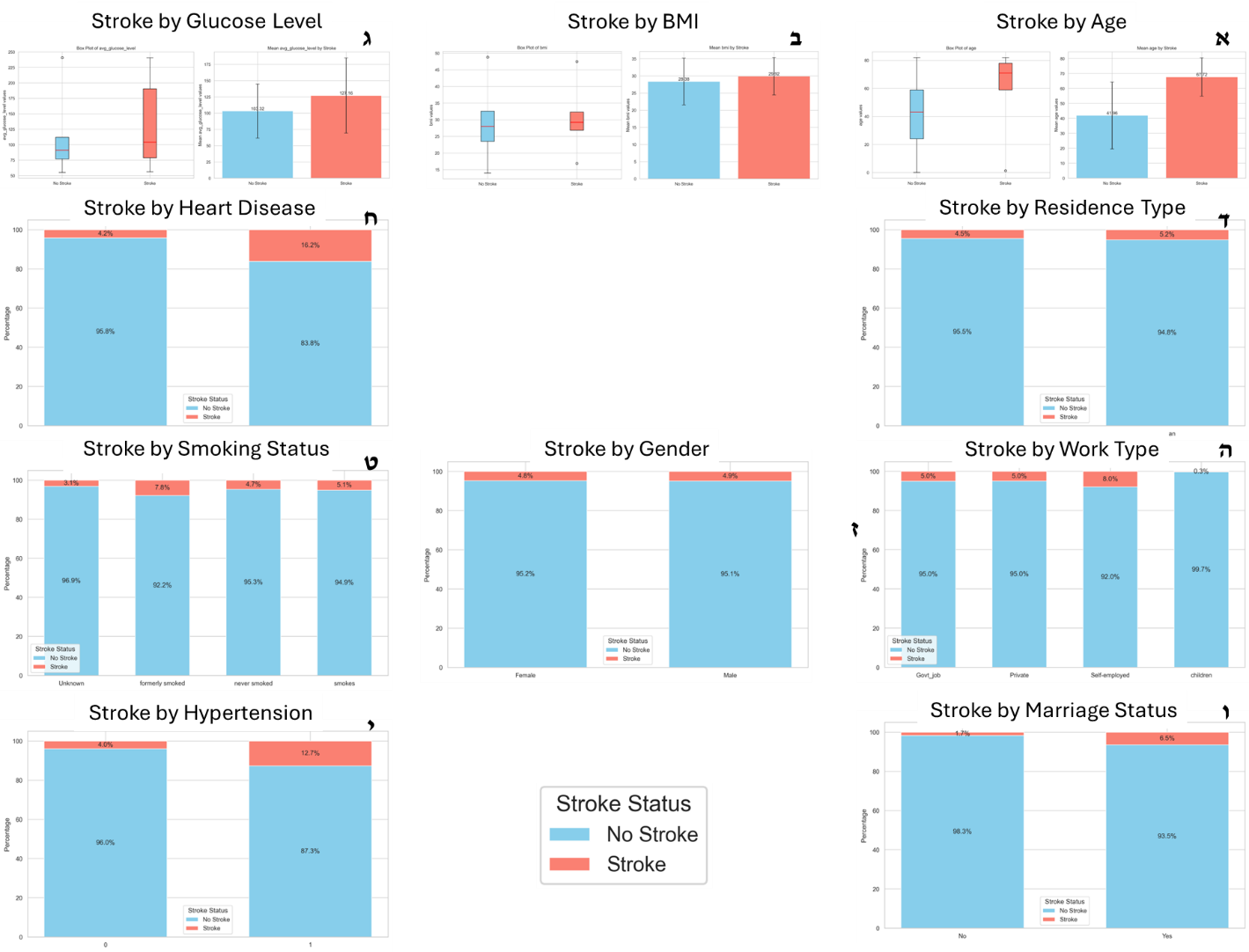
מגישים: שגיא חי 322713223, אריאל לנגה 314805201, אופק קליין 323882928

**תקציר:**  
מחקר זה מנתח גורמי סיכון לשבץ מוחי באמצעות מאגר נתונים של 4,981 מטופלים. הניתוח מתמקד בזיהוי ובכימות ההשפעה של גורמי סיכון שונים, תוך הבחנה בין גורמים הניתנים לשינוי וכאלה שאינם ניתנים לשינוי.

**מבוא:**שבץ מוחי הוא אחת הסיבות המובילות לתמותה ונכות ברחבי העולם. זיהוי מוקדם של גורמי סיכון ואפיון מאפייני החולים הנמצאים בסיכון גבוה הם קריטיים למניעה ולטיפול יעיל. מחקר זה מנתח מערך נתונים מקיף הכולל מדדים דמוגרפיים, רפואיים והתנהגותיים.  
כפי שניתן לראות באיור 1, רק 4.8% מהמקרים במאגר הנתונים הם מקרי שבץ, עובדה המציבה אתגר מתודולוגי בניתוח הנתונים ובבניית מודל החיזוי.  
סטטיסטיקה תיאורית והתפלחויות של משתני הנתונים ביחס לקבוצות השבץ מוצגות באיור 2. ניתן לראות שוני רב במשתנה ה**גיל** (2א), שם הממוצע בקבוצת השבץ הוא 67.7±12.8 שנים לעומת 42.0±22.3 בקבוצה הבריאה. ב-**BMI** (2ב) ההבדל מתון יותר, עם ממוצע של 29.9±5.5 בקבוצת השבץ לעומת 28.4±6.8 בקבוצה הבריאה. ב**רמות הגלוקוז** (2ג) נצפה הבדל עם ממוצע של 127.2±57.8 בקבוצת השבץ לעומת 103.3±41.6 בקבוצה הבריאה, כאשר השונות הגבוהה יותר בקבוצת השבץ מרמזת על השפעה מורכבת של גורם זה. מבחינת **אזור מגורים** (2ד), נצפה הבדל קטן – אחוז מקרי השבץ עבור אזור עירוני הוא 5.2%, ועבור האזור הכפרי 4.5%. עבור **סוג העבודה** (2ה)- שיעור לוקי השבץ הוא 0.3% בילדים, בעצמאים 8.0%, ובקבוצות המגזר הפרטי ועבודות ממשלתיות 5%. ב**מצב משפחתי** (2ו) עלה סיכון גבוה יותר לנשואים 6.5%, לעומת לא נשואים 1.7%. בנוגע ל**מגדר** (2ז) נצפה הבדל מינימלי - גברים 4.9%, נשים 4.8%. **מחלות לב** (2ח) הציגו הבדל משמעותי - 16.2% בחולי לב לעומת 4.2% בבריאים. **סטטוס עישון** (2ט) הראה סיכון גבוה למעשנים לשעבר 7.8% בעוד שעבור מעשנים נוכחיים שיעור זה עומד על 5.1%, ואצל אנשים לא מעשנים 4.7%. **יתר לחץ דם** (2י) הראה סיכון מוגבר לסובלים מלחץ דם גבוה - 12.7% מול 4.0% אצל אלו שלא.



**איור 1**: יחס מקרי השבץ בנתונים



**איור 2**: סטטיסטיקה תיאורית ופילוחי המשתנים השונים על פי קבוצות שבץ

**שיטות:**

ניקוי ועיבוד נתונים:

- טיפול בערכים חסרים באמצעות השלמת ממוצע או שכיח (בהתאמה לסוג העמודה) על בסיס סיווג לקבוצות גיל.

- הסרת חריגים (z-score > 3) בעמודות מספריות.

- לקראת שימוש במודל Random Forest בוצעה המרת משתנים קטגוריאליים למספריים ומסד הנתונים פוצל ל features ו target.

ניתוח נתונים וסטטיסטיקה:

עבור כל משתנה בוצע חישוב מתאם בינו ובין עמודת המטרה – 'שבץ'.

- מבחני חי בריבוע למשתנים קטגוריאליים.

- מתאם נקודתי דו-סידרתי למשתנים מספריים (משמש למדידת הקשר בין משתנה דיכוטומי למשתנה רציף).

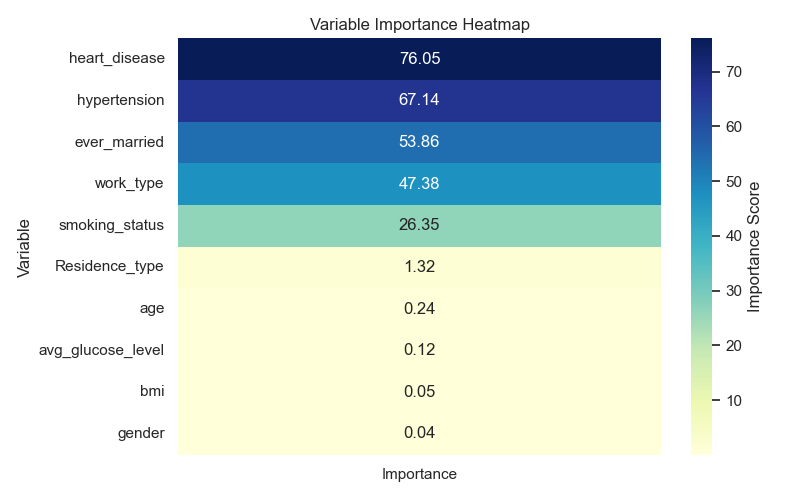
- בוצע סיווג המשתנים כגורמי סיכון ניתנים לשינוי (רמת גלוקוז, bmi, סטטוס עישון, סוג עבודה ומקום מגורים) או בלתי ניתנים לשינוי (גיל, מגדר, יתר לחץ דם, מחלות לב). המתאמים שחושבו עבור משתנים אלו נסכמו להשוואה בין ההשפעה המצטברת של כל קבוצה על מצב השבץ.

- חושבו מבחני t להשוואת קבוצות מקרי השבץ (0\1) עבור כל עמודה מספרית.

- מספר משתנים בלתי תלויים (גיל, סטטוס עישון ו BMI) נותחו לפי מקרי השבץ בפילוח מגדרי.

- בוצע שימוש במודל למידת המכונה Random Forest, יחד עם התאמת SMOTE לדאטה לא מאוזן, על מנת ליצור מודל לחיזוי תוצאת שבץ על פי הנתונים.

**תוצאות:**

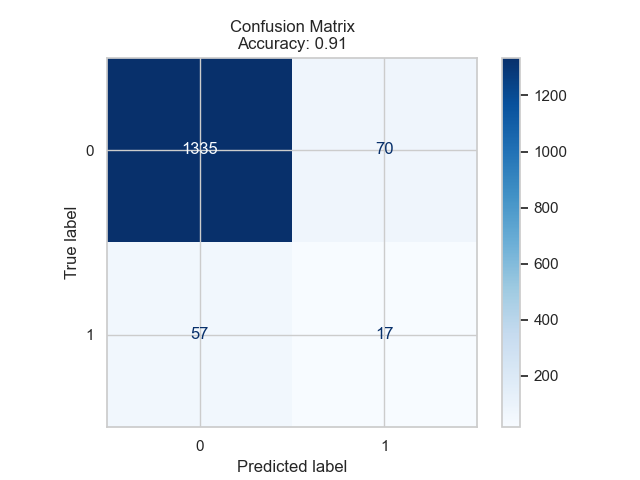


**איור 3**: מפת חום של החשיבות היחסית של המשתנים הבלתי תלויים, המבוטאת על ידי המתאם שלהם לשבץ

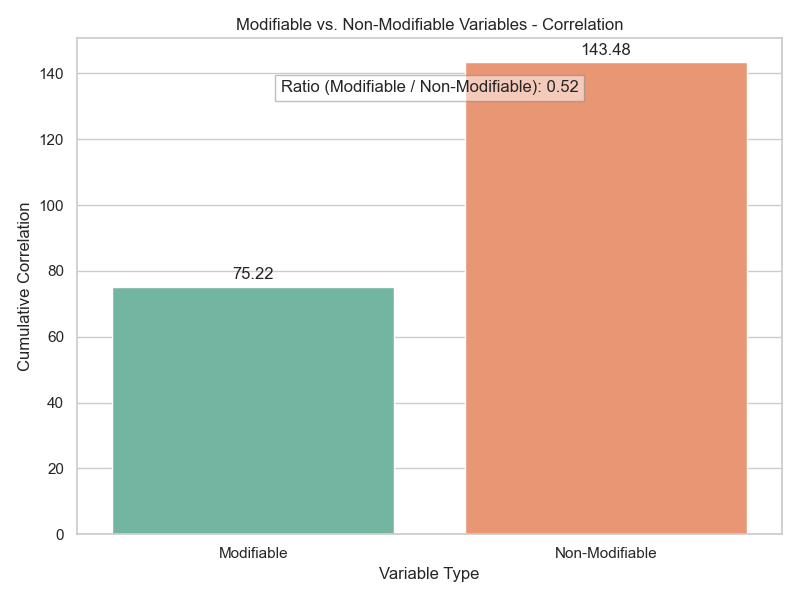
נמצא כי מחלות לב ויתר לחץ דם הם המשתנים המשמעותיים ביותר מבחינת החשיבות היחסית שלהם לשבץ (המוגדרת כמתאם שלהם למשתנה זה), ומציגים ערכים של 76.05 ו-67.14 בהתאמה. זאת בעוד שמשתנים כמו BMI (0.05) ומגדר (0.04) אותרו כבעלי השפעה מינימלית.

מפת החום באיור 3 מציגה את החשיבות היחסית של כלל המשתנים הבלתי תלויים.

משתנים שאינם ניתנים לשינוי משפיעים כמעט פי שניים מהניתנים לשינוי (143.48 מול 75.22) כפי שמוצג באיור 4.



**איור 5**: מטריצת הבלבול של מודל הניבוי.

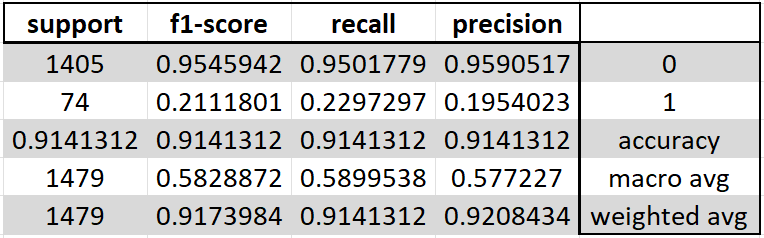


**איור 4**: מידת ההשפעה המצטברת של משתנים ניתנים ושאינם ניתנים לשינוי על שבץ. חושב על ידי סכימת המתאמים.

מטריצת הבלבול של מודל הניבוי באיור 5 מציגה דיוק (accuracy) של 0.91, אך בשל מדגם לא מאוזן בפועל המודל התקשה לסווג מקרי שבץ. המודל זיהה נכונה 1,385 מקרים של היעדר-שבץ ו-17 מקרי שבץ, אך החמיץ 57 מקרי שבץ (false negative) וסיווג בטעות 70 מקרים כשבץ שלא לצורך (false positive). שיעור השגיאה הגבוה במקרי השבץ (57 מתוך 74 מקרים, כ-77%) מדגיש את הבעייתיות בשימוש במדד הדיוק הכללי במדגמים לא מאוזנים. בעוד שהמודל מצליח היטב בזיהוי מקרים שליליים (specificity גבוה), הרגישות (sensitivity) שלו נמוכה משמעותית, מה שמעיד על צורך בשיפור יכולת הזיהוי של מקרי השבץ האמיתיים.

בטבלה 1 מוצג דו"ח הסיווג של המודל.

**טבלה 1**: דו"ח הסיווג של מודל הניבוי.



איור 6 מציג את ניתוח הקשר בין משתני הגיל, העישון ומדד מסת הגוף (BMI) לשבץ בפילוח מגדרי. בצידו השמאלי של האיור מוצגים מקרי השבץ, ובצידו הימני האחוז היחסי שלהם בכל תת-קבוצה. בהיבט ה**גיל** (5א), נצפתה עלייה משמעותית בשכיחות השבץ לאחר גיל 45 בשני המגדרים, כאשר נשים מראות שיעור גבוה יותר במיוחד בגיל 75 ומעלה. דפוס ה**עישון** (5ב) מציג תמונה מורכבת: בעוד שגברים שעישנו בעבר נמצאים בסיכון מוגבר (כ-9%), נשים שמעולם לא עישנו מציגות שיעור גבוה יותר של מקרי שבץ. בניגוד למצופה, מעשנים פעילים בשני המגדרים מציגים שיעורי שבץ נמוכים יחסית. בבחינת **BMI** (5ג), לא אותרו הבדלים משמעותיים בין גברים לנשים, למעט מקרי שבץ גבוהים יותר עבור נשים בקטגוריית המשקל הנורמלית. עבור שני המגדרים, השכיחות המרבית של שיעור השבץ נצפתה בקטגוריית עודף המשקל.

**איור 6**: ניתוח הקשר בין משתני הגיל, העישון ומדד מסת הגוף (BMI) לשבץ בפילוח מגדרי.

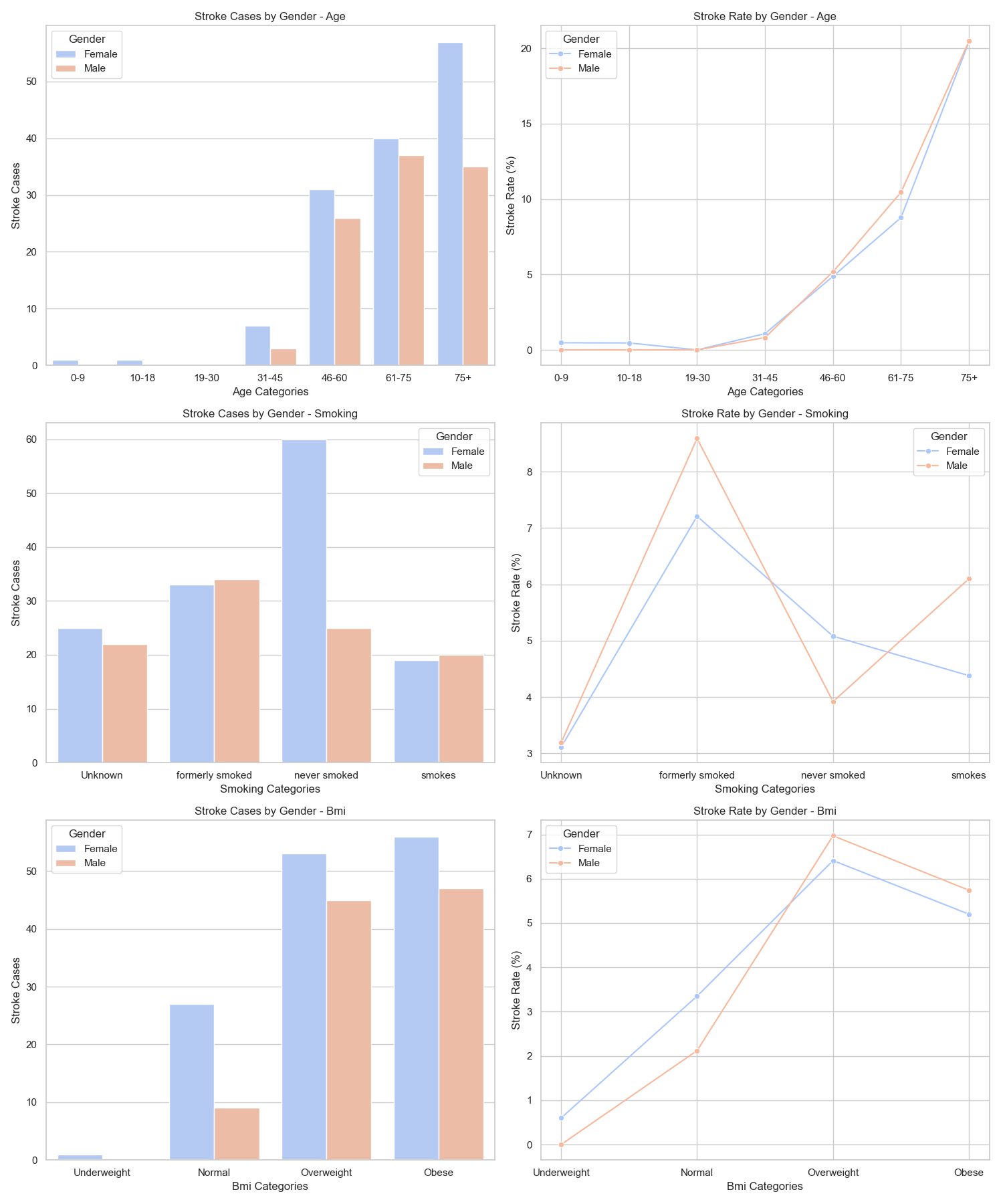
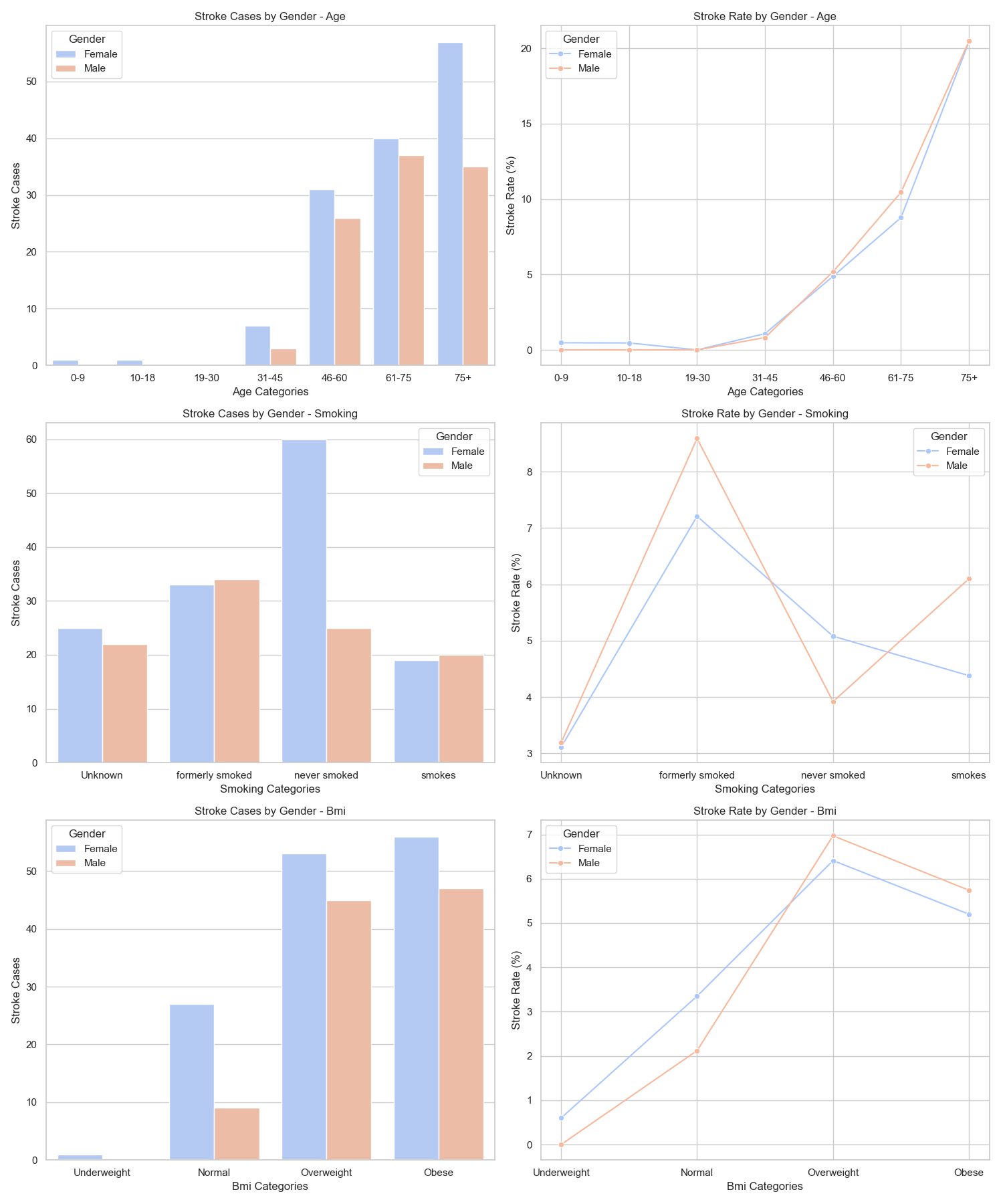
Smoking Status

BMI

Stroke Cases

Stroke Rate

Age



**א**

**ב**

**ג**

**דיון:**

ממצאי המחקר מצביעים על מספר תובנות מרכזיות:

1. גורמים שאינם ניתנים לשינוי:

- גיל הוא גורם הסיכון המשמעותי ביותר

- מחלות רקע מגדילות משמעותית את הסיכון

2. גורמים הניתנים לשינוי:

- רמת גלוקוז היא המשתנה המשמעותי ביותר

- BMI משפיע באופן שונה על גברים ונשים

3. אינטראקציות:

- השפעת הגיל מתחזקת בנוכחות גורמי סיכון אחרים

- דפוסי סיכון שונים בין המגדרים

**מסקנות**:

1. המלצות קליניות:

- ניטור מוגבר למטופלים עם מחלות רקע

- דגש על איזון רמות גלוקוז

- התאמת אסטרטגיות מניעה לגיל ומגדר

2. מגבלות המחקר:

- חוסר איזון בנתונים

- הטיה אפשרית במדגם

- מידע מוגבל על התפתחות לאורך זמן