פרויקט סיום - למידת מכונה

שבץ מוחי

קישור לgithub הפרויקט:
https://github.com/KoralElbaz/Machine-Learning.git

מגישות:
קורל אלבז 318477684
סיון יהב 314805235

תיאור המאגר

- על פי ארגון הבריאות העולמי שבץ מוחי הוא הגורם השני למוות בעולם.
- מערך נתונים זה משמש לניבוי אם מטופל צפוי לקבל שבץ מוחי בהתבסס על פרמטרי קלט כמו מין, גיל, מחלות שונות ומצב עישון.
- כל שורה בנתונים מספקת מידע רלוונטי על המטופל.
- ערכי העמודות מגוונים : מחרוזות, ערכים בינאריים וערכים מספריים.
- חשוב לציין כי המאגר שלנו מכיל 5111 מטופלים.

חלק קטן מהמאגר שלנו

	A	B	C	D	E	F	G	H	I	J	K
1	gender	age	hypertension	heart_disease	ever_married	work_type	Residence	avg_glucose_level	bmi	smoking_status	stroke
2	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
3	Female	61	0	0	Yes	Self-employed	Rural	202.21	None	never smoked	1
4	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
5	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
6	Female	79	1	0	Yes	Self-employed	Rural	174.12	24	never smoked	1
7	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly smoked	1
8	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
9	Female	69	0	0	No	Private	Urban	94.39	22.8	never smoked	1
10	Female	59	0	0	Yes	Private	Rural	76.15	None	Unknown	1
11	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
12	Female	81	1	0	Yes	Private	Rural	80.43	29.7	never smoked	1
13	Female	61	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1
14	Female	54	0	0	Yes	Private	Urban	104.51	27.3	smokes	1
15	Male	78	0	1	Yes	Private	Urban	219.84	None	Unknown	1
16	Female	79	0	1	Yes	Private	Urban	214.09	28.2	never smoked	1
17	Female	50	1	0	Yes	Self-employed	Rural	167.41	30.9	never smoked	1
18	Male	64	0	1	Yes	Private	Urban	191.61	37.5	smokes	1

הסבר על מערך הנתונים בו השתמשנו

- **Id** – מזהה ייחודי
- **Gender** – מין "זכר" או "נקבה"
- **Age** – גיל
- **Hypertension** – לחץ דם (0 אם המטופל ללא לחץ דם, אחרת 1)
- **heart_disease** – מחלת לב (0 אם המטופל ללא מחלות לב, אחרת 1)
- **ever_married** – אם המטופל התחתן בעבר או לא
- **work_type** – מקום עבודת המטופל (עבודה עם ילדים , משרה ממשלתית , אף פעם לא עבד , פרטי או עצמאי)
- **Residence_type** – סוג מגורים (כפרי או עירוני)
- **avg_glucose_level** – רמת גלוקוז ממוצעת בדם
- **Bmi** – bmi של מטופל
- **Smoking_status** – האם המטופל מעשן (יש כאלה שמוגדרים כלא ידוע, כלומר האינפורמציה לא זמינה עבור אותם המטופלים)
- **Stroke** – האם המטופל קיבל שבץ מוחי. (0 אם המטופל לא קיבל, אחרת 1)

שאלות שנרצה לענות עליהם

בפרויקט שלנו הצגנו ארבע שאלות שאותן ניסינו לחזות בעזרת המאגר באמצעות 4 שיטות למידת מכונה שונות.

עבור כל שאלה חילקנו את המאגר לשני חלקים- חלק אחד הוא לאימון והחלק השני לבחינה/ בדיקה.

בחלק זה של הפרויקט נציג את השאלות, את דיוק החיזוי של כל שיטת לימוד עבור המאגר, בעיות שנתקלנו בהן ודרך ההתמודדות שלנו איתם, ניתוח שלנו על השאלה ותוצאות הלמידה.

על פי מאפיינים (גיל, מגדר, יתר לחץ דם, מחלות לב, האם נישא, רמת גלוקוז ממוצעת בדם, סטטוס עישון),
האם המטופל קיבל שבץ מוחי?

על פי מאפיינים (מין, מחלות לב, האם נישא, סטטוס עישון, BMI) ,
האם למטופל יש/אין יתר לחץ דם?

על פי מאפיינים (גיל, סטטוס עישון, מגדר, מקום מגורים, מקום עבודה) ,
האם המטופל התחתן בעבר?

על פי מאפיינים (יתר לחץ דם, מחלות לב, שבץ, bmi,ורמת גלוקוז ממוצעת בדם)
האם המטופל מעל או מתחת גיל 43.22 ?

מהי השפעת כל גורם (באחוזים) על הסיכוי לחטיפת שבץ מוחי.

הסבר על האלגוריתמים בהם השתמשנו

- **Adaboost**

אלגוריתם זה הוא ממשפחת אלגוריתמי boosting כלומר משתמש באוסף של מודלים חלשים על מנת ליצור מודל אחד חזק.
זהו אלגוריתם למידה המחפש מספר קטן של מסווגים "חזקים" מתוך קבוצה של מסווגים "חלשים".
האלגוריתם מעניק משקל גדול לשגיאות בזיהוי (בכך מגדיל את סיכוייהן לסיווג מתאים בהמשך).
המשקל מסמל את חשיבות התכונה.

- **K-Nearest Neighbors**

באלגוריתם זה ניקח את K השכנים הקרובים ביותר לנקודה החדשה ונסווג את הנקודה לפי השכנים עם הרוב של אותו סיווג.
כלומר, הקלט תלוי ב- k התצפיות הקרובות במרחב התכונות (פיצ'רים).

Logistic Regression

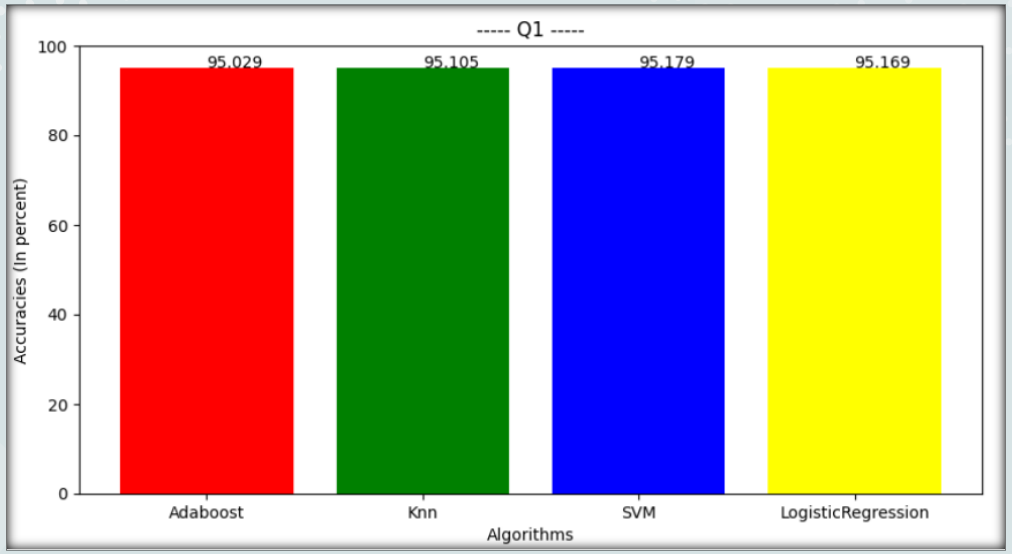
זהו מודל סטטיסטי המשמש כמודל ההסתברות של מעמד או אירוע מסוים קיים.
למשל: עובר/נכשל, ניצחון/הפסד, חי/מת או בריא/חולה.
במקום לחזות סיווג של בדיוק 0 או 1 ,
רגרסיה לוגיסטית מייצרת הסתברות - ערך בין 0 ל -1.

SVM

מכונת וקטורים תומכים עבור בעיות סיווג, בשלב האימון מתאימים מסווג שמפריד נכון ככל האפשר בין דוגמאות אימון חיוביות ושליליות.
המסווג שנוצר ב- SVM הוא המפריד הליניארי אשר יוצר מרווח גדול ככל האפשר בינו לבין הדוגמאות הקרובות לו ביותר בשתי הקטגוריות. כאשר מוצגת נקודה חדשה, האלגוריתם יזהה האם היא ממוקמת בתוך הקו המגדיר את הקבוצה, או מחוצה לו.

ניתוח שאלה ראשונה

האם המטופל קיבל שבץ מוחי?

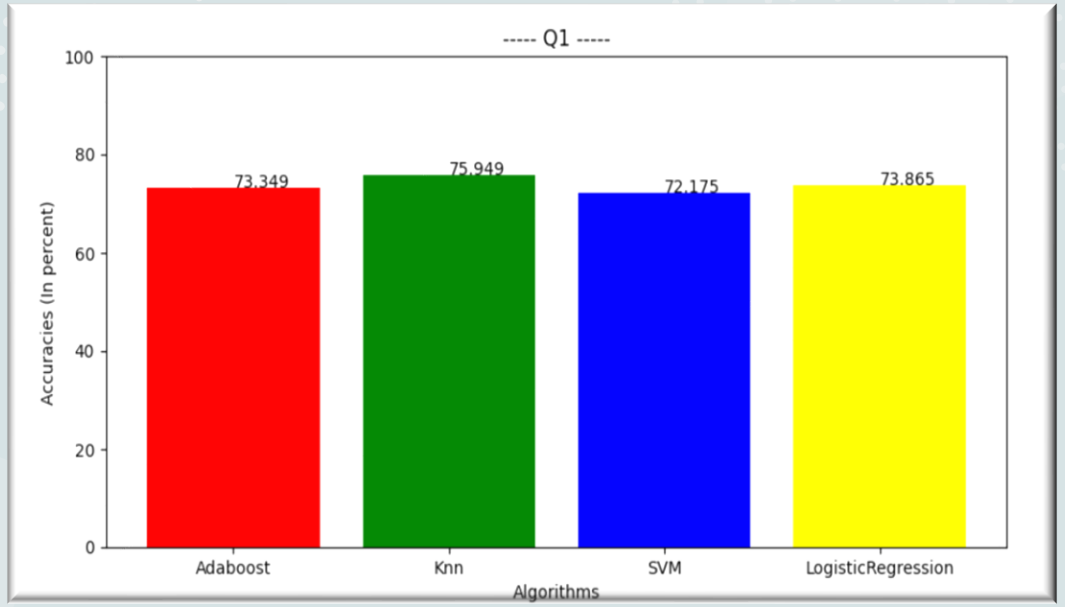



```
----- Q1: Accuracy chance of stroke -----

    Algo                |      Accuracies
Adaboost:               |       95.029%
Knn:                    |       95.105%
SVM:                    |       95.179%
Logistic Regression:    |       95.169%
```

ניתן לראות שהדיוק יחסית גבוה וקרוב בין שיטות הלימוד.

תיקון עבור שאלה 1

האם המטופל קיבל שבץ מוחי?

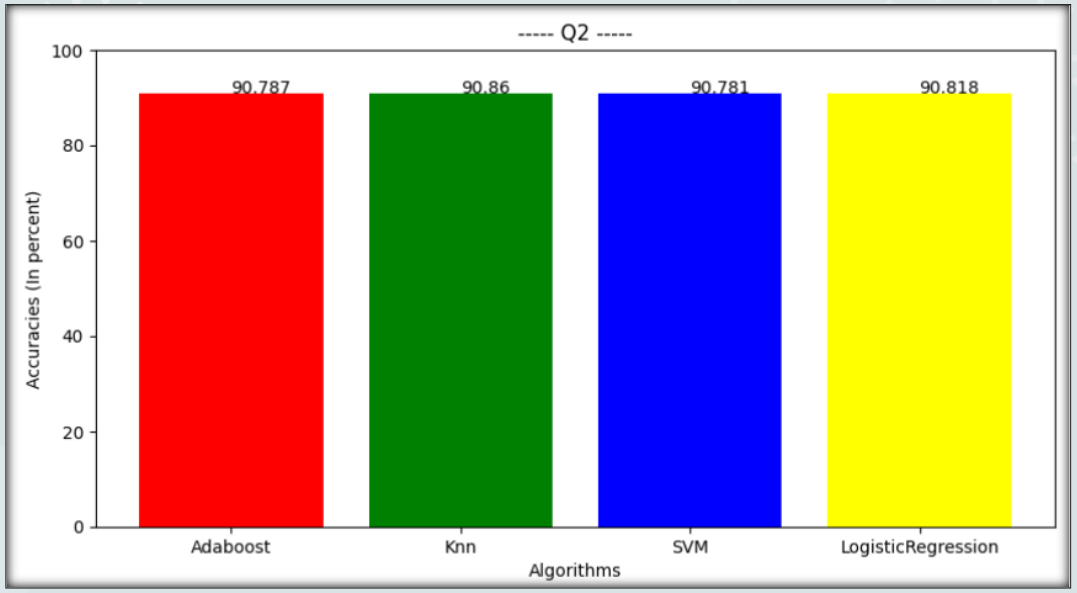



```
----- Q1: Accuracy chance of stroke -----

    Algo                |      Accuracies
Adaboost:               |       73.349%
Knn:                    |       75.949%
SVM:                    |       72.175%
Logistic Regression:    |       73.865%
```

הבנו שהמאגר הנתונים שלנו הוא אינו מאוזן ולכן לאחר איזונו ניתן לראות שתוצאות החיזוי השתנו.

ניתוח שאלה שניה

האם למטופל יש/אין יתר לחץ דם?



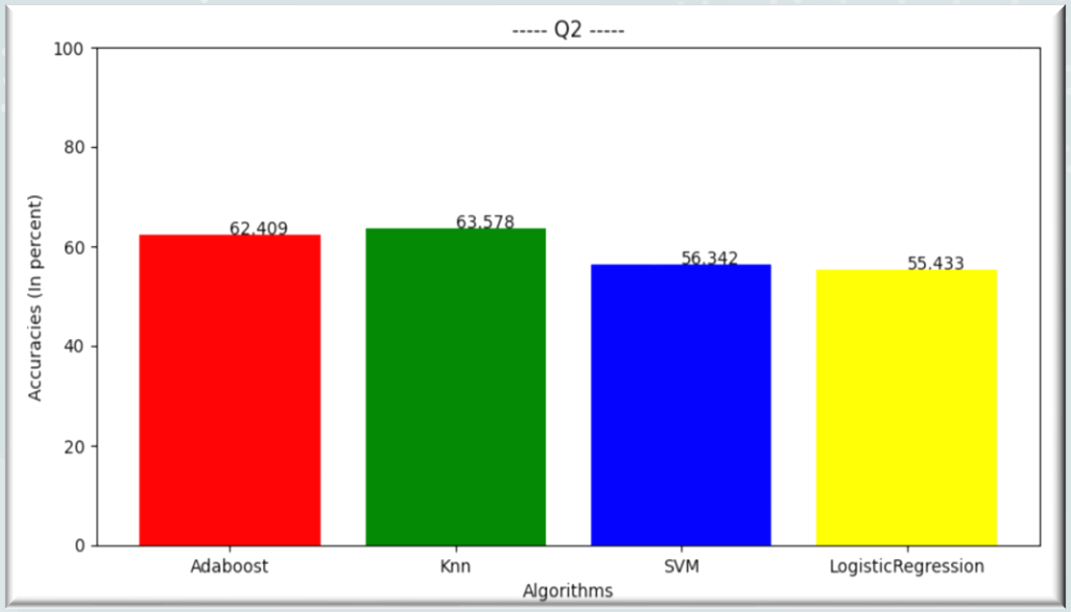

```
----- Q2: Accuracy chance of hypertension -----

    Algo              |      Accuracies
Adaboost:             |       90.812%
Knn:                  |       90.63%
SVM:                  |       90.864%
Logistic Regression:  |       90.868%
```

```
Number of patients who do not have hypertension:  4612
Number of patients:  5110
Several percent of people without hypertension:  90.254 %
```

- המסקנה הראשונית שלנו הייתה שיש קורלציה גבוהה בין הפרמטרים שלנו למצב לחץ דם של מטופל .
- לאחר מכן ניסנו לשנות את הפרמטרים לפרמטרים אחרים וקיבלנו תוצאות זהות, מכך הסקנו שהמסקנה הראשונה שלנו שגוייה .
- ההשערה שנייה שלנו הייתה, שהאלגוריתם רואה אחוזים גבוהים בקרב מטופלים שאין להם יתר לחץ דם, ולכן כמעט תמיד יחזה את האפשרות הזאת ולא משנה איזה פיצ'רים ניתן לו התוצאות דיוק יהיו גבוהות.
- כדי לחזק את הטענה בדקנו את אחוזי המטופלים שאין להם יתר לחץ דם וכפי שניתן לראות מעל 90%. שזה גם קרוב לאחוזי הדיוק שלנו בכל השיטות השונות.

תיקון עבור שאלה 2

האם למטופל יש/אין יתר לחץ דם?

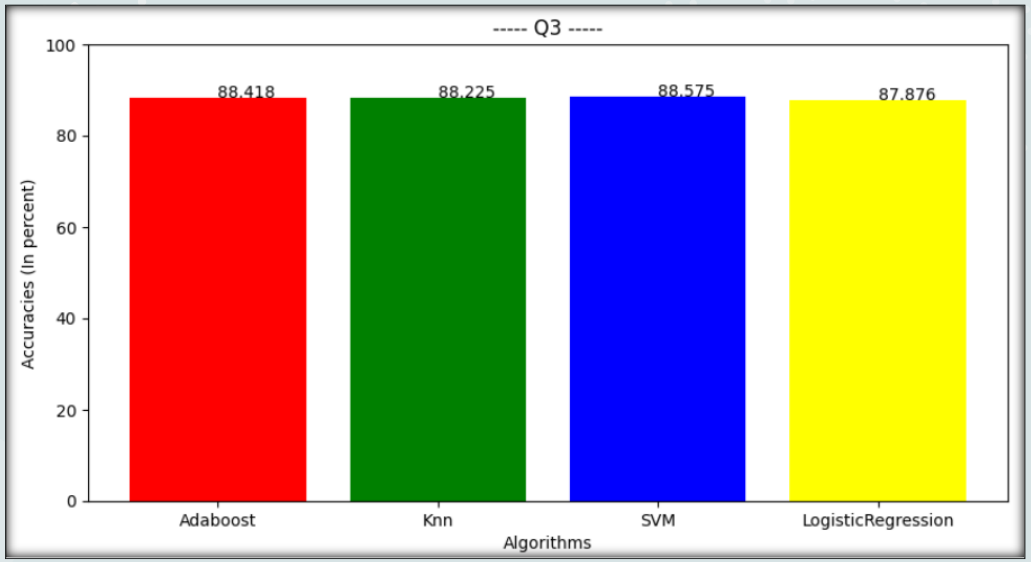



```
----- Q2: Accuracy chance of hypertension -----

    Algo              |      Accuracies
Adaboost:             |       62.409%
Knn:                  |       63.578%
SVM:                  |       56.342%
Logistic Regression:  |       55.433%
```

הבנו שהמאגר הנתונים שלנו הוא אינו מאוזן ולכן לאחר איזונו ניתן לראות שתוצאות החיזוי השתנו.

ניתוח שאלה שלישית

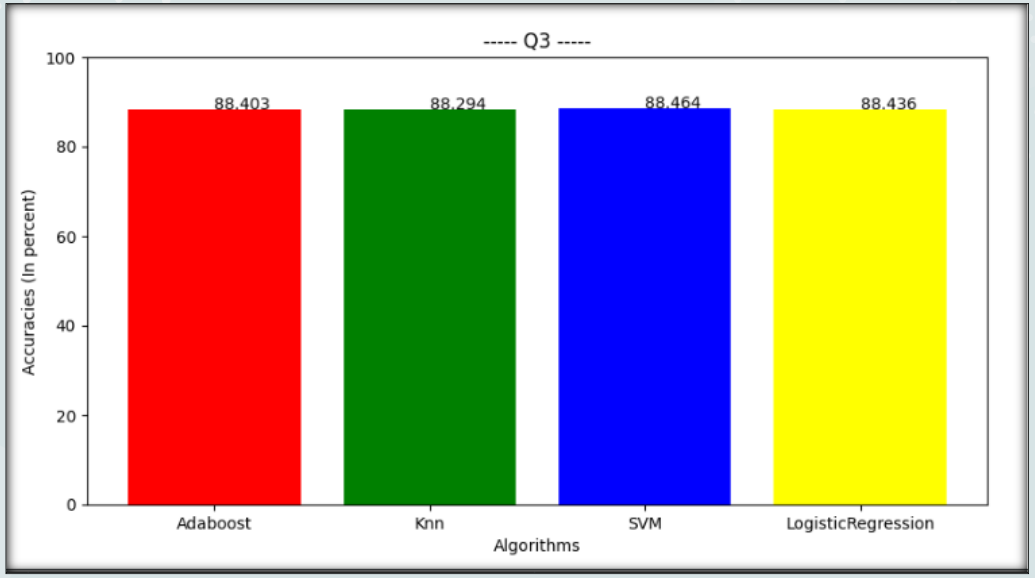
האם המטופל התחתן בעבר?




```
----- Q3: Accuracy chance of ever married -----

    Algo                |      Accuracies
Adaboost:               |       88.418%
Knn:                    |       88.225%
SVM:                    |       88.575%
Logistic Regression:    |       87.876%
```

- בתחילה המרנו את הגיל לערכים בינאריים (0 או 1 לפי הממוצע גילאים), שמנו לב שכאשר ביטלנו את ההמרה הציונים אכן גדלו (בדיעבד הבנו שגרמנו לאיבוד מידע, מה שהשפיע על אחוזי הדיוק).
- ישנם חמישה סוגי עבודה שונים.
כאשר טענו את מערך הנתונים מספרנו אותם מ–0 עד 4 בהתאם לסוג העבודה.
דבר זה גרם לנו לשער שגרמנו לעדיפויות מסוימות אשר השפיעו על תוצאות החיזוי.
כדי לבדוק השערה זו הפעלנו One-Hot-Encoder על מערך הנתונים שלנו בעמודה שמתארת את סוג העבודה, על מנת לבדוק אם יהיו שינויים באחוזי הדיוק.

לאחר השינוי...

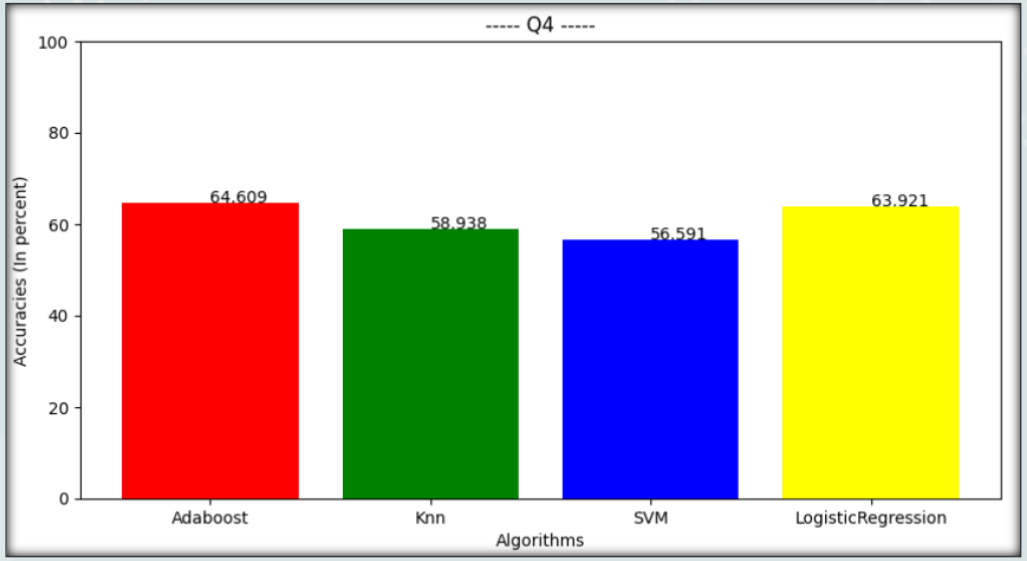


- ניתן לראות שאין שינויים משמעותיים באחוזי הדיוק.

```
----- Q3: Accuracy chance of ever married -----

    Algo              |      Accuracies
Adaboost:             |       88.403%
Knn:                  |       88.294%
SVM:                  |       88.464%
Logistic Regression:  |       88.436%
```

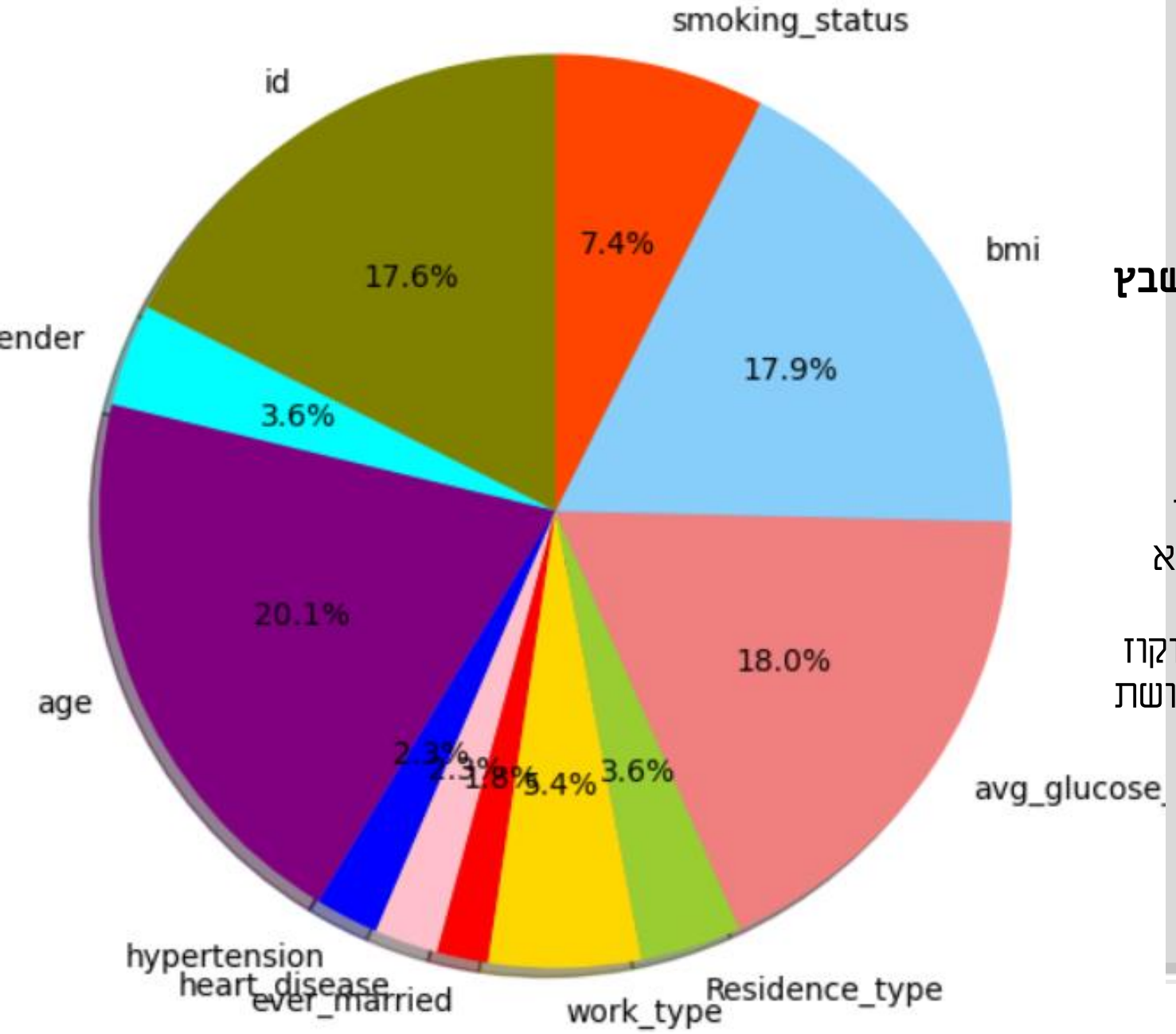
ניתוח שאלה רביעית

האם המטופל מעל ומתחת לגיל 43.22?




```
----- Q4: Accuracy chance of age > 43.22 -----

    Algo                |      Accuracies
Adaboost:               |       68.372%
Knn:                    |       63.632%
SVM:                    |       60.633%
Logistic Regression:    |       65.224%
```

- ניתן לראות שאחוזי ההצלחה שלנו לחיזוי היו נמוכים יחסית לשאלות קודמות.
- בתחילה שיערנו כי אין קורלציה בין חלק מהמשתנים לגיל (Glucose ,BMI), אז ניסינו לבדוק את תוצאות החיזוי ללא פיצ'רים אלו כדי לראות אם נוכל לקבל חיזוי טוב יותר לגיל המטופל.
אבל לא הצלחנו להגיע לתוצאה מרשימה.
- לאחר מכן ההשערה שלנו הייתה כי שינוי הגיל לערכים בינאריים לפי ממוצעי הגיל, גורם לנו לאיבוד מידע סביב טווחי גילאים מסוימים, דבר שגורם לכך שאחוזי הדיוק לא יהיו כל כך גבוהים.



ניתוח שאלה חמישית

השפעת כל גורם (באחוזים) על הסיכוי לחטיפת שבץ מוחי.

ניתן לראות בגרף שניתחנו את חשיבות כל גורם ממערך הנתונים שלנו על מנת לראות כמה הוא משפיע על הסיכוי לקבל שבץ מוחי.
ניתן לראות שהגיל, ה-bmi וממוצע רמת הגלוקוז בעלי האחוזים הגבוהים ביותר, כלומר הם שלושת הגורמים העיקריים לשבץ מוחי.

תודה על ההקשבה