סעיף ד'

בהמשך לסעיפים א ו-ב ערכו סקירה השוואתית לכלל החלופות האפשריות (לפחות 4 חלופות) לביצוע כריית מידע. בתשובתכם יש להתייחס ליתרונות/חסרונות כל אחת מהחלופות בהקשר לבעיה הנתונה.

	יתרונות הסוג	חסרונות הסוג	מדד הפיצול		הסבר כללי	יתרונות	חסרונות	ספציפית בנוגע לבעיה הנתונה
עץ החלטה	• יכול לקבל ערכים קטגוריאלים גם ערכים רציפים • לא דורש נרמול של הdataset • בעזרת זה אפשר לזהות בקלות מי מהמאפיינים לא רלוונטי ולכן לדוגמה במקרה שלנו יהיה קל לחשב לנורמלים לא מהמחלה ולכן לדוגמה השמנת יתר ולמנוע לקבל מידע זה לצבור בין הסבירות לבין שהקשר לא ברור לי. • בעזרת הנתונים ניתן לזהות חלקים חריגים לחלוטין.	• חמדן בבצוע ההחלטות שלו • זמן האימון הנדרש ארוך יחסית גדול בגלל שהוא מחשב מחדש • בדרך כלל המסקנות דורשות חישוב ארוך בגלל לבצע אליהם. • יש נטייה ל overFitting בגלל שבגלל שהם מתאימים אחרים. • לא מספק טוב בגלל שבגלל שהם מספקים לבצע חיזוי לבצע לחיזוי גדול יחסית • מצריך סט אימון	**C4.5**	**Gain Ratio**	1. בכל צומת של העץ, בוחר את התכונה שמפצלת בצורה היעילה ביותר את קבוצת הסיווג לקבוצות משנה המפורטות ביותר לסיווג הנוכחי או לאחר. קריטריון הפיצול הוא רווח המידע המנורמל (הפרש באנטרופיה). 2. התכונה עם רווח המידע הנורמלי הגבוה ביותר נבחרת כדי לקבל את ההחלטה. 3. חזרה עבור העלים שנוצרו לעץ לעיל.	- מבצע גיזום ולכן פחות סבירות לoverfitting - מסוגל להתמודד עם ערכים חסרים ופחות נוקשה בהכנת הנתונים שלו. - יכול לבצע חלוקה של העץ עם תכונות שלהם יש עלויות שונות	- יוצר עצי החלטה לא מאוזנים בגלל שתכונותיהם מביאות ל"רווח" הטוב ביותר.	• הגיזום יביא לחיזוי טוב יותר לעומת id3 • יקל עלינו בהכנת המידע שכן לא נדרש להעביר את הנתונים הרציפים לבדידים ולכן יתן חיזוי טוב יותר.
			ID3	**Information gain**	1. חישוב האנטרופיה של כל תכונה של ערכת הנתונים 2. חלוקה לקבוצות משנה באמצעות התכונה שעבורה האנטרופיה המתקבלת היא אידיאלית 3. יצירת עץ החלטה המכיל תכונה זו. 4. חזרה על כל קבוצות המשנה עם התכונות הנותרות		- פתרון לא אופטימלי - אין גיזום בניגוד ל- c4.5 ול- cart - לא תומך בערכים חסרים/רציפים	• יקשה עלינו מאוד את הביצוע מכיוון שלא תומך בערכים רציפים, ולנו בנתונים יש כמה וכמה ערכים כאלו(גיל, משקל וכו')
			Cart	**GINI index**	1. בכל צומת של העץ, בוחר את התכונה שמפצלת בצורה היעילה ביותר את קבוצת הסיווג לקבוצות משנה המפורטות ביותר לסיווג הנוכחי או לאחר. קריטריון הפיצול הוא לפי מדד גיני 2. התכונה עם רווח המידע הנורמלי הגבוה ביותר נבחרת כדי לקבל את ההחלטה. 3. חזרה עבור העלים שנוצרו לעץ לעיל.	- בניגוד ל c4.5 תומך בחיזוי רציף. - לא מחשב ערכות כללים - מבצע גיזום (פחות סבירות לoverfitting) - חסין לערכים חסרים ופחות נוקשה בהכנת הנתונים שלו. - יכול לבצע חלוקה לפי עלויות תכונה שונות	- עץ ההחלטה המתקבל הוא בינארי	• הגיזום יביא לחיזוי טוב לעומת id3 • יקל עלינו בהכנת המידע שכן לא נדרש להעביר את הנתונים הרציפים לבדידים ולכן יתן חיזוי טוב יותר. • אפשרי לבצע חיזוי יותר מדויק מאשר זה שאנו מנסים ליישם בכך שנחזה השמנת יתר שתלוי בbmi וכן בנתונים נוספים שניתנו לנו, ולהגיע למסקנות רלוונטיות יותר(רציפות)
רגרסיה	• אמנם לוגיסטי מתבצע בתחום בצעדים לצעדים • בנוסף בעזרת הנלקחים בחשבון	• מניח מבנים הנקבעים מראש • רגיש מאוד לנתונים חריגים בתחום הנתונים	**לינארית**	**פונקציית רשת עצבית בריבוע**		- עובד רק אם מדובר בקשר לינארי.	- רגיש מאוד לחריגים	• מכיוון שהבעיה הנתונה היא בדידה הישום של כלי זה יכול להיות אפשרי אבל יקשה עלינו מאוד את התהליך. • מכיוון שיש לנו ערכים רציפים ובדידים התוצאות של עץ החלטה ינטו להיות טובות יותר.
יער אקראי	**יתרונות:** • יכול להתמודד עם נתונים לא מאוזנים. וכן עם מידע חסר מונע overfitting		**חסרונות**: זמן חישוב ארוך יחסית לשאר האלגוריתמים		מטא מסווג (כמה עצי סיווג אקראיים) כאשר כל עץ סיווג תוך שימוש בתת רשימה של מאפיינים(אקראיים) מתוך כלל המאפיינים			יכול להיות יעיל מאוד לבעיה הנתונה בעקבות רמת הפירוט שיש לכל רשומה וכך שבעץ אקראי יבחנו בכל שלב רק מספר תכונות לסיווג ולא כלל התכונות ובעזרת כך להביא לתוצאות מיטביות