מדעי הנתונים ובינה עסקית

תרגיל מס' 1

**הוראות כלליות:**

1. הגשת העבודה היא בזוגות.
2. יש להגיש את העבודה דרך המודל (שותף אחד מכל זוג צריך להגיש).
3. יש להגיש את העבודה עד לתאריך 9.4.2020.
4. יש לכתוב תשובות סופיות במקום המיועד ובסוף המסמך לצרף את **כל** החישובים המתאימים.
5. יש להקפיד על **3** ספרות אחרי הנקודה העשרונית, **אי הקפדה תגרור הורדת נקודות**.
6. תשובה סופית, גם אם היא נכונה, ללא הצגת דרך החישוב המלאה בנספחים, **לא תזכה בניקוד**.

**שאלה 1 (30 נק')**

הנתונים הבאים מתארים את שיעורי הפשיעה בקנדה בשנים 1931-1968 כאשר:

Tfr Total fertility rate per 1000 women.

Partic Women's labor-force participation rate per 1000.

Degrees Women's post-secondary degree rate per 10,000.

Fconvict Female indictable-offense conviction rate per 100,000.

Ftheft Female theft conviction rate per 100,000.

Mconvict Male indictable-offense conviction rate per 100,000.

Mtheft Male theft conviction rate per 100,000.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **year** | **tfr** | **partic** | **degrees** | **fconvict** | **ftheft** | **mconvict** | **mtheft** |
| 1935 | 2755 | 238 | 13.2 | 79.4 | 20.4 | 765.7 | 247.1 |
| 1936 | 2696 | 240 | 13.2 | 91 | 22.1 | 816.5 | 254.9 |
| 1937 | 2646 | 241 | 12.2 | 100.4 | 22.4 | 821.8 | 272.4 |
| 1938 | 2701 | 242 | 12.6 | 108.9 | 21.8 | 956.8 | 285.8 |
| 1939 | 2654 | 244 | 12.3 | 123.6 | 21.1 | 1035.7 | 292.2 |
| 1940 | 2766 | 245 | 12 | 157.3 | 21.4 | 951.6 | 256 |
| 1941 | 2832 | 246 | 11.7 | 154.3 | 25.3 | 850.9 | 205.8 |
| 1942 | 2964 | 268 | 11.2 | 143.9 | 27.1 | 769.7 | 188 |
| 1943 | 3041 | 333 | 11.5 | 147.5 | 29 | 811.2 | 205.8 |
| 1944 | 3010 | 335 | 11.1 | 97.3 | 24.2 | 865.4 | 207.9 |
| 1945 | 3018 | 331 | 12.5 | 76.6 | 24.7 | 866.8 | 197.8 |
| 1946 | 3374 | 253 | 15 | 72.8 | 20.5 | 968 | 195.9 |
| 1947 | 3545 | 243 | 17.6 | 68.9 | 20.7 | 894 | 198.9 |
| 1948 | 3441 | 241 | 21.2 | 66.7 | 22.8 | 830.8 | 198.6 |
| 1949 | 3456 | 242 | 22.7 | 55 | 18.5 | 811.2 | 216.1 |
| 1950 | 3455 | 237 | 21.4 | 55 | 19.3 | 775.4 | 212 |
| 1951 | 3503 | 242 | 20.7 | 55 | 20 | 739.8 | 208 |
| 1952 | 3641 | 240 | 19.5 | 54.7 | 19.4 | 740.3 | 199.3 |
| 1953 | 3721 | 233 | 36.6 | 47.9 | 16.1 | 725.1 | 176.4 |

1. יש לנרמל את המשתנה **partic** באמצעות Min-Max normalization

כאשר New min = 1, New max = 10.

1. מהו המינימום המקורי של המשתנה? \_\_\_233\_\_\_\_\_\_\_
2. מהו המקסימום המקורי של המשתנה? \_\_\_\_\_335\_\_\_\_\_
3. מהו הערך המנורמל של המשתנה בשנת 1952? \_\_\_\_\_\_1.618\_\_\_
4. יש לצרף טבלה המכילה את המשתנה המנורמל בנספחים.
5. יש לבצע החלקה למשתנה mconvict באמצעות Weighted moving average כאשר:

|  |  |
| --- | --- |
| Weight | t |
| 0.18 | -3 |
| 0.22 | -2 |
| 0.6 | -1 |

1. מהו הערך של המשתנה mconvict המוחלק בשנת 1952\_\_\_\_\_760.484\_\_\_\_\_
2. מהו הערך של המשתנה mconvict המוחלק בשנת 1946? \_\_\_\_\_\_856.484\_\_\_\_
3. נא לצרף את המשתנה המוחלק בנספחים.
4. נא לצרף גרף עם המשתנה המקורי והמשתנה לאחר ההחלקה בנספחים.

**שאלה 2 (35 נק')**

רופאים מ-Cleveland Clinic Foundation מעוניינים לחזות נוכחות של מחלת לב בחולים. לשם כך, נאסף מידע על 100 חולים. קובץ ה-CSV המצורף (hw1\_dataset) מכיל 3 משתנים מועמדים ומשתנה מטרה אחד (prediction).

הגדרות המשתנים:

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Domain** | **Type** |
| gender | 1=male, 0=female | Nominal |
| fbs | 0=false, 1=true | Nominal |
| slope | 1=upsloping, 2=flat, 3=downsloping | Nominal |
| Prediction | 0,1,2,3,4 | Nominal |

1. מהי האנטרופיה של המשתנה gender? \_\_\_\_0.869\_\_\_\_\_\_
2. מהי האנטרופיה של המשתנה fbs? (10 נק') \_\_\_\_0.577\_\_\_\_\_\_
3. מהי האנטרופיה המותנית של משתנה המטרה בהינתן המשתנה slope \_\_\_\_1.577\_\_\_\_\_\_
4. מהי האנטרופיה המקסימלית של המשתנה slope? \_\_\_\_1.585\_\_\_\_\_\_
5. יש לחשב את ה-mutual information של משתנה המטרה ו-gender בשתי דרכים:
6. לפי ההפרש בין האנטרופיות הרלוונטיות \_\_\_\_0.090\_\_\_\_\_\_
7. לפי הנוסחה של mutual information I(X;Y) \_\_\_\_\_0.090\_\_\_\_\_

**שאלה 3 (35 נק')**

יש להשתמש באותו קובץ נתונים משאלה 2.

1. יש לבנות מודל ID3 (עד רמה 2, קודקוד השורש הוא רמה 0). יש להראות חישוב ידני של לפחות משתנה אחד בכל קדקוד, לשאר המשתנים ניתן לכתוב רק את התוצאה הסופית. יש להציג את העץ שהתקבל.
2. מהו ה - majority rule accuracy? \_\_\_\_0.57\_\_\_\_\_\_\_
3. מהו ה - training accuracy של המודל? \_\_\_\_\_\_\_0.59\_\_\_\_\_
4. 5 חולים חדשים הגיעו לבית החולים. ע"פ המודל שבניתם, סווגו את הרשומות הבאות:

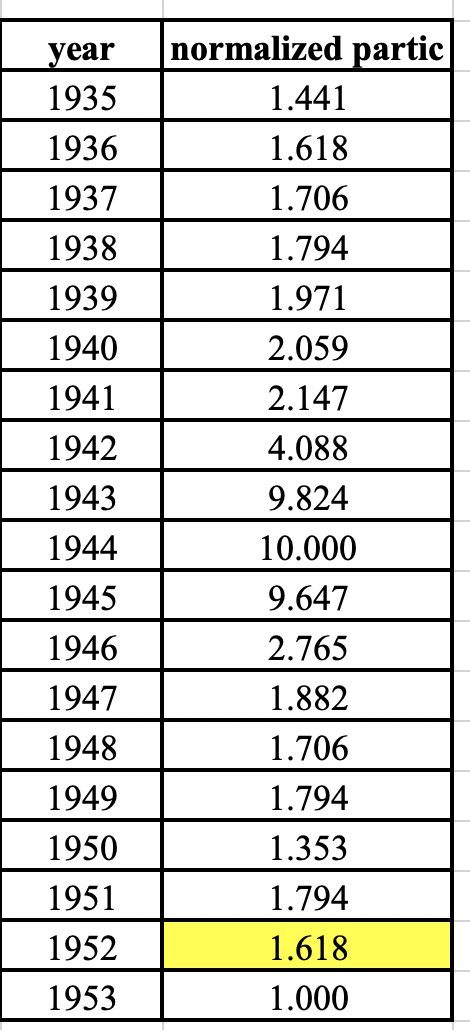
|  |  |  |  |
| --- | --- | --- | --- |
| **Gender** | **fbs** | **Slope** | **Prediction** |
| 0 | 1 | 1 | **0** |
| 1 | 0 | 2 | **1** |
| 1 | 1 | 3 | **0** |
| 0 | 0 | 3 | **3 (random choice between 3 and 0)** |
| 0 | 0 | 2 | **0** |

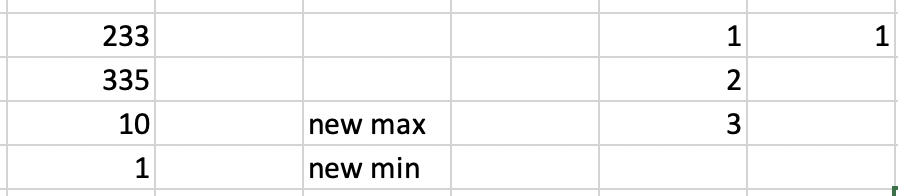
**בהצלחה!**

**נספחים**

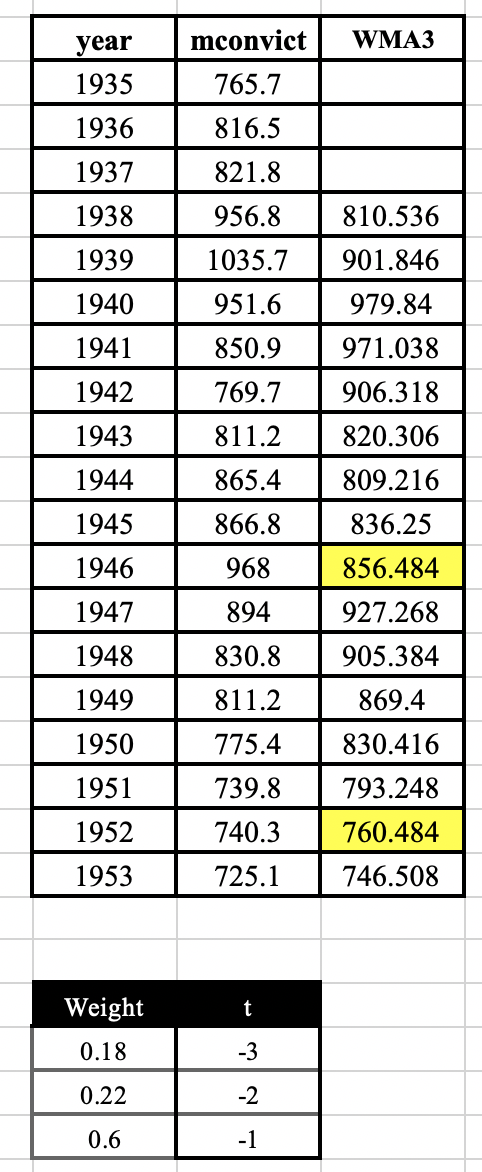
הערה: כל החישובים הנוספים הנדרשים נמצאים בקובץ האקסל המצורף.

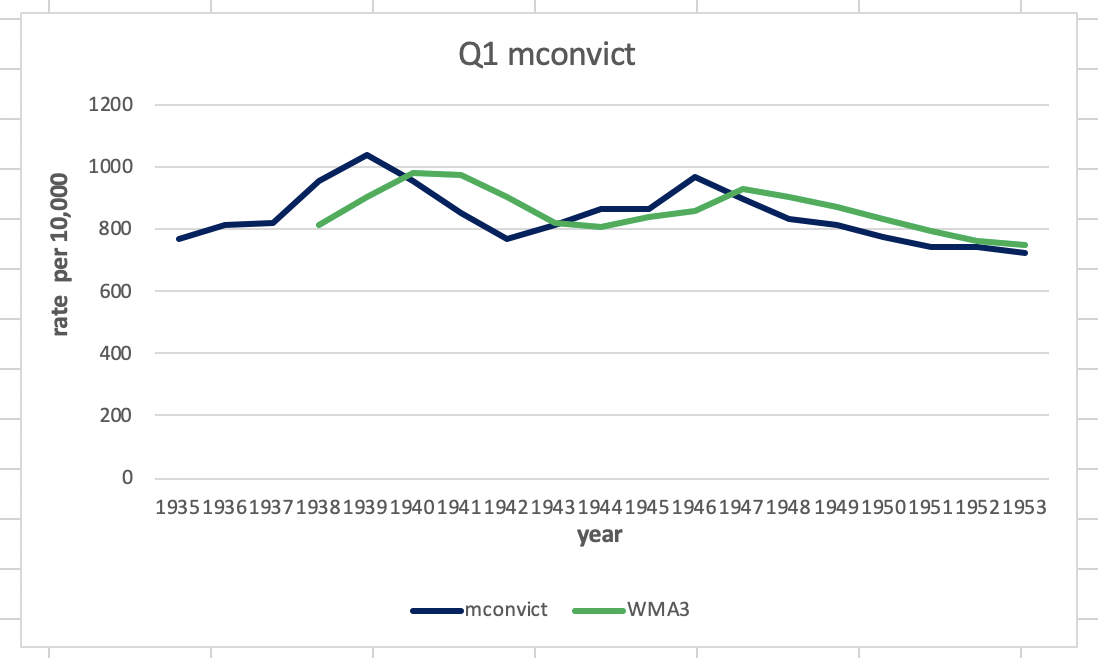
**1א.**

****

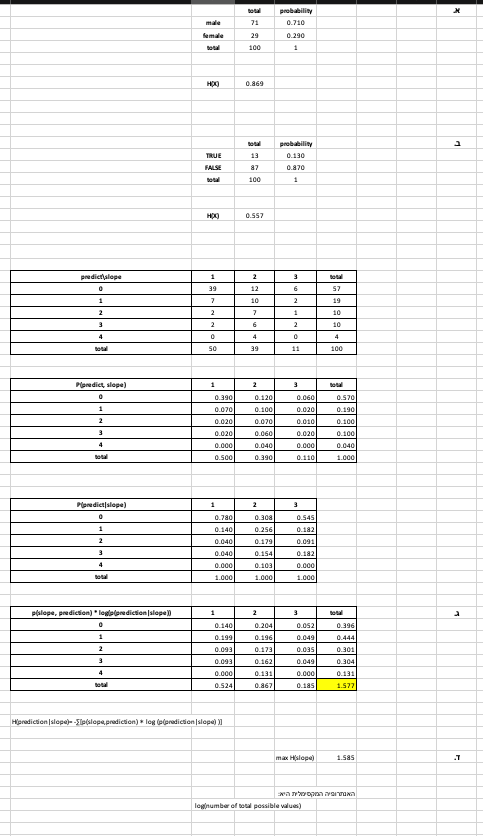


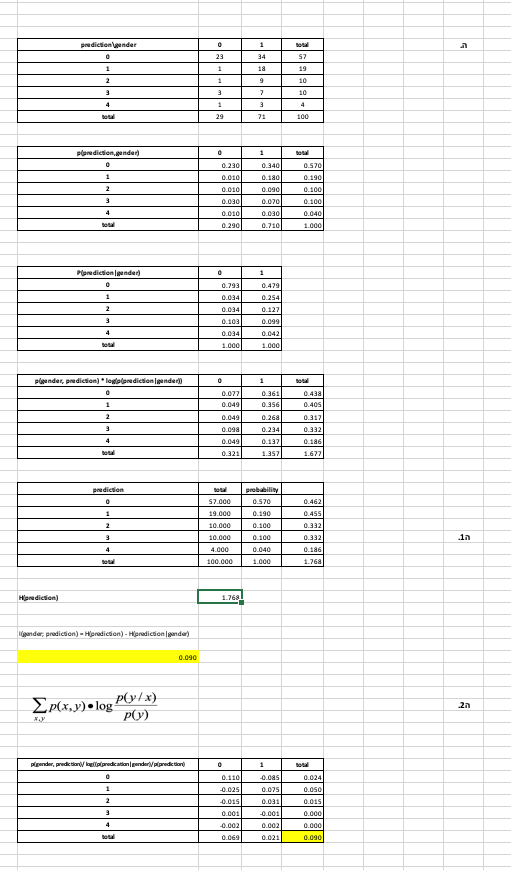
**1ב.**



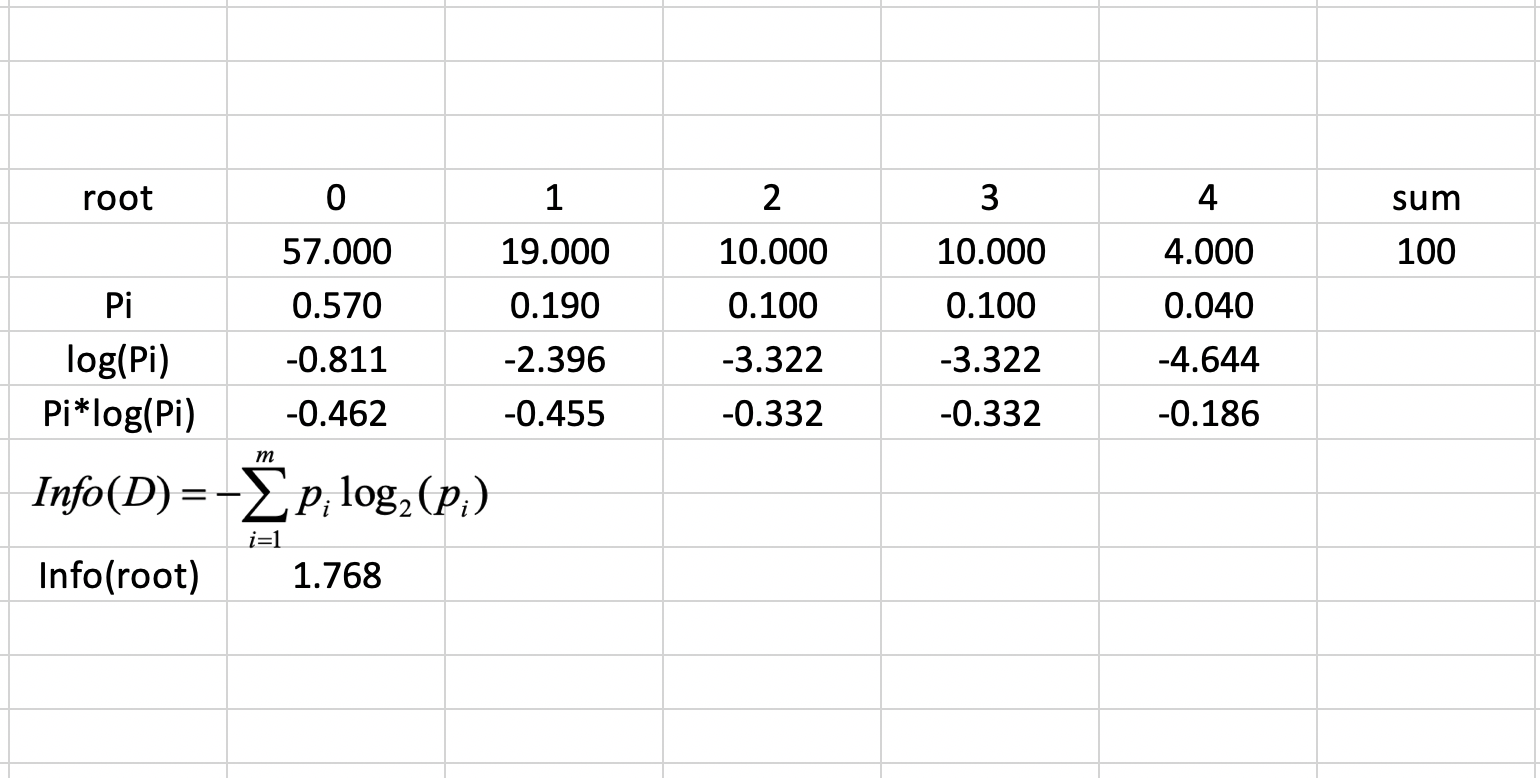


**2א.**

****

****

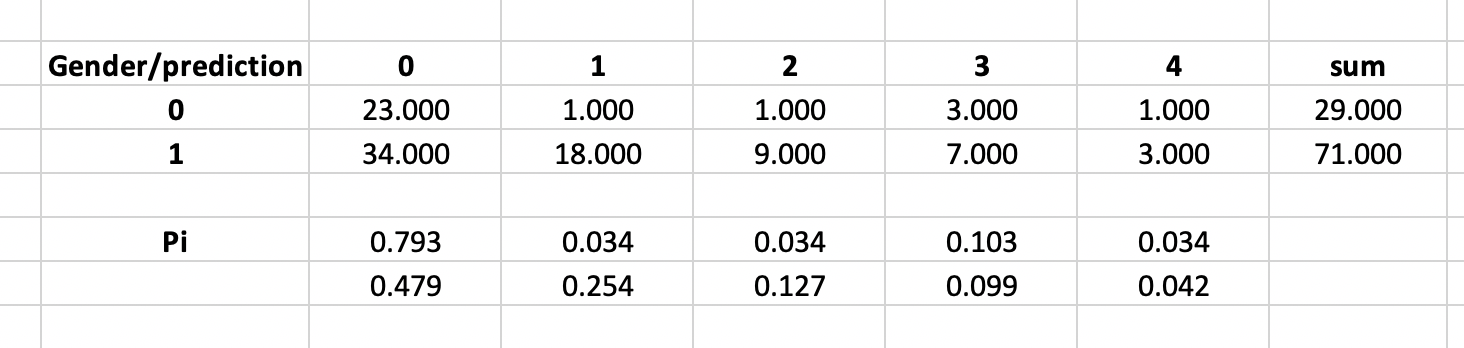
**3א.**

****

**חישוב אנטרופיה של השורש:**

**חישוב information gain של פיצול לפי gender:**

נקבע וקטור לערכי הרשומות במשתנה המטרה:



­

­

=

1.768 – 1.678 = **0.09**

**חישוב information gain של פיצול לפי fbs:**

=

1.768 – 1.745= **0.023**

**חישוב information gain של פיצול לפי slope:**

=

1.768 – 1.577= **0.191**

**נבחר לפצל לפי המשתנה slope מכיוון שהוא מספק את רווח הידע הגדול ביותר (הקטנת אי הוודאות הגדולה ביותר).**

**חישוב אנטרופיה של הקדקוד בו slope=1:**

האנטרופיה של קדקוד זה חושבה בשלב הקודם:

**חישוב information gain של פיצול לפי gender:**

­

­

=

1.048 – = **0.089**

**חישוב information gain של פיצול לפי fbs:**

=

1.048 – 0.99=**0.058**

**נבחר לפצל לפי המשתנה gender מכיוון שהוא מספק את רווח הידע הגדול ביותר (הקטנת אי הוודאות הגדולה ביותר).**

**חישוב אנטרופיה של הקדקוד בו slope=2:**

האנטרופיה של קדקוד זה חושבה בשלב הקודם:

**חישוב information gain של פיצול לפי gender:**

­

­

=

2.224 – = **0.107**

**חישוב information gain של פיצול לפי fbs:**

=

* 1. -2.122=**0.102**

**נבחר לפצל לפי המשתנה gender מכיוון שהוא מספק את רווח הידע הגדול ביותר (הקטנת אי הוודאות הגדולה ביותר).**

**חישוב אנטרופיה של הקדקוד בו slope=3:**

האנטרופיה של קדקוד זה חושבה בשלב הקודם:

**חישוב information gain של פיצול לפי gender:**

­

­

=

1.686– = **0.445**

**חישוב information gain של פיצול לפי fbs:**

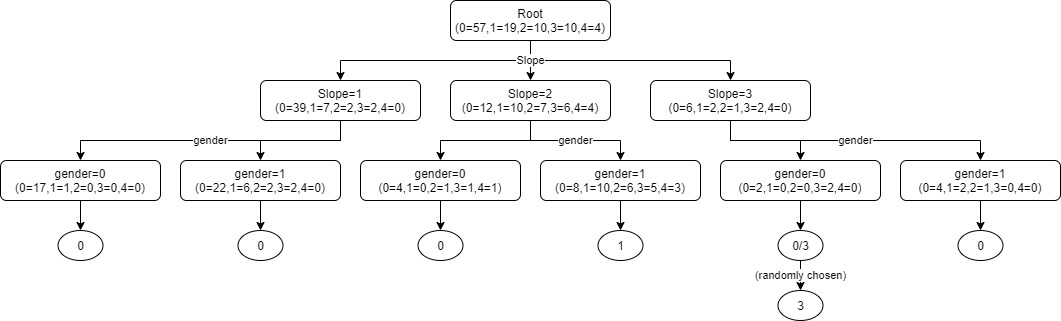
=

1.686 -1.523=**0.163**

**נבחר לפצל לפי המשתנה gender מכיוון שהוא מספק את רווח הידע הגדול ביותר (הקטנת אי הוודאות הגדולה ביותר).**

**בכל הקדקודים ברמה השנייה בחרנו לפצל לפי gender.**

העץ שהתקבל:



נשים לב כי כל רשומה בslope=1 תסווג כprediction=0, ולכן ניתן לייצג את העץ בצורה מינימאלית יותר:

