שאלה 1

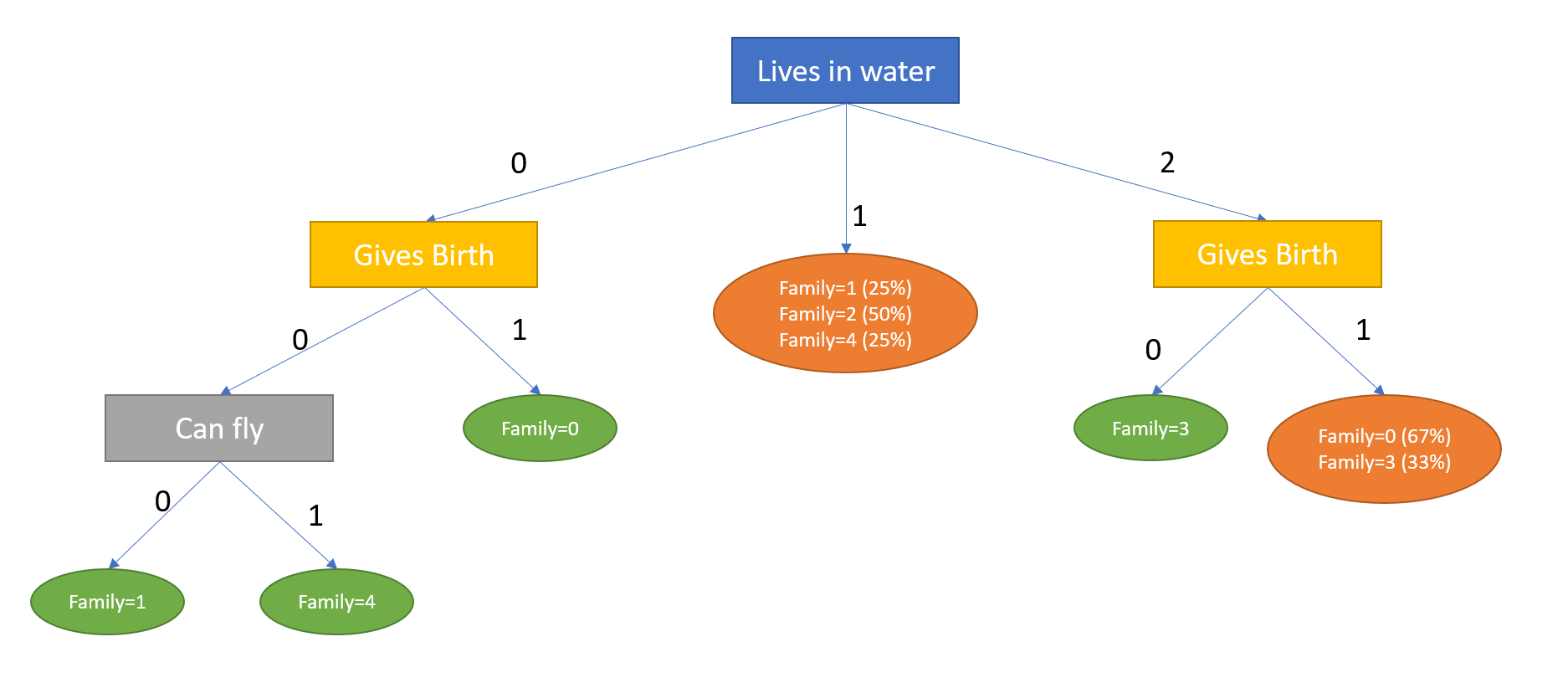
1. **מטרת כריית המידע:**מטרת כריית המידע היא לחזות האם לחולה יש CKD – Chronic Kidney Disease, על בסיס רשומות החולה ומערכת חיזוי המבוססת על 400 רשומות של פציינטים שחלקם (250) חולים והשאר (150) אינם חולים. הרשומות מכילות נתונים כגון גיל, לחץ דם, רמת המוגלובין בדם ונתונים נוספים. את כל הנתונים ופירוט עליהם ניתן לראות בסעיף ב. על מסד הנתונים שקיבלנו יש מספר הנחות יסוד.
   1. הנתונים מדויקים עד כדי תיקון ידני של תאים שזזו או ערכי שהשתנו במקצת (למשל הוספת רווח לערך yes או no).
   2. הנתונים הגיעו בפורמט arff. מכיוון שיותר נוח לעבוד עם פורמט xlsx, יש להניח שהמעבר מarff לcsv והמעבר מcsv לxlsx הוא פשוט ועמיד לטעויות.
   3. רשימת המידע על השדות מהאתר שממנו קיבלנו את מסד הנתונים מעודכנת ונכונה. (האתר מוזכר בסוף המסמך, בביבליוגרפיה).
2. **הנתונים בשימוש:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **שם השדה** | **יחידות מידה** | **סוג הנתונים** | **תחומי ערכים** | **mean** | **median** | **mode** |
| **Age** | years | נומרי | 0-100 | 51.2 | 54 | 55 |
| **Blood pressure** | Mm/Hg | נומרי | 50-180 | 75.8 | 80 | 80 |
| **Specific gravity** |  | נומרי | 1 + 0.05k (0<=k<=5) | 1.02 | 1.02 | 1.02 |
| **Albumin** |  | קטגורי | 0-5 |  |  |  |
| **Sugar** |  | קטגורי | 0-5 |  |  |  |
| **Red blood cells** |  | בינארי | 1 אם תקין |  |  |  |
| **Pus cell** |  | בינארי | 1 אם תקין |  |  |  |
| **Puc cell clumps** |  | בינארי | 1 אם קיים |  |  |  |
| **Bacteria** |  | בינארי | 1 אם קיים |  |  |  |
| **Blood glucose random** | Mgs/dl | נומרי | 70-490 | 140.6 | 117 | 107 |
| **Blood urea** | על עמודה זו נוותר, כי רק 70 מתוך 400 הרשומות (כ17%) הוא בטווח התקין של הערכים (לפי חיפוש בגוגל), ולכן סביר שהערכים המוזנים שגויים. | | | | | |
| **Serum Creatinine** | Mgs/dl | נומרי | 0.4-10 | 1.8 | 1.2 | 1.2 |
| **Sodium** | mEq/L | נומרי | 104-163 | 138.5 | 139 | 135 |
| **Potassium** | mEq/L | נומרי | 2.5-7.6 | 4.34 | 4.4 | 5 |
| **Hemoglobin** | Gms | נומרי | 6-17.8 | 13.2 | 13.7 | 15 |
| **Packed cell volume** | Percentage | נומרי | 0-100 | 40 | 41 | 52 |
| **White blood cell count** | Cells/ cumm | נומרי | 2200-16700 | 8355 | 8100 | 9200 |
| **Red blood cell count** | Millions/ cumm | נומרי | 2.1-6.5 | 4.8 | 4.8 | 4.5 |
| **Hypertension** | על עמודה זו נוותר, כי העמודה blood pressure נותן לנו את אותו המידע באופן יותר מפורט. | | | | | |
| **Diabetes Mellitus** |  | בינארי | 1 אם חולה |  |  |  |
| **Coronary Artery Disease** |  | בינארי | 1 אם חולה |  |  |  |
| **Appetite** |  | בינארי | 1 אם יש |  |  |  |
| **Pedal Edema** |  | בינארי | 1 אם חולה |  |  |  |
| **Anemia** |  | בינארי | 1 אם חולה |  |  |  |
| **Classification** |  | בינארי | 1 אם חולה |  |  |  |

1. **שלבי הKDD:**
   1. איסוף נתונים: נעשה עבורנו (הורדנו את הנתונים מהאתר שבביבליוגרפיה).
   2. ניקוי נתונים:
      1. העברה מפורמט arff לcsv ואז לxlsx.
      2. מעבר ידני על מסד הנתונים: זיהוי בעיות הבולטות בעין (תיקון ההזזה בשורה 371).
      3. הסרת העמודות bu, hypertension.
      4. החלפת ערכים מחוץ לטווח בערך חסר.
      5. הסרת שורות מרובות חוסרים (5 או יותר ערכים חסרים – שורש מכמות השדות הקיימים).
      6. השלמת ערכים מספריים חסרים בהתאם לעמודה (ממוצע).
      7. השלמת ערכים בינאריים חסרים (עצי חיזוי פשוטים).
      8. המרת ערכים קטגוריים לבינאריים (איפה שניתן).
      9. דיסקרטיזציה וניפוי הנתונים לערכים שמישים בעזרת שיטות כמו binning – **נדחה לעת חישוב העץ.**
   3. בחירת שיטות לכריית מידע:
      1. הכנה של מסד הנתונים: excel.
      2. ניקוי הנתונים קוד פייתון בפיתוח שלי בשימוש בספריית pandas.
      3. חישוב עצי החלטה בשימוש בספריית sklearn.
      4. הצגת עצי ההחלטה בשימוש בספריית graphviz.
   4. הרצת כריית המידע
      1. שימוש בקוד פייתון מוכן מספריות שהוזכרו.
      2. עטיפה בקוד משלי להתאמות לצרכי המטלה.
   5. ניתוח התוצאות
      1. ניתוח סטטיסטי מבוסס על תוצאות הריצה.
      2. ניתוח סטטיסטי של "כלל הרוב".
      3. החלטה האם כריית המידע מוצלחת.
2. חלופות לכריית המידע:
   1. מיפוי אלגוריתמים שנלמדו:
      1. עץ החלטה ID3 – Information Gain
      2. עץ החלטה C4.5 – Gain Ratio
      3. עץ החלטה CART – Gini
      4. האצה אדפטיבית AdaBoost
      5. יער אקראי Random Forest
   2. החלטה בהקשר הפרוייקט:
      1. מכיוון שכל המסווגים הקטגוריים שלנו הם בינאריים, אין טעם ממשי להשתמש בGain Ratio על פני Information Gain, ולכן ויתרתי על C4.5.
      2. את שאר המודלים מימשתי והשוויתי בקוד, מכיוון שלכל אחד מהם יש יתרונות וחסרונות שעלולים להיות מועילים, ולכן רציתי להשוות ביניהם.
      3. בחרתי לבסוף ביער אקראי והאצה אדפטיבית, כי אפשר היה לוותר על ID3 מכיוון שהוא "מוכל" בתוך היער האקראי, ועל CART מכיוון שהוא "מוכל" בהאצה אדפטיבית. המימוש של העצים האחרים נשאר בקוד.

שאלה 2

1. בניית עץ ההחלטה:  
   לצורך ביצוע המטלה, כתבתי קוד פייתון המבצע את החישובים הדרושים בעבור ID3, ומייצר את העץ. אציג את החישובים של השכבה הראשונה, והשאר אסתמך על הקוד המייצר. כמובן שהקוד מצורף גם הוא כחלק מהzip של המטלה.  
   שכבה ראשונה:
   1. **Gives birth**: ,
      1. Value=0: ,
      2. Value=1: ,
   2. **Can fly**: ,
      1. Value=0: ,
      2. Value=1: ,
   3. **Lives in water**: ,
      1. Value=0: ,
      2. Value=1: ,
      3. Value=2: ,
   4. **Has legs**: ,
      1. Value=0: ,
      2. Value=1: ,

כיוון שלעמודה "Lives in water" יש את הGain הכי גבוה, נבחר בעמודה זו להיות הבוררת הראשונה בעץ. העץ שמתקבל בסוף האלגוריתם:  


העיגולים המסומנים בכתום הם מקרים שבהם כל השורות שנותרו בטבלה בעלות אותם ערכים (פרט למשפחה), ולכן לא ניתן להבדיל בינהם עם הנתונים הקיימים.

1. ניתן לוותר על העמודות הבאות:
   1. "Lays Eggs" – בנתונים שקיבלנו יש יחס שלילה בין לידה ובין הטלת ביצים, ולכן אין טעם לשמור את שתי העמודות. מה גם שעובדה זו נכונה גם בטבע בעולם האמיתי.
   2. מספר או שם – אין צורך לשני מזהים ייחודיים, מספיק לבחור באיזה זיהוי נעדיף להשתמש ולוותר על העמודה שבה לא בחרנו.
   3. "Has Legs" – בעץ ההחלטה שקיבלנו אין שימוש בHas Legs, ולכן אפשר לומר כי הוא נתון מיותר ואפשר למחוק את העמודה (למרות שעם עוד נתונים יכול להיות שנתון זה היה משמעותי בשביל להבדיל באחד המקומות שבהם אנחנו לא יודעים להבדיל עם הנתונים.