ממ"ן 21 כריית מידע

שם: נועם שדה

תאריך: 5.5.2023

שאלה 1 – הגדרת הבעיה והכנת הנתונים

א. מטרת כריית המידע היא חיזוי מחלת כליית כרונית מתוך סט תכונות הנתונות לנו בקובץ Excel, כאשר עמודת class היא העמודה עליה נבצע חיזוי.

ב. נגדיר את הנתונים בהם נשתמש בפרויקט מבחינת תכונות, סוג נתונים, נתונים חסרים, תחומי ערכים וכו'..
כדי לעשות זאת, ראשית נבין את עמודות הנתונים בעזרת המאמר הנתון ובנתונים בקובץ Excel:

תכונה	תיאור	סוג נתונים	תחומי ערכים	ערך שכיח/ממוצע	ערכים חסרים
age	גיל	numeric	2-90	51.48	9
bp	לחץ דם	numeric	50-180	76.46	12
sg	משקל	numeric	1.005-1.025	1.01	47
al	משקל סגולי	numeric	0-5	1.01	46
su	אלבומין	category	0-5	0	49
rbc	סוכר	category	normal/abnormal	normal	152
pc	תאי דם אדומים	category	normal/abnormal	normal	65
pcc	גושי תאי מוגלה	category	present/notpresent	notpresent	4
ba	בקטריות	category	present/notpresent	notpresent	4
bgr	סוכר בדם	numeric	22-490	148.03	44
bu	דם בשתן	numeric	1.5-391	57.42	19
sc	קריאטינין	numeric	0.4-76	3.07	17
sod	סודיום	numeric	4.5-163	137.52	87
pot	אשלגן	numeric	2.5-47	4.62	88
hemo	המוגלובין	numeric	3.1-17.8	12.52	52
pcv	המטוקריט	numeric	9-54	38.88	71
wbcc	ספירת תאי דם לבנים	numeric	2200-26400	8406.12	106
rbcc	ספירת תאי דם אדומים	numeric	2.1-8	4.7	131
htn	לחץ דם גבוה	categorical	yes/no	no	2
dm	סוכרת	categorical	yes/no	no	2
cad	מחלת לב איסכמית	categorical	yes/no	no	2
appet	תיאבון	categorical	good/poor	good	1
pe	בצקת	categorical	yes/no	no	1
ane	אנמיה	categorical	yes/no	ane	1

ג. **שלב ה-KDD**

1. מטרת כריית המידע: חיזוי מחלה כליית כרונית

2. איסוף ושמירת הנתונים: הנתונים הגיעו מהקובץ שנשלח באתר הקורס מסוג arff. המרתי אותו לקובץ csv כדי לבצע עיבוד מוקדם על סט הנתונים בעזרת שפת התכנות פייתון.

3. ניקוי הנתונים ועיבוד מוקדם: בעזרת הספרייה pandas מצאתי כי ישנם ערכים חסרים וחריגים ולכן הייתי צריך לשנות אותם (יפורט בהמשך כיצד).

4. טרנספורמציות ורדוקציה על הנתונים: המרתי את העמודות מסוג "קטגוריה" ל – 0 ו-1 בעזרת המתודה LabelEncoder.
שמתי לב שמספר המופעים בעמודת המטרה – nockd נמוך משמעותית לעומת ckd, לכן השתמשתי בשיטת SMOTE על מנת לאזן את סט הנתונים.
בחרתי לא לנרמל את הנתונים ולחלק לבינים מכיוון שקיבלתי תוצאות טובות מאוד גם בלי שיטות אלו.

5. בחירת האמצעים לתהליך כריית המידע: השתמשתי בשפת התכנות פייתון ובספריות pandas, scipy, seaborn, pyplot על מנת לנתח את הנתונים. השתמשתי בספריית sklearn ובתהליך החיזוי ובחנתי מודלים שונים על מנת לחזות את עמודת המטרה.

6. ניתוח התוצאות: בדיקת הממצאים לפי המודלים שיצרתי וביצוע הערכה לפי מדדים כמו דיוק, רלוונטיות, פשטות וכו' ע"י נתונים סטטיסטים.

7. הסקת מסקנות: השתמשתי במודלים שיצרתי כדי לחזות את עמודת המטרה. לאחר מכן ניתן להציג את המודל בכל מיני דרכים : צורה ויזואלית, נוסחה מתמטית, כללי היסק או עץ החלטה, תלוי בשיטה שנבחרה.

ד. סקירה השוואתית בין החלופות האפשריות לביצוע כריית המידע

- עץ החלטה מבוסס אלגוריתם ID3 עם מדד Information Gain:
אלגוריתם חמדן ורקורסיבי שיודע לבצע חיזוי או סיווג לערכים בדידים ויחזיר עץ לא בהכרח בינארי. אלגוריתם זה דורש דיסקרטיזציה לכל הערכים הרציפים
יתרונות – בניית העץ מתבצעת בצורה פשוטה, קלה ויעילה.
חסרונות - מדד הרווח האינפורמטיבי עלול לגרום למצב של overfitting לנתוני האימון ולפגוע ברמת הדיוק של המודל.

- עץ החלטה C4.5 מבוסס Gain Ratio:
אלגוריתם שמנסה לשפר את השיטה ב- ID3 ומנסה להתגבר עם הבעיות של מדד הרווח האינפורמטיבי שהיא הטיה לבחירת התכונות בעלות ערכים רבים. לכן, אלגוריתם זה מבצע שינוי ומגדיר קריטריון חדש לפיצול והוא יחס הרווח האינפורמטיבי – gain ratio.
יתרונות– נותן תוצאות טובות יותר עבור נתונים שמכילים תכונות מרובות ערכים או תכונות שמחולקות למקטעי טווח רציפים.
חסרונות – הפיצול יביא לחלוקה פחות מאוזנת עם הרבה ענפים שקטנים ביחס לעץ.

- עץ החלטה CART מבוסס Gini Index:
אלגוריתם המייצר עץ החלטה בינארי ויודע לעבוד עם ערכים גם עם ערכים דיסקרטיים וגם עם ערכים רציפים, באמצעות שימוש במדד Gini Index.
אלגוריתם זה מעדיף פיצול לפי שתי קבוצות שיהיו יחסית שוות בגודלן ולאחר בניית העץ המלא האלגוריתם יוצר קבוצה של תתי עצים גזומים ובוחר בעץ עם פונקציית הסיבוכיות המינימלית על מנת לצמצם overfit.
יתרונות – האלגוריתם נוטה לייצר עץ קומפקטי יותר ביחס לאלגוריתמים אחרים.
חסרונות – זמן הריצה של האלגוריתם ארוך יותר ביחס לעצים האחרים.

- יער אקראי
אלגוריתם המשלב מספר עצי החלטה כאשר כל עץ תלוי בערכים של וקטור תכונות אקראי. החיזוי/סיווג נקבע על ידי רוב ההצבעות של סט העצים שממנו הוא מורכב.

יתרונות:
  - דיוק גבוה מכיוון שהוא משלב מספר של עצי החלטה
  - עמיד לרעש ולנתונים חריגים
  - יכול להתמודד עם מספר רב של פרמטרים

חסרונות:
  - לא קל להבנה ולסרטוט מפני שהוא מורכב מהרבה עצי החלטה
  - זמן הריצה של האלגוריתם ארוך יותר ביחס למודלים אחרים

- רגרסיה לינארית:
שיטה שמתאימה לחיזוי ערכים רציפים בהם קיימת קורלציה בין משתנה המטרה לאחד או יותר מהתכונות האחרות
יתרונות – קל לביצוע ומימוש ואף להבנה.
חסרונות –
  - היא לא מתאימה לחיזוי פרמטרים מורכבים
  - השיטה מתאימה רק למקרים בהם קיים קשר לינארי בין התכונות למשתנה המטרה.

ה. **<u>שלבי הכנת הנתונים</u>**

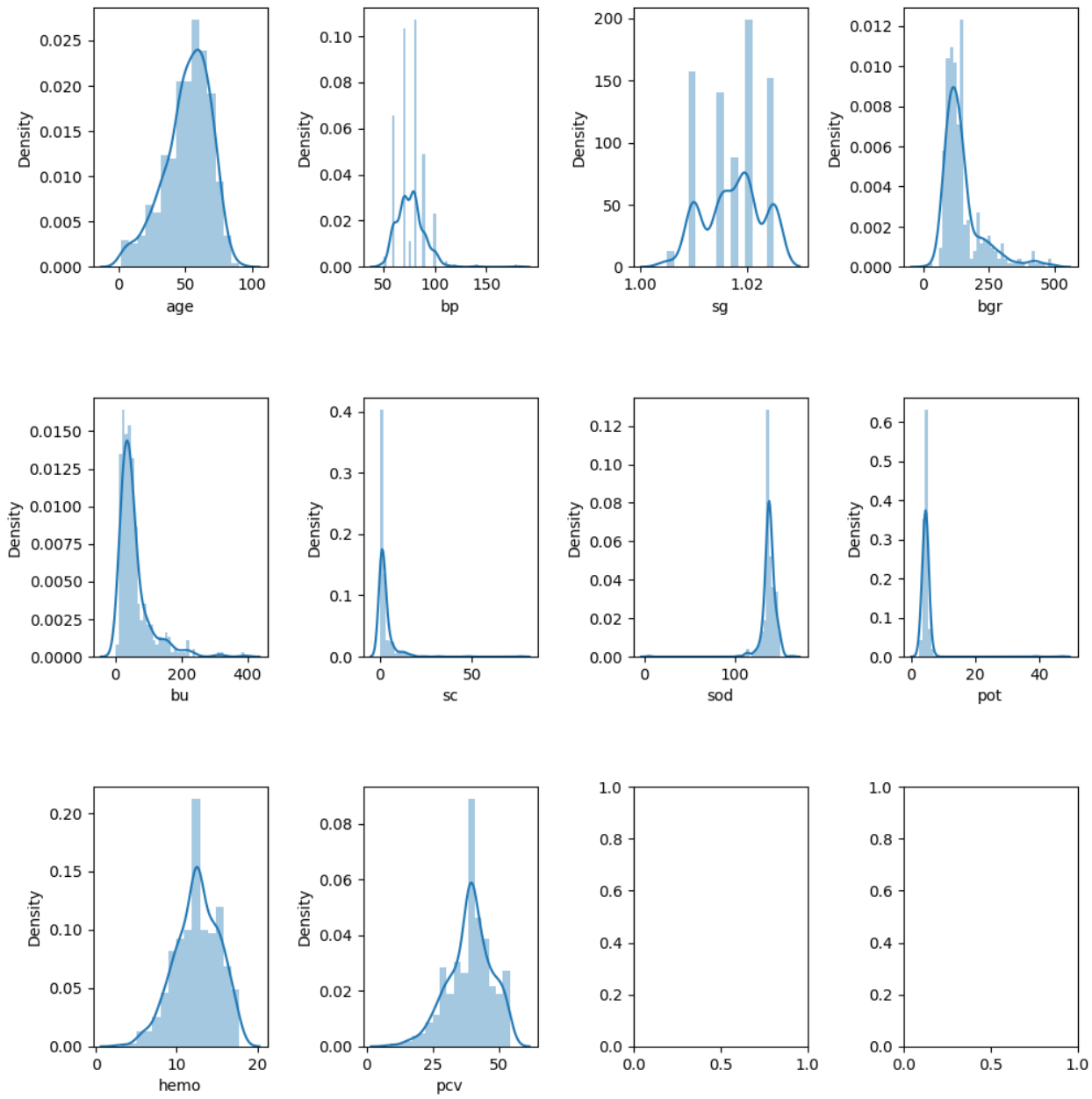
נתייחס לבעיות באיכות הנתונים כמו טיפול בערכים חסרים, ערכים חריגים, תצוגה גרפית של הנתונים, ניקוי הנתונים, שילוב והמרה של נתונים ועוד.

לסט הנתונים שקיבלנו יש הרבה נתונים חסרים, בחרתי למחוק עמודות בהם יש מעל 25% נתונים חסרים (עמודות rbc,wbcc,rbc). בחרתי לטפל בנתונים חסרים בשאר העמודות באופן הבא:

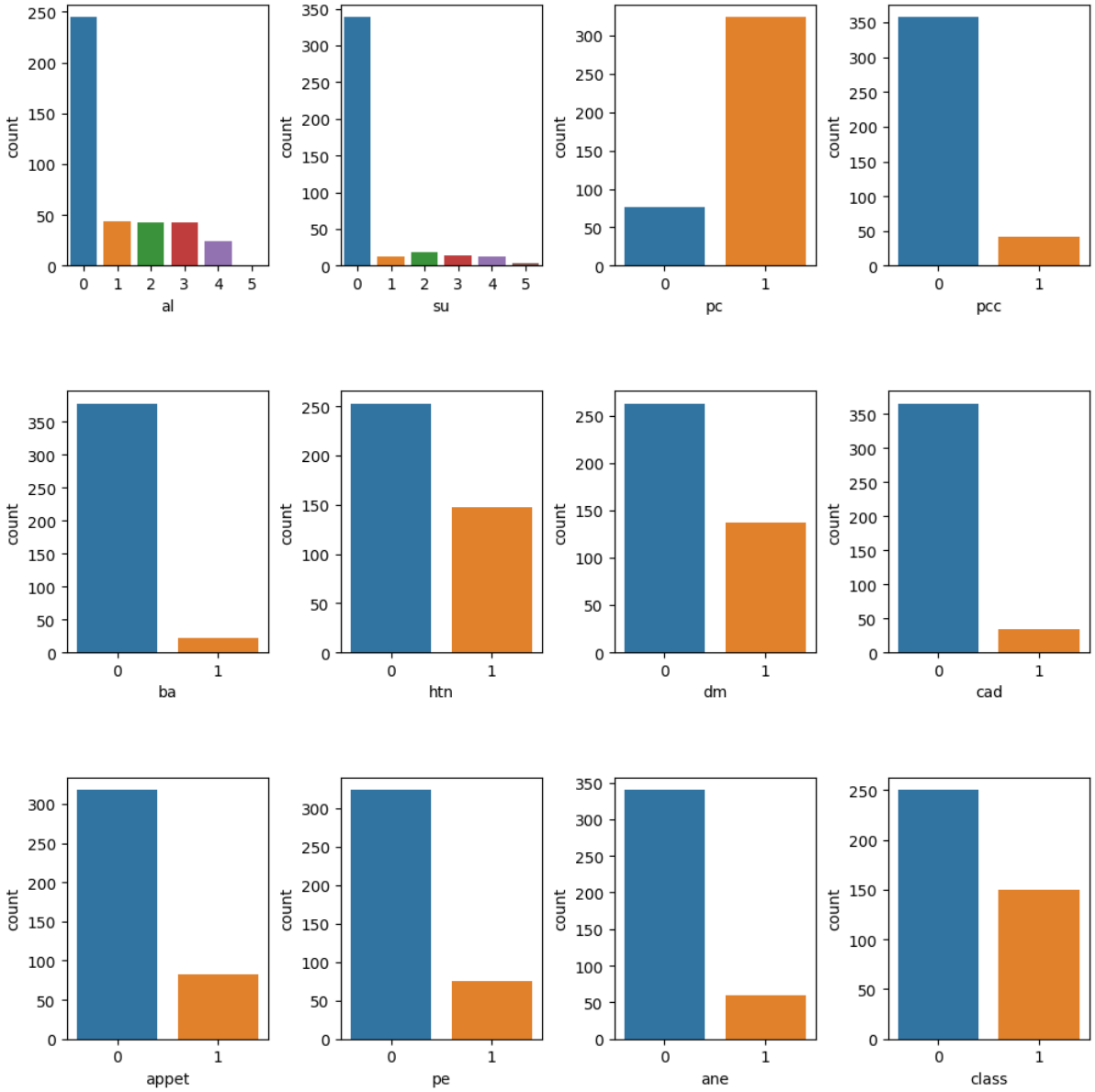
- במידה והתכונה היא קטגורית, נחליף ערכים חסרים בערך השכיח ביותר
- במידה והתכונה היא נומרית, נחליף ערכים חסרים בממוצע של המספרים של התכונה.

פירטתי בשלב ה-KDD על הטרנספורמציות שעשית לסט הנתונים.

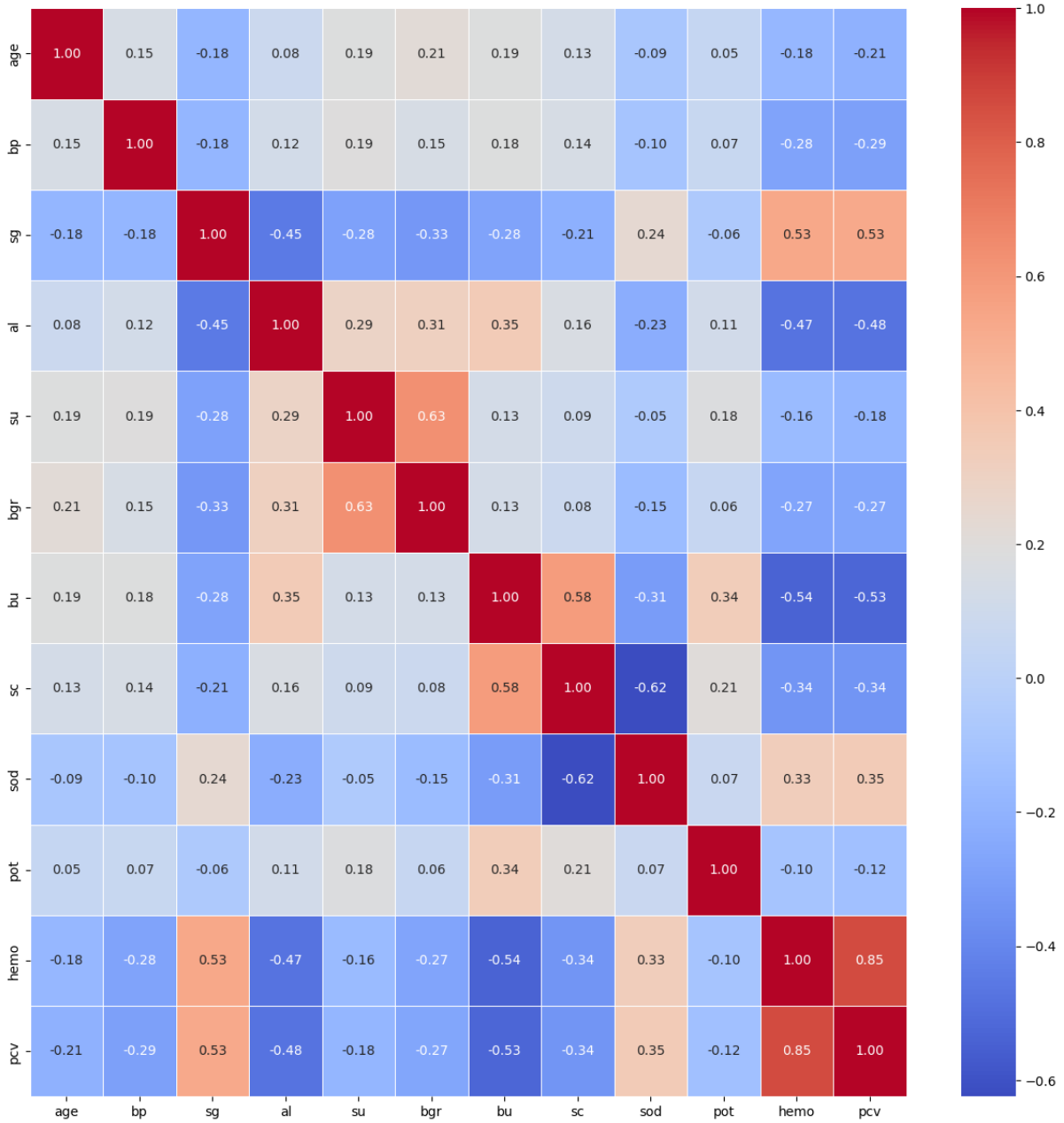
הצגת הנתונים הנומריים בעזרת היסטוגרמה:



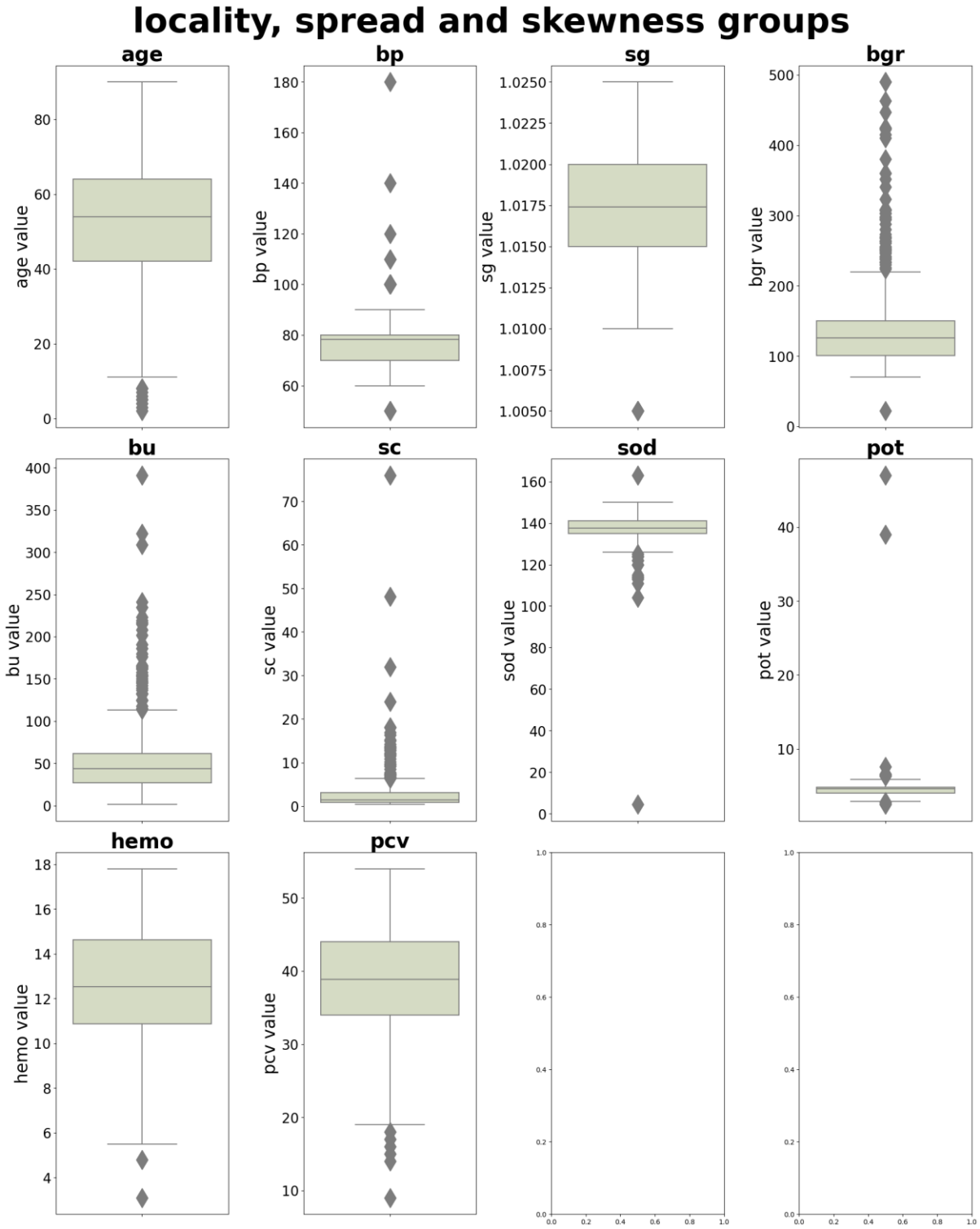
הצגת הנתונים הקטגורים באמצעות היסטוגרמה:



מפת הקורלציה:



ערכים חריגים:



נבחין כי ישנם עמודות בהם יש ערכים חריגים, לדוגמא בעמודה "bu" רוב הערכים נעים בטווח 40-60 לעומת 3 ערכים שנמצאים בטווח הערכים 300-391. בכל מקרה בחרתי לא למחוק ערכים חריגים אלה מכיוון שסט הנתונים שלנו יחסית קטן (400 שורות, וכשמפצלים לסט אימון וסט ולדיציה זה יהיה אף יותר קטן) וגם כי בעזרת ניסוי וטעיה מצאתי כי מחיקת הנתונים החריגים לא תרמו לשיפור המודל.

שאלה 2 – סיווג

א. נבחר 2 שיטות לסיווג הנתונים:

1. עץ החלטה CART המבוסס על GINI Index – בחרתי שיטה זו כי קל להבין אותה (היא יוצרת עץ מצומצם), היא טובה לסיווג נתונים קטגורים, ואף יכולה למנוע overfit.
2. יער אקראי – בחרתי שיטה זו בגלל המאמר שקיבלנו עם העבודה, שאיתו החוקרים הגיעו לביצועים של 99% דיוק בסיווג עמודת המטרה. בנוסף הוא עמיד לרעשים ונתונים חריגים.

פירטתי על שיטות אלה בסעיף ד'.

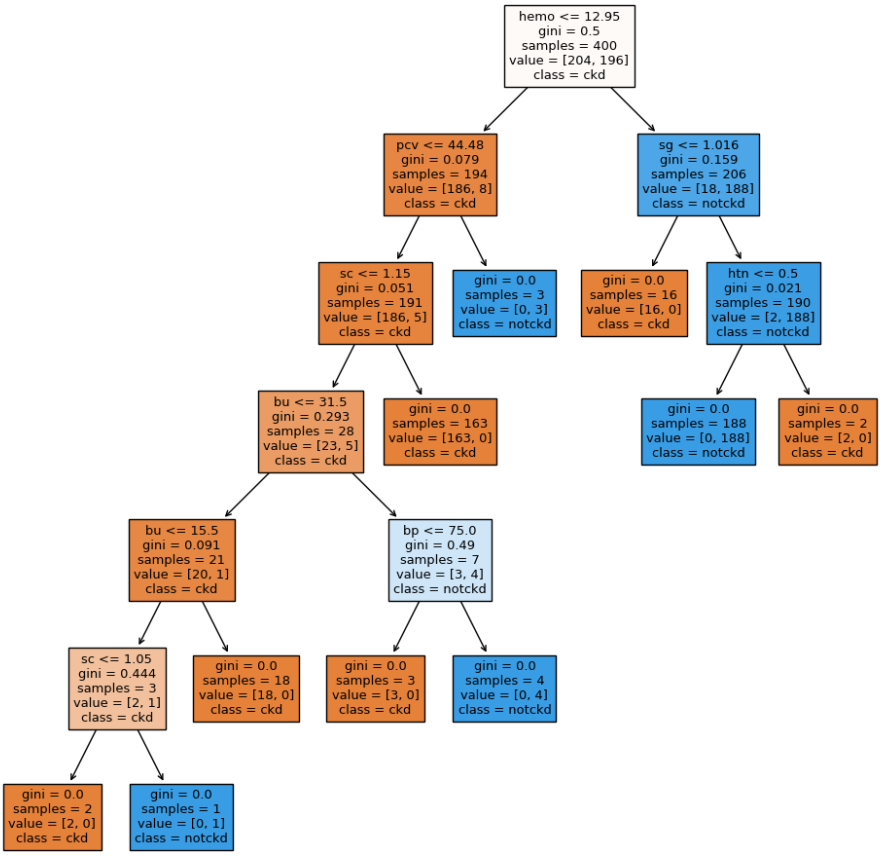
ב. נתאר את שלבי השיטות שבחרתי בסעיף א':

פיצלתי את סט הנתונים ל- 80% אימון ו-20% ולידציה באופן אקראי, ואיזנתי את סט נתונים כך שעמודת המטרה תהיה מחולקת לשתי תכונות באופן שווה. לאחר מכן הרצתי את השיטות הבאות, שעליהן פירטתי בסעיף ד':

1. עץ ההחלטה CART המבוסס על Gini Index שממומש על ידי הספרייה sklearn.
2. יער אקראי שגם הוא ממומש על ידי הספרייה sklearn.

ג. תוצאות הניתוחים של השיטות:

1. קיבלנו 10 עלים באלגוריתם זה, ו -9 צמתים. דיאגרמת עץ ההחלטה שהתקבל:

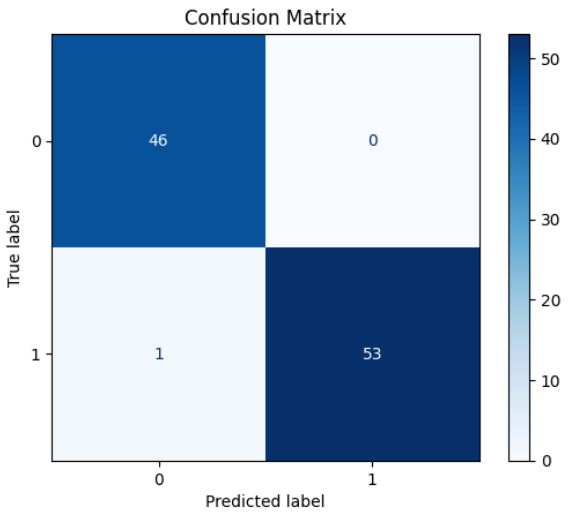


2. בשיטת יער אקראי ישנם הרבה עצים ולכן אין טעם להציג את כולם, נציג בסעיף הבא רק הערכת השיטה.

ד. <u>הערכת מידת הדיוק של השיטות</u>

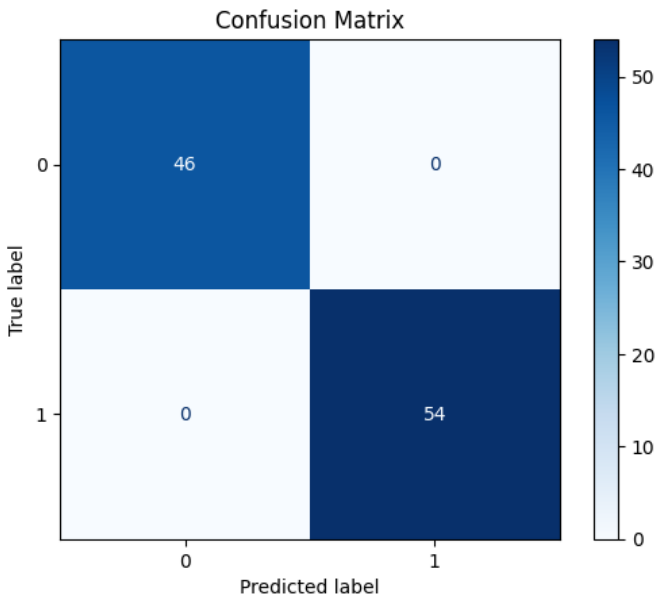
1. עץ ההחלטה CART המבוסס על Gini Index:

```
Accuracy: 0.99
Precision: 1.0
Recall: 0.9814814814814815
F1 score: 0.9906542056074767
ROC-AUC score: 0.9907407407407407
```



2. יער אקראי:

```
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1 score: 1.0
ROC-AUC score: 1.0
```



ה. <u>הסקת מסקנות:</u>

בעזרת המודל יער אקראי קיבלנו 100% דיוק, הן מבחינת accuracy והן מבחינת AUC score ולכן ברור שהוא הכי טוב. המודל השני (עץ החלטה) גם הוא מאוד טוב עם דיוק של 99%.

סט הנתונים שקיבלנו הינו סט קטן והמודלים שבחרתי עובדים באופן כללי על סטטיסטיקה, לכן דיוק השיטות הנ"ל לא בהכרח מדויקות כיוון שסטטיסטיקה עובדת על מספרים גדולים.
ניתוח וחיזוי מידע רפואי ומחלות הינו נושא רגיש ואין בו מקום לטעויות, לכן יש צורך לחקור לעומק עוד יותר כל תכונה ולהבין איך היא משפיעה על חיזוי המחלה.