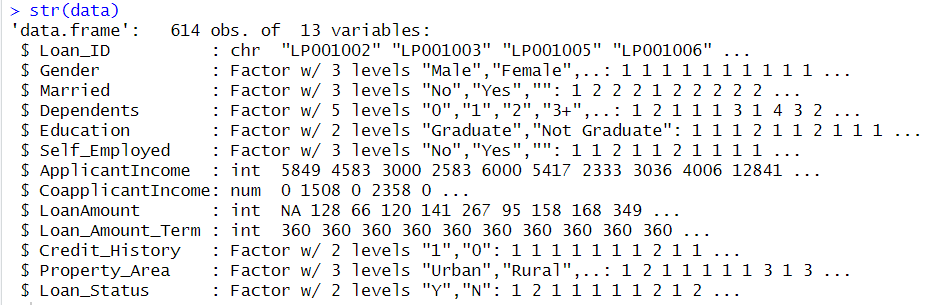
דו"ח עבודה 2- מדעי הנתונים ובינה עסקית

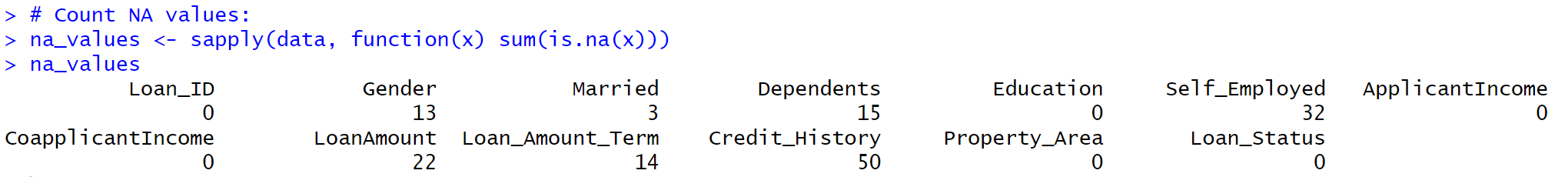
1.1

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| משתנה | Loan\_ID | Gender | Married | Dependents | Education | Self\_Employed | ApplicantIncome |
| טיפוס נתון | Chr, string | Chr,  String | Chr, string | Chr, string | Chr, string | Chr, sting | Int |
| טיפוס שיש לקבוע | Chr, String (Id) | Categorial | Categorial | Numeric | Categorial | Categorial | Numeric |
| הערות | זיהוי | Converted to factor | Converted to factor | Converted to factor | Converted to factor | Converted to factor |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| משתנה | CoapplicantIncome | LoanAmount | Loan\_Amount\_Term | Credit\_History | Property\_Area | Loan\_Status |
| טיפוס נתון | Num | Int | Chr, string | Int | Chr, sting | Chr |
| טיפוס שיש לקבוע | Numeric | Numeric | Categorial | Categorial | Categorial | Classifier |
| הערות |  |  | Converted to factor | Converted to factor | Converted to factor | משתנה החלטה |



כמות ערכי NA לכל עמודה (לאחר המרת ערכי מחרוזת ריקים לNA):



1.2

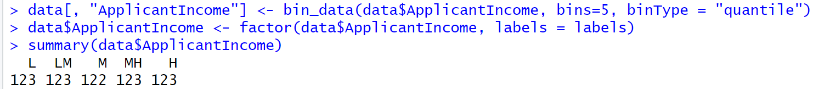
הערכים שהזנו עבור כל אחד מהעמודות:

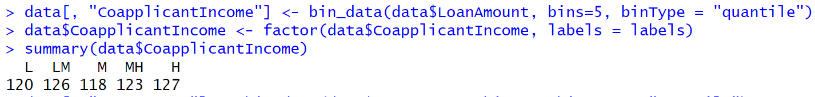
|  |  |  |
| --- | --- | --- |
| משתנה | סוג טיפוס | ערך עבור שדות NA |
| ApplicantIncome | נומרי | 5403.459 |
| CoapplicantIncome | נומרי | 1621.246 |
| LoanAmount | נומרי | 146.4122 |
| Loan\_amount\_term | קטגוריאלי | 360 |
| Gender | קטגוריאלי | Male |
| Married | קטגוריאלי | Yes |
| Dependents | קטגוריאלי | 0 |
| Education | קטגוריאלי | Graduate |
| Self\_Employed | קטגוריאלי | No |
| Credit\_History | קטגוריאלי | 1 |
| Property\_Area | קטגוריאלי | Semiurban |

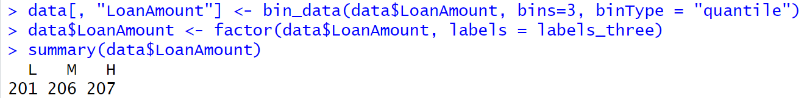
1.3

חילקנו את המידע בכל אחד מהמשתנים הנומריים לטווחים, **כאשר בכל טווח כמות אברים שווה**.  
מספר הטווחים השתנה לפי התבוננות שלנו במידע.  
את ערכי המשתנה הנומרי ApplicantIncome, חילקנו את המידע ל5 טווחים,  
את ערכי המשתנה הנומרי CoapplicantIncome, חילקנו גם כן ל5 טווחים,  
את ערכי המשתנה הנומרי LoanAmount, חילקנו ל3 טווחים.

הגדרנו Labels מתאימים עבור 5 טווחים ועבור 3:

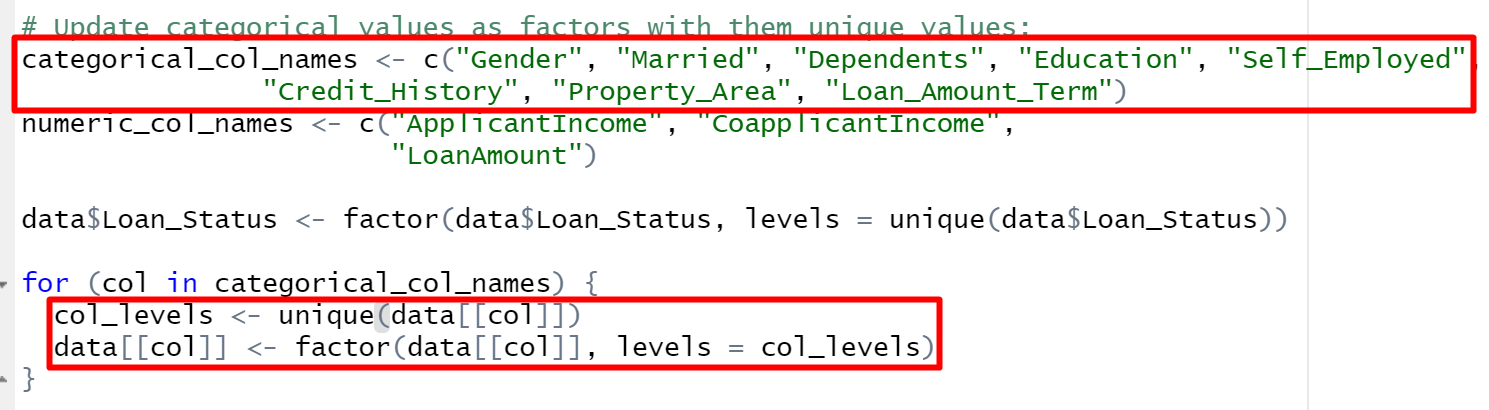
חלוקה לטווחים כאשר בכל טווח כמות שווה של ערכים (עם שונות מסוימת כמובן):





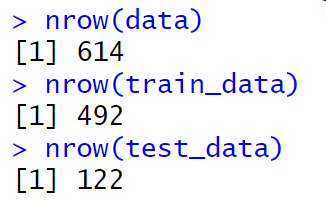
1.4

כל משתנה שהגדרנו כקטגוריאלי, הוגדר כמשתנה Factor:



הסיבה זהה לכל המשתנים הנ"ל,   
כל משתנה בעל הבחנה קטגורית הוגדר ככזה ובנוסף הוגדר כמשתנה Factor (ע"פ הנלמד במעבדה).

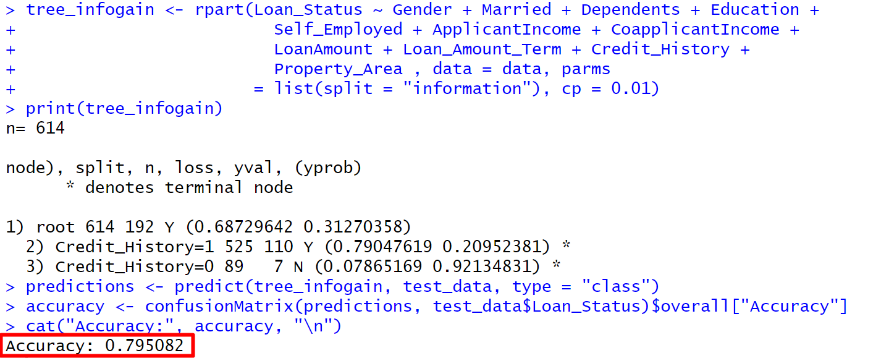
1.5

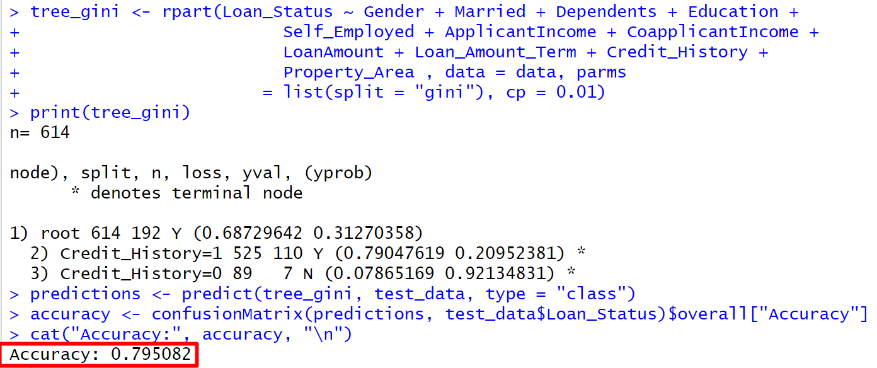
כמות הרשומות באימון ובבדיקה:

2.1

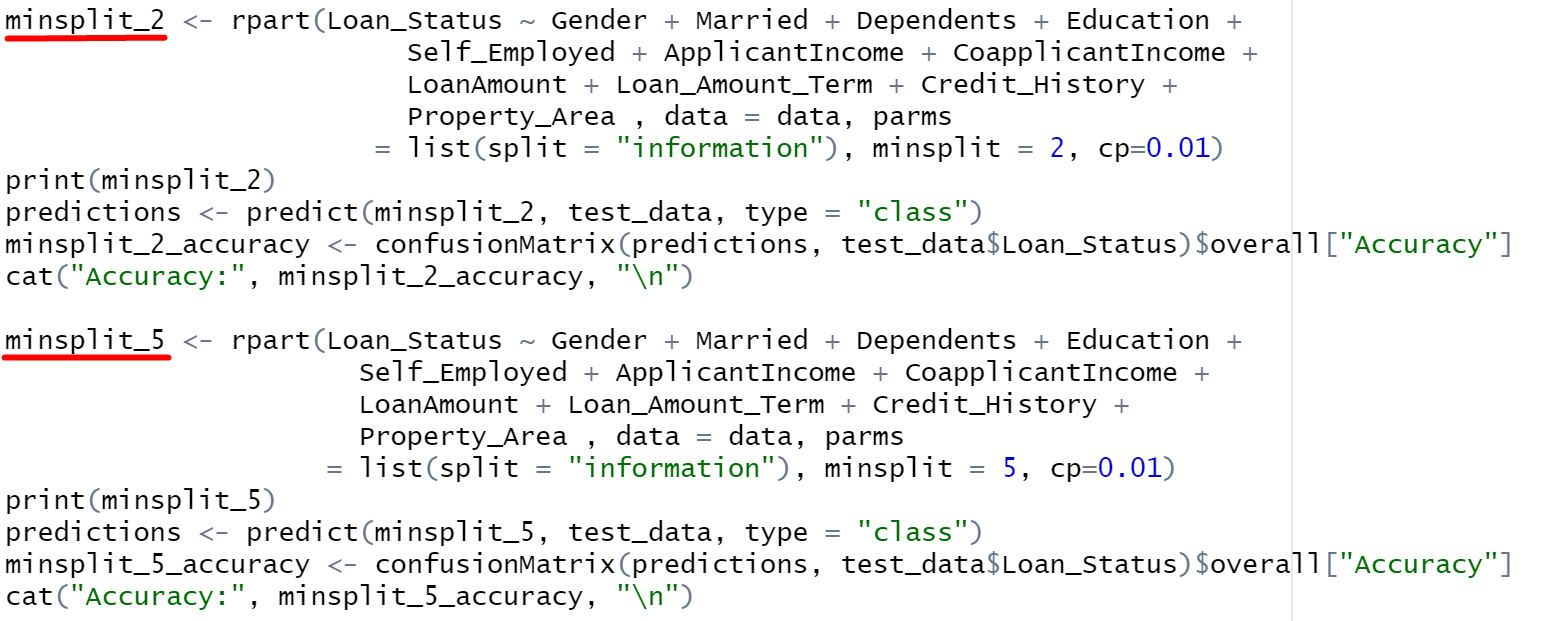
הגדרנו שני סוגי עצים, לראשון קריטיון פיצול Information Gain ולשני קריטריון פיצול Gini.   
בשני העצים הגדרנו ערךCP (Complexity Parameter) נמוך, על מנת לשלוט בסיבוכיות העץ (נדרש בשאלה).  
משתנה המטרה הוגדר להיות Loan\_Status ואילו משתני הקלט הם כלל המשתנים הרלוונטיים (ללא Loan identifier).

עבור שני העצים הדפסנו את רמת הדיוק שלהם, שיצאה זהה למרות השוני בערכי קריטריון הפיצול:





מבדיקה אמפירית שביצענו, אין הבדל משמעותי בשינוי הגדרת ערך הminsplit לערכים שבחרנו (2, 5)



2.2