סיכום התוצאות

הפרוייקט עסק בתהליכי כריית מידע על ידי שיטות שונות שנתנו זויות ראייה שונות. הוא התמקד בבסיס נתונים שקיבלנו שעיקרו איכות של יין.

הפרוייקט חולק ל-2:

ממ"ן 21 - שבו עסקנו בטיוב הנתונים והכנת המידע עבור האלגוריתמים השונים, בנוסף עבדנו בממ"ן על שיטות לחיזוי איכות היין בעזרת אלגוריתמים כגון עצי החלטה ורגרסיה לוגיסטית.

ממ"ן 22 – השתמשנו בנתונים המטוייבים מממ"ן 21 על מנת לנסות להסיק מידע נוסף על בסיס הנתונים על ידי חוקי הקשר וחלוקת הנתונים לאשכולות.

במודל החיזוי השתמשנו בשני שיטות שונות לבדיקת איכות התוצאות: ACCURACY ו- MOST COMMON. בשניהם DECISION TREE C4.5 נתן תוצאה טובה יותר מאשר LOGISTIC REGRESSION.

על סמך זה ניתן להגיד כי עדיף להשתמש בעץ ההחלטה על מנת לבצע חיזוי. יתרון נוסף לעץ ההחלטה הוא שניתן להבין בקלות יותר את הסיבה לסיווג.

בממן 22 עסקנו בקבלת תובנות נוספות על הדאטא ולא בחיזוי ישירות, השתמשנו בחוקי הקשר על ידי 2 אלגוריתמים שונים, FP\_Growth ו- Apriory.ראשית המרנו את הדאטא לאורדינאלי על מנת להתאים אותו לאלגוריתמים, המידע חולק ל-3 או 4 רמות בהתאם לאופי ההתפלגות שלו.

קיבלנו תוצאות מעניינות על הדאטא למשל הקשרים:

'ResidualSugar\_bins\_High'}) --> {'Density\_bins\_High'}

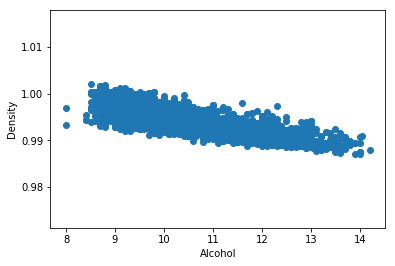
'Alcohol\_bins\_High'} --> {'Density\_bins\_Low'}}

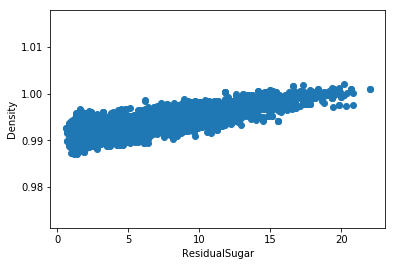
וגם קשרים מורכבים יותר

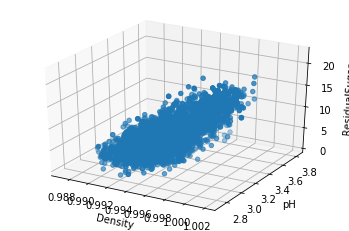
למשל האלגורית הסיק קשר נסיבתי בין עוצמה חזקה של היין וחומציות נמוכה לסוכר עודף גבוה.

|  | **antecedent** | **consequent** | **confidence** | **lift** |
| --- | --- | --- | --- | --- |
| **8** | [Density\_bins\_High, pH\_bins\_Low] | [ResidualSugar\_bins\_High] | 0.816038 | 3.286413 |
| **14** | [ResidualSugar\_bins\_LowMid, Density\_bins\_Low] | [Alcohol\_bins\_High] | 0.712644 | 3.011206 |

ניתן לראות קשרים אלו גם בעזרת דיאגרמת הפיזור:







בנוגע לחלוקה לאשכולות, האלגוריתם DBSCAN הפיק תוצאות מעניינות יותר בזיהוי חריגים על ידי חלוקה להרבה אשכולות שיכולים להביא עם מחקר נוסף אפשרות לזהות אזורים ספציפיים של יין באיכות גבוהה או נמוכה :

