**מטלה 3 – בינה עסקית:**

אביב רביד 316018175

נועה דלאל 316045335

עדן בן דוד 206957797

**חלק 1 – STTM:**

1. קובץ STTM מצורף בתיקיית Analysis
2. סכימה ויזואלית המציגה את הSTTM:

Diagram

Description automatically generated

**חלק 2 – Data Mining Techniques:**

1. תהליך ה-KDD שביצענו כלל בחירת DATA, לאחר מכן עיבוד הנתונים והכנת הטבלאות, ביצוע תהליך ETL וכעת תהליך כריית הנתונים. עבור שאלת המחקר המונחית שלנו נשתמש בשיטת Predictive משום שנרצה לסווג דירות עם פרמטרים מסוימים להיות מעל ומתחת לממוצע.
2. טכניקות למימוש:

1. טכניקה ראשונה אשר תבוצע על נתונים מספריים היא מעבר של הנתונים מקנה מידה של Square Feet

ל – Square Meters.0.0929=1, ולכן עבור דירה של לדוגמה 1180 נקבל 109.426.

2. טכניקה נוספת היא עבור חישוב מדדי ה – KPI, לדוגמה עבור מדד שטח מגורים ממוצע, אשר ערכים מהווים נתונים מספרים לא בהכרח שלמים, אנחנו נבצע סכימה וחלוקה בכמות הדירות ב - DB. לדוגמה עבור 3 דירות בשטחים (במטרים) של 100,90,120 נקבל מטר.

מבחינת תרחשי use-case, ובעבור ממוצע דירות של כ – 125 מטר מרובע ובמקביל המחיר הממוצע הוא 488,613 דולר נבדוק ניבוי עבור 2 אפשרויות:

1. עבור שורת הנתונים הבאה: נרצה לחזות האם המחיר יהיה מעל ההמוצע של מחירי הדירות

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| גודל | עיר | חדרי רחצה | חדרי שינה | שנת חידוש |
| 135 | Woodinville | 2.25 | 3 | 2014 |

עבור נתונים אלו נרצה לקבל חיזוי שאכן המחיר הוא מעל ממוצע המחירים.

1. עבור שורת הנתונים הבאה:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| גודל | עיר | חדרי רחצה | חדרי שינה | שנת חידוש |
| 138 | Renton | 1.75 | 3 | 2000 |

עבור נתונים אלו נרצה לקבל חיזוי שהמחיר מתחת לממוצע.

1. מדד הדמיון שנציע לפרויקט הינו מדד דמיון בין דירות. התכונות אותן נשווה בין כל שתי דירות יהיו התכונות המרכזיות שעונות על שאלת המחקר שבחנו, אשר הן נמצאות בטבלת Fact. בחרנו להשתמש בכל סוגי המדדים שראינו בכיתה, וליצור מדד משוקלל.

* נשתמש במדד Sorensen-Dice שהינו מדד דמיון בין משתנים נומריים בדידים עבור תכונת מספר חדרים (עמודת bedrooms), ותכונת מספר שירותים (עמודת bathrooms).
* נשתמש במדד Jaccard שהינו מדד דמיון בין משתנים קטגוריאליים עבור העיר שבה נמצאת כל דירה (עמודת city).
* נשתמש במדד Euclidean Distance שהינו מדד דמיון בין משתנים נומריים רציפים עבור גודל הדירה במטרים רבועים (עמודת sqm\_living), ועבור המחיר של הדירה (עמודת price).

נאחד את המדדים לכדי מדד אחד, נמצא את השקלול של כל המדדים, וכך נמצא את הדירות שלהן יש מדד דמיון הכי גדול ונסיק כי הן הדומות ביותר.

1. עבור שאלה עסקית 1: **Supervised**

האם דירות בעלות 3 חדרים, אשר שטח המגורים שלהן הינו מעל הממוצע של שטחי המגורים, יעלו מעל הממוצע של מחירי הדירות?

השערת H0: דירות בעלות 3 חדרים ששטח המגורים שלהן גדול מהממוצע **לא יעלו** מעל המחיר הממוצע

השערת H1: דירות בעלות 3 חדרים ששטח המגורים שלהן גדול מהממוצע **יעלו** מעל המחיר הממוצע

כדי לבחון השערה זאת, נבצע סקירה עבור הדירות בעלות **3 חדרים** ונבדוק את המחיר הממוצע של כל דירות אלו, ובנוסף נבדוק את שטח המגורים הממוצע עבור דירות אלו. עבור כל דירה שששטח המגורים שלה גדול מהממוצע נבדוק האם המחיר שלה גבוה מהמחיר הממוצע.   
במידה ונראה כי האחוז של הדירות הנ"ל (שהמחיר והגודל גדולים מהממוצע) הינו יותר מ80% נחליט כי נקבל אתH1 כלומר נדחה את H0.

עבור שאלה עסקית 2: **Unsupervised**

מה מאפיין דירות בעלות 3 חדרים ובעלות שטח מגורים מעל הממוצע?

השערת H0: דירות שמחירן מתחת לממוצע לא יתאפיינו באותן תכונות

השערת H1: דירות שמחירן מתחת לממוצע יתאפיינו באותן תכונות

כדי לבחון השערה זאת, נבצע סקירה עבור הקריטריונים הנבחרים שתיארנו קודם, ונראה שבמידה והמרחק בין האיברים בכל קבוצה (של הסיווג) יהיה קטן (כלומר תהיה סטייה של עד 15 אחוז מהתצפית הגבוהה ביותר לתצפית הנמוכה ביותר באותה הקבוצה) נוכל להסיק כי דירות שמחירן מתחת לממוצע יתאפיינו באותן תכונות. כלומר, נחליט לקבל אתH1 ונדחה את H0.

**חלק 3 – שאילתות:**

שאילתות מצורפות בקובץ הנקרא SQL Q בתיקיית Analysis.

**חלק 4 – ניהול גרסאות:**

בוצע, קישור לפרויקט: <https://github.com/EdenBenDavid1/BI_Project>