בינה עסקית – פרויקט

חלק 3

מגישות

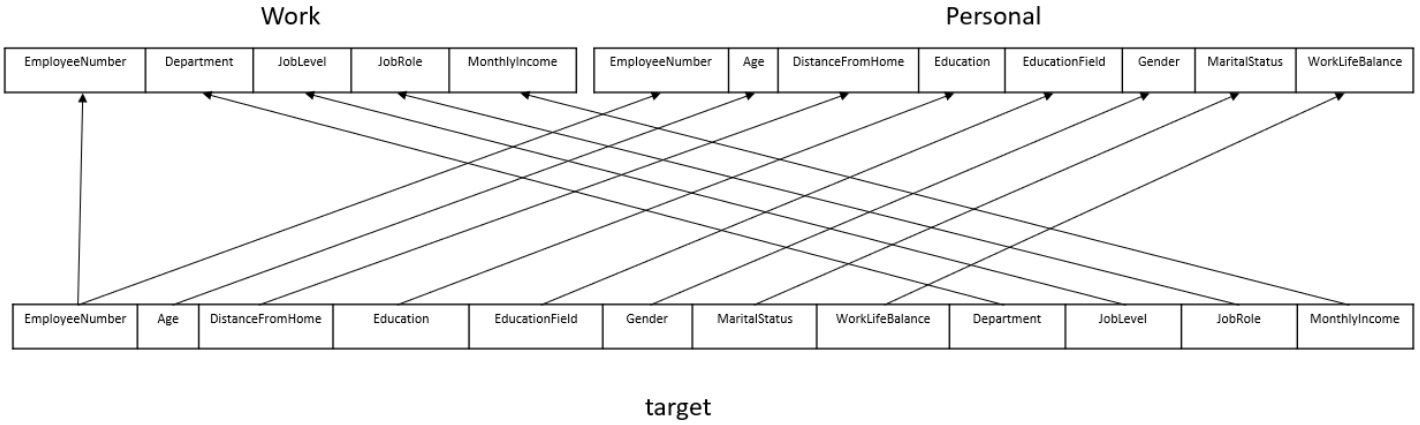
ורוניקה פרידמן

לי קישון

סמסטר ב, 2022

חלק א': STTM

1. סכימה ויזואלית אשר מציגה את ה - STTM :

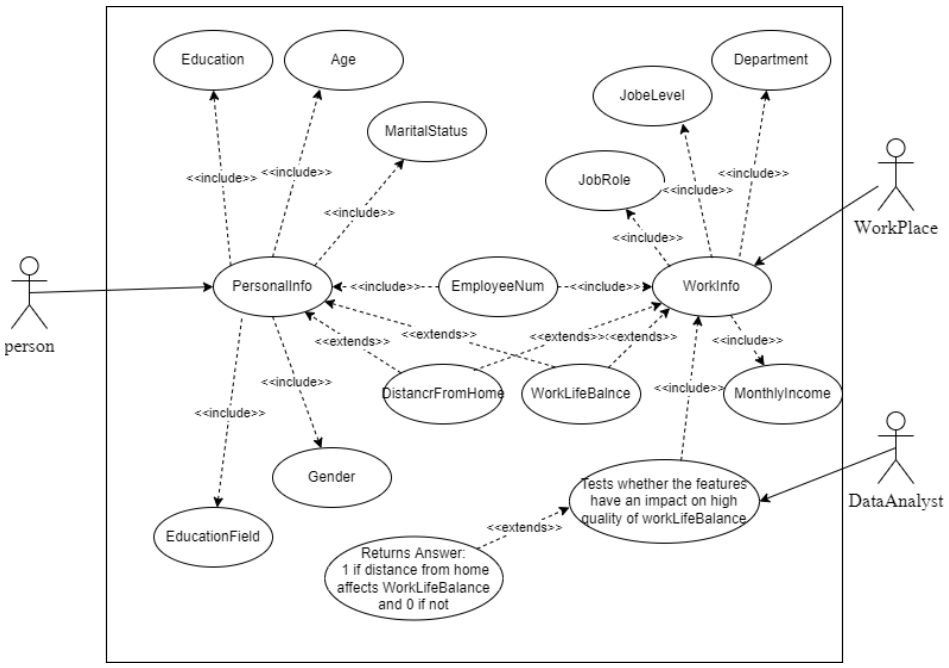


2. מצ"ב בקובץ אקסל.

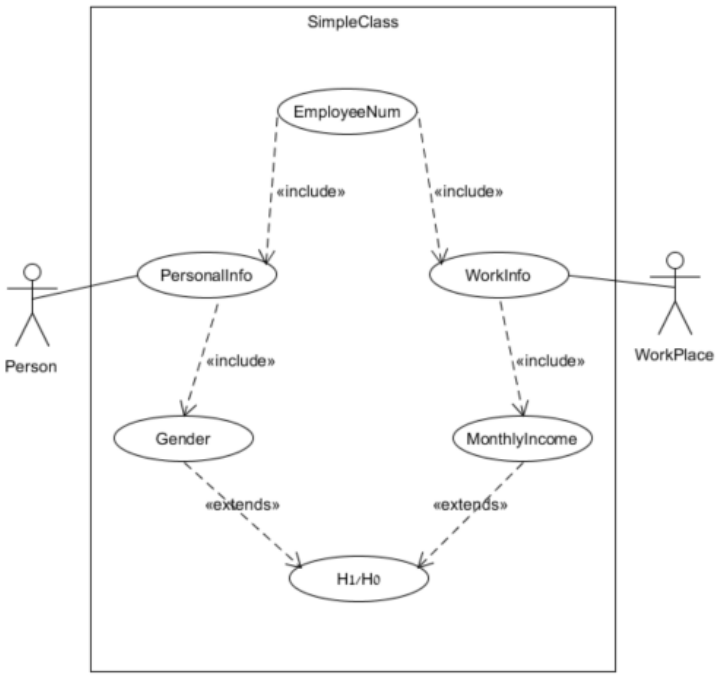
חלק ב': data mining techniques

1. תהליך ה - KDD אשר אנו מבצעים בפרויקט זה מתחיל בשיטה ה - Descriptive מאחר ורוב ה - Data חסר מידע עסקי קודם ולטובת מענה על השאלות העסקיות שלנו אנחנו נדרשים לנתח את הנתונים והמידע הקיימים לרשותינו וזאת נעשה ע"י clustering, לאחר ביצוע K- means נוכל ללמוד את המאפיינים (תכונות נתוני העסקה) של כל מגדר. לאחר ניתוח הנתונים אנחנו נשתמש בשיטת ה- Predictive השאלה שלנו דורשת binary class classification, ע"י עץ החלטה נוכל לענות על השאלה האם התכונות הנבחרות בעלות השפעה על איכות חיים-עבודה גבוה.
2. הטכניקות אשר ימומשו על אוסף הנתונים (הנתונים הם ברובם מספרים בינארים הנדרשים לחלוקה לקבוצות) הם :
   - Clustering -K- means.
   - Classification - Decision tree.

- 2 תרחישי USE CASE :
  - שאלה עסקית 1:



  - שאלה עסקית 2:



3. הגדרת מדד דמיון עבור הDW שבנינו:
   - Jaccard- עבור הנתונים הקטגוריים לרבות בינאריים כאשר המדד יוגדר בין 0 ל -1 (1 דמיון מלא, 0- אין דמיון).
   - Sorensen-Dice - עבור משתנים בדידים (בנתונים שלנו גיל משתנה בדיד).
4. השערות עבור כל אחת מהשאלות העסקיות:

שאלה עסקית: האם התכונות הבאות: גיל, מרחק מהבית, השכלה, מגדר ומצב משפחתי -בעלות השפעה על איכות חיים-עבודה גבוה (מעל 2)

- H0- Distance from home affects Worklifebalance rate
- H1- Distance from home does not affects Worklifebalance rate
- דרך קבלת ההחלטה: Sorensen-Dice
- מדוע זו הדרך: מרחק מהבית הינו נתון בדיד ולכן השתמשנו בדרך זו.

שאלה עסקית: מה ניתן ללמוד על תכונות נתוני ההעסקה - (Department,JobLevel,JobRole,MonthlyIncome) של העובדים ביחס למגדר שלהם.

- H0 - men and women get the same MonthlyIncome
- H1- Men’s MonthlyIncome is higher than a Women’s MonthlyIncome
- דרך קבלת ההחלטה: מבחן חי בריבוע
- מדוע זו הדרך: מאחר והמדד העיקרי הינו קטגוריאלי מבחן חי בריבוע הוא המתאים ביותר כאשר ממוצע שכר הגברים הינו הערך המצופה וממוצע שכר הנשים הינו הערך הנצפה.

חלק ג': שאילתות

מצורף בקובץ אקסל בתיקיית PIPELINE.

SELECT AVG (MonthlyIncome) as avg_MonthlyIncome ,Gender FROM WorkData INNER JOIN PersonalData ON EmployeeNumber=EmployeeNumber
SELECT COUNT(WorkLifeBalance) as Rate FROM WorkData WHERE WorkData=Gender
SELECT AVG (JobLevel) as avg_JobLevel ,Gender FROM WorkData INNER JOIN PersonalData ON EmployeeNumber=EmployeeNumber
SELECT MAX(Age) as max_Age,MIN(Age) as min_Age, AVG (MonthlyIncome) as avg_Age FROM WorkLifeBalance WHERE WorkLifeBalance>1
SELECT MAX(DistanceFromHome) as max_DistanceFromHome,MIN(DistanceFromHome) as min_DistanceFromHome, AVG (DistanceFromHome) as avg_DistanceFromHome FROM WorkLifeBalance WHERE WorkLifeBalance>1
SELECT MAX(Education) as max_Education,MIN(Education) as min_Education, AVG (Education) as avg_Education FROM WorkLifeBalance WHERE WorkLifeBalance>1

חלק ד': ניהול גרסאות

1. קישור לפרויקט:

https://github.com/VeronikaFridman/BI-project-HR-IBM.git