# חיזוי הצלחת בית עסק על סמך נתונים דמוגרפיים, גיאוגרפיים וניתוח שוק

**מנחה:** פרופסור שאול מרקוביץ

מגישים:

* מתן גורן 206670549 [matangoren@campus.technion.ac.il](mailto:matangoren@campus.technion.ac.il)
* דניאל גילה 314625609 [danielg293@campus.technion.ac.il](mailto:danielg293@campus.technion.ac.il)
* מתן הרשקוביץ 123456789 [matanhershko@campus.technion.ac.il](mailto:matanhershko@campus.technion.ac.il)

# תוכן עניינים

* מבוא
* תיאור הבעיה
* פתרון הבעיה בקווים כלליים
* חלק א': חיזוי בהתאם למאפייני העסק
* חלק ב': חיזוי בהתאם למאפייני העסק ומיקומו
* חלק ג': חיזוי התאם למאפייני העסק, מיקומו, ופרמטרים נוספים הקשורים במיקום הגיאוגרפי.

# מבוא

כיום בעידן המדיה החברתית לחוות דעת של משתמשים (reviews) יש השפעה רבה על מידת ההצלחה של בתי העסק. לקוחות רבים בודקים ביקורות על בתי העסק ועל סמך הביקורות מחליטים היכן לבצע את רכישתם. מאז הקמתה ב-2004, Yelp צברה תאוצה בקרב הלקוחות בתור אחת האפליקציות המובילות בתחום סקירת חוות דעת ודירוג בתי עסק. נכון ל-2017 יש ב-Yelp יותר מ148 מיליון ביקורות על מליוני עסקים, אשר כאמור משרתות את הלקוחות לבחירת בית העסק המתאים.

ממחקר שנערך ב-"Harvard Business School" בשנת 2011 נובע כי כל "כוכב" בדירוג ב-Yelp משפיע על המכירות של בית העסק בכ-5-9 אחוזים. עקב כך, לבעלי העסקים התווספה עוד פלטפורמה בה הם צריכים לשמור על סטטוס ומעמד בקרב הלקוחות, והם פועלים רבות על מנת להבטיח דירוג טוב ב-Yelp.

בפרויקט זה נרצה להבין לעומק מה הם הגורמים אשר משפיעים על דירוג בית עסק ב-Yelp ולנסות ולחזות את הדירוג העתידי של בית עסק נתון.

**הפרויקט יתמקד בתחום המסעדנות בלאס ווגאס שבמדינת נבאדה, ארה"ב. נסקור מסעדות מסוגים שונים ונרצה למצוא קשר בין מאפייני המסעדה, מיקומה, ופרמטרים נוספים הקשורים בה, לבין הצלחתה הפוטנציאלית.**

# תיאור הבעיה

יזמים רבים בשוק החופשי אשר מעוניינים לפתוח בית עסק מתחבטים רבות בסוגיית מיקומו העתידי של בית העסק, דבר אשר משפיע באופן משמעותי על הצלחת העסק. על מנת לפתור סוגייה זו, לרוב נדרשים היזמים לבצע ניתוח שוק מעמיק ולהתייחס לפרמטרים רבים כגון: תחרות, נגישות אופי האוכלוסייה וכדומה. בדיקה ידנית זו הינה משימה קשה כיוון שכמות המידע ובתי העסק שאותם צריך לבדוק היא גדולה וניתוחה דורש מאמצים רבים, ולרוב, ליזם אין מספיק ידע מקדים בנוגע לאלו פרמטרים משפיעים על הצלחת העסק. על כן, בפרויקט זה נרצה להתמודד עם השאלה "מהו המיקום האידיאלי לפתיחת בית עסק במטרופולין מסויים?". כלומר, בהינתן פרופיל של בית עסק שעתיד להיפתח, נרצה לבנות כלי שיאפשר לנתח את מידת הצלחת העסק – דירוגו ב-Yelp.

# פתרון הבעיה בקווים כלליים

## תיאור כללי

על מנת לפתור את הבעיה שתוארה לעיל, נרצה להשתמש בכלים מעולם "הבינה המלאכותית" ובפרט מתחום הלמידה. נרצה לממש מסווג (Classifier) אשר בהינתן פרופיל של בית עסק מסויים שעתיד להיפתח יחזה (Prediction) את מידת ההצלחה של אותו העסק. כלומר, המסווג יחזה את דירוג העסק ב-Yelp.

## שלבי הפתרון ותיאור המערכת

נרצה לחלק את מהלך הפרויקט ל-3 חלקים עיקריים:

**חלק א': חיזוי בהתאם למאפייני המסעדה**

בחלק זה נרצה למעשה לבצע ניתוח של תכונות שמשפיעות על הצלחתה של מסעדה בלאס ווגאס. נשתמש בdata set המונה כ4900 מסעדות ברחבי העיר ובו תכונות שמאפיינות את המסעדה כמו:

-אופי המסעדה (למשל: מסעדת בשרים/ קינוחים/ בר אלכוהול/ מסעדה רומנטית וכו')

-סוגי המנות המוגשות במסעדה (למשל: בשרי, חלבי, מקסיקני, צמחוני, אלכוהול, קינוחים, אוכל איטלקי וכו').

- שעות הפתיחה השבועיות של המסעדה.

- קהל היעד של המסעדה (בליינים/ תיירים/ אנשי עסקים/ זוגות וכו')

- פרמטרים כלליים נוספים (למשל: כמות החניה, wi-fi, מותרת הכניסה לחיות, נוחות לילדים, התאמה לקבוצות ואירועים פרטיים וכו')

למעשה בחלק זה אנו רוצים לבצע חיזוי על בסיס הידע שלנו על מסעדות קיימות. אנו מניחים שמסעדה חדשה שתוקם בעיר תצליח יותר או פחות כתלות בפרמטרים של מסעדות דומות

**חלק ב': חיזוי בהתאם למאפייני המסעדה ומיקומה**

בחלק זה נרצה להשתמש במידע שציינו בחלק א' ובנוסף להתייחס גם למיקומה של המסעדה. השוני המשמעותי לעומת חלק א' הוא שכאן אנו גם רוצים לבדוק את מידת ההתאמה של המסעדה לאזור שבו היא עתידה להיפתח.

לדעתנו, יש חשיבות גדולה למיקום של המסעדה מכיוון שמיקום זה קובע מי יהיו המסעדות המתחרות ומהו איכותן.

**חלק ג': חיזוי בהתאם למאפייני המסעדה, מיקומה** **ופרמטרים נוספים הקשורים במיקום הגיאוגרפי**

בחלק זה נרצה להתייחס לכל האמור בחלקים הקודמים וגם לקחת בחשבון פרמטים נוספים שלדעתנו עשויים להשפיע במידה ניכרת על הצלחת המסעדה.

בין השאר נבחן את הפרמטרים:

* המצב הסוציו אקונומי של האוכלוסייה באזור .
* אטרקטיביות האזור מבחינה תיירותית.
* מספר בתי המלון באזור.
* איכות מצב התחבורה הציבורית באזור.

להערכתנו בחלק זה נצליח להגיע לרמה אופטימלית של דיוק מכיוון שהצלחתה של מסעדה תלויה הן בפרמטרים שקשורים למסעדה עצמה ולאיכותה ,הן במתחרים באזור והן בגורמים נוספים הקשורים לאזור הגאוגרפי.

בכל חלק שצויין לעיל נבצע על הdata sets את הפעולות הבאות:

1. **איסוף דוגמאות ועיבוד מקדים (preprocessing) שלהן**

כדי לבצע את האמור לעיל יש ראשית כל למצוא data set מתאים המכיל תכונות רלוונטיות. לאחר מכן נבצע עיבוד מקדים של התכונות, למשל:

* נשלים ערכים חסרים במסעדות שלא צוין עבורן ערך.
* נהפוך מחרוזות המכילות את שעות הפתיחה למספר מייצג כלשהו וננרמל אותו לטווח הגיוני.
* הוספת פרמטרים בינאריים הנלקחים מתוך מחרוזת ארוכה שהופיעה בתיאור המסעדה

1. **בחירת תכונות**

לאחר שננתח את מקורות המידע ואת ה-datasets השונים נרצה למעשה לבצע תהליך של בחירת מאפיינים (Feature Selection) על מנת לזהות את תת-הקבוצה האופטימלית של התכונות הרלוונטיות ביותר עבור המערכת הלומדת.

1. **אימון המסווג**

בשלב זה נרצה לאמן את המסווג על קבוצת אימון (Training Set) על ידי הרצת תוך שימוש בכלים אשר ראינו בקורס "מבוא לבינה מלאכותית" במטרה להגיע לאחוזי דיוק גבוהים על קבוצת הוידוא (Validation Set) ובפרט עבור בית העסק אותו המשתמש רוצה לפתוח.

# תיאור הניסויים

על מנת למקסם את ביצועי המערכת הלומדת, עלינו לבחור תת קבוצה אופטימלית של תכונות רלוונטיות מכל המידע שנשיג. לשם כך עלינו לבצע תהליך feature selection אשר במהלכו נאמוד את השפעת המאפיינים השונים על תוצאת המסווג. בנוסף, נצטרך לכוונן פרמטרים הנוגעים לאיסוף המידע שלנו. לדוגמה, מהו הרדיוס המעניין אותנו בהקשר של ״בתי עסק באיזור״ האם רדיוס זה מתאים לכל התכונות, או שלדוגמה עבור בתי מלון באיזור עלינו להשתמש ברדיוס אחר. לבסוף, נצטרך לכוונן את הפרמטרים הספציפיים עבור המסווג בו נבחר להשתמש.

**חלק א' : חיזוי בהתאם למאפייני המסעדה**

כפי שציינו לעיל, חלק זה יעסוק בניתוח פרמטרים הקשורים במסעדה עצמה.

1. **איסוף דוגמאות ועיבוד מקדים (preprocessing) שלהן**

בחלק זה השתמשנו בdata set ובו כ 4900 מסעדות מרחבי העיר לאס ווגאס. הפרמטרים המקוריים שהופיעו הינם:

1. Attributes -מחרוזת ובה המון פרמטרים הנוגעים למסעדה כמו: חנייה לאופניים, חנייה בתשלום, wifi, אזור ישיבה בחוץ, מבחר האלכוהול, התאמה לילדים, התאמה לחיות, רמת מחירים, קינוחים ועוד.
2. Categories – מחרוזת ובה פרטים על סוג המסעדה למשל: מקסיקני, פיצה, איטלקי, תאילנדי, נודלס, מרקים, סנדוויבצ'ים, ארוחות בוקר וצהריים, צמחוני, בר אלכוהול ועוד.
3. Hours- שעות פתיחה. מחרוזת מהצורה:

…" "Monday:10:00-21:00, Tuesday: 10:00-22:00

1. Review count -מספר המדרגים של בית העסק.
2. Stars- הדירוג של בתי העסק בכוכבים (בין 1 ל5)

\*\*\* בכוונה לא הזכרתי כאן את הפרמטרים כתובת ונ"צ, נשאיר את זה לחלק ב\*\*\*

**העיבוד המקדים התבצע באופן הבא:**

* **Parsing**:

1. מהמחרוזת attributes לקחנו את כל הפרמטרים שהופיעו ב90% מהמסעדות והוספנו אותן כפרמטרים לdata set.
2. מהמחרוזת categories לקחנו את 20 הערכים הכי שכיחים והוספנו אותם כפרמטרים לdata set.
3. מהמחרוזת hours יצרנו פרמטר שהינו סכום שעות הפתיחה השבועיות של המסעדה.

* **נרמול והשלמת ערכים חסרים:**

1. רוב הפרמטרים מהמחרוזת attributes הינם בינאריים לכן לא נדרש נרמול עבורם. היו גם פרמטרים שהכילו טקסט חופשי (כמו wifi- no, yes, free) ופרמטרים עם מספר גבוה יותר של ערכים (רמת המחירים בין 1 ל5) . עבור פרמטרים אלו ביצענו נרמול כך ש- \*\*\*\*
2. הנירמול התבצע באופן דומה ל1
3. כאן נרמלנו את הערכים למספר כלשהו בין 0 ל1 כאשר הערך המקסימלי קיבל 1, המינימלי קיבל 0 וכל השאר קיבלו את היחס מתוך ההפרש בין המקסימום למינימום.

השלמת הערכים החסרים התבצעה כך שכל ערך חסר הושלם בחציון של שאר המסעדות עבור הפרמטר הזה. למשל אם למסעדה מסויימת היה חסר הפרמטר של טווח מחירים אז היא קיבלה את ערך החציון של שאר המסעדות. פעלנו כך מכיוון שלדעתנו החציון הוא הערך שיזיק הכי פחות לתיוג עתידי. החציון גם מתאים לערך חסר במקרה בוליאני כי למשל אם 95% מהמסעדות הינן עם אפשרות לקבל כרטיס אשראי סביר להניח שהמסעדה בה ערך זה חסר תהיה יכולה לקבל אשראי.