

**המחלקה להנדסת תעשייה וניהול**

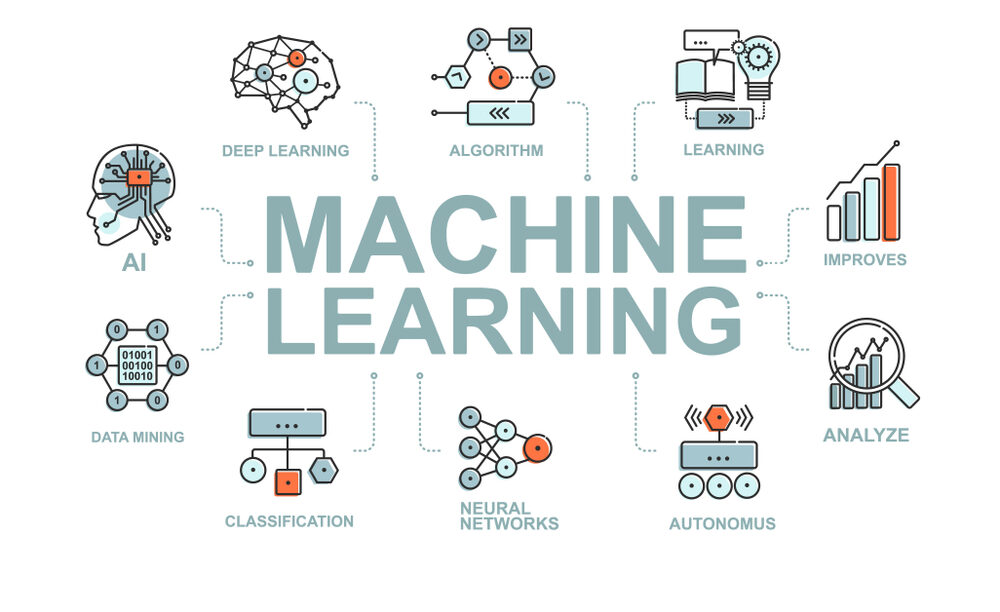
**קורס לימוד מכונה 364.1.1811**

**Machine Learning – Part 2**

**מגישים:**

**עומרי זאבי – 313327041**

**אסטר מרדכי – 208285973**



**Decision Trees**

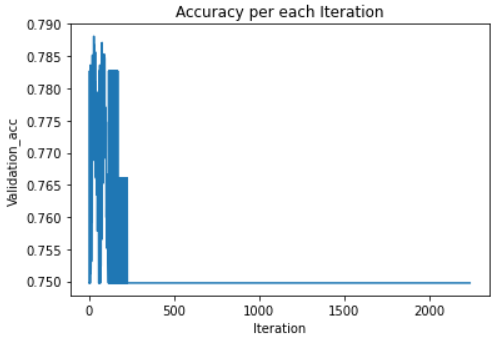
תחילה, הגדרנו מודל עץ החלטה ראשוני ללא כיוונון פרמטרים על סט האימון שהכנו בחלק א' של הפרויקט. התקבלו התוצאות הבאות:

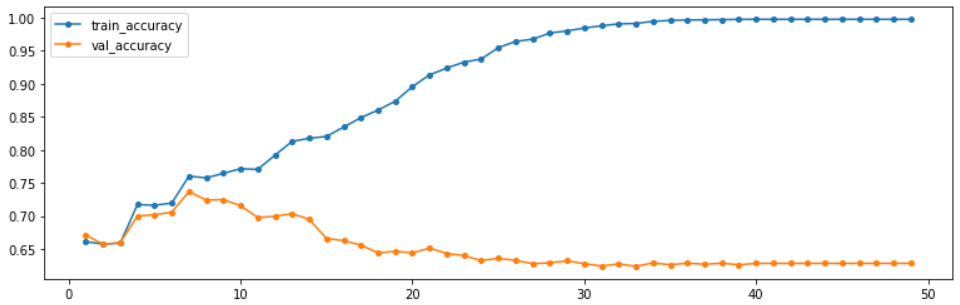
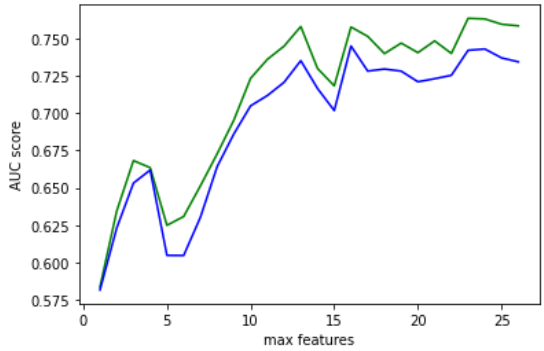


אנו מאמינים כי דיוק של כ-100% התקבל עקב Over Fitting של המודל, משום שהתבצעו הרצות רבות על אותו סט הנתונים. בנוסף, קיבלנו דיוק נמוך יחסית של 62% על סט הוולידציה ולדעתנו הדבר נובע מהפרמטרים הראשוניים של המודל. לכן, ציפינו לראות שינוי באחוזי הדיוק לאחר ביצוע כיוונון פרמטרים על המודל. לאחר מכן, ביצענו את כיוונון הפרמטרים בעזרת אלגוריתם Grid Search כדי לבחון מהם ההיפר – פרמטרים שנותנים את אחוזי הדיוק הגבוהים ביותר עבור המודל שלנו. האלגוריתם בחן 2,240 קומבינציות של 4 היפר – פרמטרים אותם בחרנו לכוונן (מצורף [בנספחים](#_אחוזי_דיוק_של) פלט הדיוק של כל הקומבינציות האפשריות). יש לציין כי האלגוריתם גם ביצע 10-fold cross validation כך שהאימון בוצע על 9 סטים, והוולידציה בוצעה על סט אחד. להלן טבלה המרכזת את הפרמטרים שכיווננו עבור מודל עץ ההחלטה:

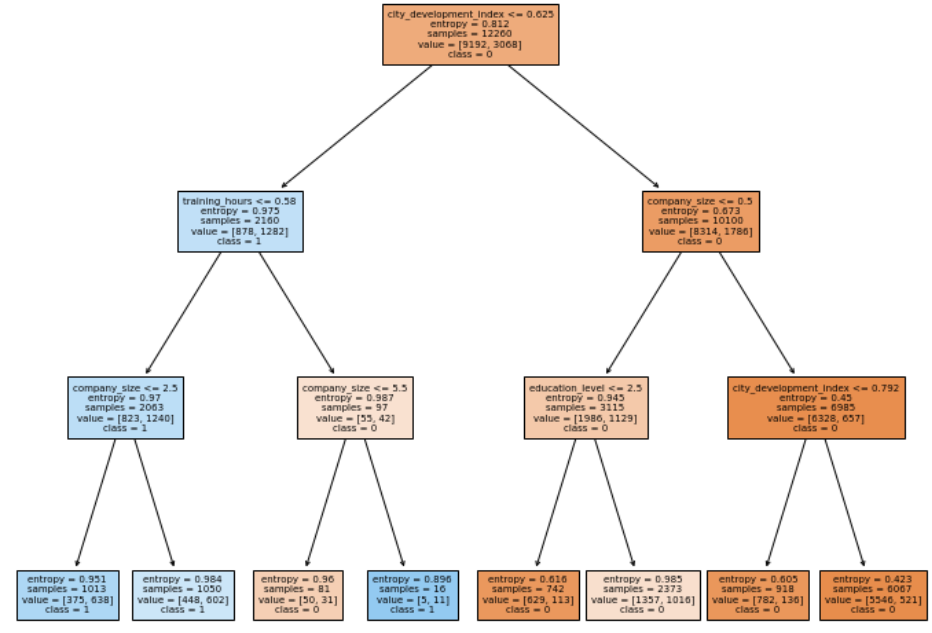
|  |  |  |
| --- | --- | --- |
| **פרמטר** | **משמעות** | **השפעה על המודל** |
| **Max\_depth** | מס' הרמות המקסימלי בעץ. המשמעות של כיוונון פרמטר זה היא שאנו מבצעים trade of בין דיוק המודל לבין זמן הריצה וכמות החישובים שאנו דורשים ממנו. בבעיית הסיווג שלנו ובכלל, אם נדרוש שהעץ ייצור עוד ועוד רמות, ניתן להניח שנפחית את הסיווגים הלא נכונים של ה-samples ל-classes, אך ניאלץ ליצור עוד ועוד פיצולים בעץ שעולים לנו בכוח חישוב והופכים את העץ למורכב יותר. | ניתן לראות בגרף בהמשך כי עבור אחוזי הדיוק על סט האימון והוולידציה עולים, ולאחר מכן הדיוק על האימון עולה בעוד שהדיוק על הוולידציה יורד עד להתכנסות בערך כש- *.* |
| **Max\_features** | פרמטר הבוחן את מספר הפיצ'רים המקסימלי שיש לאמן איתם את המודל. המשמעות בבחירת כוונון פרמטר זה היא לבדוק אם יש לאמן את המודל עם כל הפיצ'רים או רק עם חלקם כדי לקבל את התוצאות הטובות ביותר. | עבור המודל עם כלל הפיצ'רים, התקבלו אחוזי דיוק טובים יותר לעומת מודל חלקי. |
| **Criterion** | פרמטר למדידת איכות הפיצול של העץ. | נבדקו שני מדדים – "Entropy" ו-"Gini", וקיבלנו כי מדד Entropy נתן את התוצאות הטובות ביותר. |
| **cp\_alpha** | אנו משתמשים בגיזום וקטימה כדי להימנע מ-over fitting. בהינתן אלפא, מדד זה בוחן את עלות הגיזום המינימאלית. | התוצאות הכי טובות התקבלו עבור *. כלומר, הגדלת ערך של פרמטר זה מניבה אחוזי דיוק נמוכים יותר.* |

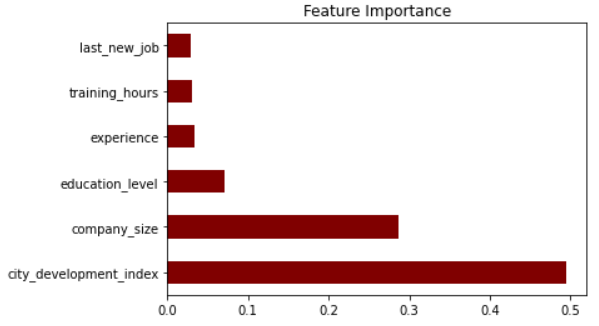
ערכי הפרמטרים אשר הניבו את אחוזי הדיוק הגבוהים ביותר עבור סט הנתונים שלנו הם (לפי אלגוריתם Grid-Search):

ניתן לראות כי לאחר כ-250 הרצות, שינוי ההיפר פרמטרים לא השפיע על דיוק הוולידציה:

בנוסף, ניתן לראות את השפעת עומק העץ המקסימלי על אחוזי דיוק המודל בגרף להלן. ברגע שהעומק עובר את 7, דיוק האימון והוולידציה מתחילים להתרחק. כמו כן, בגרף משמאל ניתן לראות את השפעת מספר הפיצ'רים על אחוזי הדיוק. אחוזי הדיוק הגבוהים ביותר מתקבלים עם המודל המלא, ע"פ מדד AUC. מדד זה מתחשב בשטח מתחת לעקומה ומתאים לבעיות סיווג בינארי כמו זו שלנו.

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטילבסוף, עבור המודל הטוב ביותר באמצעות אלגוריתם Grid Search שביצע 10-fold cross validation התקבלו אחוזי הדיוק משמאל. ניכר שיש שיפור ביחס לתוצאות המודל ההתחלתי בו הדיוק על סט הוולידציה היה 62%, כי לאחר כיוונון ההיפר פרמטרים הגענו לדיוק של 74% כך ששיפרנו את דיוק המודל ב-12%. השיפור הנ"ל התאפשר הודות לבחירת היפר-הפרמטרים באופן המותאם לסט הנתונים ובהתבסס על כל הקומבינציות האפשריות של ארבעת הפרמטרים אותם בחרנו לכוונן.  
יכולת ה- interpretabilityשל עץ החלטה מהווה יתרון משום שע"י בחינת האלגוריתם ניתן להסביר כיצד קיבל את הה חלטה שהיה עליו לקבל בהתאם למשימת הלמידה. ניתן לייצג עץ החלטה באמצעות סטים של תנאי "if-then" ובהתאם לצפות פיצ'רים שעל פיהם נקבעו הפיצולים השונים וכיצד הם בוצעו בכל רמה בעץ. באמצעות ויזואליזציה של התהליך אפשר להבין איך הלמידה מתבצעת שלב אחר שלב. כפי שכבר נאמר, היכולת הזו של המודל מסייעת במשימת הלימוד בכך שניתן בקלות יותר לפרש את תהליך הלמידה של העץ.   
כעת, נציג את העץ הטוב ביותר שהתקבל, אך מטעמי נראות נגדיר כי מס' הרמות המקסימלי הוא 3 (ולא 7 שהוא האופטימלי במקרה שלנו):



תמונה שמכילה טקסט

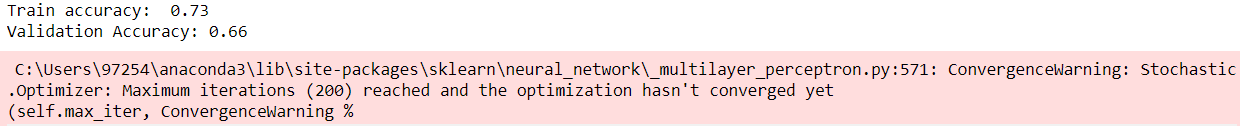
התיאור נוצר באופן אוטומטיעל ידי התבוננות בפיצולי העץ ובמאפיינים השונים לפיהם הוא מבצע את הפיצולים, ניתן ללמוד על חשיבות הפיצ'רים וטווח הערכים שלהם לצורך החלטה על המאפיינים הבאים המתאימים לסיווג. למשל, עבור sample בה אינדקס ההתפתחות של העיר קטן מ-0.625, המאפיין הבא הנבחן הוא שעות התרגול, בעוד שאם הוא גדול מ-0.625 אז הסיווג יעבור דרך הפיצ'ר של גודל החברה בה נמצא העובד. כדי להעמיק את ההבנה שלנושל עץ ההחלטה, נציג כעת את המאפיינים בעלי החשיבות הגבוהה ביותר:

ניתן לראות כי הפיצולים הראשונים בעץ מתבצעים לפי חשיבות הפיצ'רים לעיל. למשל, שורש העץ מבצע פיצול על אינדקס ההתפתחות של העיר שהוא המאפיין בעל החשיבות הגדולה ביותר. הפיצול הבא אחריו מתבצע לפי המשתנים company size ו-training hours שחשיבותם גבוהה גם כן ביחס לשאר המשתנים. בהתחשב גם בידע המקדים שלנו בתחום, ניתן להסיק כי עובד שמתגורר בעיר מפותחת מאוד ועובד בחברה גדולה יחסית יהיה בעל נטייה להישאר בחברה לאחר תוכנית ההכשרה ל-DS. ניקח לדוגמה את תל אביב, שמאוד מפותחת ובהתאם האוכלוסייה בה ממעמד כלכלי חברתי גבוה יחסית, וכדי לשמר את המעמד יש להחזיק בעבודה שמכניסה הרבה כמו ב-DS. בנוסף, אם החברה היא חברה גדולה ושולחת את עובדיה להתמחות ב-DS, ניתן להסיק שישנן אופציות קידום ופיתוח בתוך החברה ושווה לעובד להישאר בחברה שמשקיעה בו. כמובן שרמת השכלה מסוימת וניסיון בעבודה מהווים תנאי מקדים לכניסה לחברות DS ובפרט לתוכניות הכשרה בתחום, ושעות התרגול מכריעות גם כן, אך ללא עיר מפותחת מספיק וללא חברה גדולה בה יש את האופציה להתפתח, כל זה לא יכול לקרות. לראייה, ה-majority voting בפיצול הימני בעץ נוטה באופן ברור לכך שהעובד יישאר בחברה (class 0) על סמך מאפיינים אלו.

**Artificial Neural Networks**

תחילה, ביצענו נרמול לנתונים באמצעות StandartScaler העושה תקנון לערכי הפיצ'רים. להלן טבלת הקונפיגורציה לפי ערכי ברירת המחדל:

|  |  |
| --- | --- |
| **פרמטר** | **ערך** |
| מספר נוירונים בשכבת הכניסה | כמספר הפיצ'רים. זהו הקלט של הרשת |
| מספר שכבות חבויות | 1 |
| מספר נוירונים בשכבה החבויה | 100 (ערך ברירת המחדל ב-MLPClassifier) |
| מספר נוירונים בשכבת היציאה | 2 (כמספר ה-classes) |

הפלט שהתקבל עבור הרצת המודל עבור קונפיגורציית ברירת המחדל הינו:

|  |  |  |
| --- | --- | --- |
| **פרמטר** | **מוטיבציה** | **השפעה על המודל** |
| Hidden layer size | מספר השכבות ומספר הנוירונים בשכבה החבויה. בחרנו לבדוק על 1-3 שכבות חבויות עם נוירונים כמס' הפיצ'רים בכל שכבה כדי לבחון רמות שונות של מורכבות הרשת ולראות את השפעתן על הדיוק | ככל שכמות הנוירונים בשכבה החבויה גדלה, כך גם אחוז הדיוק גדל עד למצב של התכנסות. כלומר, במצב זה הגדלת מס' הנוירונים בשכבה החבויה לא תשפיע כמעט על אחוזי הדיוק (ערך ההתכנסות תלוי בסט הנתונים). בנוסף, קיים trade off בין מס' הנוירונים לבין זמן ריצת המודל |
| Activation | פונ' האקטיבציה אמונה על הפלט שכל נוירון מפיק בשכבות החבויות ועל כן עשויה להשפיע על אחוזי דיוק הרשת | נבדקו 3 פונקציות אקטיבציה: relu, logistic ו- tanh(החלטנו להישאר עם 2 האחרונות מכיוון שהן הניבו תוצאות טובות יותר). פונקציות האקטיבציה הללו מסייעות בהמרת הקלט של הרשת אל עבר הפלט הרצוי (במקרה שלנו 0/1). לאחר תהליך הכיוונון גילינו כי פונ' relu הפיקה אחוזי דיוק גבוהים יותר מאשר האחרות. |
| Solver | שיטת עדכון המשקלות ברשת. השתמשנו ב-sgd על מנת כדי לעדכן את המשקלות בכדי להגדיל את דיוק המודל | שיטה איטרטיבית זו נועדה לסייע למודל למצוא אופטימום (מינימום) לוקלי של פונקציית השגיאה ובכך לגרום לשגיאה להתכנס עד שעדכון המשקלות יהיה זניח וניתן יהיה לעצור. |
| Learning rate | מגדיר את קצב עדכון המשקלות של הרשת. קצב הלמידה עשוי להשפיע על אחוזי דיוק הרשת | השפעת קצב הלמידה מתבטאת באחוזי הדיוק של הרשת, אך לא ניתן לזהות מגמה ברורה. לכן, בכל איטרציה בחנו קצב למידה שונה באופן שרירותי כך שקשה להעריך את ההשפעה באופן מדויק על מודל הלמידה |

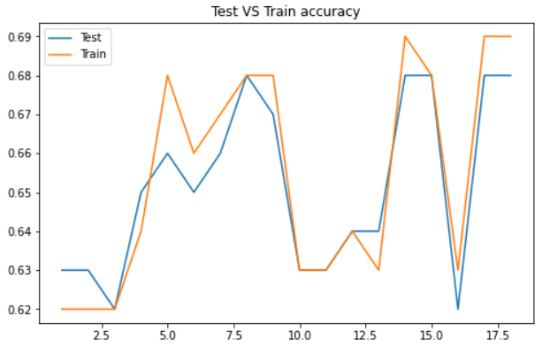
אחוזי הדיוק שהתקבלו לא גבוהים, ונראה שבעקבות אימון המודל על ערכי ברירת המחדל הרשת לא התכנסה לאופטימום. ככל הנראה, נדרשות יותר מ-200 איטרציות על מנת למצות את פוטנציאל אחוזי הדיוק של המודל על סט האימון וסט הוולידציה. בחלק הקודם אף גילינו כי להרבה פיצ'רים בצורת דמי אין כמעט חשיבות עבור מודל הלמידה של עץ ההחלטה, ולכן ניתן להניח שגם סט הנתונים במצבו הנוכחי משפיע לרעה על אחוזי הדיוק.

השתמשנו בתהליך האימון ב-10-fold cross validation, ובחנו סך הכל 18 קומבינציות אפשריות של היפר פרמטרים לצורך שיפור המודל. משמאל, ישנו גרף המתאר את אחוזי הדיוק על האימון ועל סט הבחינה כפונקציה של קומבינציות הפרמטרים אותם כווננו. בנוסף, נציג טבלה של אחוזי דיוק המודל עבור כל קומבינציה שנבחנה:

תמונה שמכילה שולחן

התיאור נוצר באופן אוטומטינראה כי הקונפיגורציה של המודל הטוב ביותר מכילה 27 נוירונים בשכבת הכניסה, 2 בשכבת היציאה ו-2 שכבות חבויות בעלות 28 נוירונים כל אחת. בהתאם, אימנו רשת בעלת ההיפר פרמטרים שהניבו את אחוזי הדיוק הגבוהים ביותר. ההיפר פרמטרים שלה הם:

תמונה שמכילה טקסט, כתום, צילום מסך

התיאור נוצר באופן אוטומטי

ניתן לראות כי שיפרנו את אחוזי הדיוק על סט האימון בכ-6% לעומת המודל הראשוני. עם זאת, אחוזי הדיוק על סט הבחינה השתפרו במעט, רק בכ-2%.   
נסביר כעת את המשמעות של הקונפיגורציה הטובה ביותר. ראשית, היא נבחרה בתהליך של Cross-Validation ונבחנו מספר קומבינציות שונה של היפר פרמטרים, אך לבסוף נשארנו עם 18 הקומבינציות לעיל. אנו סבורים שהפרמטרים נבחרו ממספר סיבות:

1. מס' השכבות החבויות - שכבה אחת עשויה לא לעבד כנדרש את הדאטה ולהיות פשוטה יחסית, בעוד שמס' שכבות מספקות הרבה יחידות עיבוד ובהתאם חישוב המשקלים מורכב ומדויק יותר. תהליך עדכון המשקלות עבור כל פיצ'ר הוא קריטי לשיפור דיוק המודל. לכן, סביר שיש עדיפות ל-2 שכבות חבויות לעומת 1 או 3.
2. מס' הנוירונים בכל שכבה – מס' הנוירונים כמעט זהה למס' הפיצ'רים, וזה מאפשר למידה מותאמת לכל פיצ'ר בדאטה סט. כמות נוירונים גדולה יותר מסוגלת להעלות את אחוזי הדיוק מכיוון שהמודל לומד יותר "לעומק", אך יכולת ההכללה של המודל עשויה להיפגע מכך (over fitting).
3. פונקציית האקטיבציה - מגדירה את הפלט של הנוירון בהינתן קלט כלשהו. הפונ' שנבחרה היא Relu"", אשר מחזירה את המקסימום מבין ערך הקלט לבין אפס. לדעתנו, הפונקציה הזו נבחרה משום שהיא פולטת עבור כל נוירון את הערך המקסימאלי מבין כלל הקלטים שלו, ובגלל שברשת שלנו הקלט לכל נוירון בשכבה החבויה הראשונה הוא כמספר הפיצ'רים, הפונקציה הזו מפיקה את אחוזי הדיוק הטובים ביותר עבורה.

תמונה שמכילה שולחן

התיאור נוצר באופן אוטומטילסיכום, כיוונון ההיפר פרמטרים היה יעיל עבור המודל, אך פחות מכפי שצפינו.   
לפי מטריצת המבוכה, נראה שלמודל יש נטייה לסווג נכונה את רוב ה-samples החדשים להיות 0, אך הוא מסווג כמחצית מה-samples שה-label שלהן 1 לא נכון.

ניתן להסיק כי המודל שלנו מסווג בצורה טובה ל-label של 0 מכיוון שיש רוב לטובת תצפיות אלו בסט האימון, לעומת תצפיות בעלות label של 1. הוא לומד הרבה יותר samples של 0 מאשר של 1 (בסעיף זה אנו מתייחסים לדיוק המודל שלמד דאטה סט לא מאוזן, על אף שבשאר העבודה אנו מנטרלים את העניין באמצעות דיוק Roc\_Auc בהתאם להנחיות).

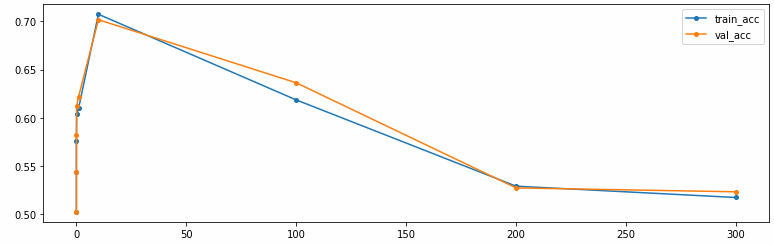
**SVM**

תחילה, הרצנו מודל דיפולטיבי על הדאטה סט והתקבלו התוצאות הבאות:

|  |  |  |
| --- | --- | --- |
| **פרמטר** | **משמעות** | **השפעה על המודל** |
| C | משמעות פרמטר זה היא כמות ה"ענישה" של כל סמפל שסווג בצורה שגויה על ידי המודל. במילים אחרות, מדבור בעובי מישור ההפרדה הליניארי שמשפיע על הסיווג ל-classes. ככל שהערך של C גדול יותר, מישור ההפרדה הליניארי יהיה צר יותר ויסווג טוב יותר samples הקרובים לאזור ההפרדה, ולהפך. בחרנו לכוונן ערכים בין ל- לאחר מחקר בנושא באינטרנט | קיבלנו שהמודל שלנו מניב את התוצאות הכי טובות כאשר  *. שיפרנו בכ-10% את תוצאות הדיוק עבור הערך הנ"ל של הפרמטר* |

אחוזי דיוק אלו נמוכים מאחוזי הדיוק של שאר האלגוריתמים שבדקנו קודם לכן. נבצע כיוונון היפר פרמטרים בכדי לשפר את האחוזים הללו. לפני כן, נפרט על הפרמטרים אותם נרצה לכוונן במודל זה:

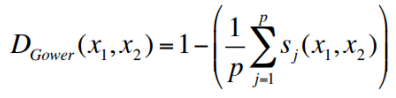
לאחר כוונון ההיפר פרמטרים, קיבלנו את התוצאות הבאות. נראה שהקונפיגורציה שהניבה את אחוזי הדיוק הגבוהים ביותר עבור מודל זה מקיימת . הבחירה בערכים הללו של C הייתה מתוך מטרה לבחון איך סדר הגודל שלו משפיע על דיוק המודל. הוצאנו אף גרף של השפעת הפרמטר על דיוק המודל.

תמונה שמכילה שולחן

התיאור נוצר באופן אוטומטי



בנוסף לכך, קיבלנו את המקדמים של משוואת הישר המפריד. תהליך הסיווג של המודל מתבצע באמצעות מקדמים אלו באופן הדומה לרגרסיה. עבור sample נתון, ישנה הכפלה של כל ערך של פיצ'ר במקדם שלו ולאחר מכן מחברים את כל הערכים האלו, כך שאם המספר שיוצא הוא שלילי אז המודל מסווג ל-class 0, אחרת מסווג ל-class 1. ניתן לראות למשל כי אינדקס ההתפתחות של העיר מאוד "מושך" את ה-sample לכיוון class 0 משום שהמקדם שלו בעל הערך הכי גדול (בתוספת הסימן השלילי). כמו כן, הפיצ'רים בעלי ערך המקדם שגבוה ביחס לשאר הם training\_hours, relevant\_experience, וחלק ממשתי הדמי של gender ו-company\_type. גם עבור משתנים אלו ניתן להבין שהם בעלי השפעה רבה יותר מאחרים בנוגע ליכולת סיווג המודל ל-class המתאים, וכל אחד מהם תורם ל-sample להשתייך ל-class מסוים בהתאם לסימן המקדם שלו. זה תואם באופן מסוים למסקנות הקודמות שלנו, שהרי בחלק של עץ ההחלטה ראינו בעץ עצמו שהוא מתחיל לחלק לפי הפיצ'ר של אינדקס ההתפתחות של העיר ובהתאם נראה שהוא הפיצ'ר בעל המקדם הכי גדול כאן. הפיצולים הבאים בעץ מערבים גם את training\_hours, שגם לו מקדם גבוה במודל ה-SVM. לסיכום, ישנם פיצ'רים מסוימים שאפשר להבחין בקורלציה בין החשיבות שלהם לעץ ההחלטה לבין הגודל האבסולוטי של המקדם שלהם במודל ה-SVM.

**Clustering**k-medoid הוא אלגוריתם עבור למידה שאינה מונחית, כלומר ללא תיוגי מחלקות ומטרתו לבצע אשכול ולא חיזוי או קלסיפיקציה. על כן, הוצאנו את משתנה המטרה וכך התאמנו את סט הנתונים לתצורת משימת הלמידה. בנוסף, בסיס הנתונים נורמל ומשתנים קטגוריאליים הומרו למשתני דמה. נצפה לקבל מודל עם שני אשכולות שכן במודל המקורי ישנם שני classes -0 ו-1. נשתמש בסט פיצ'רים המכיל את כל הפיצ'רים שקיבלנו למעט CITY ו-Enrolee id בדומה לסט הנתונים עליו הרצנו את אלגוריתמי הלמידה האחרים. נשתמש במטריקת Gower למדידת מרחק כיוון שכל רשומה בסט הנתונים שלנו מכילה שילוב של נתונים קטגוריאליים, רציפים ובינאריים. מרחק Gower הוא מספר בין 0 (זהה) ל-1 (שונה באופן מקסימלי). המרחק של Gower מחושב כממוצע של הבדלים חלקיים בין פרטים. הצורה הכללית של המקדם היא כדלקמן:

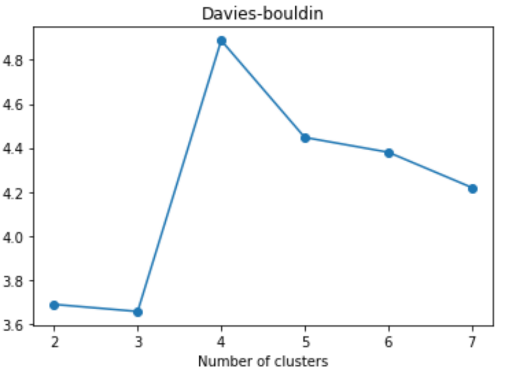
הרצנו את האלגוריתם ובחרנו K שונים, בטווח של ]2,8]k=.

שמרנו את התוצאות שהתקבלו עבור כל אחד מה-K שנבחנו, ובדקנו את התכונות הבאות:

* הומוגניות (נקודות צפופות וקרובות ל-cluster).
* הפרדה (יש הפרדה טובה בין מרכזי ה-cluster).

נבדוק את קיום התכונות על ידי המדדים הבאים:

**Silhouette -** נרצה למקסם, ציון גבוה מעיד על הפרדה טובה ולכידות הנתונים בקלאסטר מסוים.

**Davies-bouldin-משלב הומוגניות והפרדה:** מדד הלוקח בחשבון גם את ההומוגניות בתוך האשכול וגם מרחקים בין האשכולות-נרצה למזערו. ציון נמוך מעיד על הפרדה טובה ועל הומוגניות.

לפי המדד Davies-bouldin מספר ה-clusters האופטימלי הינו 2-3 ולפי המדד Silhouette מספר ה-clusters האופטימלי הינו 2. לכן החלטנו כי כמות הclusters- האופטימלית היא K=2. מבחינת משמעות, כיוון שקיבלנו את כמות הclusters- לה ציפינו (2) ניתן להניח כי כל cluster מייצג 0 או 1 בהתאם lables שהותאמו לכל sample בסט הנתונים המקורי.

**השוואה בין המודלים**להלן טבלת השוואה בין ביצועי המודלים השונים, לפי קריטריון roc\_auc:

|  |  |  |  |
| --- | --- | --- | --- |
| **דיוק/מודל** | **עץ החלטה** | **רשת נוירונים** | **SVM** |
| **סט האימון** | 81% | 79% | 71% |
| **סט הוולידציה** | **78%** | **78%** | **70%** |
| **סט הבחינה** | 74% | 68% | - |

לאור תוצאות הדיוק על סט הוולידציה והבחינה המופיעים בטבלה, נמליץ על מודל עץ ההחלטה למשימת הסיווג, וזאת כדי להגיע לאחוזי הדיוק הגבוהים ביותר על סט הבחינה. השתמשנו ב-20% מהדאטה סט המקורי בתור סט בחינה, ועליו בחנו את תוצאות המודלים השונים. מכיוון שעץ ההחלטה הניב את אחוזי הדיוק הגבוהים ביותר, זהו המודל הטוב ביותר עבור משימת הלימוד שלנו. לא בדקנו סט בחינה על מודל SVM, שהרי התוצאות הטובות ביותר שלו בין כה נמוכות מאלו של עץ ההחלטה ושל רשת הנוירונים.

**שיפור המודל הנבחר**בהתאם לממצאי הסעיף הקודם, המודל הנבחר שלנו הוא עץ החלטה. נבחן כעת כיצד נוכל לשפר את המודל הנ"ל.

|  |  |  |  |
| --- | --- | --- | --- |
| **בעיה** | **הסבר** | **כיצד לפתור** | **מה נצפה לשפר** |
| מספר רב של פיצ'רים שהם דמי בעלי חשיבות קטנה | אנו חושדים שהדבר עלול ליצור "רעש" במודל הלמידה ובכך לא למצות את פוטנציאל אחוזי הדיוק | נספק לו את המשתנים בעלי החשיבות הגדולה ביותר עליהם ילמד | אחוזי דיוק הוולידציה של המודל |
| גודל סט הוולידציה | גדלים שונים עשויים לספק אחוזי דיוק שונים לאותו מודל | לבצע ניסוי וטעיה בנוגע למס' אפשרויות של גודל סט הוולידציה |

|  |  |  |
| --- | --- | --- |
| **תיאור ביצוע השיפור** | **שיפר/לא שיפר** | **מדוע שיפר/לא שיפר** |
| 1. בחינת מספר גדלים אפשריים של validation set (0.1, 0.15, 0.2, 0.25, 0.3). בחרנו להשתמש בגדלים מקובלים של סט ולידציה | **לא שיפר** | לא שיפרנו את דיוק המודל מכיוון שלדעתנו אין הבדל משמעותי בכמות הדאטה הנלמדת ע"י המודל ועל כן ביכולתו לסווג |
| 2. בחירת K הפיצ'רים הטובים ביותר על סמך אלגוריתם SelectKBest. האלגוריתם בחן את k הפיצ'רים הטובים ביותר לפי קריטריון חי בריבוע, והרצנו גרף ([נספח](#_שיטות_שיפור_דיוק)) של אחוזי דיוק הוולידציה על מנת לראות איזה K לבחור | **לא שיפר** | ניכר מהגרף שדווקא מספר הפיצ'רים הגדול ביותר מוביל לאחוזי הדיוק הגבוהים ביותר, והדבר תואם לממצאים בחלק של עץ ההחלטה |
| 3. בחירת 9 הפיצ'רים הטובים ביותר על סמך תכונת feature\_importance. בחנו את המודל הטוב ביותר על הדאטה סט המכיל את 9 הפיצ'רים עם החשיבות הגדולה ביותר | **לא שיפר** | באופן דומה להסבר לעיל, זהו לא המס' האופטימלי של הפיצ'רים עבור המודל |
| 4. שילוב שיפורים 1 ו-3: בחינת מספר גדלים אפשריים על הדאטה סט המכיל את 9 הפיצ'רים בעלי ה- feature\_importanceהטובים ביותר | **לא שיפר** | בהתאם לסיבות חוסר השיפור של הצעה 1 ו-3 |
| 5. שימוש במודל Random Forest עם כיוונון היפר פרמטרים למציאת קומבינציית הפרמטרים הטובים ביותר | **שיפר** | מיצינו את יכולת המודל והדאטה שלנו והשתמשנו במס' רב של עצים (100) וב-10-Fold Cross Validation לצורך הלמידה, וקיבלנו שממוצע אחוזי דיוק הוולידציה הוא 80% (שיפור של כ-2%) |

על מנת להתמודד עם הבעיות הנ"ל, ביצענו 5 ניסיונות שיפור עבור המודל שלנו. להלן הפירוט:

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטיתמונה שמכילה טקסט, כתום

התיאור נוצר באופן אוטומטיתמונה שמכילה טקסט

התיאור נוצר באופן אוטומטייש לציין כי השיפורים המתייחסים לגודל סט הוולידציה באים לשפר את הנתונים, ואלו המתייחסים לכמות הפיצ'רים המשתנה באים לשפר את כלל המודל. בנוסף, גם שיפור מס' 5 הוא שיפור מודל – כלומר יצירת מודל משופר המתבסס על הקיים. להלן הפלטים של אחוזי הדיוק עבור כל אחד מהשיפורים (לפי הסדר מימין לשמאל למעט 2 שנמצא בנספחים):

את הפלטים הנוספים עבור שיטות השיפור ניתן לראות ב[נספחים](#_שיטות_שיפור_דיוק).  
**הגשת חיזויים סופיים**קובץ ההגשה מכיל את תוצאות תיוג מודל ה-RF הטוב ביותר עבור סט הבחינה שקיבלנו. הקובץ מכיל עמודה אחת המציינת את תיוג הרשומות הסופי לאחר הכנסתן למודל. על מנת לבצע את התיוג לרשומות בסט הבחינה, ביצענו עיבוד מקדים לנתונים כפי שביצענו לסט האימון שעבדנו עימו במהלך הפרויקט. לאחר ביצוע העיבוד המקדים , וידאנו כי לא אבדו רשומות מסט הבחינה ולבסוף הרצנו את מודל ה-RF עם הקונפיגורציה הנבחרת לקבלת קובץ החיזויים הסופי.

**נספחים:**

## **אחוזי דיוק של כל קומבינציות הפרמטרים האפשריות**

**תמונה שמכילה שולחן

התיאור נוצר באופן אוטומטי**

**תמונה שמכילה שולחן

התיאור נוצר באופן אוטומטי**

**תמונה שמכילה שולחן

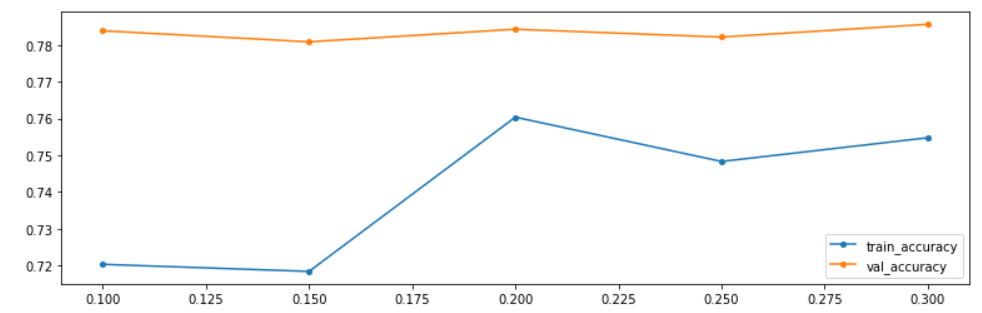
התיאור נוצר באופן אוטומטי**

## **שיטות שיפור דיוק המודל**

שיפור 1 – גודל סט בחינה שונה:

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי



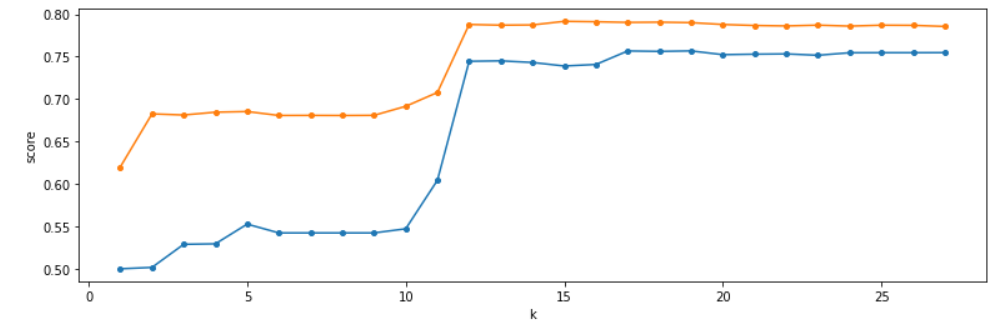
תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

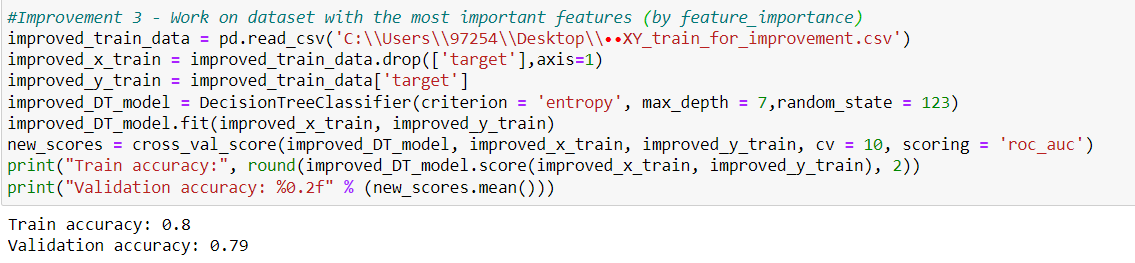
שיפור 2 – Kהפיצ'רים הטובים ביותר:

תמונה שמכילה טקסט

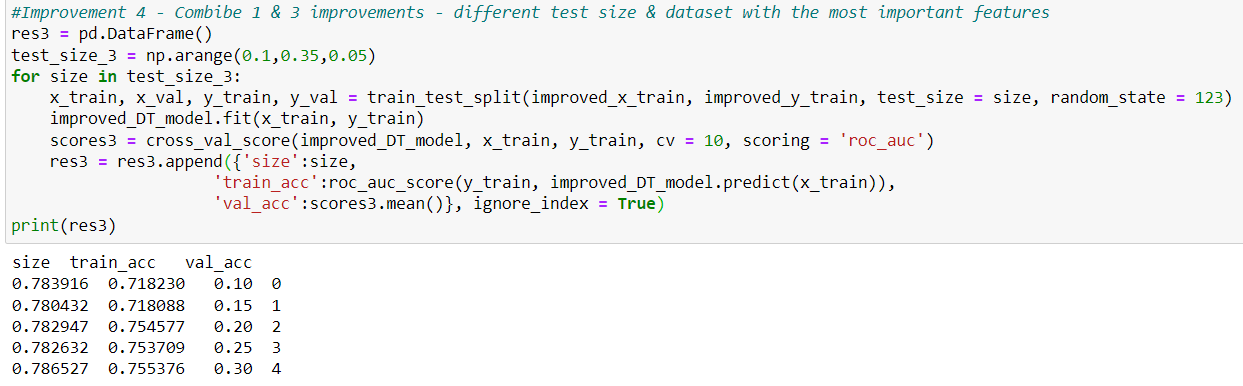
התיאור נוצר באופן אוטומטי

שיפור 3 – 9 הפיצ'רים בעלי החשיבות הגדולה ביותר

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

שיפור 4 – שילוב של 1 ו-3:



שיפור 5 – Random Forest וכוונון היפר פרמטרים

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי