**שיבוץ שחקנים ראשיים לסרט תוך   
מקסום רווחים מהסרט**

**פרויקט בבינה מלאכותית**

**מגישות:**

חן חבקוק ת"ז 312320997

טל עובד ת"ז 206129710

**מנחה:**

גיא קושילביץ

**הקדמה**

**רקע**

תעשיית הסרטים

קולנוע הוא תחום [אמנות](https://he.wikipedia.org/wiki/%D7%90%D7%9E%D7%A0%D7%95%D7%AA) העוסק ביצירתם ובהקרנתם של [סרטי קולנוע](https://he.wikipedia.org/wiki/%D7%A1%D7%A8%D7%98_%D7%A7%D7%95%D7%9C%D7%A0%D7%95%D7%A2).   
הקולנוע הוא מערכת תרבותית, במסגרתה נוצרים מאות ואלפי תוצרי תרבות הנצפים על ידי מיליונים ברחבי העולם. הקולנוע הוא בין השאר כלי [תקשורת המונים](https://he.wikipedia.org/wiki/%D7%AA%D7%A7%D7%A9%D7%95%D7%A8%D7%AA_%D7%94%D7%9E%D7%95%D7%A0%D7%99%D7%9D), ובעל השפעות גדולות על החברה והתרבות.   
תעשיית הקולנוע והסרטים היא המצרך האמנותי הפופולרי ביותר.

המילה "קולנוע" היא הלחם של המילים [קול](https://he.wikipedia.org/wiki/%D7%A7%D7%95%D7%9C) ו[תנועה](https://he.wikipedia.org/wiki/%D7%AA%D7%A0%D7%95%D7%A2%D7%94_(%D7%A4%D7%99%D7%96%D7%99%D7%A7%D7%94)) כאשר הדגש הוא על המילה הקול .   
מקורו בהתפתחות הטכנולוגית - הקולנוע בא לאחר ה[ראינוע](https://he.wikipedia.org/wiki/%D7%A8%D7%90%D7%99%D7%A0%D7%95%D7%A2) - הסרט האילם.

בזמן ה[ראינוע](https://he.wikipedia.org/wiki/%D7%A8%D7%90%D7%99%D7%A0%D7%95%D7%A2), הבדלי שפה שבין ארץ לאחרת לא היוו מגבלה מכיוון שהסרטים היו אילמים. מדי פעם, כשנוצר צורך, היו מקרינים במקביל לסרט כותרות (במעין שלט נלווה) וכשרצו להקרין סרט בארץ אחרת, החליפו את הכותרות לשפה המדוברת באותה ארץ.

הפקת סרטים היא יקרה מאד לכן ההשקעות הגדולות מתבצעות לרוב על ידי אנשים בעלי ניסיון. הסרטים מדורגים לרמות שונות, כשרמות גבוהות יותר פירושן השקעה גדולה יותר ויקרה יותר.

הקולנוע ההוליוודי

הקולנוע ה[הוליוודי](https://he.wikipedia.org/wiki/%D7%94%D7%95%D7%9C%D7%99%D7%95%D7%95%D7%93) החל את דרכו יחד עם הקולנוע עצמו.   
הוליווד עצמה הוקמה בתחילת שנות העשרים, כשיוצרי הראינוע נדדו מערבה. שם רכשה קבוצה מהם אדמה מבעליה, ששמם היה הוליווד, והחליטו להשאיר את השם. הוליווד הייתה במקום מושלם למגוון של לוקיישנים שונים. החל מהנוף ההררי מצפון, הים במערב, ההרים והמדבר במזרח, ועמקי קליפורניה הפוריים.

הקולנוע זכה לדחיפה גדולה בשל יצירתו של הקולנוען האמריקני ד.וו.גריפית שהתחיל את הקריירה הקולנועית שלו בתור שחקן, אך מהר מאוד התחיל גם לצלם ותוך זמן קצר גם קיבל הזדמנות לביים.   
בשנת[1915](https://he.wikipedia.org/wiki/1915)  ביים את סרטו הראשון "לידתה של אומה", סרט שיצר למעשה את שפת הקולנוע כפי שאנו מכירים אותה היום כמו: עריכה בתוך הסצנה, שימוש ב[קלוז-אפ](https://he.wikipedia.org/wiki/%D7%A7%D7%9C%D7%95%D7%96_%D7%90%D7%A4) וקצב עריכה שונה בהתאם לדרמטיות שבסצנה.   
סרטו השני, "אי סובלנות" כלל סצנות המונים ותפאורות בגובה עשרות מטרים, מצלמות שהועלו עם מעליות והמצאות רבות שהפכו את הסרט, שאורכו 163 דקות, לתצוגת איכות קולנועית כבר ב-1916.

קולנוען נוסף וידוע שהביא את הראינוע לפסגות חדשות היה [צ'ארלי צ'פלין](https://he.wikipedia.org/wiki/%D7%A6%27%D7%90%D7%A8%D7%9C%D7%99_%D7%A6%27%D7%A4%D7%9C%D7%99%D7%9F), אשר נחשב לאחת מהדמויות החשובות והמשפיעות ביותר בהיסטוריה של [תעשיית הקולנוע](https://he.wikipedia.org/wiki/%D7%AA%D7%A2%D7%A9%D7%99%D7%99%D7%AA_%D7%94%D7%A7%D7%95%D7%9C%D7%A0%D7%95%D7%A2). סרטיו של צ'פלין הביאו לעולם קולנוע חדש שלא נראה כמוהו, כוריאוגרפיה ייחודית ושימוש מבריק באביזרים ותפאורה.

תור הזהב של הוליווד

הוליווד ידועה כמרכז הפקות הסרטים העולמי. האולפן הראשון בהוליווד הוקם ב1911.  
בשנות השלושים ובשנות הארבעים, במהלך התקופה אשר זכתה לכינוי "עידן הזהב של הוליווד", הפיקו אולפני הסרטים בהוליווד בעיקר סרטי [מערבונים](https://he.wikipedia.org/wiki/%D7%9E%D7%A2%D7%A8%D7%91%D7%95%D7%9F), סרטי סלפסטיק (קומדיה משולבת בתנועות גופניות מכאיבות ומרעישות), סרטי "פילם נואר" (סרטי פשע, בעלי אווירה אפלה), סרטים מוזיקליים (מחזמר), סרטי אנימציה מצוירים ו[סרטים ביוגרפיים](https://he.wikipedia.org/wiki/%D7%A1%D7%A8%D7%98_%D7%91%D7%99%D7%95%D7%92%D7%A8%D7%A4%D7%99). באותה העת אלפי אנשים הועסקו על ידי אולפני הסרטים - שחקנים, מפיקים, במאים, כותבים, פעלולנים, אנשי מלאכה וטכנאים. בבעלותם של אולפני הסרטים היו מאות בתי קולנוע בערים שונות ברחבי ארצות הברית בהם הוצגו הסרטים שלהם. נקודת השיא של תעשיית הסרטים האמריקאית הייתה במהלך אמצע [שנות הארבעים](https://he.wikipedia.org/wiki/%D7%A9%D7%A0%D7%95%D7%AA_%D7%94-40_%D7%A9%D7%9C_%D7%94%D7%9E%D7%90%D7%94_%D7%94-20) כאשר אולפני הקולנוע הפיקו באותה העת כ-400 סרטים בשנה אשר הוצגו בפני קהל של 90 מיליון צופים אמריקאים מדי שבוע.

בשנת 1941 מוציא לאקרנים הבמאי [אורסון ולס](https://he.wikipedia.org/wiki/%D7%90%D7%95%D7%A8%D7%A1%D7%95%D7%9F_%D7%95%D7%9C%D7%A1) את הסרט [האזרח קיין](https://he.wikipedia.org/wiki/%D7%94%D7%90%D7%96%D7%A8%D7%97_%D7%A7%D7%99%D7%99%D7%9F) - סרט מופתי שנחשב לאחד החשובים בהיסטוריה של הקולנוע. חידושיו כללו שפה קולנועית עשירה, שבירה של מסגרת הזמנים המוכרת ובנייה של דמויות מורכבות ובעלות מספר צדדים.

לאורך השנים, התעשייה התפתחה, צברה ניסיון ושידרגה את היכולות הטכנולוגיות, המקצועיות והאומנותיות.

**רקע מהקורס**

ראינו בקורס כי תהליך החיזוי הוא תוצאה של תהליכי למידה שונים.  
למידה הינה תהליך, המקבל התנסויות כקלט, ומבצע שינויים בבסיס ידע במטרה לשפר, על פי מדד נתון, את היכולת של פותר הבעיות, המשתמש בבסיס הידע לפתור קבוצת בעיות.  
ישנם סוגי למידה שונים: כמו למידה מודרכת ((supervised learning, למידה לא מודרכת ((unsupervised learning, ולמידה מחיזוקים (reinforcement learning).  
בקורס התעסקנו בלמידה מודרכת.  
  
שלבי הלמידה המרכזיים:  
1. מידול הבעיה  
2. איסוף דוגמאות וייצוגן  
3. עיבוד מידע  
4. שימוש באלגוריתם למידה – מסווג  
5. הערכת ביצועים

מושגים:  
אובייקט – אלמנט מהסוג בו הבעיה הנדונה עוסקת  
תכונה של אובייקט – הינה מידע שמאפיין אותו

מושג – הינו תת קבוצה של מרחב האובייקטים  
מושג מטרה – המושג אותו נרצה ללמוד

סיווג של אובייקט – האם הוא שייך למושג המטרה

בשלב הראשון של הלמידה, מידול הבעיה נרצה להתאים את המושגים לבעייתנו הנתונה, לאחר מכן בשלב השני נרצה לאסוף דוגמאות רבות מהן נרצה ללמוד את מושג המטרה נמצא כמה שיותר דוגמאות מתויגות (דוגמא בצירוף הסיווג הנכון שלה) ונשתדל להימנע כמה שיותר מדוגמאות רועשות כלומר דוגמאות המופיעות בצירוף שגוי. את הדוגמאות ננסה לפרק לתכונות ולא להסתכל עליהן כעל אובייקטים אטומים כדי שנוכל להכליל מהדוגמאות שראינו, כלומר כל דוגמא תיוצג כווקטור תכונות.  
השלב השלישי, עיבוד מקדים של המידע, הינו שלב לא פחות חשוב משלב יצירת המסווג.  
בשלב זה נרצה ל"קרצף" ולערוך את הנתונים ככה שיתאימו לפורמט שאותו המסווג מצפה לקבל, וכן נרצה לתקן בעיות איכות הנובעות מחוסר מידע או ממידע לא מעובד מספיק.  
שלב זה בדרך כלל דורש הבנה והסקה טובה של הנתונים, וכולל למשל: השלמת ערכים חסרים, הסרת דוגמאות רועשות, נרמול נתונים, הורדת מימד ועוד.  
בשלב הרביעי, שימוש באלגוריתם למידה, נייצר את המסווג עצמו המקבל מאגר דוגמאות, לומד מהן על מנת לחזות ולפעול על דוגמאות שלא נמצאות במאגר.  
השלב האחרון נועד לבדוק את הדיוק של המסווג, כלומר עד כמה ניתן לסמוך על המסווג. את דיוק המסווג נבדוק ע"י קבוצת מבחן מופרדת לחלוטין ובה נשתמש רק פעם אחת.  
  
קיימות בעיות שונות, כמו בעיות בינאריות שמטרתן לקבוע האם אובייקט שייך לקבוצה או לא, ובעיות רגרסיה שמטרתן לנבא מספר ממשי (כפי שנתעסק בעבודתנו).

**הרעיון הכללי של הפרויקט (תיאור הבעיה):**

בעידן של היום, לאחר התפתחות רבה של תעשיית הסרטים, חשוב לתכנן כל סרט מראש במטרה למקסם את הרווחים שניתן להפיק ממנו.

ולכן, המטרה העיקרית בפרויקט: בהינתן סרט יחד עם תכונות (ז'אנר, במאי, תסריטאי, תקציב סרט..) המאפיינות אותו וכמות שחקנים נדרשת, נרצה לייצר מערכת המחזירה את שילוב השחקנים האופטימאלי כך שרווחי הסרט בקופות בתי הקולנוע ((box office יהיו מקסימאליים.

**דרך הפתרון**

את הבעיה שהצגנו החלטנו לפתור בשני שלבים מרכזיים:

1. מסווג -   
   המסווג יקבל כקלט סרט יחד עם התכונות הרלוונטיות המאפיינות את הסרט ויחזיר את הצפי לרווח הסרט בקופות בתי הקולנוע.
2. אלגוריתם -   
   האלגוריתם יקבל סרט עם תכונות המאפיינות אותו, וכמות שחקנים נדרשת ויחזיר את קבוצת השחקנים עבורה צפי ההכנסות מהסרט הוא מקסימאלי.

כעת נפרט:

שלב א' – בניית המסווג/רגרסור

1. השגת הנתונים

מאגר הנתונים הוא המידע עליו מתבסס כל הפרוייקט ולכן חשוב מאוד שיכיל נתונים טובים, ובנוסף שיכיל כמה שיותר סרטים – מה שיגדיל את רמת הדיוק המתבקשת.

ולכן, בשלב הראשון חיפשנו מאגרי נתונים מספיק גדולים שיכילו כמה שיותר תכונות שנראו לנו רלוונטיות לקביעת הרווחים העתידיים של סרט.   
לאחר חיפוש נרחב ברשת בחרנו בשני מאגרי נתונים מאתר Kaggle, שהכילו תכונות שונות ואיחדנו אותם למאגר נתונים אחד.   
בשלב זה מאגר הנתונים הכיל כ4800 סרטים שונים עם תכונות כגון:

* שפת מקור
* פופולריות הסרט
* חברת הפקות
* תאריך יציאה של הסרט
* רווחי הסרט
* אורך הסרט (זמן)
* ממוצע הציון שקיבל הסרט
* כותרת הסרט
* שחקנים
* הצוות (במאי, מפיק, תסריטאים...)
* תקציב הסרט
* ז'אנרים
* אתר הסרט
* מילות מפתח

בשלב זה עברנו על מאגר הנתונים, וידאנו בצורה רנדומלית חלק מהסרטים לאמת את המידע וניסינו להבין כמה מידע יהיה עלינו להשלים.  
בעיות שנתקלנו בהן בחלק זה:

* + מציאת מאגר נתונים עם כמות גדולה של סרטים שיכיל כמה שיותר נתונים שלדעתנו יעזרו לחזות את רווחי הסרט – קשה למצוא מאגר יחיד שיכיל את "כל מה שאנחנו צריכים" ולכן בסוף מצאנו שני מאגרים שהיה ניתן לשלב אותם יחסית בקלות וביצענו את השילוב.
  + מציאת מאגר נתונים עם כמה שפחות "חוסרי מידע" – במעבר ידני מאוד קשה לראות בכמה חוסרים מדובר, נרחיב על ההתמודדות עם שלב זה בחלק הבא.

1. קרצוף נתונים

מאגר הנתונים שהיה בידינו בשלב זה היה רחוק מלהיות מושלם לעבודה עם מסווג/רגרסור.   
לכן בשלב זה התחלנו לחפש ולכתוב קטעי קוד שיעזרו לנו להשלים את המידע החסר או לתקן מידע קיים (למשל – שמות של סרטים/שחקנים שהכילו תווים בלתי קריאים).

דברים שביצענו:

* השלמת ערכים חסרים של קרוב ל1000 סרטים של תקציב ורווח.
* נרמול ערכי התקציב, הרווח.
* בניית מאגרי שחקנים, במאים, מפיקים ותסריטאים מהמידע הקיים.
* בכל מאגר, נתינת "ציון" מספרי לכל בעל מקצוע. ציון המהווה ממוצע כל הרווחים של הסרטים בהם אותו בעל מקצוע השתתף.
* לבסוף איחדנו את כל המאגרים לכדי DB המכיל עבור כל סרט את:  
  שמו, "ציון" עבור השחקנים, הבמאים, המפיקים והכותבים, תקציב מנורמל ורווח מנורמל. כאשר לכל סרט יש ציון על פי ממוצע 3 השחקנים הראשיים ולכל שאר בעלי המקצוע יש ציון ממוצע של כלל בעלי אותו מקצוע שהשתתפו בסרט.

1. בחירת אלגוריתם + שלבי הניסויים

במטרה להגיע לרגרסור המיטבי מבחינת תוצאות עבור מאגר הנתונים שאיתו אנו עובדות, נרצה לבחון מספר רגרסורים כאשר על כל רגרסור נרצה לבחון התייחסות שונה לתכונות הקיימות, למשל:

* עבור תכונות כמו שחקן, במאי – האם כדאי לתת ערך לפי הסרט הרווחי ביותר שלקח בו חלק או האם כדאי לעשות ממוצע בין הסרטים בהם השתתף.
* עבור סרטים עם ערכים קיצוניים – האם כדאי להסיר אותם (outliers).
* שימוש בgrid – שתפקידו להריץ את הרגרסור עם פרמטרים שונים ולחפש את שילוב הפרמטרים שיתנו את הערכים האופטימליים.
* עבור תכונות דומות – האם כדאי לוותר על תכונה כלשהי (correlation matrix).
* ממוצע רגרסורים
* קבוצת ואלידציה

לשם בחינת התוצאות השתמשנו בשתי פונקציות שיתנו לנו אומדן להעריך את התוצאות  
שהתקבלו.

1. mean\_squared\_error – מחשבת את תוחלת ריבועי ההפרשים בערכי הרווחים בין הpredicted לבין הactual.

2. avg\_absolute\_percentage\_error – מחשבת את ממוצע אחוז ההפרש בין כל זוג ערכי predicted וactual.

בחרנו לעבוד עם ארבעה רגרסורים במטרה למצוא את הרגרסור בעל התוצאות הטובות ביותר:

1. RandomForestRegressor –

אלגוריתם זה מבוסס על שילוב של מס' עצי החלטה (מס' העצים הוא פרמטר לרגרסור).   
האלגוריתם דוגם אקראית תתי קבוצות שונות מתוך קב' האימון (עם החזרות/חפיפה בין דוגמאות בקב' השונות) ומאמן את כל אחת מהקבוצות על עץ החלטה. בסוף האלגוריתם משתמש בממוצע התוצאות שהתקבלו כדי לתת חוזי מדויק יותר לערך המתבקש.   
יתרונותיו של האלגוריתם הם:

* הקטנת רעשים (overfitting) – באמצעות כך שדוגמאות רועשות לא משפיעות על כל אחד מעצי ההחלטה ביער, מה שמגדיל את דיוק האלגוריתם.

1. DecisionTreeRegressor –   
   עץ החלטה הינו אלגוריתם המתבסס על חיפוש במבנה של עץ כאשר כל צומת בעץ מהווה פיצול
2. KNeighborsRegressor –   
   "אלגוריתם השכן הקרוב ביותר", הוא אלגוריתם לסיווג ולרגרסיה מקומית.   
   בשני המקרים הקלט תלוי ב-K התצפיות הקרובות במרחב התכונות.עבור רגרסור, בהינתן דוגמא חדשה, האלגוריתם מחזיר את ממוצע הערכים של K השכנים הקרובים ביותר. שיטת שקלול נוספת עבור K השכנים הקרובים ביותר היא באמצעות נתינת משקל של d\1 כאשר d הינו המרחק לאותו שכן. בנוסף קיימות מספר שיטות למדידת המרחק בין שני משנים שונים, כגון מרחק אוקלידי, מרחק המינג ועוד. וכך דוגמאות האימון מיוצגות כווקטור תכונות המרחב רב ממדי, כאשר כל אחד עם תווית סיווג כלומר בעבודתנו בא לידי ביטוי בערך של רווח הסרט. שלב האימון מתבסס על אחסון דוגמאות האימון במבנה נתונים שיאפשר חיפוש מהיר. והקלט הינו וקטור ללא סיווג (סרט לא רווח) ותפקיד הרגרסור לקבוע את רווח הסרט ע"פ ממוצע הרווח של K הסרטים הקרובים ביותר לסרט הנבדק.חיסרון בשיטה זו היא כאשר התפלגות דוגמאות האימון מוטה למשל במקרה שלנו אם לצור העניין רוב הסרטים בעלי רווח גדול יחסית, אזי רווח דוגמאות המבחן יחושב ע"פ ממוצע של הקרובים ביותר (כאשר רובם בעלי רווח גדול) וכך רוב הסרטים בדוגמאות האימון יקבלו ערך רווחי גבוה מידי.
3. LinearRegression -   
   רגרסור לינארי – משתמש בשיטת הרגרסיה הלינארית. זו שיטה מתמטית למציאת קשר בין משתנה בלתי תלוי X, למשתנה תלוי ,Yבהנחה שהקשר בניהם לינארי.בשיטה מחושב הקו הישר העובר הנקודות שבמדגם. במצב של קשר ישיר מדויק, כל נקודות המדגם ימצאו על הקו עצמו, אך בפועל גורמים נוספים משפיעים על המדגם והנקודות לרוב מפוזרות מסביב לקן ולא עליו.במקרה זה הקו מחושב בצורה כזו שסכום ריבועי המרחקים של נקודות המדגם מהקו יהיה הקטן ביותר. כמו כן, קיימת רגרסיה לינארית מרובה המחשבת קשר בין מספר משתנים בלתי תלויים יחד, למשתנה תלוי אחד, כפי שבעבודתנו הרגרסור הלינארי יחפש את הקו עבורו סכום ריבועי המרחקים של דוגמאות האימון ממנו, יהיה הקטן ביותר, וע"י כך ינסה להיות כמה שיותר מדויק ועקבי לדוגמאות האימון כאשר יקבל דוגמת מבחן ויצטרך להחזיר עבורה את הערך מספרי של חיזוי הרווחים מאותו הסרט.

**ניסוי ראשון**

בשלב זה היה בידינו מאגר הנתונים לאחר שלב סידור והשלמת כל הנתונים החסרים, להלן התוצאות לכל אחד מהרגרסורים. נזכיר כי בשלב זה ערכי השחקנים, במאים, מפיקים, ותסריטאים חושבו לפי ממוצע ערכים של הסרטים בהם השתתפו (בהמשך הניסויים נבחן גם את האפשרות לmax בין הסרטים).

כפי שניתן לראות בגרפים הpredict - אכן עוקב את ערכי ה- actual, כלומר ניתן לראות שסרט בעל רווח גבוה אכן מתאפיין בחוזי של רווח גבוה יחסית, אך גם ניתן לראות שהצפי רחוק מרמת הדיוק שהיינו רוצים לראות בשלב זה, כלומר הסטייה באחוזים אכן גדולה.  
לכן עברנו בחזרה לשלב "קרצוף הנתונים" לבדוק כי לא טעינו במהלך השלב, ולאחר שווידאנו זאת, הסקנו כי יכול להיות שקיימת בעיה בחוסר דיוק במאגר הנתונים אליה נתייחס בניסויים הבאים.

נסביר כעת על הגרפי כאשר ערכי ה-x מייצגים א הסרט עצמו (דוגמת האימון) וערך ה- y מייצג ת ערך הרווח המנורמל כאשר הגרף הכתום מייצג את מה שהרגרסור חוזה ואילו הגרף התכלת מייצג את מה שבפועל קורה, כלומר את התיוג של דוגמת האימון – ערך הרווח של הסרט בפועל.

נתאר כעת את התוצאות שהתקבלו לפי סדר הניסויים שביצענו:

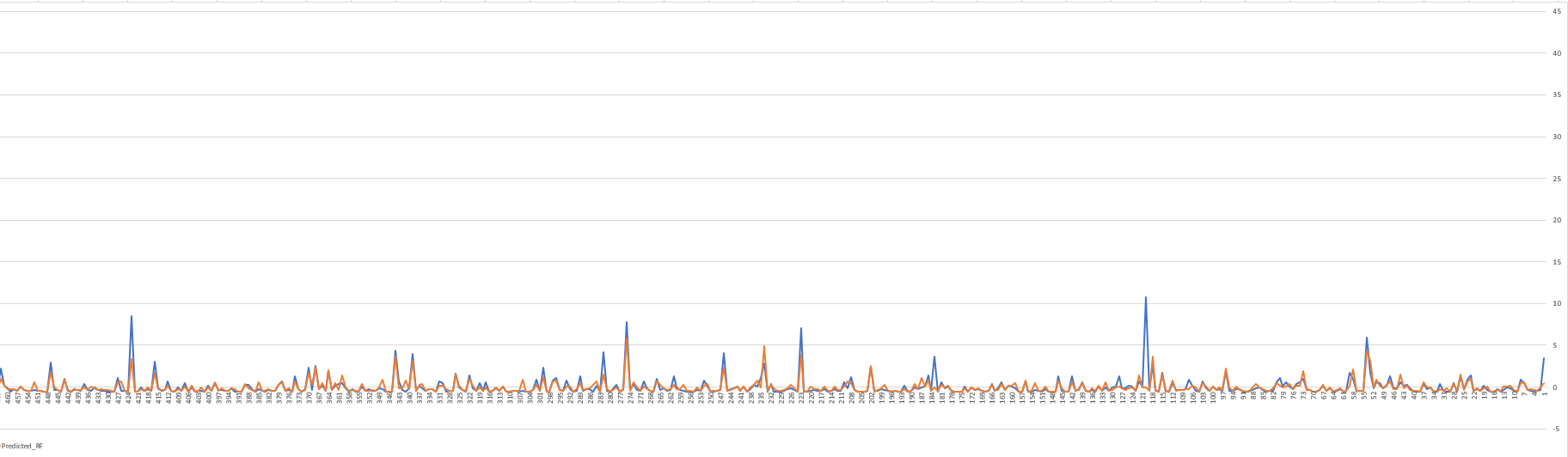
**Random Forest**

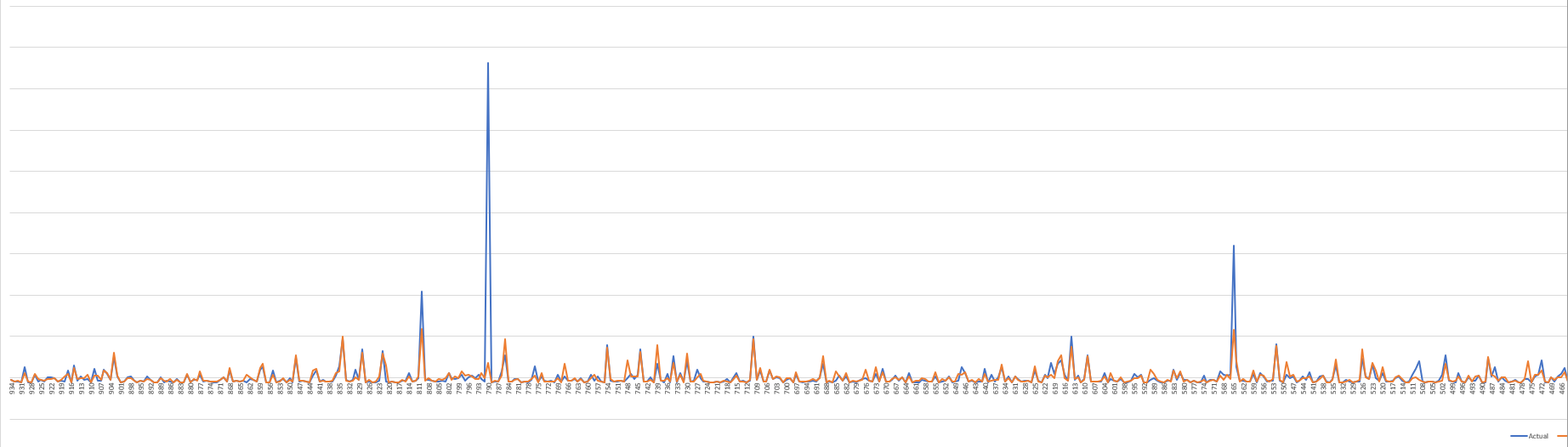
**Actual value  
  
Predicted RF**



ממוצע ריבועי ההפרשים בין התוצאות

ממוצע אחוזי הסטייה





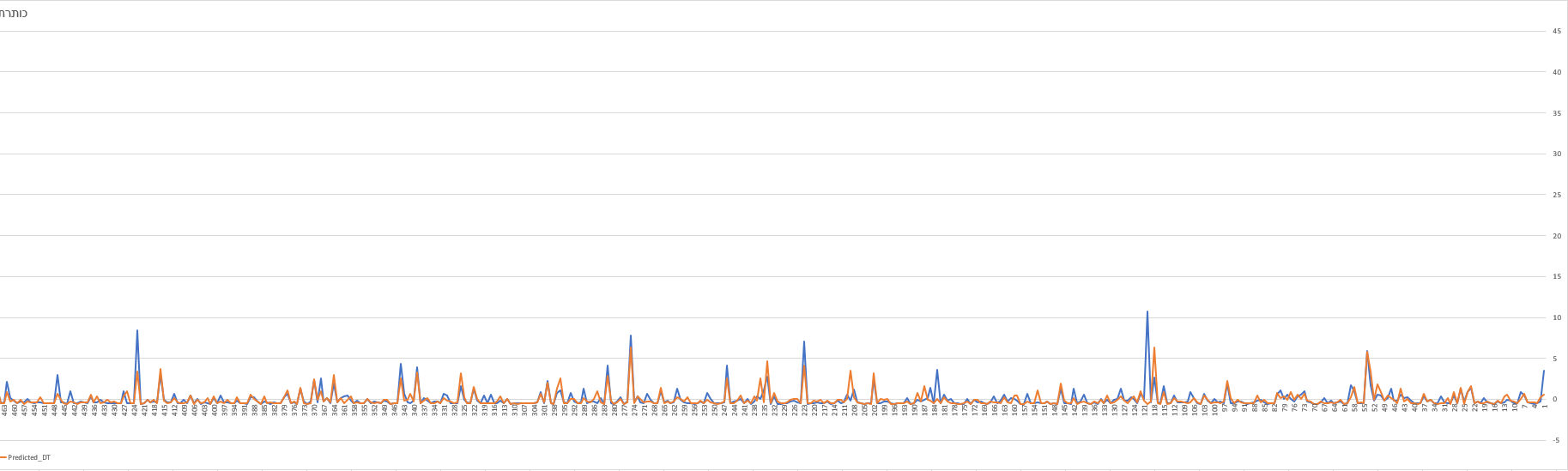
**Decision Tree**

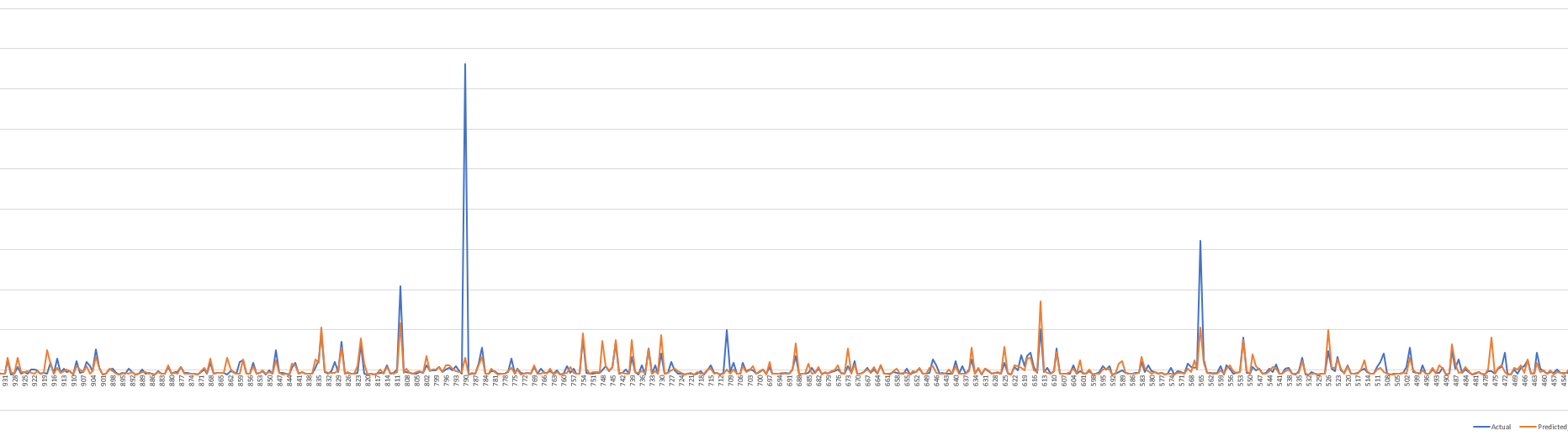


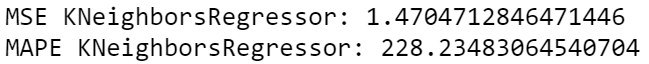
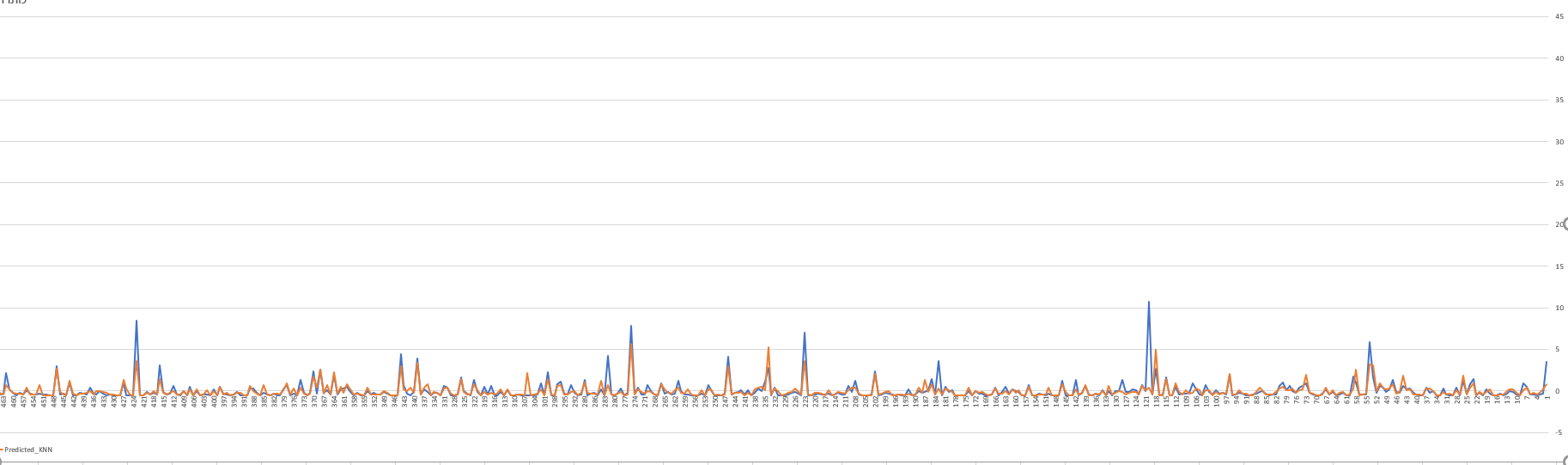
**Actual value  
  
Predicted DT**

ממוצע ריבועי ההפרשים בין התוצאות

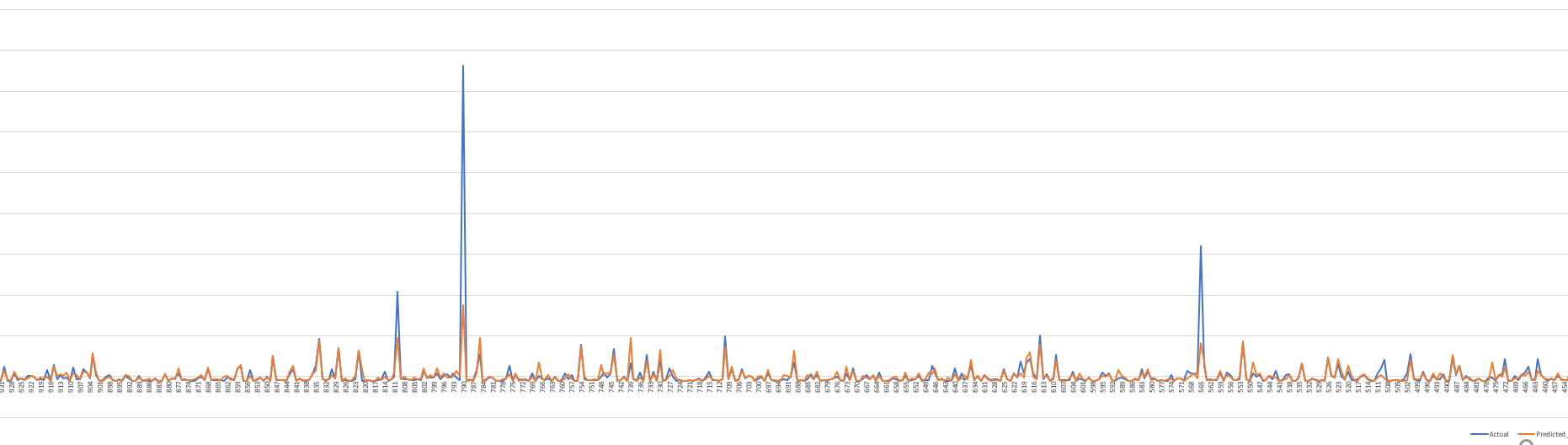
ממוצע אחוזי הסטייה







**Actual value  
  
Predicted RF**



ממוצע ריבועי ההפרשים בין התוצאות

ממוצע אחוזי הסטייה

**KNN**

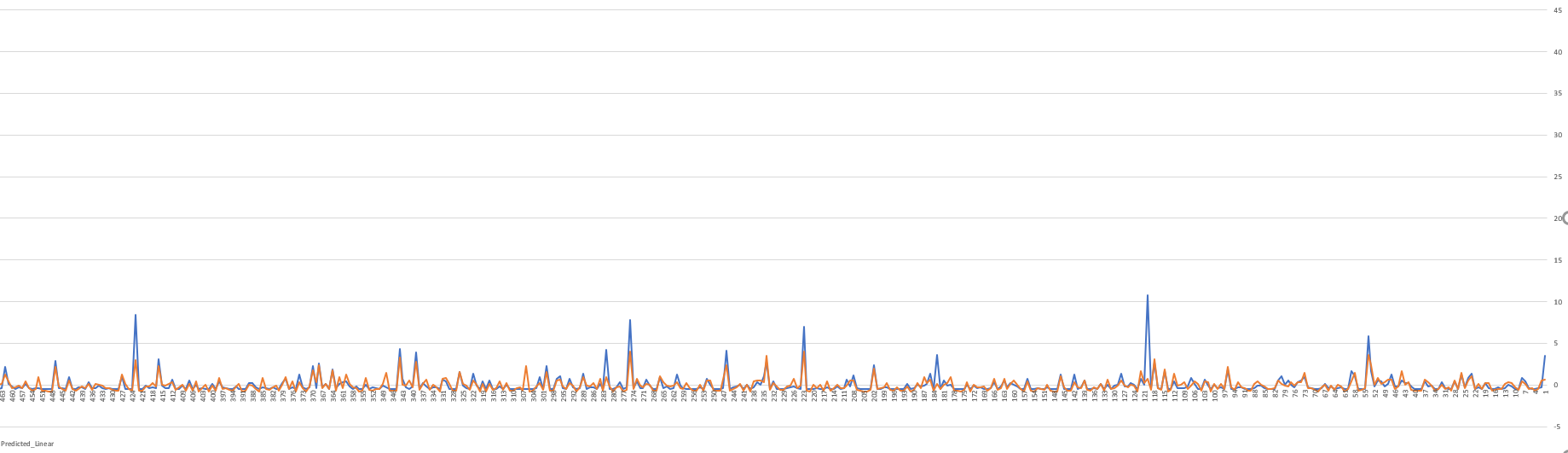
**Linear**

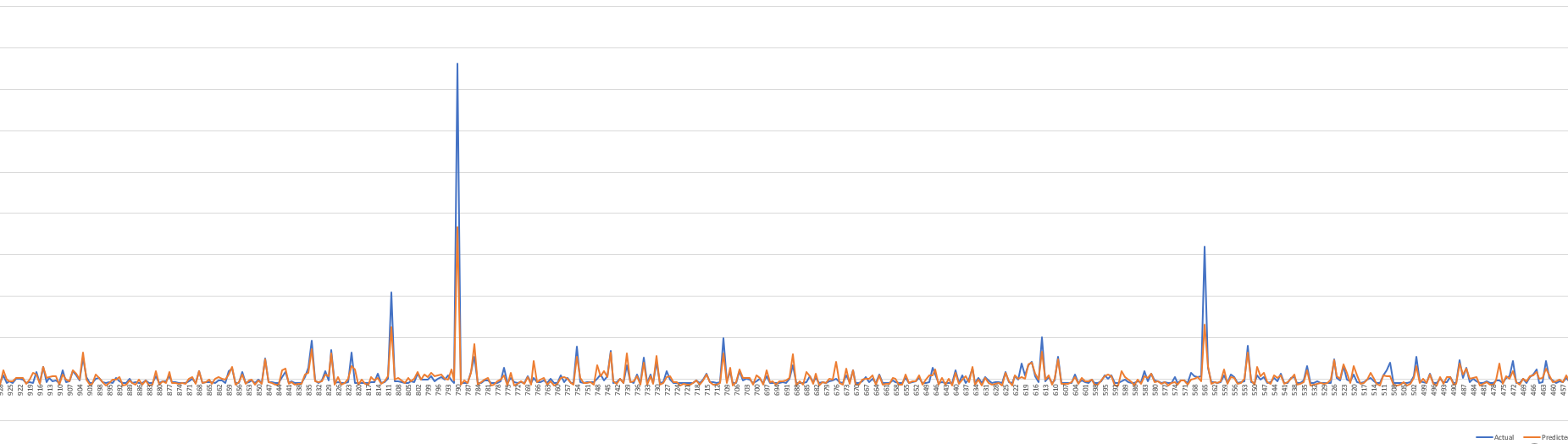


ממוצע ריבועי ההפרשים בין התוצאות

ממוצע אחוזי הסטייה

**Actual value  
  
Predicted RF**





**ניסוי שני**

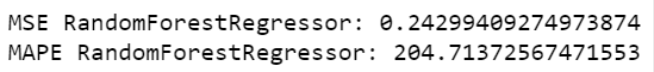
בשלב זה החלטנו לנסות להתמודד עם חוסר הדיוק באמצעות outliers, כלומר הסרת דוגמאות "רועשות", בהן ערך הרווח שהתקבל שונה באופן קיצוני מהרוב הגדול של הערכים או שחשוד שהתקבל עקב שגיאת מדידה או שגיאה בהבנת התופעה הנמדדת. לכן, הפעלנו את פונקציית הoutliers ובעקבות כך הורדו סרטים שהערכים שלהם היו חריגים. לאחר ביצוע הoutliers ניתן לראות ע"פ הגרפים כי ערכי הpredict – כפי שצפינו עדיין מתנהגים דומה לערכי ה- actual (הרגרסורים חוזים את התוצאות נכון מבחינת ערך גבוה יותר או נמוך יותר) כמו כן ע"פ נתוני הפונקציות שתיארנו לעיל ניתן להסיק כי אחוזי השגיאה ירדו משמעותית אך בכל זאת עדיין אינם קרובים להיות מדויקים.   
\*נשים לב לשינוי שחל בערכי הy של הגרפים לעומת בניסוי הראשון)

**תזכורת לנתונים מהניסוי הקודם:**

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:





ממוצע ריבועי ההפרשים בין התוצאות

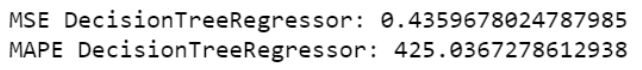
ממוצע אחוזי הסטייה

**תזכורת לנתונים מהניסוי הקודם:**



ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:



ממוצע ריבועי ההפרשים בין התוצאות

ממוצע אחוזי הסטייה



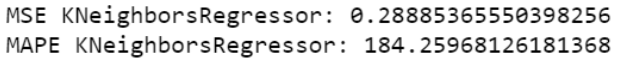
**תזכורת לנתונים מהניסוי הקודם:**

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

ממוצע ריבועי ההפרשים בין התוצאות

ממוצע אחוזי הסטייה



**תזכורת לנתונים מהניסוי הקודם:**

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:





ממוצע ריבועי ההפרשים בין התוצאות

ממוצע אחוזי הסטייה

**ניסוי שלישי**בשלב זה ביצענו grid שתפקידו להריץ את הרגרסורים עם פרמטרים שונים ולחפש את שילוב הפרמטרים שיתנו את הערכים האופטימליים, נציג את ההבדלים בתוצאות:

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

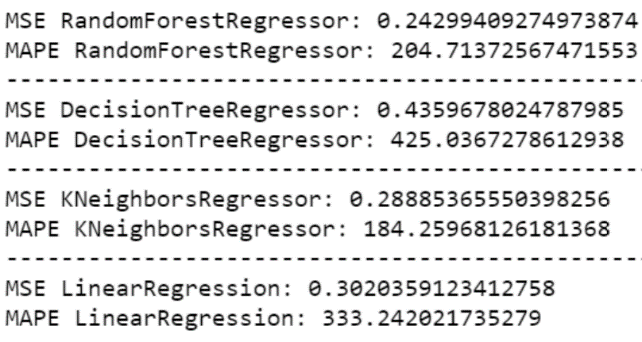
ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

**לאחר הgrid:**

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:



ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

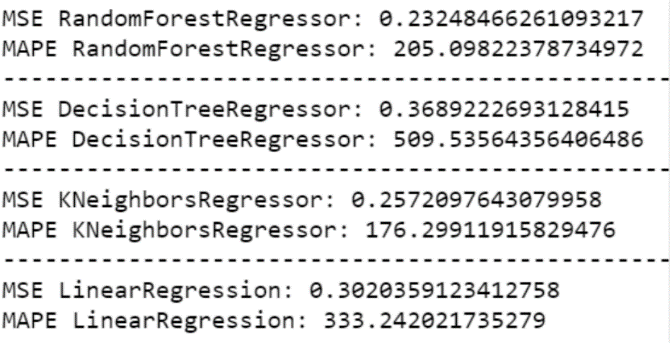
ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

**תזכורת לנתונים מהניסוי הקודם(2):**



ניתן לראות כי אחוזי השגיאה אכן ירדו, אך לא בהפרש גדול מספיק, ולכן המשכנו לניסוי הבא:

**ניסוי רביעי**

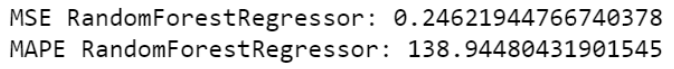
בשלב זה החלטנו לשנות את אופן חישוב הערכים בDB, כפי שהסברנו, יצרנו מאגר של שחקנים, במאים, מפיקים וכותבים.   
בהתחלה כל אחד קיבל ציון לפי ממוצע הרווחים של הסרטים בהם השתתפו. בשלב זה החלטנו לשנות את צורת החישוב הזו ולתת לכל בעל מקצוע את הציון שלו לפי הסרט בעל הרווח הגבוה בו השתתף וזאת מכיוון שלאחר מעבר על מאגר הנתונים, שמנו לב כי קיימים במאים, מפיקים וכותבים מפורסמים מצליחים המשתתפים בסרטים רווחים מאוד אך במאגר הנתונים גם מופיעים המון סרטים פחות מצליחים שלהם ולכן לעיתים הציון שקיבלנו קטן לעומת שחקנים אחרים המופיעים בפחות סרטים במאגר שהיו פחות מוצלחים – דבר שהוביל לחוסר דיוק.  
לאחר ששינינו את הDB קיבלנו את התוצאות שלהלן, ניתן ראות כי כפי שחזינו אכן הגדלנו את הדיוק ואחוזי הסטייה ירדו משמעותית.

**תזכורת לנתונים מהניסוי הקודם:**

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:





ממוצע ריבועי ההפרשים בין התוצאות

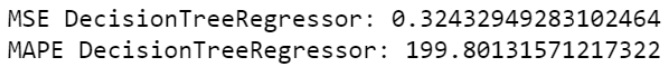
ממוצע אחוזי הסטייה

**תזכורת לנתונים מהניסוי הקודם:**

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:





ממוצע ריבועי ההפרשים בין התוצאות

ממוצע אחוזי הסטייה



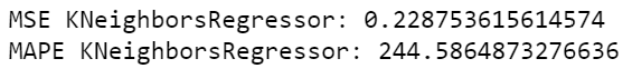
**תזכורת לנתונים מהניסוי הקודם:**

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

ממוצע ריבועי ההפרשים בין התוצאות

ממוצע אחוזי הסטייה

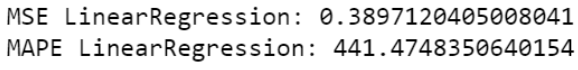


**תזכורת לנתונים מהניסוי הקודם:**

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:





ממוצע ריבועי ההפרשים בין התוצאות

ממוצע אחוזי הסטייה

**ניסוי חמישי**בשלב זה ביצענו שוב grid לבחירת הפרמטרים המיטביים, נציג את ההבדלים בתוצאות:

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

**תזכורת לנתונים מהניסוי הקודם (4):**

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

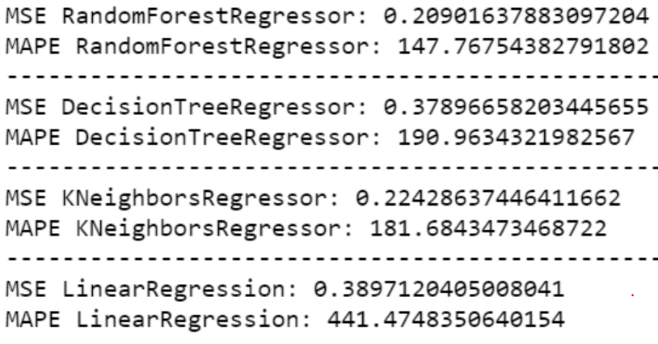
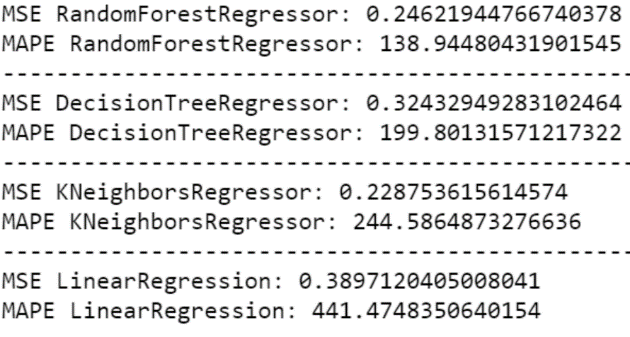
ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

**לאחר הgrid:**



ניתן לראות כי אחוזי השגיאה אכן ירדו, אך לא בהפרש גדול מספיק, ולכן המשכנו לניסוי הבא:

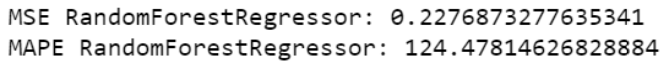
**ניסוי שישי**בשלב זה חזרנו שוב לשלב קרצוף הנתונים במטרה לבדוק מה יכול לשפר משמעותית את תוצאות הניסוי. הבחנו כי ערכי הבמאים, מפיקים, וכותבים גבוה משמעותית מערכי השחקנים. (בהתחשב בעובדה כי ערך השחקנים אינו פחות חשוב ואף יותר). ולכן החלטנו לנסות לשנות את ערכי השחקנים שיהיו גם הם ע"פ הסרט הרווחי ביותר ולא ע"פ ממוצע כל הסרטים שאותו שחקן השתתף בהם. ולכן ביצענו באופן דומה לניסוי הרביעי את לקיחת ערך הmax, רק הפעם לשחקנים – כלומר, כל שחקן קיבל את הערך שלו לפי הסרט הרווחי ביותר בו השתתף.  
ניתן לראות כי אכן אחוזי השגיאה ירדו משמעותית מה שעולה בקנה אחד עם העובדה שהשחקנים מאוד משפיעים על רווח הסרט, במציאות בכלל וע"פ למידת הרגרסורים בפרט.

**תזכורת לנתונים מהניסוי הקודם:**

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:





ממוצע ריבועי ההפרשים בין התוצאות

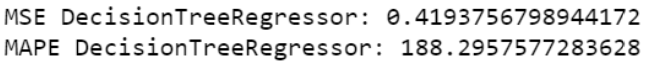
ממוצע אחוזי הסטייה

**תזכורת לנתונים מהניסוי הקודם:**

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:





ממוצע ריבועי ההפרשים בין התוצאות

ממוצע אחוזי הסטייה



**תזכורת לנתונים מהניסוי הקודם:**

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

ממוצע ריבועי ההפרשים בין התוצאות

ממוצע אחוזי הסטייה



**תזכורת לנתונים מהניסוי הקודם:**

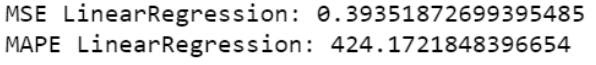
ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:



ממוצע ריבועי ההפרשים בין התוצאות

ממוצע אחוזי הסטייה



**ניסוי שביעי**בשלב זה ביצענו grid לבחירת הפרמטרים המיטביים, נציג את ההבדלים בתוצאות:

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

**תזכורת לנתונים מהניסוי הקודם (6):**

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

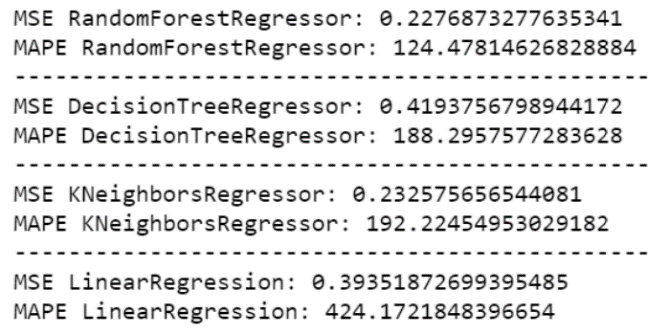
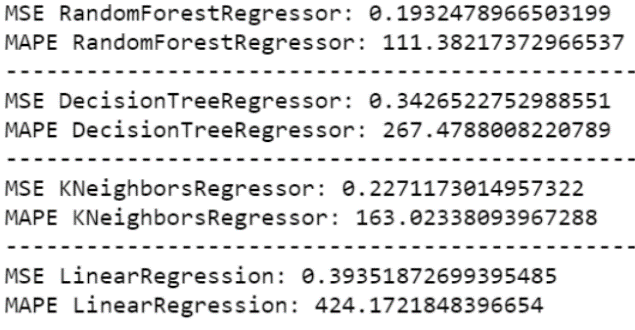
ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

**לאחר הgrid:**



**ניתן לראות שיפור בחלק מהרגרסורים**

**לאחר שלב זה חשבנו על ניסוי נוסף:**כפי שהסברנו לעיל ערך המפיקים, כותבים ובמאים, מורכב ממוצע/max של כל אותם בלי מקצוע המשתתפים באותו הסרט, ואילו ציון השחקנים לסרט ניתן ע"י ממוצע של שלושת השחקנים הראשיים, לכן חשבנו להגדיל או להקטין את כמות השחקנים איתם נרצה לחשב את ערך השחקן עבור כל סרט, אך נתקלנו בבעיה ששחקנים שהופיעו ברשימה כמספר 4 הורידו משמעותית את ערך השחקנים עבור סרטי קלאסיקות המצליחות בכל הזמנים ולכן החלטנו שזה מוריד את הדיוק וויתרנו על אופציה זו.

**ניסוי שמיני**

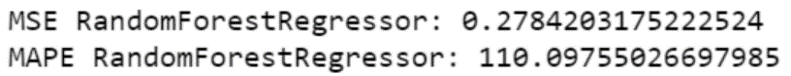
בשלב זה חקרנו שוב את מאגר הנתונים וראינו כי ערך הבמאים והמפיקים מאוד קרוב אחד לשני ולכן השתמשנו ב- correlation matrix על מנת לבדוק האם קיימות תכונות מיותרות (מיותרות בגלל שהן דומות אחת לשנייה ונותנות השפעה כפולה, מה שעלול לגרום לחוסר דיוק).  
אמנם כמות התכונות אינה רבה מראש, אך החלטנו כן לנסות ולראות אם יש שיפור באחוזי הדיוק.   
ניתן לראות כי אחוזי השגיאה קטנו כפי שצפינו ולאחר הgrid הגענו לאחוזי הסטייה המינימאליים ברגרסור \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
לכן החלטנו ללכת על \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**תזכורת לנתונים מהניסוי הקודם:**

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:





ממוצע ריבועי ההפרשים בין התוצאות

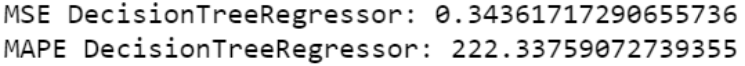
ממוצע אחוזי הסטייה

**תזכורת לנתונים מהניסוי הקודם:**

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:





ממוצע ריבועי ההפרשים בין התוצאות

ממוצע אחוזי הסטייה

**תזכורת לנתונים מהניסוי הקודם:**

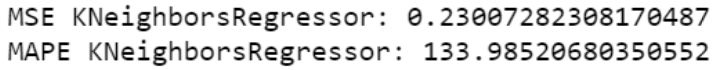
ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:



ממוצע ריבועי ההפרשים בין התוצאות

ממוצע אחוזי הסטייה

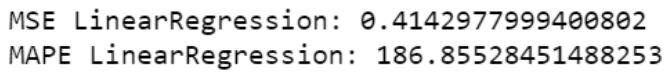


**תזכורת לנתונים מהניסוי הקודם:**

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:





ממוצע ריבועי ההפרשים בין התוצאות

ממוצע אחוזי הסטייה

**ניסוי תשיעי (ואחרון!)**בשלב זה ביצענו grid לבחירת הפרמטרים המיטביים, נציג את ההבדלים בתוצאות:

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

**תזכורת לנתונים מהניסוי הקודם (8):**

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

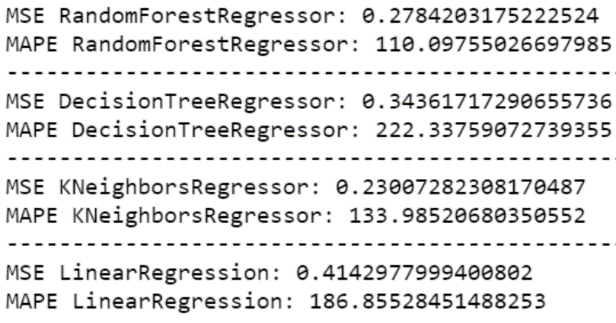
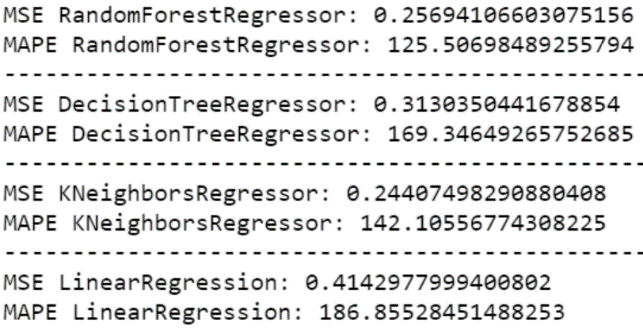
ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

ממוצע ריבועי ההפרשים בין התוצאות:

ממוצע אחוזי הסטייה:

**לאחר הgrid:**



לסיכום הסקנו כי אכן הניסויים שביצענו הצליחו לדייק את הרגרסור יותר ויותר וכי התצפיות היו תאומות לתוצאות, אך עם זאת הדיוק של הרגרסור מלכתחילה היה רחוק מהמציאות, השערה שלנו היא שמאגר הנתונים לא היה מדויק דיו, וכי הופיעו המון דוגמאות רועשות שהגדילו משמעותית את חוסר הדיוק.

שלב הניסויים

* משקלים שונים לתכונות
* הסרת דוגמאות קיצוניות (outlier)
* הסרת דוגמאות זהות (correlation matrix)
  + תוצאות, טבלאות

1. בחינת המודל

שלב ב' - אלגוריתם

1. בדיקה מיטבית של שלשות שחקנים  
   דברים שיתנו בשררררררר –
2. ז'אנרים – למאגר שחקנים להוסיף לכל שחקן את הז'אנרים הרלוונטיים אליו.
3. תקציב – בהתאם לתקציב ללהק שחקן, כלומר במאגר שחקנים לאמוד את התקציב לשחקן.

**סיכום**

**ביבליוגרפיה**

* המושג [קולנוע](https://he.wikipedia.org/wiki/%D7%A7%D7%95%D7%9C%D7%A0%D7%95%D7%A2) (ויקיפדיה).
* מאגר נתונים הנתון כקובץ .csv של 4800 סרטים ומאפיינים עליהם כולל שחקנים הלקוח מאתר Kaggle - <https://www.kaggle.com/tmdb/tmdb-movie-metadata/data>
* מאגר נתונים הנתון כקובץ .csvשל 2500 סרטים ומאפיינים עליהם כולל ז'אנר, חברה מפיצה, רווחים (Box office) הלקוח מאתר Kaggle - <https://www.kaggle.com/yjeong5126/box-office-data-20172019>
* השלמת מידע חסר למאגר הנתונים – ויקיפדיה, IMDB.
* מידע על האלגוריתמים:
  + [Random forest](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html)
  + [עצי החלטה](https://gdcoder.com/decision-tree-regressor-explained-in-depth/)
  + KNN, רגרסור לינארי - ויקיפדיה