**Support Vector Regression (SVR)**

SVR הוא מודל רגרסיה המבוסס על Support Vector Machines (SVM), אשר מתעלם משגיאות קטנות בתוך טווח מסוים (Epsilon Insensitive Tube) ומתמקד במציאת קו רגרסיה מיטבי. רק נקודות מחוץ לטווח ϵ\epsilonϵ הופכות ל-Support Vectors, והן אלו שמשפיעות על המודל.

המודל מעניש רק שגיאות גדולות, מה שמאפשר לו להיות עמיד בפני רעש.

למה צריך לנרמל את הנתונים?

אלגוריתם SVR רגיש מאוד לסקאלה של הנתונים, ולכן מומלץ לנרמל אותם באמצעות StandardScaler, כך שיתקבל: ממוצע = 0 ו- שונות = 1. כך כל התכונות יקבלו משקל שווה ולא תהיה השפעה חזקה יותר לתכונות מסוימות בשל יחידות המדידה שלהן.

איך SVR עובד?

1. נבנה היפר-מישור (Hyperplane) שמתאים לנתונים.
2. נוצר "צינור" סביב קו הרגרסיה שמתעלם משגיאות קטנות.
3. רק נקודות מחוץ לצינור משפיעות על מבנה המודל.
4. ניתן להשתמש בגרעינים (Kernels) כדי להתמודד עם קשרים לא ליניאריים.

פרמטרים מרכזיים של SVR

1. Epsilon (𝜀) – קובע את "עובי הצינור"

עבור 𝜀 קטן -> הצינור יהיה צר מאוד, והמודל ינסה להתאים את עצמו לנתונים בצורה יותר קשיחה.

עבור 𝜀 גדול -> הצינור יהיה רחב יותר, והמודל יתעלם משגיאות קטנות יותר.

2. C – קובע את מידת הענישה על נקודות מחוץ לצינור

עבור C גבוה -> המודל מעניש בחומרה חריגות, מה שעלול לגרום לאוברפיטינג.

עבור C נמוך -> המודל יהיה גמיש יותר אך עשוי לא לתאר את הנתונים במדויק.

3. Kernel – גרעין לקביעת צורת הצינור קובע כיצד הנתונים יוצגו במרחב אחר, מה שמשפיע ישירות על צורת הרגרסיה. לכל גרעין יתרונות שונים בהתאם למבנה הנתונים:

* Linear Kernel – מתאים כאשר יש קשר ליניארי חזק בין הנתונים. יוצר קו רגרסיה ישר ומדויק לבעיות פשוטות.
* Polynomial Kernel – מאפשר יצירת רגרסיה גמישה יותר, מתאימה לקשרים מורכבים ולא ליניאריים בין התכונות.
* RBF (Radial Basis Function) Kernel – מתחשב בקשרים לא ליניאריים בצורה חזקה ומשתמש בהשלכה למימד גבוה. לרוב מתאים לרוב הבעיות הלא ליניאריות, אך דורש כוונון זהיר כדי למנוע התאמת יתר (overfitting).

ניסיונות לשימוש באלגוריתם:

במהלך הפרויקט ניסינו שני אלגוריתמים שונים לחיזוי מחירי הטיסות: SVR עם גרעין RBF ו- LinearSVR. למרות הניסיונות לכוונן את המודלים ולשפר אותם, התוצאות שהתקבלו היו גרועות.

ניסיון ראשון - SVR עם גרעין RBF: האימון נמשך יותר מ-50 דקות, ולכן בדקנו את סיבוכיות הזמן ומצאנו שהיא בין O(N^2) ל O(N^3) – לא פרקטי לדאטה גדול.

ניסיון שני - ניסינו לשפר את הביצועים באימון ב-Batches – חלוקה למקטעים, אימון כל מקטע בנפרד ושמירת וקטורי התמיכה בלבד.

גם אימון ב-Batches לא פתר את הבעיה, ולכן הפסקנו את התהליך באמצע.

#הסיבה שזה לא עבד היא שגרעין RBF מתאים לזיהוי קשרים לא ליניאריים מורכבים, אבל הוא אינו מתאים לדאטה גדול בגלל סיבוכיות חישובית גבוהה.

ניסיון שלישי - :LinearSVR עברנו למודל LinearSVR, שהוא פשוט יותר, עם גידול ליניארי בסיבוכיות. אימנו את המודל וביצענו חיזוי, אך קיבלנו ביצועים חלשים מאוד שמראים שהמודל כמעט לא מסביר את השונות במחירי הטיסות ויש פערים גדולים בין התחזיות לערכים האמיתיים.

ניסיון רביעי - כוונון היפר-פרמטרים (Hyperparameter Tuning):

השתמשנו ב- GridSearchCV לחיפוש הערכים האופטימליים של C ו-אפסילון אך גם כאן לאחר אימון מחדש עם הפרמטרים האופטימלים, הביצועים כמעט לא השתפרו.

אמנם LinearSVR עובד טוב עבור דאטה גדול, אבל המודל הליניארי מתקשה להתמודד עם קשרים לא ליניאריים בין המשתנים. ובגלל שהתכונות שלנו כנראה לא מסבירות היטב את המחירים, המודל מתקשה לחזות.