scikit learn				our code				ACC/MSE and Time
After optimizing (fit)		Before optimizing		After optimizing		Before optimizing		
Acc/mse	time	Acc/mse	time	Acc/mse	time	Acc/mse	time	
88.4%	42 sec	82.7%	0.11 sec	86.1%	10.9 sec	83%	2.4 sec	**DecisionTreeClassifier**
5.7	13.1 sec	8.4	0.16 sec	5.9	8.4 sec	6.78	2.52 sec	**DecisionTreeRegressor**
87.6%	33.5 sec	85.2%	0.21 sec	86.5%	96 sec	82.3%	2.85 sec	**RandomForestClassifier**
5.2	92.4 sec	5.9	0.12 sec	5.31	83.5 sec	6.4	2.1 sec	**RandomForestRegressior**

מסכנות והסברים :

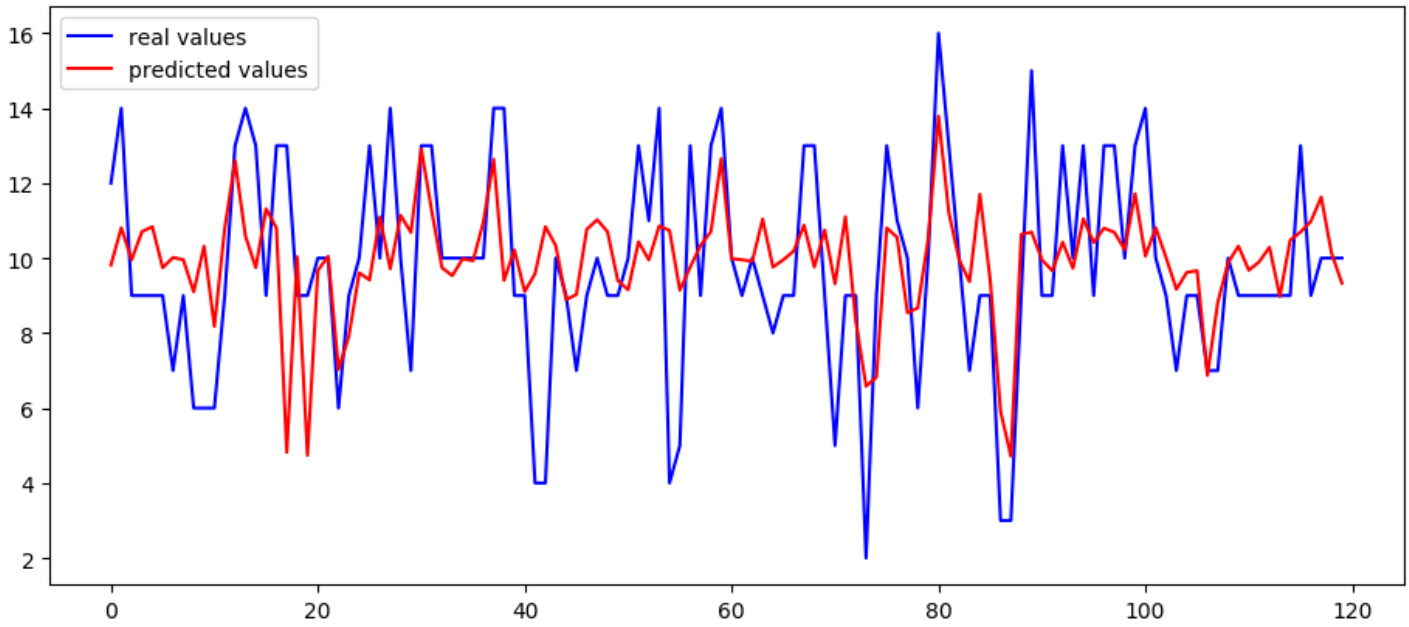
1. הבדל זמני ריצה :

- הקוד ב SKLEARN כתוב ב C ו C++ לכן סביר להניח שזמן הריצה יהיה יותר טוב .
- מכיוון שהקוד כתוב ב C יש אפשרות יותר לגשת לזיכרון ולהקצות מקום בזיכרון . למשתנים , מטריצות , מערכים , טבלאות גיבוב וזה חוסך בעלויות החיפוש , המחיקה או כל מניפולציה על המשתנים , ולפעמים מקצר או מבטל שימוש בלולאות .
- יש לאלגוריתמים בספריית SKLEARN אפשרות לעבוד באופן מקבילי שזה מייעל את זמן הריצה .
- בכמה מקרים בדידים (בעצים) ניתן לראות שיש לנו רמת דיוק קצת יותר טובה וזה בוודאות כי הפעלנו את הקוד כמה פעמים והשתמשנו מההתחלה בערכים של פרמטרים שנתנו תוצאות יותר טובות (למידת בני אדם) סוג של אופטימיזציה ידנית .

2. בגדול רואים שאנחנו מפסידים מול SKLEARN מבחינת דיוק וזה יכול לנבוע מכמה סיבות :

- לא מימשנו CROSS VALIDATION בשלב האופטימיזציה בקוד שלנו .
- כמות הפרמטרים שהשתמשנו בהם באופטימיזציה של הקוד שלנו היו פחות מהפרמטרים שהשתמשנו בהם באופטימיזציה של קוד הקוד ב SKLEARN בגלל שאצלנו ה GRIDSEARCHCV עבד יותר לאת והיה קשה לנסות עוד ערכים שונים של הפרמטרים (היה חשוב לנו להציג שהאופטימיזציה אצלנו עובדת) .
- השתמשנו בעומקים קצת גבוהים ברוב המודלים שלנו כיוון שנתנו תוצאות טובות , יכול להיות שבשלב מסוים היה לנו קצת OVERFIT לנתונים והיה כדאי לעשות גיזום .

Our random forest regressor (first 120 test) :



הערות :

- האופטימיזציה עובדת בזמן סביר אצלנו כי הפעלנו אותה לשעות וסיכמנו את התוצאות של הפרמטרים הכי טובים וכרגע מפעילים אופטימיזציה על מדגם מהפרמטרים שנתנו את התוצאות הכי טובות כדי שירוץ לך בשמן סביר

- עבור כל מודל שבנינו בחלק B יש 3 שלבים שרצים אוטומטית כשמריצים את הקוד (לכל שלב יש דיווח זמן ריצה ותוצאות) :
  1. בניית המודל הבסיסי - training the model
  2. אופטימזציה למודל בעזרת GridSearchCV וסט הווליידציה ודיווח על הפרמטרים הכי יעילים מימשנו באופן עצמאי את GridSearchCV , ניתן למצוא את הקלאס בחלק A(כמובן ששונה מהמימוש של SKLEARN אבל עושה את העבודה שלו)
  3. בניית מודל עם הפרמטרים שמצאנו בסעיף קודם

- זמני הריצה תלויים ביכולת עיבוד המחשב ובעוד גורמים רנדומליים לכן סביר להניח שיהיה שינוי
- מצורף קובץ בשם algo_log שבו ניתן לראות קלט דוגמה לריצת הקוד
- כשמריצים את C מופיעות לנו כמה הערות באדום משום מה אבל זה לא מפריע לריצת הקוד נשמח אם לא תתייחס לזה .

רוב הקוד נכתב באופן עצמי , חלק מפונקציות העזר נלקחו מ :
https://github.com/SebastianMantey/Decision-Tree-from-Scratch

נשמח לענות לכל שאלה בנוגע ללוגיקה של הקוד \ המחלקות השונות

Lashka03@campus.haifa.ac.il

Samar.kardosh@gmail.com