הצעת פרויקט גמר - קורס למידת מכונה בהנחיית פרופ' ליעד גוטליב

איילת קטקוב 325408409, רואי סיבוני 214662439

בסיס הנתונים: EuroSAT

סיווג תמונות לווייניות מהלוויין Sentinel2 ל10 קטגוריות (יער, נהר, שטח מגורים עירוני, שדות חקלאיים, אזור תעשיה (מפעלים) וכו'). מכיל כ27,000 תמונות מקוטלגות.

קישור לעמוד github בו יש קובץ README המסביר את בסיס הנתונים:

phelber/EuroSAT: EuroSAT: Land Use and Land Cover Classification with Sentinel-2

קישור ההורדה של שתי הגרסאות נמצא פה באתר:

EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification

למה בחרנו את בסיס הנתונים הזה?

יש אלמנט מרתק בתמונות לוויניות, שאופן העבודה הלוגית איתן צריך להיות שונה מעט מעם natural images, כי בין פיקסל לפיקסל יש מרחק של עשרות מטרים לכן צריך להסתמך פחות על מידע מהפיקסלים השכנים, ויותר אינדיבידואלית ולוגית.

שאלות:

1. האם גם מודלים פשוטים יכולים להגיע לתוצאות טובות על בסיס הנתונים?
2. אנחנו רוצים לראות איך צריך לשנות ולעדכן מודלים כדי להתאים אותם לעבודה על תמונות לוויניות (גם טכנית גם לוגית).
3. לראות אילו מודלים נותנים ביצועים טובים יותר כשהם מופעלים על תמונות, ומי מהם יעילים יותר.
4. לראות האם בסיס הנתונים מאוזן יחסית בין הקטגוריות ואיך זה משפיע על פעילות המודלים השונים.
5. האם ישנם קלאסים מסוימים (קטגוריות) שהמודלים מתקשים בזיהוי שלהם באופן חריג? אם כן מה זה אומר לנו עליהם ואיזה גישות ניתן לנסות כדי לשפר את הביצועים למרות הכל.
6. האם כל מודל לגופו, יצליח להבדיל בין תמונות מקטגוריות שההבדל ביניהן יכול להיות מאתגר אפילו לעין האנושית? (אזור עירוני לעומת תעשייתי, נהר לעומת אגם, שדות חד שנתיים לעומת שדות רב שנתיים וכו').
7. האם יש צורך בהורדת מימדים dimensionality reduction עבור בסיס הנתונים, האם זה משפר ביצועים?
8. האם רואים סימנים לoverfitting בהרצת המודלים השונים? אם כן מה זה יכול להגיד לנו באותו מצב.

כלים - מודלים:

1. Logistic regression רגרסיה לוגיסטית, מותאמת לעבודה עם קטגוריות מרובות (שימוש בsoftmax).
2. Random Forest.
3. SVM – support vector machine.
4. KNN – שכן קרוב.
5. רשת נוירונים מבוססת קונבולוציה – CNN.