

תרגיל 2 – שיטות בלמידת מכונה

מגישה: דנה אבירן

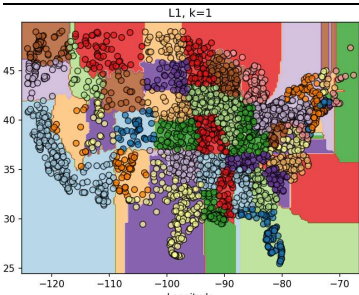
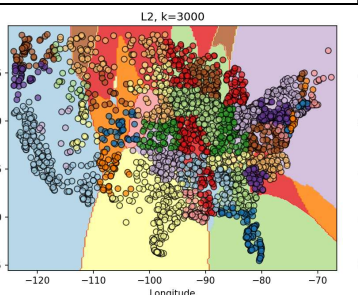
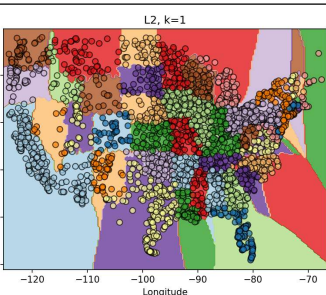
5.1 K Nearest Neighbor

להלן טבלת תוצאות הדיוק לכל k שכנים ולכל פונקציית מרחק:

K	L1 distance metric	L2 distance metric
1.0	0.967	0.9667
10.0	0.9617	0.9577
100.0	0.9231	0.9201
1000.0	0.745	0.7417
3000.0	0.4018	0.3981

5.2 - תשובות לשאלות:

- עבור שני המודלים מתקיימת המגמה כי ככל שהערך של מספר השכנים k גדל, כך גדל הדיוק. מגמה זו עקבית בשני המודלים הנ"ל, אשר להם פונקציות מרחק שונות. הסבר אפשרי הוא שככל שהמודל בעל ערך k גבוה יותר, כך הוא זקוק ל- train set גדול יותר על מנת שהייצוג של ה-labels יהיה פרופורציונלי ל- k הנבחר. אחרת, המודל יבצע התאמת-יתר כך שיסווג רוב כולל של הדוגמאות ב-test ל-labels שלהם דוגמאות רבות ב-train.
- ויזואליזציה של גבולות ההחלטה באמצעות הפונקציה הנתונה לנו:

העץ בעל הדיוק הגבוה ביותר כאשר משתמשים בפונקציית המרחק $L1$: $k=1$. הדיוק שלו הוא 0.967.	העץ בעל הדיוק הגבוה ביותר בפונקציית המרחק $L2$: $k=3000$. הדיוק שלו הוא 0.3981.	העץ בעל הדיוק הגבוה ביותר כאשר משתמשים בפונקציית המרחק $L2$: $k=1$. הדיוק שלו הוא 0.9667.
		

- הסתכלו על שתי הויזואליזציות עבור מודלים של פונקציית מרחק $L2$, עבור $k=1$, $k=3000$ בהתאמה.

a. מה ההבדל בדרך שבה כל אחד מהמודלים מבצע חלוקה של המרחב?

נסמן את המודל $k=1$ ב-A ואת המודל $k=3000$ ב-B. ניתן לראות שבמודל A המרחב מתחלק לאזורי צבעים רבים יותר. כמו כן, כל אזור מוגדר על שטח קטן וריבועי יותר. בנוסף, הנקודות על המפה בצבע מסוים נמצאות לעיתים קרובות יותר תחת שטח באותו הצבע. לעומת זאת, מודל B מבצע הכללה גסה של חלוקה למספר קטן משמעותית של

אזורים שונים, הריבועים שמחלקים את המרחב גדולים מאוד, פחות מגוונים ומפספסים אזורים רבים על המפה שמודל A כן הצליח ללמוד.

b. למה במודל עבורו הדיוק הוא מקסימלי יש דיוק גבוה יותר?

למודל בעל הדיוק המקסימלי ($L2, k=1$) יש דיוק גבוה יותר מאשר למודל בעל הדיוק המינימלי ($L2, k=3000$) מכיוון שכאשר $k=3000$, בכל פעם שהמודל יחפש את 3000 השכנים הקרובים ביותר ב-Train לדוגמה מסוימת ב-Test, נקבל בסבירות גבוהה את התווית של המדינה שבה יש הכי הרבה דוגמאות מה-Train שנמצאות באזור של הדוגמה הספציפית הזו. בעצם, כאשר מגדירים באלגוריתם מספר גדול מדי של שכנים k , המדינות שבהן יש רוב של דוגמאות ב-Train בקרב 3000 השכנים הללו יהיו התוויות של רוב מוחלט של דוגמאות רבות אחרות שנמצאות בקרבתן. לכן, רוב מוחלט של דוגמאות שאמורות להיות מסווגות למדינות שבהן היו פחות דוגמאות ב-Train לא יסווגו במדינה הנכונה, ולכן הדיוק ירד. לעומת זאת, כאשר $k=1$ אנו מבטיחים שרוב הדוגמאות ב-Test, ובמיוחד דוגמאות שלא נופלות בגבולות המדינות, אלא דוגמאות שנמצאות במקומות מרכזיים יותר בכל כל אזור, יסווגו בצורה נכונה.

4. הסתכלו על שני הגרפים שעבורם $k=1$, עבור פונקציות המרחק השונות $L1, L2$.

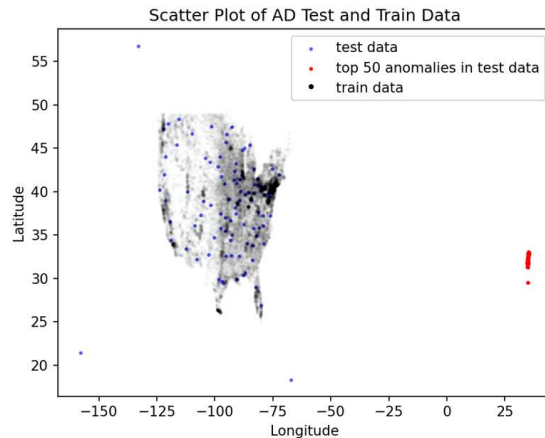
איך פונקציות המרחק השונות משפיעות על החלוקה של המרחב?

במרחב דו-ממדי כמו בדוגמה הנ"ל, פונקציית המרחק $L1$ מחשבת את המרחק בין הנקודות (x_1, y_1) ל- (x_2, y_2) על ידי הנוסחה $|x_1 - x_2| + |y_1 - y_2|$, כלומר מחשבת את סכום ההפרשים המוחלטים על כל ממד. לעומת זאת, פונקציית המרחק $L2$ מחשבת את המרחק על ידי הנוסחה $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$, כלומר מחשבת את האוקלידי, המרחק הקצר ביותר בין שתי נקודות במרחב. מבחינה מרחבית, פונקציית המרחק $L1$ יוצרת רשת ריבועית של נקודות ברווח שווה לאורך הצירים. לעומת זאת, $L2$ יוצרת גיאומטריה מעגלית או כדורית, עם נקודות במרחק שווה מהמרכז. מבחינת חריגות, $L1$ פחות רגישה לחריגות כי היא מתייחסת רק להבדלים מוחלטים. לכן, לערכים קיצוניים יש השפעה פרופורציונלית וליניארית על המרחק. לעומת זאת, $L2$ רגישה יותר לחריגות כי הפרשים בריבוע מגדילים את ההשפעה של סטיות גדולות יותר.

5.3 - גילוי אנומליות באמצעות KNN

משימה: בשימוש בפונקציית מרחק $L2$, מצאו את 5 השכנים הקרובים ביותר מה-train לכל אחת מהדוגמאות ב-test, באמצעות שימוש בספרייה Faiss, ושמרו את המרחקים לשכנים. סכמו את 5 המרחקים לשכנים הקרובים ביותר עבור כל דוגמה בטסט. מצאו את 50 דוגמאות עם המרחקים הגדולים ביותר. נגדיר נקודות אלו כאנומליות כאשר שאר הנקודות נחשבות נורמליות.

ויזואליזציה של האנומליות ב-Test לעומת שאר הדאטה:



5.4 - מה ניתן להגיד על האנומליות שהמודל מצא? איך הדוגמאות הללו שונות משאר הדוגמאות הנורמליות?

דוגמאות נורמליות אמורות ליצור תבניות של קבוצות במרחב לפי ההקשר הרלוונטי. במקרה שלנו, הקבוצות אמורות לשקף התפלגות על פני מרחב גאוגרפי שבו נמצאות נקודות הציון בארצות הברית. הדוגמאות החריגות בהקשר זה הן חמישים הנקודות ב-test data שמרוחקות ביותר מהדוגמאות ב-train data. אלו חמישים הנקודות שסכום מרחקן מכל חמש הדוגמאות הקרובות ביותר אליהן הוא הגדול ביותר. הדוגמאות החריגות הללו עשויות לייצג נקודות הממוקמות באזורים רחוקים מכל קבוצת נקודות אחרת מה-train או להצביע על שגיאות בתיוג. במקרה שלנו, ניתן לראות מבדיקה של הנקודות בגוגל מפות ש-50 הנקודות הללו הן מסביבת קפריסין.

6 - עצי החלטות

משימה: אמנו 24 עצי החלטה שונים בשימוש ב-train data ובצעו סיווג לנקודות הנצ' למדינות. 24 העצים הם קומבינציות של הערכים לשני ההיפר-פרמטרים: עומק עץ מירבי: (1, 2, 4, 6, 10, 20, 50, 100) ומספר עלים מירבי: (50, 100, 1000).

6.1 - תוצאות דיוקי המודלים השונים:

Train Accuracy Table

		Max Depth							
Max Leaves		1	2	4	6	10	20	50	100
	50	0.1133	0.1893	0.3629	0.5928	0.8476	0.8551	0.8551	0.8551
	100	0.1133	0.1893	0.3629	0.5928	0.9266	0.9511	0.9511	0.9511
	1000	0.1133	0.1893	0.3629	0.5928	0.9424	1.0	1.0	1.0

Validation Accuracy Table

		Max Depth							
Max Leaves		1	2	4	6	10	20	50	100
	50	0.1135	0.1861	0.3572	0.5802	0.8356	0.8445	0.8445	0.8445
	100	0.1135	0.1861	0.3572	0.5802	0.9181	0.9424	0.9424	0.9424
	1000	0.1135	0.1861	0.3572	0.5802	0.9311	0.9804	0.9804	0.9804

Test Accuracy Table

Max Leaves	Max Depth							
	1	2	4	6	10	20	50	100
50	0.1135	0.1891	0.3572	0.5822	0.8312	0.8359	0.8359	0.8359
100	0.1135	0.1891	0.3572	0.5822	0.9115	0.9314	0.9314	0.9314
1000	0.1135	0.1891	0.3572	0.5822	0.9224	0.9784	0.9784	0.9784

6.2 - תשובות לשאלות:

1. העץ בעל דיוק ה-validation הגבוה ביותר: $\text{max_depths} = 20$, $\text{max_leaves} = 1000$.
(לעצים בעלי עומק מירבי של 50 או 100 היו אותן תוצאות, לכן בחרתי את העומק המירבי המינימלי מבניהם).

a. דיוק ה-Train שלו: 1

b. דיוק ה-Validation שלו: 0.9804

c. דיוק ה-Test שלו: 0.9784

2. האם העץ הנ"ל מצליח לבצע הכללה מדוגמאות ה-Train? דיוק ה-Train של 1

מצביע על כך שהעץ השיג דיוק מושלם ב-Train set. עם זאת, דיוק Train גבוה לא בהכרח מבטיח הכללה טובה לדוגמאות חדשות. דיוק ה-Validation של 0.9804 ודיוק ה-test של 0.9784 קרובים מאוד אחד לשני ולדיוק ה-train הגבוה ולכן נראה שהמודל מצליח להכליל עבור דוגמאות חדשות.

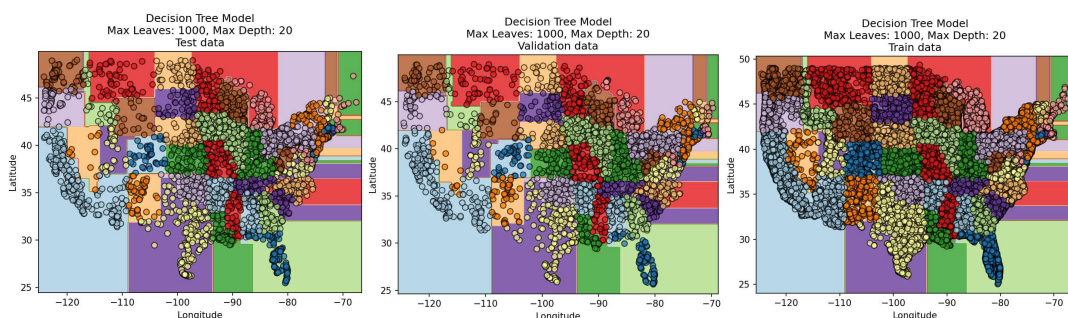
ראו דיוקי test של שאר העצים. האם ה-validation set מאפשר לבחור באופן

מספק את העץ האופטימלי? נראה שה-validation set אכן מאפשר בחירה של העץ האופטימלי כי אכן יש התאמה בין העץ עם הדיוק האופטימלי על ה-validation set לבין העץ עם הדיוק האופטימלי על ה-test set.

3. האם 50 עלים מספיקים כדי לקבל דיוק מושלם? כפי שניתן לראות מהתוצאות 50

עלים לא מספיקים כי היה ניתן להגיע לדיוק טוב יותר ואף לדיוק מושלם בשימוש במספר עלים גדול יותר. הסיבה לכך היא שעץ החלטות עם מספר גדול יותר של עלים עשוי לאפשר גמישות רבה יותר בזיהוי דפוסים עדינים בנתונים ולהיות מצויד יותר להתמודד עם המורכבות של הנתונים ובכך להשיג דיוק גבוה יותר.

4. ויזואליזציה:

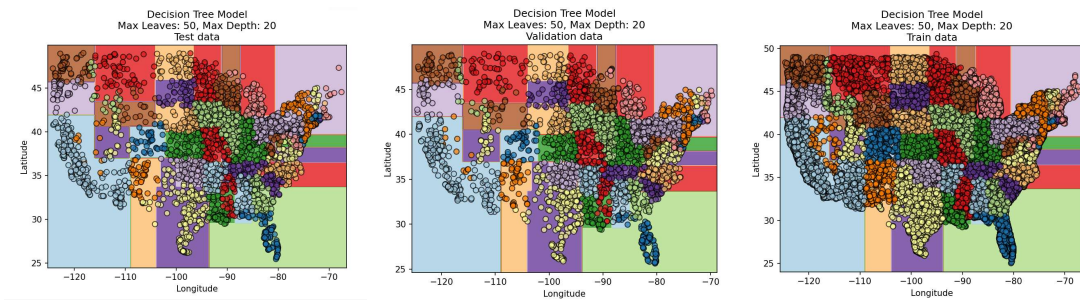


כל פיצול בעץ מתאים להחלטה המבוססת על תכונה ספציפית בדוגמאות, המובילה לחלוקות המישרות לפי הצירים, כלומר הם מקבילים לצירים (אנכיים או אופקיים). כתוצאה מכך הצורה של כל מחלקה היא צורה מלבנית, כלומר נוצרו גבולות החלטה מלבניים המקבילים לצירים.

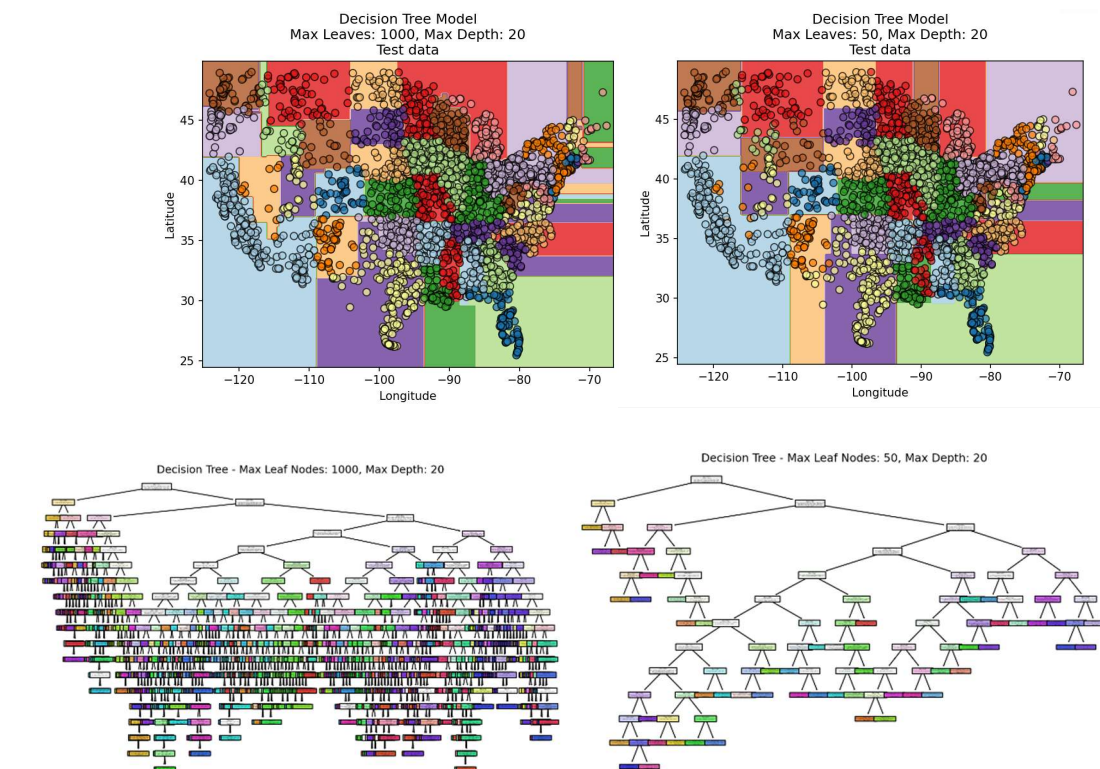
5. בחרו את העץ עם דיוק האימות הכי טוב שיש לו רק 50 עלים. מה השתנה?

העץ בעל דיוק האימות הגבוה ביותר שיש לו 50 עלים: $\text{max_leaves}=50$, $\text{max_depths}=20$, $\text{accuracy}=0.8445$. היינו מצפים שבעבור עומק מקסימלי של 50 או 100 יהיה דיוק גבוה יותר, אך מסתבר ששלושה העצים היו זהים עבור עומק מקסימלי של 20, 50, 100 עלים, וכך גם הדיוק של העץ.

ויזואליזציה של ביצועי העץ:



השוואה בין מודל היער הרנדומלי למודל העץ האופטימלי:

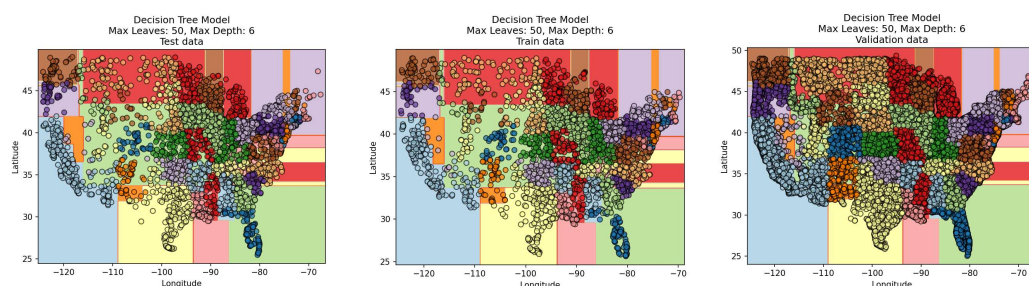


הוויזואליזציה של העץ לעיל בהשוואה לוויזואליזציה של העץ האופטימלי, בהינתן כי העץ האופטימלי בעל אחוזי דיוק גבוהים יותר (0.9804 לעומת 0.8445) נראית דומה אך מדויקת פחות. ניתן לראות שהחלוקה למחלקות בעץ שבו יש מקסימום 50 עלים היא חלוקה גסה יותר, כלומר הריבועים שמחלקים את המרחב גדולים יותר, פחות מגוונים ומפספסים אזורים קטנים יותר על המפה שהעץ האופטימלי כן הצליח ללמוד. בנוסף לכך, ניתן לראות את ההבדל בכמות ההחלטות בעץ כך שהעץ האופטימלי כולל מורכבות גדולה משמעותית. מעניין לראות שלמרות שהעץ בעל מקסימום של 50 nodes פשוט משמעותית מהאופטימלי, הוא עדיין מצליח להניב תוצאות דיוק גבוהות שדומות מאוד לתוצאות העץ האופטימלי.

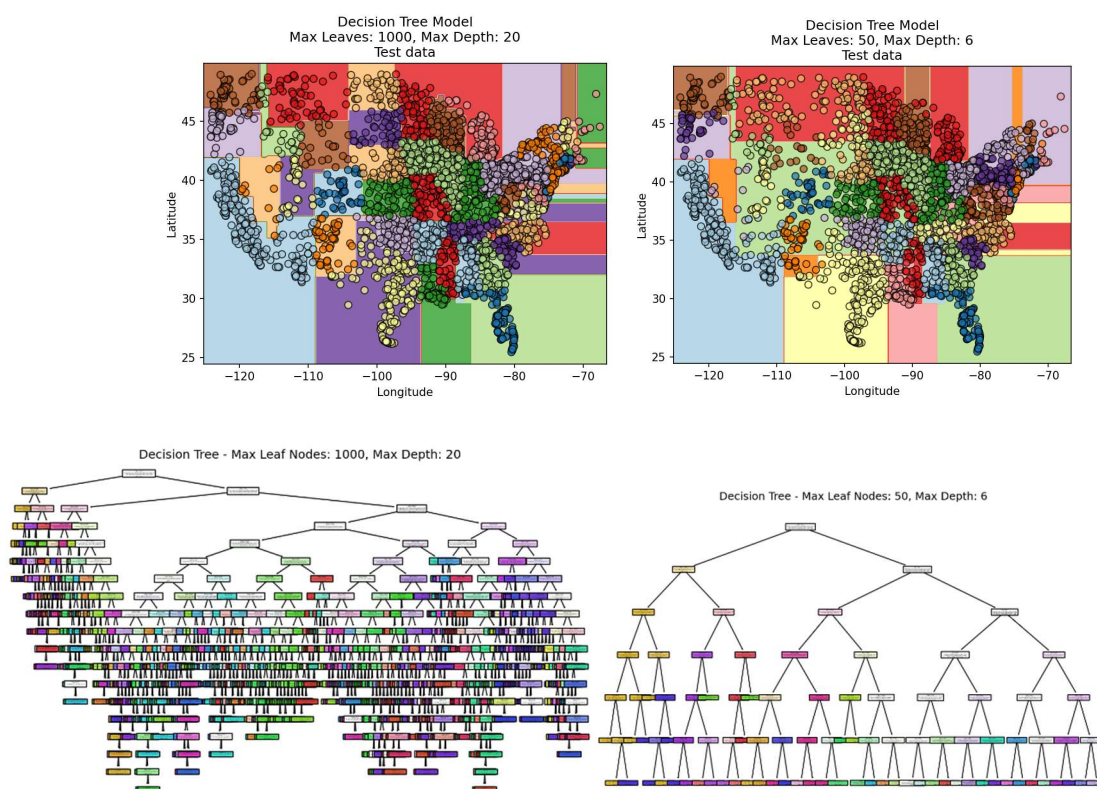
6. בחרו את העץ עם דיוק האימות הכי טוב שיש לו עומק מקסימלי של 6. מה השתנה?

העץ בעל דיוק האימות הגבוה ביותר שיש לו עומק מקסימלי 6: $\text{max_leaves}=50$, $\text{max_depths}=6$, $\text{accuracy}=0.5802$. היינו מצפים שבעבור מספר מקסימלי של 1000 עלים יניב דיוק גבוה יותר מאשר מספר מקסימלי של 50 עלים, אך מסתבר ששלושה העצים היו זהים עבור מספר מקסימלי של 50, 100, 1000 עלים, וכך גם הדיוק של העץ.

ויזואליזציה של ביצועי העץ:



השוואה בין מודל היער הרנדומלי למודל העץ האופטימלי:



הוויזואליזציה של העץ לעיל בהשוואה לוויזואליזציה של העץ האופטימלי, בהינתן כי העץ האופטימלי בעל אחוזי דיוק גבוהים יותר (0.9804 לעומת 0.5802) נראית גסה יותר. ניתן לראות שהחלוקה למחלקות בעץ שבו יש מקסימום 6 רמות היא חלוקה גסה, כלומר הריבועים שמחלקים את המרחב גדולים יותר באופן משמעותי, פחות מגוונים ומפספסים אזורים רבים על המפה שהעץ האופטימלי כן הצליח ללמוד. בנוסף לכך, בולטים ההבדלים בצבעים וריבוע גדול

ירוק בצבע ירוק שנמצא במרכז-שמאל המפה לעיל לעומת המפה בעץ האופטימלי שמראה חלוקה לריבועים קטנים רבים באותו האזור. מעניין לראות שהעץ בעומק מירבי 6 עם מספר מקסימלי של 50 עלים משאלה זו הניב תוצאות דיוק משמעותית פחות טובות מהעץ בעל מספר מקסימלי של 50 עלים ועומק מירבי של 20 מהשאלה הקודמת. הסבר אפשרי הוא שעץ עמוק יותר יכול ללכוד דפוסים ויחסים מורכבים יותר בתוך הנתונים מכיוון שהוא יכול לקבל יותר החלטות לפני שהוא מגיע לצומת עלים.

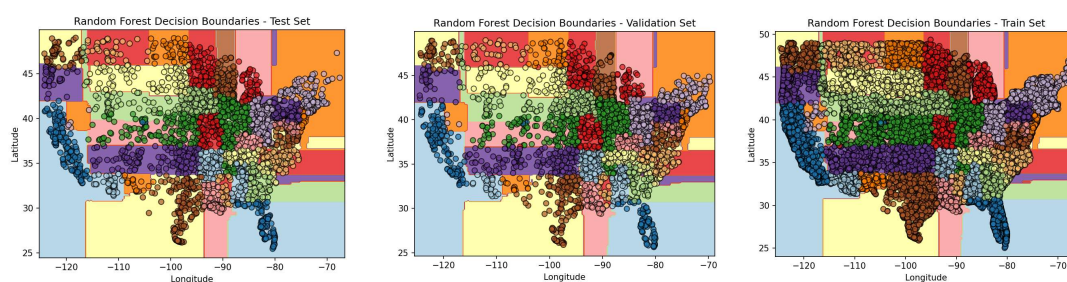
7. אמנו יער רנדומלי של עצים בשימוש ב-300 עצים עם עומק מקסימלי של 6.

האם המודל הזה אקספרסיבי יותר מזה של העץ האופטימלי? איך הוויזואליזציה עזרה להגיע למסקנה הזו?

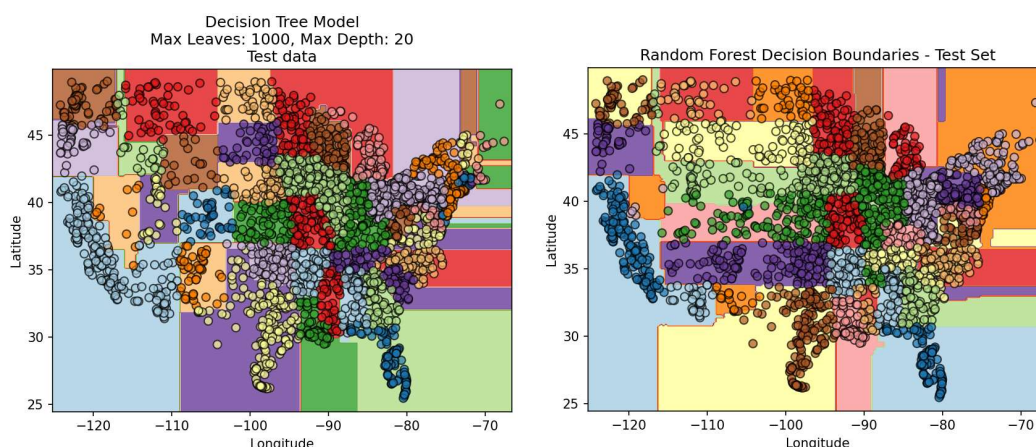
ביצועי המודל:

```
train accuracy: 0.8179018767425409
validation accuracy: 0.8042609853528628
test accuracy: 0.8045938748335553
```

ויזואליזציה של ביצועי המודל:



השוואה בין מודל היער הרנדומלי למודל העץ האופטימלי:



הדיוק של מודל היער הרנדומלי על ה-test set הוא בערך 0.804, לעומת העץ האופטימלי שהדיוק שלו על ה-test set הוא בערך 0.9784. ניתן לראות מהוויזואליזציה שמודל היער הרנדומלי אקספרסיבי פחות כי יש אזורים בוויזואליזציה של מודל העץ האופטימלי שמראים מורכבויות בחלוקה למדינות, למשל בפינה ימנית-עליונה, כאשר נראה שהמודל של היער

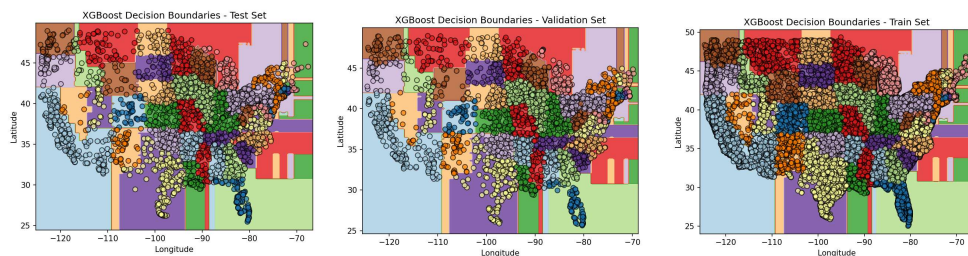
הרנדומלי מסווג את כל האזור בצבע אחד, כלומר כקבוצה יחידה, ובכך מפספס סיווגים של נקודות למדינות שבהן יש פחות דוגמאות ב-test set.

8. אמנו דגם XGBoost עם אותם פרמטרים כמו היער האקראי ועם פרמטר $\text{learning rate}=0.1$. מה דיוק ה-Test שלו? במה שונות התחזיות של XGBoost מאלה של יער רנדומלי? איזה אלגוריתם מצליח יותר במשימה זו?

ביצועי המודל:

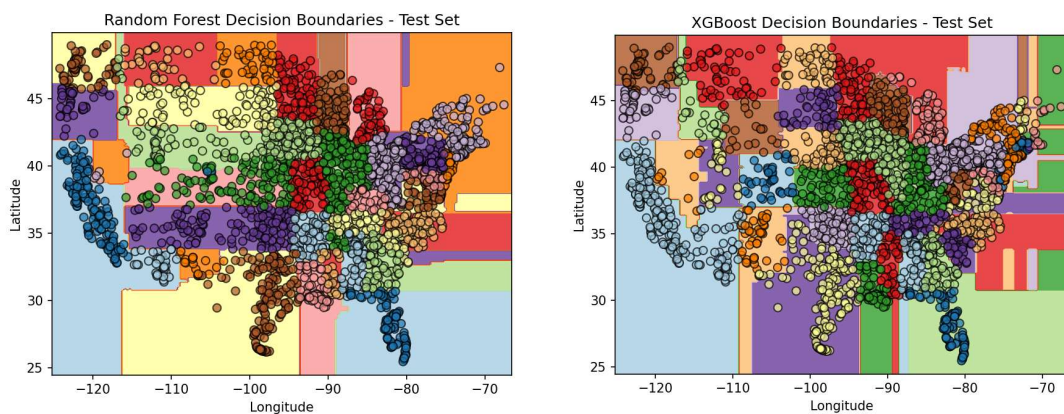
```
train accuracy: 0.9917606425034331
validation accuracy: 0.974034620505992
test accuracy: 0.9653794940079894
```

ויזואליזציה של ביצועי המודל:



דיוק ה-test של המודל הוא 0.9653.

השוואה בין מודל XGBoost למודל היער הרנדומלי:



התחזיות של מודל XGBoost מדויקות יותר מהתחזיות של מודל היער הרנדומלי. מהוויזואליזציה נראה ש-XGBoost חילק את המרחב ליותר מלבנים, ובהתחשב בדיוק שלו ניתן להבין שחלוקה זו מדויקת יותר. בדומה להשוואה הקודמת, ניתן לראות שמודל היער הרנדומלי אקספרסיבי פחות כי יש אזורים בוויזואליזציה של מודל XGBoost שמראים מורכבויות בחלוקה למדינות, בעוד שנראה שמודל היער הרנדומלי מסווג את כל האזור בצבע אחד, כלומר כקבוצה יחידה, ובכך מפספס סיווגים של נקודות למדינות שבהן יש פחות דוגמאות ב-test set.