

האוניברסיטה העברית

**דפוס צריכת תבלינים עולמיים וניתוח קשריהם לצורך אפיון
מטבחים קולינריים**

מגישות:

נועם מעוז

ליאור עידו

שירה ברקוביץ'

יובל רוזנטל

נושא המחקר

מאז ומתמיד מילאו התבלינים תפקיד מרכזי בעשייה הקולינרית, והשפעתם חורגת מעבר להיבטים חושיים של טעם וריח. הם מעצבים את זהותם של מטבחים אזוריים ומשקפים מאפיינים תרבותיים, חברתיים והיסטוריים, עד כדי יכולת לשמש סמן למוצא גיאוגרפי של מתכון או תרבות שלמה.

בעבודתנו ביקשנו לבחון את הדפוסים והקשרים בצריכת תבלינים, ואת מידת יכולתם לשקף מאפיינים קולינריים ייחודיים, תרבותיים ואזוריים. לשם כך ניתחנו מגמות בצריכה לאורך זמן, זיהינו שילובים חוזרים של תבלינים, והערכנו את הפוטנציאל לתרגם דפוסים אלו לכדי אלגוריתם המסוגל לאפיין מטבחים שונים בעולם. בכך אנו מציעות שימוש בתבלינים כמדד מבוסס-נתונים לניתוח התרבות הקולינרית הגלובלית.

מאגרי מידע

לצורך המחקר נעשה שימוש בשלושה מאגרי מידע עיקריים:

1. FAOSTAT¹ – מאגר הנתונים הרשמי של ארגון המזון והחקלאות של האו"ם (FAO)

FAOSTAT הוא מאגר הנתונים הרשמי של ארגון המזון והחקלאות של האו"ם, המרכז מידע רחב היקף בתחומי מזון, תזונה, חקלאות ומשאבים טבעיים. המאגר כולל נתונים מ-2010 ועד 2022 עבור כ-245 מדינות וטריטוריות, ובו מיליוני רשומות על ייצור, מסחר, צריכה, מחירים ועוד, ובכללן גם נתונים מפורטים על צריכת תבלינים. בעבודה זו נעשה שימוש בנתונים מתחום ה-Food Balance Sheets², המספקים תמונה אינטגרטיבית של היצע המזון ברמה הלאומית והעולמית (ייצור, יבוא, יצוא, מלאי ושימושים), וזמינים גם במונחי צריכה לנפש – דבר המאפשר ניתוח מגמות והשוואות לאורך זמן.

2. AllRecipes³ – מאגר מתכונים עולמי

Allrecipes הוא אחד מאתרי המתכונים הגדולים בעולם. הוא פועל מאז 1997 ומבוסס על שיתוף מתכונים ותוכן גולשים. האתר מושך כ-60 מיליון מבקרים חודשיים, מתעדכן בכ-200 מתכונים חדשים מדי שבוע, וכולל אלפי דירוגים ותמונות שהופכים אותו לקהילה קולינרית פעילה. המתכונים מחולקים גם לפי מטבחים אזוריים ותרבותיים, מה שמאפשר לבחון דפוסי בישול והשוואה בין מסורות שונות. בעבודה זו נעשה שימוש במאגר זה לצורך ניתוח דפוסי השילוב בין תבלינים, במטרה לזהות אילו מהם מופיעים יחד לעיתים קרובות.

3. Kaggle Recipe Ingredients Dataset⁴ – מאגר מתכונים מתוייגים לפי מטבח

מאגר נתונים מהאתר Kaggle שהוצג בתחרות "What's Cooking". מאגר זה כולל מתכונים מגוונים עם רשימת רכיבים ותיוג מטבח (איטלקי, הודי, קוריאני ועוד). מטרת התחרות הייתה לחזות את קטגוריית המטבח לפי רשימת הרכיבים. בעבודה זו נשתמש בקובץ train.json המכיל 39,774 מתכונים, אותו נחלק ל-train ול-test, כדי לבחון אלגוריתמי סיווג המבוססים על תבלינים בלבד ולבדוק את דיוקם.

¹ FAOSTAT

² FBS – Food Balance Sheets

³ AllRecipes

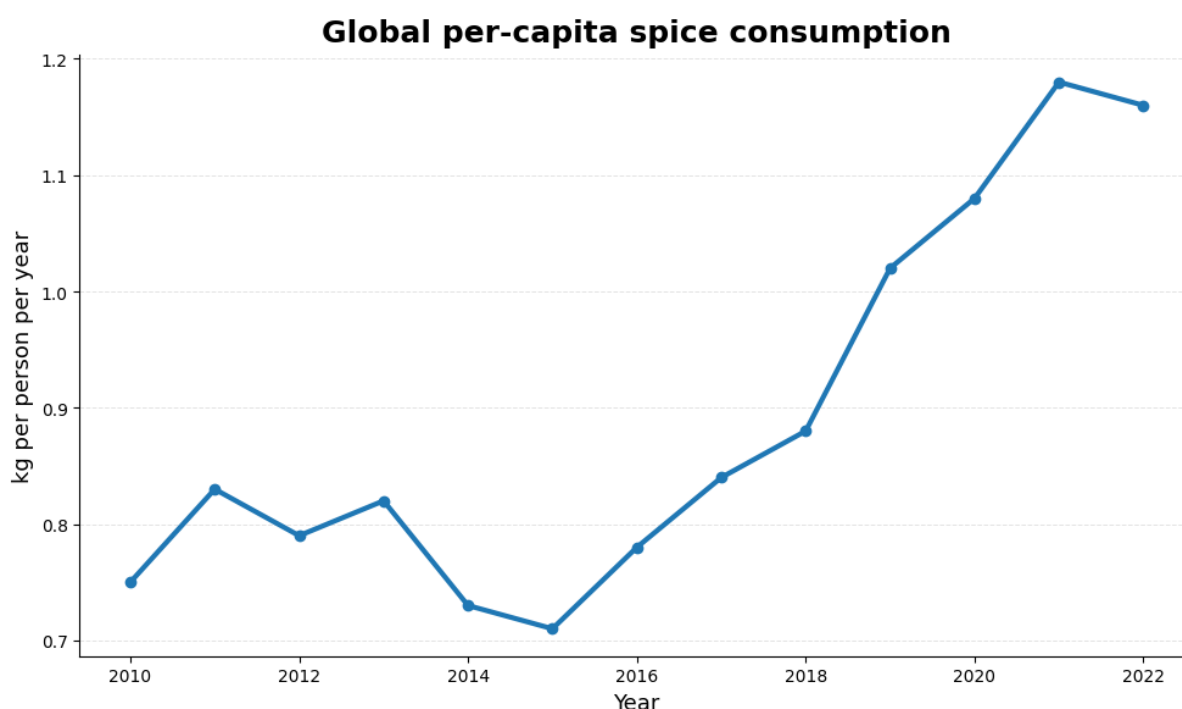
⁴ Kaggle –Recipe Ingredients Dataset

תיאור הפתרון

מטרת עבודה זו היא לבחון את דפוסי השימוש בתבלינים במגוון מנות קולינריות, במטרה להעריך את מידת תרומתם כמדד לזיהוי מטבחים שונים, ולבדוק את הפוטנציאל לשימוש בהם כבסיס לפיתוח מודלים לסיווג אוטומטי של מתכונים לפי מוצאם הקולינרי.

כצעד ראשון, בחנו את מגמות הצריכה הגלובליות של תבלינים בעשור האחרון, במטרה להעריך את היקף השימוש בהם ואת מקומם בתזונה העולמית. לשם כך השתמשנו בנתוני FAOSTAT, תוך בחירה בפרט "Spices + (Total)" הכולל את כלל קבוצות התבלינים המדווחות, ובייחוד את קטגוריית המשנה "Spices, Other" שבה נכללים תבלינים מרכזיים כמו וניל, קינמון, אגוז מוסקט והל. אף שהקטגוריה מדווחת ברמת קבוצה ולא עבור כל תבלין בנפרד, היא מספקת מדד עקבי ואחיד לאורך זמן.

נוסף לכך, לצורך הניתוח נבחר מדד Food supply quantity (kg/capita/yr), המשקף את הכמות הממוצעת של התבלינים הזמינה לצריכה לנפש בשנה. מדד זה מתחשב ביבוא, יצוא והפסדים, ובכך מציע תמונה מהימנה של הרגלי הצריכה בפועל. בנוסף, התבססנו על יחידת הניתוח "World", המייצגת ממוצע גלובלי משוקלל של כלל המדינות. לבסוף, עיבדנו את הנתונים לסדרת זמן המתארת את צריכת התבלינים לנפש (בק"ג לשנה) ברמה גלובלית בין השנים 2010–2022.



הסדרה הוצגה בגרף מגמות, הממחיש עלייה מצטברת בצריכת התבלינים העולמית לאורך התקופה, עם האצה חדה במיוחד בשנים 2019–2021. מגמה זו עשויה להיות מושפעת, בין היתר, מעלייה בבישול הביתי בתקופת מגפת הקורונה, מהתחזקות מגמות בריאות עולמיות, ומהעדפה גוברת לשימוש בתבלינים טבעיים. ממצאים אלו מחזקים את ההנחה המרכזית של עבודתנו: צריכת התבלינים היא תופעה רחבת היקף ורלוונטית בקנה מידה עולמי, ולכן יש היגיון מחקרי לבחון אם דפוסי השימוש בהם יכולים לשמש בסיס לסיווג מתכונים ולהבחנה בין מטבחים קולינריים שונים.

בהמשך, פנינו לאתר AllRecipes על מנת לנתח קשרים בין תבלינים כפי שהם מופיעים במתכונים. אספנו קישורי מתכונים דרך מפות האתר (sitemaps), חילצנו שם מתכון, מטבח ורשימת רכיבים, וניצלנו את המבנה המובנה בעמודים (JSON-LD) בשל עקביותו; כאשר הוא לא היה זמין השתמשנו ברשימות הרכיבים שב-HTML כגיבוי. לאחר השליפה יושם ניקוי נתונים: נרמול טקסט, הסרת כפילויות, פיצול צירופים כמו "salt and pepper", ואיחוד שמות מטבחים תוך הסרת תוויות גנריות. התוצר המכיל 5,000 מתכונים נשמר בקובץ recipes_ingredients.csv⁵ כאשר כל שורה מייצגת רכיב יחיד וכוללת מזהה מתכון ייחודי, שם המטבח, שם המתכון, כתובת ה-URL ושם הרכיב.

כדי להתמקד בתבלינים בלבד יצרנו מילון תבלינים ייחודי. לשם כך הורדנו את הקובץ spices.json⁶ מהאתר SpiceNice⁷, שהכיל רשימה רחבה של תבלינים עם שמות חלופיים ותצורות קולינריות שונות. הרחבנו את הקובץ כך שלכל תבלין יופיעו גרסאות שונות לפי צורתו (למשל seeds לעומת powder וכו'), ולאחר שיפור נוסף ידנית התקבל מילון מסודר של **106 תבלינים** שנשמר בקובץ spices_map.json⁸.

```
"sumac": [
  "sumac",
  "sumac powder",
  "ground sumac"
],
"cinnamon": [
  "cinnamon",
  "cinnamon powder",
  "ground cinnamon"
],
"clove": [
  "clove",
  "dried clove"
],
```

באמצעות מילון זה זיהינו את התבלינים מתוך רכיבי המתכונים ששמרנו קודם לכן, תוך טיפול במקרים מיוחדים (למשל הבחנה בין "garlic cloves" ל"clove" או דיוק של "pepper" לפי ההקשר). התוצאות נשמרו בקובץ recipes_spices.csv⁹, ובסך הכל זוהו **72 תבלינים שונים ב-3902 מתכונים**.

id	cuisine	recipe_name	recipe_url	ingredient	spice
4	american	Strawberry-Orange Bundt Cake	https://www.allrecipes.com/recipe/83710	2.5 teaspoons vanilla extract	vanilla
5	german	Knefla Soup II	https://www.allrecipes.com/recipe/24402	pepper to taste	black pepper
6	vietnamese	Chef John 's Fresh Spring Rolls	https://www.allrecipes.com/recipe/24491	4 sprigs chopped fresh basil	basil
6	vietnamese	Chef John 's Fresh Spring Rolls	https://www.allrecipes.com/recipe/24491	4 sprigs chopped fresh mint	mint
6	vietnamese	Chef John 's Fresh Spring Rolls	https://www.allrecipes.com/recipe/24491	4 sprigs chopped fresh cilantro	coriander
8	asian_inspi	Asian Potato Salad	https://www.allrecipes.com/recipe/84408	1 teaspoon sesame oil	sesame
8	asian_inspi	Asian Potato Salad	https://www.allrecipes.com/recipe/84408	0.125 teaspoon dry hot mustard	yellow mustard
8	asian_inspi	Asian Potato Salad	https://www.allrecipes.com/recipe/84408	0.25 cup chopped fresh cilantro	coriander
9	american	Lentil Burgers	https://www.allrecipes.com/recipe/23679	1 pinch ground black pepper	black pepper
11	american	Vegan "Lobster" Rolls	https://www.allrecipes.com/recipe/28143	ground black pepper to taste	black pepper
11	american	Vegan "Lobster" Rolls	https://www.allrecipes.com/recipe/28143	1 stalk celery, minced	celery
11	american	Vegan "Lobster" Rolls	https://www.allrecipes.com/recipe/28143	0.5 teaspoon dried dill weed	dill
11	american	Vegan "Lobster" Rolls	https://www.allrecipes.com/recipe/28143	0.5 teaspoon sesame seeds	sesame
14	american	Blue Cheese Dip	https://www.allrecipes.com/recipe/24901	2 tablespoons dried parsley	parsley

כשהושלמה עבודת ההכנה, עברנו לזיהוי ולמיפוי הקשרים בין התבלינים. בשלב הראשון טענו את מסד הנתונים שיצרנו, שכלל זוגות של מתכון ותבלין. לאחר סינון כפילויות, בנינו עבור כל תבלין את קבוצת

⁵ [recipes_ingredients.csv](#)

⁶ [spices.json](#)

⁷ [SpiceNice](#)

⁸ [spices_map.json](#)

⁹ [recipes_spices.csv](#)

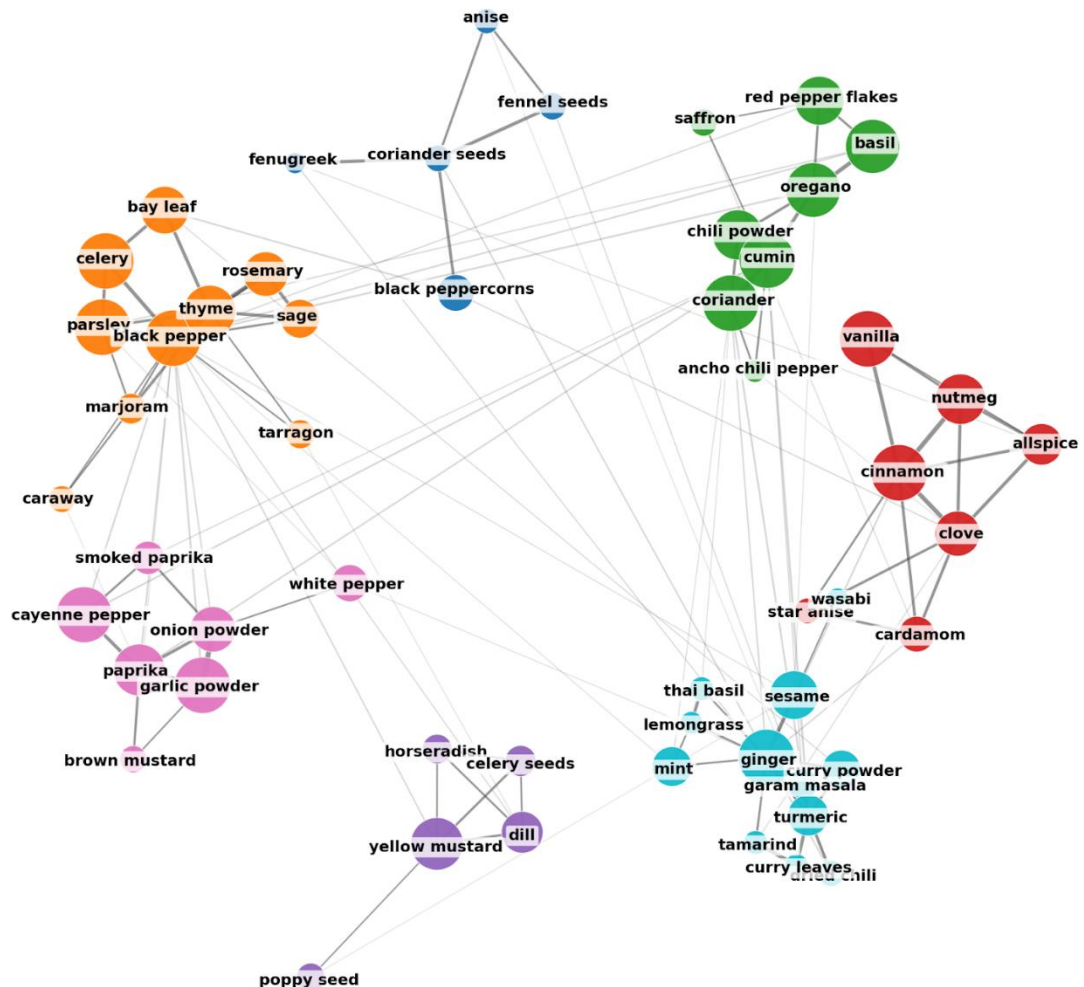
המתכונים שבהם הופיע. כדי להבטיח יציבות סטטיסטית ולהפחית רעשים, כללנו רק תבלינים שהופיעו לפחות בחמישה מתכונים, בעוד תבלינים נדירים הוסרו. לאחר סינון זה, נותרו 54 תבלינים לניתוח. צעד זה איפשר להתמקד בתבלינים בעלי שכיחות מספקת, וכך להבטיח שהקשרים שיתגלו בהמשך יהיו מובהקים יותר.

בשלב הבא, בנינו רשת קשרים בין תבלינים. כל תבלין ייצג צומת בגרף, וקשת בין שני תבלינים נוספה אם הופיעו יחד לפחות בשני מתכונים. בדיקות רגישות הראו כי העלאת סף זה גרמה לירידה חדה במספר התבלינים והקשרים, ולכן בחרנו בו כנקודת איזון המבליטה קשרים משמעותיים מבלי לפגוע בכיסוי ובמבנה הקהילתי. חוזק הקשר חושב באמצעות מדד דמיון קוסינוס: מספר ההופעות המשותפות חולק בשורש המכפלה של שכיחות כל אחד מהתבלינים. בחירה במדד זה נועדה לאזן בין שכיחות כללית גבוהה לבין הופעות משותפות בפועל, ולמנוע מתבלינים נפוצים לקבל באופן אוטומטי משקל יתר.

לאחר יצירת הרשת בוצע סינון נוסף של הקשרים לשם הפחתת רעש ושיפור הקריאות. עבור כל צומת נשמרו שלושת החיבורים החזקים ביותר (Top-K), ונבחרו רק הקשרים החזקים ביותר מתוך כלל הקשרים. דבר זה איפשר למקד את הגרף בקשרים מובהקים תוך שמירה על המבנה הגלובלי של הרשת. חלוקת הרשת לקהילות התבצעה באמצעות אלגוריתם Louvain לזיהוי קהילות, אשר מותאם לגרפים משוקללים ודלילים, אינו מחייב לקבוע מספר קהילות מראש, וממקסם מודולריות – קריטריון המתיישב ישירות עם מטרתנו לזהות אשכולות תבלינים בעלי הופעה משותפת גבוהה. תבלינים שאותרו באותה קהילה מייצגים קבוצות קולינריות טבעיות, שכן הם נוטים להופיע יחד במתכונים.

לבסוף, יצרנו ייצוג חזותי של הגרף. כל קהילה מוקמה באזור נפרד במרחב באמצעות פריסה טבעית, דבר שהקל על הבחנה בין קבוצות. הקשתות הפנימיות הובלטו, בעוד הקשרים הבין-קהילתיים סומנו בקווים דקים ובהירים כדי להמחיש זיקות משניות. גודל הצמתים משקף את מספר המתכונים שבהם הופיע כל תבלין, וצבעי הצמתים מבטאים את השתייכותם הקהילתית.

Spice Communities in Recipes — Co-occurrence Network

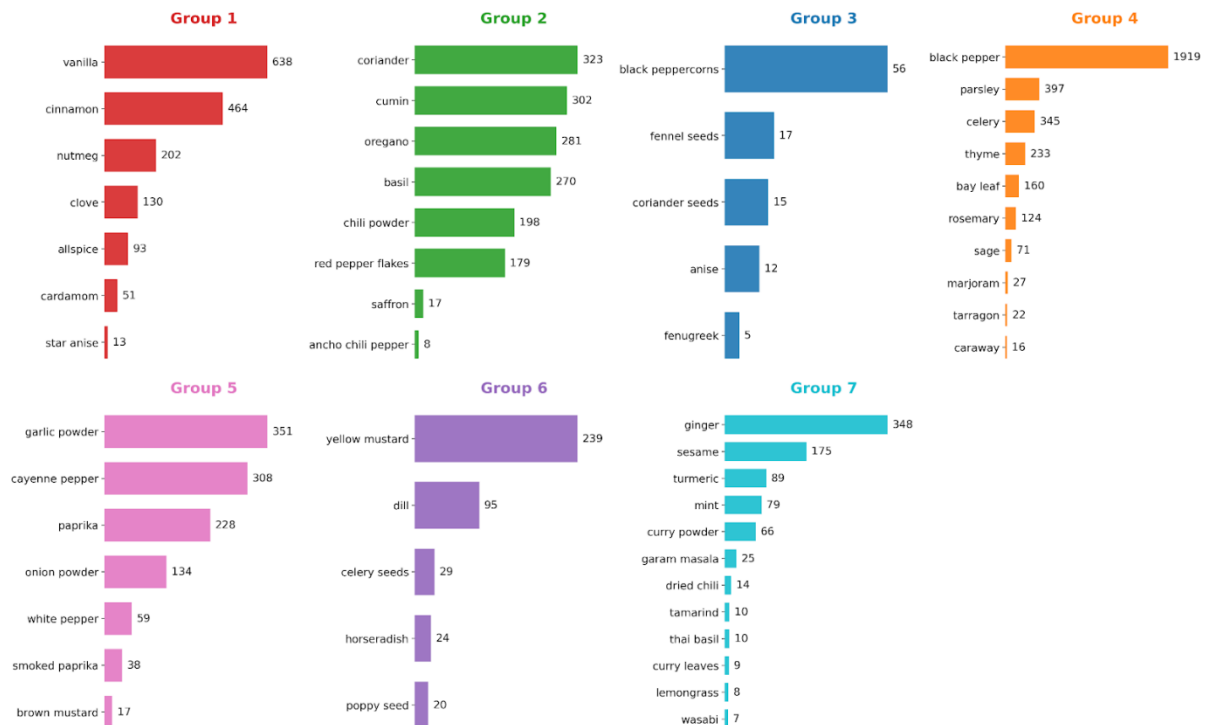


ברשת שנבנתה נמצאו **124 קשרים** בין זוגות תבלינים. הדרגה הממוצעת עמדה על 4.59, כלומר כל תבלין מחובר בממוצע לכ-**4-5 תבלינים** אחרים. הצפיפות שהתקבלה (0.087) מראה שמדובר ברשת דלילה יחסית, שבה פחות מ-9% מהקשרים האפשריים מתקיימים בפועל – תופעה אופיינית לרשתות טבעיות מורכבות. במסגרת מנגנון ה-auto-prune, נבחר סף משקל של 0.0415 כך שנשמרו רק קשרים חזקים דיים, כלומר, זוגות תבלינים בעלי הופעה משותפת מספקת ביחס לשכיחות הכוללת שלהם. סף זה הבטיח שהגרף יישאר קריא וממוקד בקשרים מובהקים.

פירוק הרשת לקהילות, באמצעות אלגוריתם Louvain, הניב **7 קהילות**, שכל אחת מהן כוללת לפחות שני תבלינים. ערך המודולריות לפיצול שהתקבל ($Q \approx 0.534$) מעיד על **מבנה קהילתי מובהק**. המשמעות היא שהתבלינים מתחלקים לקבוצות פנימיות עם קשרים חזקים בתוך הקבוצה וקשרים חלשים יותר בין קבוצות שונות.

הגרף שהתקבל אכן מציג חלוקה טבעית והגיונית של תבלינים לקהילות קולינריות. ניתן לראות כי תבלינים שבדרך כלל מופיעים יחד במתכונים התכנסו לאותן קבוצות: למשל, וניל, קינמון, אגוז מוסקט וציפורן יוצרים קהילה המזוהה עם אפייה וקינוחים; כוסברה, כמון, אורגנו ובזיליקום מופיעים בקבוצה

המשקפת את המטבח הים-תיכוני והלטינו-אמריקאי ; פפריקה, פלפל קאיין ואבקות שום ובצל התאגדו לקהילה חריפה ומעושנת ; ואילו ג'ינג'ר, כורכום, למון גראס ועלי קארי מאפיינים קהילה אסייתית-הודית מובהקת. בכך, מבנה הרשת מחזק את ההנחה כי ניתוח מבוסס קהילות מצליח ללכוד דפוסים קולינריים משמעותיים : **תבלינים שסביר לצפות כי יופיעו יחד אכן מקובצים לאותן קהילות.**



באופן זה, הרשת מאפשרת לא רק זיהוי קשרים בין תבלינים בודדים אלא גם הבנה של דפוסי תיבול קולקטיביים המשקפים מסורות קולינריות שונות. מהניתוח עולה כי הופעת תבלינים משותפים במתכונים קשורה להשתייכותם לקהילות קולינריות, כך **שהתבלינים עצמם מהווים אינדיקציה מרכזית לשיוך קולינרי**. על בסיס ממצא זה ביקשנו לבדוק אמפירית באמצעות מודלים את יכולת החיזוי של שיוך מטבחים לפי תבלינים בלבד.

תחילה, כדי שאימון המודלים יהיה יעיל, נדרשנו למאגר רחב יותר של מתכונים. לשם כך, השתמשנו בקובץ הנתונים train.json מהמאגר "Recipe Ingredients Dataset", טענו אותו ואת מפת התבלינים שפיתחנו, וביצענו נרמול טקסט, הרחבה לצורות שכיחות ושימוש בתבניות התאמה (Regex) שאיפשרו לאתר תבלינים ברשימות הרכיבים, בדומה לשיטה שנקטנו בניתוח הקהילות. לאחר חילוץ התבלינים לכל מתכון, ערכנו sanity check ומצאנו שבכ-20.3% מהמתכונים לא זוהה אף תבלין. מתכונים אלה סונונו, כדי שהמודלים יתבססו על מידע מהותי בלבד.

בהמשך, ניתחנו את מבנה הקטגוריות ומצאנו **20 מטבחים ייחודיים**. כדי להפחית רעש וריבוי קטגוריות, מיפינו מטבחים ספציפיים **לחמש קבוצות אזוריות**: אירופה, מזרח אסיה, דרום אסיה, הקריביים ואמריקה הלטינית, וצפון אפריקה והמזרח התיכון. בדקנו את חלוקת המתכונים בכל אזור כדי למנוע הטיה משמעותית. במקביל, יצרנו קובץ מתועד (¹⁰recipes_with_spices.csv) הכולל מזהה מתכון,

¹⁰ [recipes_with_spices.csv](#)

רשימת רכיבים, רשימת תבלינים שנמצאו, המטבח המקורי והשיוך האזורי.

recipe_id	ingredients	spices_list	cuisine	region
10259	['romaine lettuce',	black pepper	greek	europa
25693	['plain flour', 'grou	black pepper, thyl	southern_us	latin_america_and_caribbean
20130	['eggs', 'pepper', 's	black pepper, garl	filipino	east_asian
13162	['black pepper', 'st	bay leaf, black pe	indian	south_asian
6602	['plain flour', 'suga	cinnamon, ginger	jamaican	latin_america_and_caribbean
42779	['olive oil', 'salt', 'n	bay leaf, black pe	spanish	europa
3735	['sugar', 'pistachio	vanilla	italian	europa

לאחר שהגדרנו בבירור את המאפיינים (Features) – רשימת התבלינים, ואת התיוגים (Labels) – האזורים הקולינריים, בנינו מטריצת תכונות בינארית: כל עמודה מייצגת תבלין, וכל שורה מייצגת מתכון ("1" אם התבלין מופיע, "0" אחרת). מתוך 106 התבלינים שלנו, נמצאו בפועל 83 תבלינים שהופיעו ב- 31,698 מתכונים. את הנתונים חילקנו ל-80% סט אימון ו-20% סט בדיקה (כמקובל).

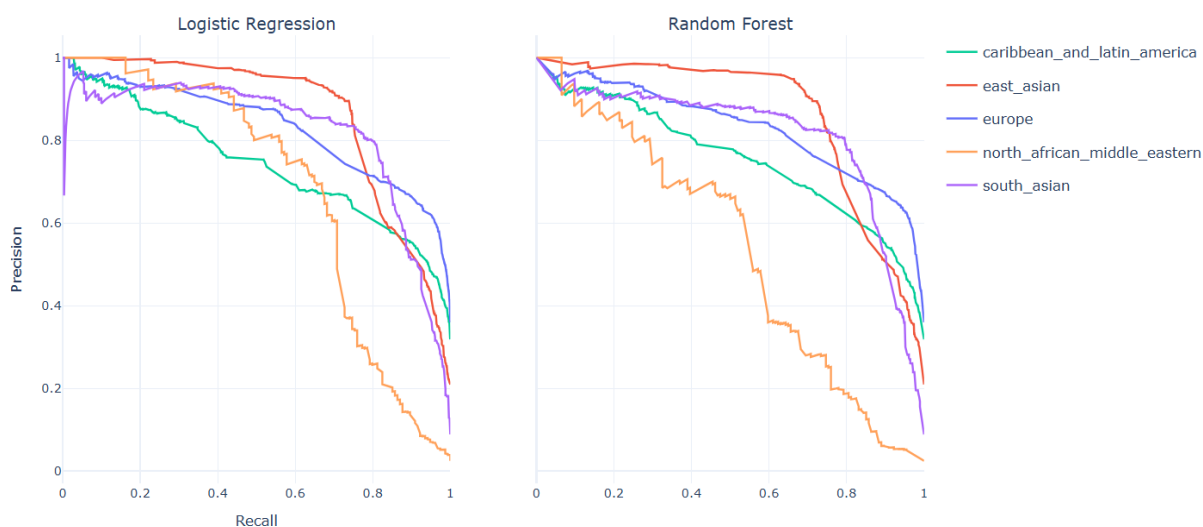
בשלב הבא, אימנו שני מודלים משלימים: **Logistic Regression** ו-**Random Forest**. Logistic Regression כמודל ליניארי ופשוט, המאפשר להבין את התרומה הישירה של כל תבלין לסיווג, ו-Random Forest כמודל לא ליניארי וגמיש, המאפשר ללכוד אינטראקציות מורכבות יותר בין תבלינים. ההשוואה ביניהם תרמה להבנת טיב הקשר בין תבלינים לשיוך קולינרי: אם שני המודלים משיגים ביצועים דומים, הדבר מעיד שהקשר בין תבלינים לשיוך קולינרי הוא ברור וחזק. בנוסף, לנוכח חוסר איזון בגודל הדגימות בין אזורים שונים, השתמשנו במשקלי מחלקות (class weights) מאוזנים כדי למנוע הטיה לטובת אזורים עם יותר דגימות ולוודא שגם הקטגוריות הקטנות יילקחו בחשבון.

כעת, נעבור להערכת המודלים. במקרה שלנו מדובר בסיווג רב-מחלקות (Multi-class), שכן המודלים נדרשים לנבא את שיוכו של מתכון לאחת מחמש קטגוריות אזוריות. כדי להעריך את ביצועי המודלים, השתמשנו במדדים מותאמים למצב זה: macro avg כדי לתת לכל קטגוריה משקל שווה ו-weighted avg – ממוצע המשקל את גודל כל קטגוריה. כ-baseline בחנו מהי רמת הדיוק שהיינו מקבלים אילו המודל היה מנבא תמיד את הקטגוריה הנפוצה ביותר, ונמצא דיוק של כ-36% בלבד – מה שהדגיש בהמשך את התרומה של המודלים שלנו.

מדדנו עבור כל מודל את ה-Accuracy (הדיוק הכללי), וה-ROC AUC (בגישת one-vs-rest). התוצאות הראו כי שני המודלים הציגו ביצועים דומים למדי: דיוק כללי של כ-73% המצביע על שיעור סיווג נכון גבוה יחסית, ו-ROC AUC גבוה מאוד (כ-0.92), המעיד על יכולת הבחנה טובה בין הקטגוריות. המודלים התמודדו די טוב עם חוסר האיזון בדגימות, כך ש-Logistic Regression הראה יכולת מעט גבוהה יותר בהבחנה בין קבוצות מסוימות ו-Random Forest הצליח להתמודד טוב יותר עם שונות בין אזורים קולינריים. מהתוצאות נובעת מסקנה ברורה: **התבלינים מהווים משתנה חזק לשיוך אזור קולינרי**, ומאפשרים גם למודל פשוט וגם למודל גמיש להשיג ביצועים גבוהים.

על מנת לבחון כיצד המודלים מתמודדים עם האתגר של זיהוי מטבחים שונים, יצרנו עקומות Precision-Recall לפי קטגוריות עבור כל מודל, המתארות את הקשר בין אחוז התחזיות החיוביות שהיו נכונות בפועל (Precision), לבין אחוז הדגימות החיוביות שזוהו נכון (Recall).

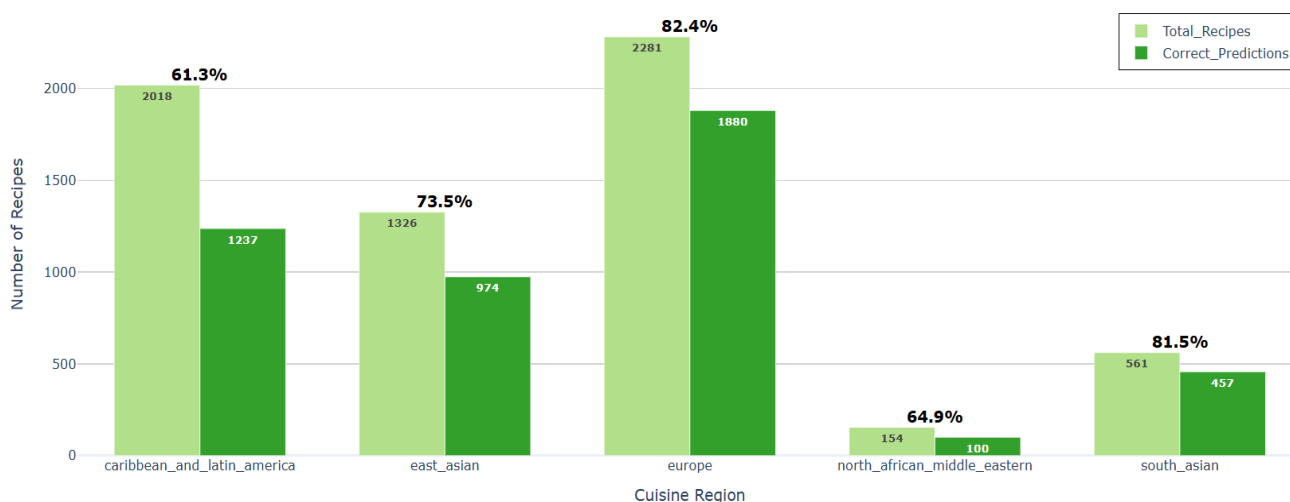
Precision-Recall Curves Comparison



ניתן לראות ששני המודלים מצליחים במיוחד בזיהוי מטבחים כמו מזרח אסיה ואירופה, שם מתקבלות עקומות גבוהות (ו-AP גבוה, כ-0.85 בשניהם). כלומר, המודלים מצליחים לזהות את רוב האזורים הנכונים תוך שמירה על שיעור נמוך יחסית של טעויות סיווג. לעומת זאת, באזורים כמו צפון אפריקה והמזרח התיכון, נראית ירידה חדה בביצועים (ו-AP נמוך, כ-0.53 ב-Random Forest). בהשוואה בין המודלים, Logistic Regression שומר על עקביות טובה יותר בין הקבוצות, בעוד Random Forest מפגין יתרון קל בזיהוי הקטגוריות הגדולות אך מתקשה יותר באזורים הקטנים.

לאחר שבחנו את ביצועי שני המודלים וראינו כי הם מספקים תוצאות דומות למדי, התמקדנו במודל Random Forest, כדי לקבל תמונה ממוקדת יותר של שיעורי ההצלחה. השווינו בין מספר המתכונים האמיתיים בכל קטגוריה לבין מספר המתכונים שהמודל הצליח לזהות נכון.

Random Forest: True vs Correctly Predicted Recipes per Region



ניתן לראות שהתוצאות תואמות למה שקיבלנו קודם לכן – ישנם שיעורי הצלחה גבוהים במטבחים עם מתכונים רבים כמו אירופה ודרום אסיה (82%-74%). לעומתם, שיעורי הצלחה נמוכים יותר במטבחים עם פחות דגימות כמו צפון אפריקה והמזרח התיכון (65%), מה שממחיש את ההשפעה של מידע מוגבל וחוסר האיזון בגודל הדגימות. נשים לב כי בקריביים ואמריקה הלטינית נרשמת הצלחה יחסית נמוכה של כ-61%, על אף שהיו יחסית הרבה מתכונים, ככל הנראה בשל ההטרוגניות הרבה בקבוצה זו.

אתגרים טכניים

האתגר הראשוני שנתקלנו בו היה למצוא אתר מתכונים שממנו נוכל לבצע Web Crawling בצורה מוצלחת. בדקנו אתרי מתכונים רבים, אך ברובם נתקלנו במגבלות שונות שמנעו שליפה סדורה של נתונים. בסופו של דבר הצלחנו להתבסס על אתר Allrecipes, שבזכות ה-sitemaps שלו איפשר לנו לבנות תהליך חילוץ מסודר של תבלינים.

אתגר נוסף, היה לחלץ את התבלינים מרכיבי המתכון בצורה מדויקת. לשם כך, נדרשה בחינה מעמיקה של הדאטה כדי לטפל במקרי קצה. טיפול ממוקד במקרים אלה שיפר באופן ניכר את דיוק החילוץ. לצורך כך נדרשנו לבנות את הקוד בשלבים, לבצע בדיקות ידניות מדגמיות, לאתר טעויות ולבצע התאמות חוזרות עד שהגענו לרמת תוצאות מספקת.

מחקרי המשך

1. **ניתוח טעויות:** בחינת מתכונים שסווגו שגוי תאפשר לזהות אם השגיאה נובעת מפספוס בזיהוי תבלינים, מחוסר איזון בין קטגוריות מטבח או ממנות רב־תרבותיות. תהליך זה יסייע בטיוב מילון התבלינים ובכוונון קריטריוני הסינון.
2. **ממשק "מה-אם" אינטראקטיבי:** פיתוח כלי אינטראקטיבי שיאפשר להוסיף, להסיר או להחליף תבלינים במתכון ולראות את השפעתם על תחזית המטבח, כולל הצעת תבלינים חלופיים "קרובים" המבוססים על רשת התבלינים או דמיון בין מתכונים.
3. **ניתוח דינמי של מגמות:** שילוב נתוני FAOSTAT עם ארכיוני מתכונים היסטוריים יאפשר לבחון כיצד דפוסי שימוש בתבלינים השתנו לאורך זמן, ולהעריך את השפעתם של גלובליזציה, טרנדים בריאותיים ושינויים כלכליים על יכולת המודל לסווג מתכונים.

סיכום

בעבודה זו בחנו את התבלינים ככלי מחקר להבנת מסורות קולינריות. ראשית, ניתחנו מגמות צריכה גלובליות שהדגישו את התרחבות השימוש בתבלינים ואת חשיבותם ההולכת וגוברת בתזונה העולמית, מה שביסס את הרלוונטיות המחקרית. לאחר מכן, שילוב של כריית נתונים, עיבוד טקסטואלי ומיפוי למילון תבלינים ייעודי, איפשר לנו לבנות גרף קהילות שחשף אשכולות ברורים המשקפים דפוסי בישול אזוריים. לבסוף, אימנו מודלי סיווג שהשיגו דיוק גבוה בניבוי שיוך קולינרי של מתכונים – בהתבסס על נוכחות התבלינים בלבד. ממצאים אלו מצביעים על כך שתבלינים יכולים לשמש אינדיקטור יעיל להבנת זהות קולינרית ולסיווג מנות בין מטבחים שונים.

נספחים

- [סרטון](#) הדגמה לכלי החיזוי.
- [תיקייה](#) הכוללת את קבצי הנתונים העיקריים איתם עבדנו במהלך המחקר.