גנומיקה חישובית - Challenge

לירוז אברמוביץ 318723897 | נגה גרוס 322316183

המודלים שבחרנו להשתמש בהם לצורך הפרדיקציה הם - **XGboost** ו- **Random Forest**. בכדי לחזות את הפרמטרים הדרושים במטלה (ערך הממוצע וערך מקסימלי), עלינו להשתמש במודלים שחוזים משתנה רציף ועל כן נשתמש במודלים שרלוונטים לבעיות רגרסיה.

מודל XGBoost הוא כלי המשפר את ביצועי המודל על ידי שילוב של מספר מודלים חלשים ליצירת מודל חזק יותר. הוא מבוסס על עקרונות של Gradient Boosting ומוסיף אופטימיזציות כמו שימוש בעצים ורגולריזציה לצורך התאמה מדויקת של פרמטרים ולמניעת Overfitting. XGBoost מתאים ליישומים בגנומיקה חישובית בזכות יכולתו לספק רמת דיוק גבוה ויכולת עיבוד נתונים גדולים. המודל מצליח להתמודד עם נתונים לא ליניאריים ומורכבים, דבר שמסייע בניתוח קשרים בין מוטציות גנטיות לתכונות או מחלות.

מודל Random Forest בונה מספר עצי החלטה וממזג אותם על מנת לקבל תחזיות מדויקות ויציבות יותר. שיטה זו מציעה התאמה גבוהה יותר בהשוואה לעצי החלטה בודדים, מה שהופך אותה למתאימה לנתונים בעלי מימד גבוה כמו רצפי גנטיקה. בנוסף, Random Forest מספק מדדים לחשיבות התכונות, המאפשרים להבין אילו תכונות גנטיות הן החשובות ביותר. המודל פחות רגיש לערכים קיצוניים, דבר שיכול להיות מועיל כאשר עובדים עם נתונים ביולוגיים מורכבים.

המודל XGBoost בעל קורלצית ספירמן של 0.33 לכן חזה טוב יותר את הנתונים ביחס למודל Random Forest (קורלציה של 0.22).

<u>בנוסף לפיצרים המוצעים במטלה, בחרנו להשתמש בפיצרים משלנו:</u>

- כמות הנוקלאוטידים שהשתנתה בויריאנט ביחס ל Control. יכול לספק מידע על ההשפעה של הויריאנט על תפקוד הגן או על המבנה הכללי של הDNA.
- מספר ההופעות של כל זוג נוקלאוטידי אפשרי (16 אפשרויות). יכול לסייע בזיהוי דפוסים גנטיים והשפעתם על תפקוד או מבנה של DNA ולעזור בהבנת הקשרים בין זוגות נוקלאוטידים למחלות או לתכונות גנטיות מסוימות.
- Melting Temperature. זו הטמפרטורה בה נפרדים שני גדילים של DNA זה מזה, ולכן מהווה אינדיקטור חשוב לכמה טוב הם מתאימים זה לזה. הפיצ'ר עוזר להבין את יציבות הקשרים בין זוגות הבסיסים של הויריאנט ובכך ניתן ללמוד על יציבות המבנה הגנומי של הויריאנט.

- Robustness. מתייחסת ליכולת של הפיצ'ר להתמודד עם שינויים או רעשים בנתונים. פיצ'ר זה חשוב כדי להבטיח שהמודל יוכל להתמודד עם שונות בנתונים ובמקרים של חוסר עקביות, ולהיות מדויק גם בתנאים משתנים.

הפיצרים שנבחרו בסוף לצורך אימון שני המודלים שלנו הם:

- Folding Energy of window 13. את אנרגיית קיפול חישבנו על ידי חלונות בגודל של 40 נאוקלטודים כל פעם, ולאחר מכן ביצענו מיצוע מקומי על כל 20 חלונות.
- מספר הקודונים שהשתנו בויריאנט ביחס לcontrol.
- מספר ההופעות של צמד הנוקלאוטידים "TA"
- ציון PSSM מקסימלי של MOTIF 9
- האינדקס בו התקבל הציון הPSSM המקסימלי בMOTIF 5

פיצרים אלו נבחרו בקוד על ידי שימוש באלגוריתם Forward feature selection ועל ידי קורלציית ספירמן.

אנרגיית הקיפול היא קריטית להבנת יציבות המבנה של DNA. על ידי חישוב אנרגיית הקיפול בחלונות של 40 נוקלאוטידים וחישוב מיצוע מקומי, ניתן לקבוע כיצד שינויים בויריאנט משפיעים על יציבות המבנה. שינויים בקודונים יכולים להשפיע על תרגום החלבון ולגרום לשינויים בתפקוד החלבון, ולכן מספר הקודונים שהשתנו הוא פרמטר חשוב להבנת ההשפעה של הויריאנט. צמד נוקלאוטידים ספציפי כמו "TA" יכול להיות חשוב בגלל השפעתו על מבנה הגנום, רמות ביטוי גנים, או אינטראקציות ביולוגיות אחרות. ציון PSSM מספק מידע על התאמה בין רצפים ועל ההסתברות. ציון מקסימלי של MOTIF יכול להצביע על השפעה משמעותית של הויריאנט על פעולות ביולוגיות. ובנוסף, האינדקס של הציון המקסימלי מספק מידע על המיקום המדויק של MOTIF משמעותי ברצף. האינדקס יכול להצביע על השפעה או שינוי באזורים קריטיים במבנה הגנום שעלולים להשפיע על תפקודו או ביטויו של הגן.