מגישים:

ניב עוזר- 208993329

גיא קופל- 207232240

**נושא**: רשימת מועמדים לזכייה בתואר ה-MVP

**שאלת מחקר**: האם בעזרת למידת מכונה ניתן לקבוע מי יהיו המועמדים לזכייה בתואר ה-MVP בליגת ה-NBA בעונה הנוכחית, על ידי הרכשת נתונים של שחקנים משנים קודמות והמועמדים לזכייה בתואר באותם שנים?

שאלת המחקר מנוסחת כך שעבודתנו תלויה פחות באירועים אקראיים בליגת ה-NBA אלא השאלה מאפשרת לחזות מי הם המועמדים לזכייה ולא מי יהיה הזוכה היחיד בתואר.

בעולם הכדורסל, פרס השחקן המוערך ביותר (MVP) הוא אחד מהשבחים היוקרתיים ביותר ששחקן יכול לקבל. פרס ה-MVP מוענק לשחקן בעל ההשפעה המשמעותית ביותר על הצלחת קבוצתו לאורך העונה הסדירה, כפי שהוצבע על ידי פאנל של כותבי ספורט ושדרנים. פרס ה-MVP הוא עדות למיומנות האישית של השחקן וליכולתו להוביל את קבוצתו לניצחון. במאמר זה, נחקור את התהליך של למידת מכונה אשר תוכל לנבא על ידי מידע שקיבלה על שחקני הליגה בעונות הקודמות את בחירת המועמדים לזכות בתואר ה-MVP בליגת ה-NBA.

**למה בחרנו בנושא הזה:**

בחרנו בנושא זה משום שמדובר בפרויקט המשלב מספר עניינים המעניינים אותנו ובתור שני שחקני עבר לנסות לראות האם שילוב של הידע שלנו וטכנולוגיה חדישה כמו למידת מכונה הוא אפשרי .

בפרויקט שלנו, אנו עוסקים בניתוח נתוני שחקני NBA והצגת התוצאות באמצעות עץ ההחלטות. מדובר בנושא פופולרי ומעניין, וזה ייתן לנו אפשרות להתעמק בו ולבחון את המאפיינים השונים של השחקנים המועמדים לתואר ה-MVP

הניתוח של נתוני השחקנים והצגת התוצאות באמצעות עץ ההחלטות מאפשרים לנו להתעמק במיומנויות ובכישורים הדרושים לפרויקט כזה, כגון העברת נתונים למכונת למידה, בחירת פיצ'רים וניתוח התוצאות. בנוסף, הפרויקט יכול לסייע לנו להתפתח כמפתחי תוכנה ולחקור חדשנות וטכנולוגיות חדשות בתחום מדעי המחשב.

כדי לחזות את בחירת המועמדים לפרס ה-MVP, נשתמש באלגוריתמים של למידת מכונה. למידת מכונה היא ענף של למידת מכונה המתמקדת בזיהוי דפוסים וקשרים בנתונים כדי לבצע תחזיות. באופן ספציפי, נשתמש בסוג של למידת מכנה הנקראת ניתוח רגרסיה כדי ליצור מודל המנבא אילו שחקנים צפויים לזכות בפרס ה-MVP על סמך הביצועים שלהם לאורך העונה הרגילה.

השלב הראשון בתהליך יצירת התוכנית החוזה את בחירת המועמדים לפרס ה-MVP הוא איסוף נתונים. אנחנו צריכים נתונים על ביצועים של שחקנים בודדים לאורך העונה הסדירה, כמו גם נתונים על ביצועי הקבוצה. אספנו נתונים אלה ממגוון מקורות, כולל האתר הרשמי של ה-NBA, Basketball Reference, .Espn

ביצענו את הרכשת הנתונים על ידי הספרייה Beautiful Soup, אך נתקלנו בבעיית חסימת parse באמצעות bs4 מאתרים מסוימים ובעיה נוספת בעת לחיצה על כפתור לחשיפת פרטים נוספים וגלילת של האתר באמצעות גלגלת העכבר. הפתרון לבעיה זו היה באמצעות ספריית Selenium.

בעבודה שלנו אנחנו נבצע למידת מכונה כך שאזור האימון של המכונה יהיה מידע משנת 2020-21, ומידע משנת 2021-22. דבר ראשון יצרנו מס”ד נתונים שמחולק **למספר קטגוריות:**

1. שם השחקן.
2. עמדת השחקן.
3. גיל השחקן.
4. קבוצה.
5. האם זכה בגביע האזורי? (בינארי)
6. כמות זכיות בעונה.
7. סה"כ נקודות.
8. ממוצע נקודות למשחק.
9. דירוג הקבוצה.
10. האם השתתף במשחקי האול-סטאר? (בינארי)
11. מאזן פלוס/מינוס.
12. ממוצע ריבאונד התקפה למשחק.
13. ממוצע ריבאונד הגנה למשחק.
14. ממוצע אסיססטים למשחק.
15. כמות משחקים ששיחק בעונה.
16. ממוצע דקות למשחק.
17. ממוצע חסימות למשחק.
18. ממוצע חטיפות למשחק.
19. אחוז קליעות לשלוש.
20. אחוז קליעות לשתיים.
21. שנה נוכחית.
22. האם מועמד לזכייה בתואר ה-MVP.)בינארי)

כפי שנאמר אלו הקטגוריות (עמודות) שבעזרתן ניתן להבחין אילו שחקנים בולטים יותר מהשאר. השורות הן כל השחקנים בעונה הנוכחית ב-NBA.

בשביל ליצור מרחב מדגם גדול יותר בשביל למידת מכונה רכשנו את המידע על שלושת העונות האחרונות ושרשרנו את שלושת מסדי הנתונים למסד נתונים אחד גדול.

לאחר שאספנו את הנתונים שלנו, נצטרך לטפל בהם, כמו השלמת ערכים מספריים חסרים על ידי fillna, והשמטת שורות בהם יש מידע לא קריטי על ידי dropna.

בלמידת המכונה השתמשנו בשיטת עץ החלטה, שנתן לנו התאמה של 98%, דבר המסמן על overfitting. בניסיוננו לשנות את רמת הדיוק ניסינו לגשת לעניין באמצעות cross validation בדיקה של הקטגוריות המשמעותיות ביותר , הגדלת שטח האימון של האלגוריתם ושימוש באלגוריתמי למידת מכונה אחרים, לאחר הבדיקה התברר ששיטת עץ החלטה היא השיטה הטובה ביותר למימוש למידת המכונה.

בשיטת עץ ההחלטה ראינו שמקיים overfitting לכן נאלצנו להגדיל את שטח האימון של המכונה בשביל להוריד את אחוז הדיוק. לשם כך בנינו עוד מסד נתונים שעוזר לנו בהבחנה בשחקנים בולטים בליגת ה-NBA וזהו מסד נתונים של כל זוכי תואר ה-MVP במרוצת השנים ונתונים על אותם שחקנים בסטטיסטיקות המשחק. בעזרת מפת חום אנחנו ראינו באיזה קטגוריות יש קורלציה אופטימלית ועל בסיס אותם קטגוריות חישבנו את הממוצע וביצענו השוואה לשחקני העונה הסדירה. בהתאם לאותם   
בהינתן קטגוריה של שחקן בטבלה שטובה יותר מקטגוריה של זוכה MVP , הוספנו אותו באופן מכאני לתואר מועמד לMVP , דבר שהגדיל את מרחב המדגם ובעת הגדלה ל 33% של שטח האימון (עבור 3 שנים) נקבל פחות או יותר מרחב אימון של עונה סדירה ומרחב ניבוי של עונה שלמה של מועמדים  
ייטב את האלגוריתם ל75% ופתרנו את בעיית הoverfitting.

כדי להטיב את איכות הטבלה שלנו היינו צריכים להבין יותר לגבי מה הופך שחקן רגיל לשחקן MVP ובשביל להבין זאת רכשנו טבלה של כל זוכי הMVP בכל הזמנים וחקרנו אותה שוב פעם בעזרת טבלת חום כדי להבין מהם הפרמטרים המשפיעים ביותר בתור זוכה MVP, לאחר מכן נבצע עליהם ממוצע ונשווה לשחקני המדגם.

**וויזואליזציה:**

במהלך הפרויקט השתמשנו בוויזואליזציה וEDA עמ' לבחון את הנתונים ולהסיק מהם מסקנות אשר נוכל להשתמש בהם כן לשפר את איכות התוכן של המידע שלנו, ובכך נוכל לבחון טוב יותר איזה שחקנים בולטים מהשאר ובכך להיות ממוקמים יותר גבוה לקבל תואר מועמד לMVP.

עשינו זאת באמצעות מפת חום ומטריצות קורלציות כדי לבחור עבור איזה קטגוריות יש את ההתאמה הגבוהה ביותר , בהינתן שני קטגוריות עם התאמה גבוהה הצגנו אותם בscatter plot לראות איך הם מושפעים אחד מן השני בצורה מדויקת יותר בניסיון לקבל על קטגוריות אלו עוד קצת ידע,  
ניצלנו את הידע הזה ע"י יצירת מפה של ארצות הברית והתאמה ויזואלית של "נעיצת" נקודה אדומה כך שעבור כל שחקן שאי פעם קיבל את תואר הMVP , בדקנו מאיזו קבוצה הוא הגיע , השוונו אותה לאיזה ארץ היא ממוקמת בארה"ב וע"פ כך גם נעצנו במפה את האזור והעברנו קו קלסיפיקציה המראה היכן יש יותר נתונים ולראיה קיבלנו שהרבה יותר מנצחי התואר MVP הגיעו מאזור המזרח ועל כן קיבלנו פיסת מידע נוספת שהסיכוי לנצח בתואר בהינתן הגעה מהמזרח גבוהה יותר.  
דבר שמצביע לנו גם שבדרך כלל הקבוצות שבאות מאזור המזרח הן כמעט בהכרח קבוצות חזקות יותר ועל כן יש לבדוק בקפידת יתר את השחקנים המובלים בקבוצות מהחוף המזרחי.

**סיכום**

בפרויקט שלי אני משתמש בעץ ההחלטות ככלי למידה ממוחשבת. הפרויקט שלנו מתמקד בנתוני שחקני NBA בשלוש השנים האחרונות כדי לבדוק מי הם המועמדים לתואר ה-MVP ומי לא, ולחזות את המועמדים לתואר ה-MVP לשנה הבאה. ישנם מספר סיבות שגרמו לי לבחור בעץ ההחלטות ככלי למידה ממוחשבת עבור הפרויקט שלי.

למעשה, עצי ההחלטות הם כלי מאוד קל לפענוח ולהצגה. עצי ההחלטות הם אלגוריתם למידה מנוהלת ומתאימים לבעיות סיווג ורגרסיה. הם בונים מודל על ידי חלוקת הנתונים לקבוצות קטנות יותר בהתבסס על המאפיין המשמעותי ביותר בכל צומת. כל צומת פנימי מייצג בדיקה על מאפיין וכל ענף מייצג את התוצאה של הבדיקה, וכל צומת עלה מייצג החלטה. המודל המתקבל הוא עץ דומה לעץ, והוא מאפשר פענוח קל ואינטואיטיבי של הנתונים ושל חוקי הקבלת ההחלטות שנוצרו.

בנוסף, עצי ההחלטות יכולים להתאים למגוון רחב של נתונים כגון נתונים מספריים וקטגוריים. בפרויקט שלי.

כשהתחלנו לעבוד עם עץ ההחלטות, נתקלנו במספר מורכבויות וקשיים. הבעיה העיקרית הייתה בנתון עצום של שחקני NBA והפרטים המספריים שקשה היה להתאים את העץ המתאים למטרות הפרויקט.

לכן, החלטנו להשתמש בטכניקה שנקראת "backtracking". כשעושים tracking אחורית, אנו ניצור את העץ כאילו אנו מניחים את הכלים לפענוח המידע על ידי המשתמשים בסקרים והחלטות עצמם. המתאר המצטבר של העץ הוא המתאר הכי מדויק ומתקבל על ידי חיתוך המשתמשים בכל צומת. זו הייתה טכניקה מוצלחת שהשתמשנו בה ליישום עץ ההחלטות על הנתונים שלנו.

בנוסף, היה צורך בסט נתונים גדול ומגוון כדי להשתמש בbacktracking, וכדי לשפר את סיכויי הצלחת המודל. הפעם, היינו צריכים לאסוף יותר נתונים על כל שחקן, ולסנן את הנתונים המיותרים. זה היה עבודה מורכבת ולא פשוטה, אך הצלחנו לסנן את הנתונים ולאסוף נתונים מפורטים על כל שחקן.

הפרויקט מתמקד בחיזוי השחקן הבולט ביותר (MVP) בליגת ה-NBA לעונה הנוכחית, באמצעות נתונים מהשלוש השנים הקודמות. בחרנו בנושא זה מכיוון שהנתונים זמינים באינטרנט וכי ה-NBA היא ליגה פופולרית ומעניינת למבחר גדול של אנשים. לצורך תהליך החיזוי, בחרנו בקטגוריות הבאות: נקודות, ריבאונדים, אסיסטים, חטיפות וחסימות. המטרה הייתה לבחור את השחקן המתאים ביותר להיות MVP.

כדי להבטיח איכות נתונים טובה יותר, השתמשנו בכלים כמו fillna ו-dropna על מנת לנקות ולתקן את הדאטה פריימים. בנוסף, השתמשנו במפות חום ומטריצות קורלציה כדי להשוות עמודות בהתאמה גבוהה ולהשתמש בהן לצורך החיזוי.

בחרנו בעץ ההחלטה כאמצעי הלמידה המתאים ביותר לצורך החיזוי. אחת הסיבות לבחירה זו הייתה כי האלגוריתם הזה מתאים במיוחד לנתונים חסרים או לנתונים עם רעש.

סיכום אישי למהלך העבודה:

העבודה זרקה אותנו להמון מקרים כאשר היינו צריכים להשתמש בחומר שניתן לנו בקורס בשילוב עם האינטרנט ע"מ להביא את הכל למקשה אחת שתעבוד יחד

נחשפנו לשימוש בדברים שבתחילה נראו לנו כמו "סינית" אבל לאט לאט מצאנו את עצמנו לומדים יותר ויותר איך להשתמש בכלים שלמדנו בשימוש עם העזרה שמהאינטרנט שבחוץ.

חלק מהדברים דווקא היו לנו הרבה יותר קלים , בקורס עצמו לדוגמא למדנו על מטריצת קורלציות והאינטרנט הציע לנו להשתמש במפת החום שלשנינו הייתה הרבה יותר אינטואיטיבית אך לדוגמא הרגשנו קצת אובדי עצות כשאר ניסינו לייצר מפה ולסמן ע"פ ערים את עיר הקבוצות המנצחות , אך שוב פעם הפרויקט גרם לנו הן "לשבור את הראש" והן לשפר את יכולתנו עם התמודדות בעיות ועל כן אנחנו אפילו שמחים על לקיחת חלקנו בקורס הזה.