חיזוי דירוג שחקני טניס על סמך אישיותם

מבוא כללי

מחקר רב נעשה על האישיות של ספורטאים מקצועניים (Allen et al., 2013). מספר מחקרים ניסו לגלות אילו מאפיינים אישיותיים מבחינים בין ספורטאים לספורטאי-עילית (Maksum & Indahwati, 2023; Mitić et al., 2021; Steinbrink et al., 2020). במחקר זה נבחן האם ניתן לחזות האם ספורטאי יגיע לרמה עילית על סמך מאפייני אישיות שלו. עבודה זו מנסה לנבא זאת, במקרה הזה לגבי שחקני ושחקניות טניס, באמצעות ניתוח ראיונות שלהם לתקשורת.

שיטה ותוצאות

1. נתונים

הראיונות נאספו מהאתר ASAP Sports[1]. נאספו הראיונות של כל השחקנים שאי פעם דורגו בדירוג הטניס העולמי (ATP לגברים ו-WPA לנשים), ושיש לפחות שני ראיונות שלהם לאחר שניצחו במשחק ושני ראיונות לאחר שהפסידו. כמו כן בחרתי לנתח את ארבעת (שני הניצחונות ושני ההפסדים) הראיונות הראשונים כרונולוגית (כלומר, שהתרחשו בראשית דרכו של השחקן) כדי לנסות למדוד את אישיות השחקן בתחילת דרכו ולראות האם נוכל לנבא בעזרתה את מידת הצלחתו בהמשך.

בחרתי לחלק את השחקנים לשתי קבוצות: שחקני עילית – שחקנים שדירוגם המקסימלי היה לפחות בשמינייה העליונה, ושחקנים "בינוניים" – אלה שדירוגם המקסימלי לא היה בשמינייה העליונה. בחרתי לעשות זאת כך כדי שיהיה חלוקה יחסית מאוזנת בין שתי הקבוצות, וכי מקובל להתייחס לשחקני טניס שהגיעו לדירוג בשמינייה העליונה כשחקני עילית.

לאחר איסוף הנתונים ומיונם, התקבלה רשימה של 417 שחקנים, שלכל אחד מיוחסים ארבעה ראיונות (הראיונות שינותחו).

2. Feature Engineering

השתמשתי במספר כלים על מנת להוציא מאפיינים אישיותיים מן ראיונות השחקנים. רשימת המשתנים השלמה מופיעה בנספח.

2.1. Large Language Model – LLM

נמצא שמודלי שפה גדולים (LLMs) מבצעים עבודה די טובה בהערכת תכונות ה-big 5 (Derner et al., 2024), ומאפיינים פסיכולוגיים אחרים (Rathje et al., 2024). תחת ההנחה כי מודלי השפה השונים מבצעים את הפעולות האלה במידת דיוק דומה, נעשה שימוש במודל Gemini של Google בגלל האופציה החינמית שלו. המאפיינים שנאספו באמצעות Gemini היו כדלהלן:

- **תכונות ה-big 5:** מספר מחקרים ניסו לנבא הצלחה בספורט באמצעות תכונות ה-big 5 (Ghaderi & Ghaderi, 2012; Piedmont et al., 1999). מחקר אחד מצא שספורטאי עילית היו גבוהים יותר בנעימות (agreeableness) ומצפוניות (conscientiousness) ונמוכים יותר בנוירוטיות (neuroticism) (Steca et al., 2018). הקלט (prompt) ל-Gemini נבנה לפי הקווים המנחים לקלט של מודל שפה שמתוארים במאמר של שואנגר (Schoenegger et al., n.d.) על מנת להגיע לדיוק מקסימלי ושונות מינימלית. הקלט היה בנוי מהוראות למטלה והנחיות איך לבצע אותה, ולאחר מכן כל הראיונות של שחקן מסוים.

[1] https://www.asapsports.com/showcat.php?id=7&event=yes

- **תכונות אישיותיות:** לפי המאמר מאת ואן רוסום (van Rossum, 2006), מאמנים תופסים שהתכונות החשובות ביותר לספורטאי עילית הן ביטחון עצמי (self-confidence), רצון לנצח (will to win), ריכוז (concentration), דבקות (persistence) ותחרותיות (competitiveness). הקלט שניתן ל-Gemini על מנת שיעריך תכונות אלה היה זהה לקלט של תכונות ה-big 5 מלבד התכונות השונות שעליו לנתח והגדרתן.

2.2. LIWC

Linguistic Inquiry and Word Count (LIWC) היא תוכנה הסופרת מילים מקטגוריות (מילונים) מסוימות. בהתבסס על המאמר של מיטיץ' ועמיתים (Mitić et al., 2021), נבחרו קטגוריות בהקשר של זמן (עבר, הווה ועתיד), מסוגלות עצמית (self-efficacy) (כגון רגש חיובי ורגש שלילי) (Malloch & Feng, 2022) ואינטליגנציה רגשית (רגש חיובי, רגש שלילי, נימוס, קונפליקט) (Dover & Amichai-Hamburger, 2023). כמו כן, נבדקו גם קטגוריות נוספות כגון גוף (ראשון, שני ושלישי, יחיד ורבים), מניעים (הישגים, כוח, שייכות), פיזיולוגיה (בריאות, מחלה, רווחה, רווחה מנטלית) וכו'. כאן חשוב להזכיר שלכל שחקן נותחו אותה כמות של ראיונות אחרי ניצחון, שצפויים להכיל מילים חיוביות, וראיונות אחרי הפסדים, שצפויים להכיל מילים שליליות, במטרה למנוע מעגליות ביכולת ההסקה ממשתנים חיוביים ושליליים.

2.3. מאפיינים טקסטואליים נוספים

- **אורך דיבור השחקן ביחס לאורך הריאיון:** רציתי לבדוק האם ליחס בין כמה השחקן מדבר בראיון לבין אורך הריאיון כולו יכולה להיות יכולת חיזוי, בעקבות הקורלציה החיובית בין אורך תשובה לבין משתני אישיות כמו מוחצנות ומצפוניות (Dai et al., 2022). לכן הוספתי גם משתנה שמתאר את היחס בין אורך תשובות הנבדק לבין אורך הריאיון כולו.

2.4. מאפיינים דמוגרפיים

הוספתי לניתוח גם מספר נתונים דמוגרפיים כמו שנת וחודש לידה, מין, יד דומיננטית וגובה. הנתונים הדמוגרפיים כמו גם דירוגי השחקנים נלקחו מה-GitHub של ג'ף סאקמן (Sackmann, 2024b, 2024a).

3. ניבוי

3.1. המודלים

אימנתי שלושה מודלים שונים. בכל המודלים השתמשתי ב-10-fold CV כשיטת ה-resampling כאשר כל המודלים מאומנים עם אותם folds. כמו כן ה-preprocessing של כל המודלים כלל תקנון משתנים מספריים והוספת עמודות אינדיקטורים (one-hot) למשתנים קטגוריאליים. הפרטים לגבי כל מודל מתוארים בטבלה 1.

טבלה 1: סיכום המודלים שאומנו

מודל	כללי	Tuning	ערך נבחר ומשמעות
SVM	SVM פולינומיאלי	C (פרמטר ענישה): נבדקו ערכים בין אפס (לא כולל – כי אין וקטור שמפריד בין הקבוצות באופן "מושלם") ל16 שמפוזרים באופן לוגריתמי (יותר ערכים קטנים) מעלה (של הפולינום): נבדקו המעלות אחד, שתיים ושלוש	**C: 0.58** – ערך C נמוך יחסית, כלומר ה"עונש" על סיווגים

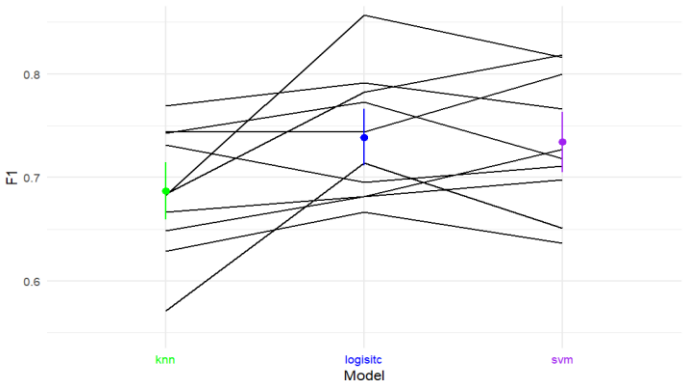
			לא נכונים הוא קטן יחסית **מעלה: 1** – SVM לינארי
רגרסיה לוגיסטית	מודל רגרסיה לוגיסטית שמשתמש ב-elastic net כדי לצמצם את מספר המשתנים	אלפא: נבדקו ערכים בין אפס לאחד (כולל אפס ואחד, במטרה לבדוק גם ridge ו-lasso) שמפוזרים באופן אחיד למדא (פרמטר ענישה): נבדקו ערכים בין אפס לאחד (כולל אפס, במטרה לבדוק גם OLS), שמפוזרים באופן לוגריתמי (לא נבדקו ערכים מעל אחד בעקבות ניסויים קודמים שהראו שהתוצאה אינה משתפרת)	**אלפא: 0.16** – נטייה ליותר עונש ridge **למדא: 0.29** – עונש יחסית קטן
KNN	מודל KNN	K: נבדקו ערכים אי-זוגיים (כדי למנוע "תיקו-ים" בעקבות הימצאות רק שני קטגוריות) בין 1 ל-20	**13** – מודל קצת פחות גמיש (ביחס לשאר ערכי ה-K המועמדים)

3.2. השוואה בין המודלים

השוואתי בין המודלים באמצעות שימוש ב -10 fold CV עם מדד להערכת טיב המודל של F1. השתמשתי ב-F1 מכיוון שהדבר המרכזי שאני מעוניין לבחון הוא האם השחקן יהיה ברמה עילית, וכי חשוב במקרה שלנו להימנע מטעויות מסוג false negative ו-false positive באותה מידה. המודלים אומנו על אותם folds. את תוצאות המודלים ניתן לראות בטבלה. הקורלציה בין המודלים לא הייתה גבוהה במיוחד – לאף מודל לא הייתה קורלציה עם מודל אחר הגבוהה מ0.4. השונות בין ה-folds השונים בכל מודל הייתה יחסית נמוכה לכן ניתן להסיק שביצועי המודל יציבים למדגמים שונים. התוצאות לכל fold מוצגות בגרף 1.

טבלה 2: ביצועי המודלים

	ממוצע F1 בין folds	סטיית תקן של F1 בין folds	מקדם שונות (Coefficient of Variation)
SVM	0.73	0.025	0.034
רגרסיה לוגיסטית	0.74	0.027	0.037
KNN	0.69	0.027	0.04



גרף 1: ביצועי המודלים לפי ה-folds השונים

1.1. מודל סופי

המודל הסופי שנבחר הוא מודל הרגרסיה הלוגיסטית הן בגלל ציון ה-F1 הגבוה שלו והשונות הדומה בין כל המודלים. ביצועי מודל הרגרסיה הלוגיסטית מוצגים בטבלה 3. נשים לב כי ציון ה-F1 נמוך במעט מהציון על ה-training set ולכן המודל עושה overfitting במידה מסוימת. כמו כן, בהתבוננות על ה-confusion matrix

טבלה 4: confusion matrix

		אמת	
		לא	**כן**
ניבוי	**לא**	48	43
	כן	4	10

(טבלה 4) נשים לב כי המודל נוטה לנבא כי השחקן לא יהיה ספורטאי עילית, וכשהוא מנבא ששחקן כן יהיה ספורטאי עילית הוא צודק באחוזים גבוהים.

טבלה 3: ביצועי המודל על ה-test set	
מדד	תוצאה
F1	0.67
דיוק (Accuracy)	0.55
רגישות (Sensitivity)	0.92
ספציפיות (Specificity)	0.19

דיון

לאחר ביצוע elastic net על מודל הרגרסיה הלוגיסטית נשארו 11 משתנים שלא התאפסו. כמובן שאי אפשר להסיק מכך דבר, לכן בניתי מודל רגרסיה לוגיסטית נוסף על סמך ה-test set המשתמש רק ב-11 המשתנים הללו. המקדמים של המשתנים במודל זה וה-p-value שלהם מוצגים בטבלה. מתוך 11 המשתנים, חמישה יצאו מובהקים סטטיסטית במודל החדש – LIWC_power , N5, year, LIWC_Analytic ו-LIWC_Clout ואחד כמעט מובהק - LIWC_mental, שאמנם הוא ה-p-value שלו אינו קטן מ-5%, אך הוא קרוב לכך מאוד וההשפעה שלו מאוד גדולה (הוא בעל המקדם הגדול ביותר למרות שכל שאר משתני ה-LIWC הם באותה סקאלה כמוהו). את ההסברים על המשתנים ניתן למצוא בנספח.

טבלה 5: המשתנים, מקדמיהם וה-P-value		
	Estimate	Pr(>|z|)
LIWC_power	-1.99	0.0008
N5	-0.21	0.0198
year	-0.03	0.0231
LIWC_Analytic	-0.05	0.0333
LIWC_Clout	-0.04	0.0458
LIWC_mental	-44.28	0.0516
answers_ratio	-2.51	0.1119
LIWC_function	0.10	0.2202
LIWC_lack	0.72	0.4005
LIWC_fatigue	0.75	0.5441
LIWC_prosocial	0.09	0.9017

התוצאות הללו יכולות להעיד שאי-ביטוי מניעים כוחניים, רמת נוירוטיות נמוכה, אי-ביטוי חשיבה אנליטית או מנהיגותית ואי-דיבור על בריאות נפשית, יחד עם שנת לידה מוקדמת (נמוכה), בעלי יכולת ניבוי לכך שהשחקן יהיה ספורטאי עילית.

מעניין לראות כי התוצאות הללו תואמות את ממצאיהם של סטקה ועמיתים (Steca et al., 2018) בכך שספורטאי עילית נמוכים בנוירוטיות. אפשר גם לשים לב להטיה בכך שכל ששנת הלידה של השחקן מוקדמת יותר כך יש לו סיכוי גבוה יותר להיות ספורטאי עילית, זאת כנראה בעקבות כך שבעבר ראיינו בעיקר ספורטאי עילית וכיום נוטים לעשות זאת פחות. לא הצלחתי למצוא ספרות נוספת התומכת בתוצאות האחרות.

טבלה 6: תוצאות חקירת פוסט-הוק				
Model	F1	Accuracy	Sensitivity	Specificity
knn	0.68	0.64	0.79	0.49
logistic	0.67	0.55	0.92	0.19
svm	0.63	0.54	0.79	0.30

בעקבות התוצאות הדומות של שלושת המודלים על ה-training set, הוחלט לבצע גם חקירת פוסט-הוק של ביצועיהם על ה-training set. תוצאות אלו מוצגות בטבלה 6. ניתן לראות שדווקא מודל ה-KNN הפשוט לכאורה מציג את הביצועים הטובים ביותר גם ב-F1 וגם בדיוק. זה בעוד מודל הרגרסיה הלוגיסטית הוא בעל רגישות גבוהה יותר. ניתן גם לראות שביצועי מודל הרגרסיה וה-SVM דומים מאוד. שני המודלים האלה יותר שמרנים ביחס למודל ה-KNN.

לסיכום, שלושת המודלים שאימנתי היו בעלי ביצועים יחסית דומים, כאשר מודל הרגרסיה הלוגיסטית ומודל ה-SVM דומים בביצועיהם וב-overfitting שלהם, בעוד מודל ה-KNN הציג ביצועים פחות טובים על ה-training set אך ביצועים יותר טובים על ה-test set. מודל הרגרסיה הלוגיסטית, באמצעות שימוש ב-elastic net אִפשר לנו לבחון את חשיבות והשפעת המשתנים השונים, על מנת להבין יותר טוב את הגורמים המשפיעים על הפיכת ספורטאי לספורטאי עילית.

נספחים

משתנה	הגדרה
top_level	"yes" אם הדירוג הגבוה ביותר של השחקן אי-פעם הוא בין שמונה לאחד (כולל), ו-"no" אחרת
Sex	"M" לזכר, "F" לנקבה
year	שנת לידה
month	חודש לידה
hand	יד דומיננטית. "L" לשמאל, "R" לימין
height	גובה השחקן בס"מ
O5	מידת פתיחות (Openness) בין אחד לעשר לפי ה-big-5. 10 משמעו מאוד פתוח, אחד משמעו מאוד סגור
C5	מידת מצפוניות (Conscientiousness) בין אחד לעשר לפי ה-big-5. עשר משמעו מאוד מצפוני
E5	מידת מוחצנות (Extraversion) בין אחד לעשר לפי ה-big-5. עשר משמעו מאוד מוחצן
A5	מידת נעימות (Agreeableness) בין אחד לעשר לפי ה-big-5. עשר משמעו מאוד נעים
N5	מידת נוירוטיות (Neuroticism) בין אחד לעשר לפי ה-big-5. עשר משמעו מאוד נוירוטי
confi	מידת ביטחון עצמי בין אחד לעשר. עשר משמעו מאוד בטוח בעצמו
will	מידת 'רצון לנצח' בין אחד לעשר. עשר משמעו מאוד רוצה לנצח
concer	מידת ריכוז בין אחד לעשר. עשר משמעו מאוד מרוכז
persis	מידת עקשנות בין אחד לעשר. עשר משמעו מאוד עקשן
comp	מידת תחרותיות בין אחד לעשר. עשר משמעו מאוד תחרותי
answers_ratio	היחס בין אורך תשובות השחקן לבין אורך הריאיון כולו
LIWC_Analytic	אחוז המילים בתשובות השחקן המבטאות חשיבה אנליטית
LIWC_Clout	אחוז המילים בתשובות השחקן המבטאות מנהיגות\ סטטוס
LIWC_Authentic	אחוז המילים בתשובות השחקן המבטאות אותנטיות
LIWC_Tone	אחוז המילים בתשובות השחקן המבטאות טון ריגשי
LIWC_function	אחוז מילות הפעולה בתשובות השחקן
LIWC_pronoun	אחוז מילות הגוף בתשובות השחקן
LIWC_ppron	אחוז מילות גוף אישי בתשובות השחקן
LIWC_i	אחוז מילות גוף ראשון יחיד בתשובות השחקן
LIWC_we	אחוז מילות גוף ראשון רבים בתשובות השחקן
LIWC_you	אחוז מילות גוף שני בתשובות השחקן
LIWC_shehe	אחוז מילות גוף שלישי יחיד בתשובות השחקן
LIWC_they	אחוז מילות גוף שלישי רבים בתשובות השחקן

LIWC_Drives	אחוז המילים בתשובות השחקן המבטאות מניעים
LIWC_affiliation	אחוז המילים בתשובות השחקן המבטאות שיוכיות
LIWC_achieve	אחוז המילים בתשובות השחקן המבטאות הישגיות
LIWC_power	אחוז המילים בתשובות השחקן המבטאות כוח
LIWC_allnone	אחוז המילים בתשובות השחקן המבטאות גישה של 'הכל או כלום'
LIWC_tone_pos	אחוז המילים בתשובות השחקן בעלות טון חיובי
LIWC_tone_neg	אחוז המילים בתשובות השחקן בעלות טון שלילי
LIWC_emotion	אחוז המילים בתשובות השחקן המבטאות רגש
LIWC_emo_pos	אחוז המילים בתשובות השחקן המבטאות רגש חיובי
LIWC_emo_neg	אחוז המילים בתשובות השחקן המבטאות רגש שלילי
LIWC_emo_anx	אחוז המילים בתשובות השחקן המבטאות חרדה\ לחץ
LIWC_emo_anger	אחוז המילים בתשובות השחקן המבטאות כעס
LIWC_emo_sad	אחוז המילים בתשובות השחקן המבטאות עצב
LIWC_prosocial	אחוז המילים בתשובות השחקן המבטאות התנהגות פרו-חברתית
LIWC_polite	אחוז מילות הנימוס בתשובות השחקן
LIWC_conflict	אחוז המילים בתשובות השחקן המבטאות קונפליקט
LIWC_comm	אחוז המילים התקשוריות בתשובות השחקן
LIWC_family	אחוז המילים בתשובות השחקן שקשורות למשפחה
LIWC_friend	אחוז המילים בתשובות השחקן שקשורות לחברים
LIWC_health	אחוז המילים בתשובות השחקן שקשורות לבריאות
LIWC_illness	אחוז המילים בתשובות השחקן שקשורות לחולי
LIWC_wellness	אחוז המילים בתשובות השחקן שקשורות לרווחה
LIWC_mental	אחוז המילים בתשובות השחקן שקשורות לבריאות נפשית
LIWC_need	אחוז המילים בתשובות השחקן המבטאות צורך
LIWC_want	אחוז המילים בתשובות השחקן המבטאות רצון
LIWC_acquire	אחוז המילים בתשובות השחקן המבטאות השגה
LIWC_lack	אחוז המילים בתשובות השחקן המבטאות מחסור
LIWC_fulfill	אחוז המילים בתשובות השחקן המבטאות סיפוק
LIWC_fatigue	אחוז המילים בתשובות השחקן המבטאות עייפות
LIWC_reward	אחוז המילים בתשובות השחקן המבטאות הישגים
LIWC_risk	אחוז המילים בתשובות השחקן המבטאות סיכון
LIWC_curiosity	אחוז המילים בתשובות השחקן המבטאות סקרנות
LIWC_allure	אחוז המילים בתשובות השחקן המבטאות קסמיות
LIWC_time	אחוז המילים בתשובות השחקן הקשורות לזמן
LIWC_focuspast	אחוז המילים בתשובות השחקן הקשורות לזמן עבר

LIWC_focuspresent	אחוז המילים בתשובות השחקן הקשורות לזמן הווה
LIWC_focusfuture	אחוז המילים בתשובות השחקן הקשורות לזמן עתיד

ביבליוגרפיה

Allen, M. S., Greenlees, I., & Jones, M. (2013). Personality in sport: A comprehensive review. *International Review of Sport and Exercise Psychology*, *6*(1), 184–208. https://doi.org/10.1080/1750984X.2013.769614

Dai, Y., Jayaratne, M., & Jayatilleke, B. (2022). Explainable Personality Prediction Using Answers to Open-Ended Interview Questions. *Frontiers in Psychology*, *13*, 865841. https://doi.org/10.3389/fpsyg.2022.865841

Dover, Y., & Amichai-Hamburger, Y. (2023). Characteristics of online user-generated text predict the emotional intelligence of individuals. *Scientific Reports*, *13*(1), 6778. https://doi.org/10.1038/s41598-023-33907-4

Ghaderi, D., & Ghaderi, M. (2012). *Survey the relationship between big five factor, happiness and sport achievement in Iranian athletes*.

Maksum, A., & Indahwati, N. (2023). Personality traits, environment, and career stages of top athletes: An evidence from outstanding badminton players of Indonesia. *Heliyon*, *9*(3), e13779. https://doi.org/10.1016/j.heliyon.2023.e13779

Malloch, Y., & Feng, B. (2022). What Motivates People to Support?: Impacts of Message Valence and Self-Efficacy on Linguistic Features of Response. *Frontiers in Psychology*, *13*, 798205. https://doi.org/10.3389/fpsyg.2022.798205

Mitić, P., Nedeljković, J., Bojanić, Ž., Franceško, M., Milovanović, I., Bianco, A., & Drid, P. (2021). Differences in the Psychological Profiles of Elite and Non-elite Athletes. *Frontiers in Psychology*, *12*, 635651. https://doi.org/10.3389/fpsyg.2021.635651

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C. E., & Van Bavel, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, *121*(34), e2308950121. https://doi.org/10.1073/pnas.2308950121

Sackmann, J. (2024a). *Tennis_atp* [Dataset]. https://github.com/JeffSackmann/tennis_atp

Sackmann, J. (2024b). *Tennis-wta* [Dataset]. https://github.com/JeffSackmann/tennis_wta

Schoenegger, P., Greenberg, S., Grishin, A., Lewis, J., & Caviola, L. (n.d.). *Can AI Understand Human Personality? - Comparing Human Experts and AI Systems at Predicting Personality Correlations*.

Steca, P., Baretta, D., Greco, A., D'Addario, M., & Monzani, D. (2018). Associations between personality, sports participation and athletic success. A comparison of Big Five in sporting and non-sporting adults. *Personality and Individual Differences*, *121*, 176–183. https://doi.org/10.1016/j.paid.2017.09.040

Steinbrink, K. M., Berger, E. S. C., & Kuckertz, A. (2020). Top athletes' psychological characteristics and their potential for entrepreneurship. *International Entrepreneurship and Management Journal*, *16*(3), 859–878. https://doi.org/10.1007/s11365-019-00612-6

van Rossum, J. H. A. (2006). Psychological Characteristics of Elite Athletes According to Top Level Coaches. *High Ability Studies*, *7*(1), 15–23. https://doi.org/10.1080/0937445960070103