רקע אקדמי והצלחה בהכשרת מדעי הנתונים: מחקר גישוש במסלול הטכנולוגי לתואר שני במדעי המידע

אריאל רוזנפלד

המחלקה למדעי המידע, אוניברסיטת בר-אילן

ariel.rosenfeld@biu.ac.il

אבשלום אלמלח

המחלקה למדעי המידע, אוניברסיטת בר-אילן

avshalom.elmalech@biu.ac.il

תקציר

תוכניות הכשרה רבות למדעי המידע (Information Science – IS) מרחיבות בהדרגה את תוכניות הלימודים שלהן כדי לכלול קורסים למדעי הנתונים החישוביים, כגון למידת מכונה מפוקחת ולמידת מכונה לא מפוקחת. תוכניות אלו מתמקדות הן בפיתוח מיומנויות קלאסיות במדעי המידע והן בפיתוח מיומנויות ליבה של מדעי הנתונים בקרב תלמידיהן. מכיוון שמיומנויות מדעי הנתונים קשורות לחשיבה מתמטית וחישובית, הצוות האקדמי וסטודנטים פוטנציאליים מעלים לעתים חששות לגבי הרקע הנדרש לסטודנטים כדי להצליח בחלקה החישובי של ההכשרה. על מנת להתמודד עם החששות, אנו מדווחים על מחקר גישוש שבאמצעותו בחנו את מחזור הסטודנטים של הכשרת בוגרי IS באוניברסיטת בר-אילן לשנת 2020 (להלן: התוכנית שלנו), תוך התמקדות בקורסי הליבה של מדעי הנתונים החישוביים – למידת מכונה מפוקחת ולמידת מכונה לא מפוקחת. המחקר שלנו מראה כי בניגוד לחששות שהובעו, סטודנטים בעלי רקע במדעי הרוח נוטים לקבל ציון גבוה יותר בחלקה החישובי של ההכשרה לעומת אלו בעלי רקע ממדעי החברה, ואף מצליחים יותר בתוכנית ההכשרה כולה. באשר לחששות הנוגעים למגדר ולגיל, איננו מוצאים בנתונים ראיות התומכות בחששות הללו. נוסף לכך, הציון הממוצע של סטודנטים לתואר ראשון מנבא היטב הן את הצלחתם בכל תוכנית ההכשרה והן בחלקה החישובי. לבסוף, הנתונים מצביעים על הלימה משמעותית בין הצלחה בחלק החישובי של ההכשרה ושאר חלקי ההכשרה.

מילות מפתח- חינוך למדעי מידע; חינוך למדעי נתונים; הכשרה למדעי נתונים חישוביים

Academic Background and Success in Data Science Training: An Exploratory Study in the Technological Track for a Master's Degree in Information Science

Abstract

Many information science (IS) training programs are gradually expanding their curricula to include computational data science courses such as supervised and unsupervised machine learning. These programs focus on developing both classic information science competencies as well as core data science competencies among their students. Since data science competencies often associated with mathematical and computational thinking, departmental officials and prospective students often raise concerns regarding the appropriate background students should have in order to succeed in this newly introduced computational content of the program. In order to address these concerns, we report on an exploratory study through which we examined the 2020 student class of Bar-Ilan University's IS graduate training, focusing on the core computational data science courses (i.e., supervised and unsupervised machine learning). Our study shows that, to the contrary of many of the expressed concerns, students from the Humanities tend to score higher in data science competencies than those from the Social Sciences and better succeed in the training program as a whole. In addition, students' Bachelors' average grade acts as a good indicator for both their success in the training program and in the data science part thereof. In addition, we find no evidence to support concerns regarding age or sex. Finally, our study suggests that the computational data science part of students' training is very much aligned with the rest of their training program.

תחילת דבר

פרופ' יהודית בר-אילן ז"ל, בשבתה כראשת המחלקה למדעי המידע, היתה מהחלוצות להבין את הצורך וחשיבות העולם החישובי עבור מדעני המידע. כמדענית מחשב בהכשרתה וכמדענית נתונים היא דגלה בגישה משלבת המחברת בין העולם הטכנולוגי והחישובי לעולם מדעי המידע הקלאסי. שילוב זה מייצר הזדמנויות ואתגרים מיוחדים בעיקר עבור סטודנטים מרקעים מגוונים. במחקר זה אנו מתמקדים בחלק מאתגרים אלו.

מבוא

מדעי המידע (Information Science – IS) היא דיסציפלינה העוסקת בידע הקשור ליצירה, איסוף, ארגון, אחסון, שליפה, פרשנות, העברה, טרנספורמציה וניצול של מידע (Belkin & Robertson, 1976; Borko, 1968; Williams, 1988). בארגונים רבים מצטבר כיום נפח נתונים גבוה המגיע ממגוון מקורות ובצורות רבות ושונות, למשל טקסט, וידאו ואודיו. על מנת להפיק תובנות תקפות מבחינה מדעית ועסקית מנתונים אלה, קהילת IS מאמצת בהדרגה טכניקות של מדעי הנתונים (Data Science – DS).

אף שניתן לייחס את תחום DS לתחומי הסטטיסטיקה ומדעי המחשב (Davenport, 2020), בעשור האחרון נעשה שימוש בטכניקות DS כדי להתמודד עם מגוון רחב של בעיות ושאלות מחקר מדיסציפלינות שונות מחוץ למדעים המדויקים, כגון מוזיקה (Burgoyne *et al*., 2015), לימודי ספרות (Rommel, 2004), ארכיאולוגיה (Eiteljorg, 2004; Forte 2015), בלשנות (Hajic, 2004), היסטוריה (Thomas & William 2004; Zaagsma, 2013), פילוסופיה (Ess, 2004) ועוד. למטרת המחשה, הבה נבחן שתי דוגמאות מתחום דעת אשר איננו בליבת מדעי המחשב והסטטיסטיקה:

דוגמה 1: בהינתן סט של כתבי יד, חוקרת מעוניינת בתיארוך אוטומטי של כתבי היד, זיהוי המחבר, המקור וכו'.

דוגמה 2: בהינתן סט של תמונות, חוקר מעוניין לסווג את התמונות באופן אוטומטי לקבוצות דומות (למשל לפי ז'אנר מסוים או סגנון ציור), לזהות דפוסים בין מאפיינים שונים של התמונות (למשל, בתמונות מאירופה נעשה שימוש רב יותר בצבעים כהים), זיהוי חריגות בסט התמונות (למשל, תמונה בעלת מאפיינים שונים מהאחרות) וכו'.

דוגמאות אלו אינן תאורטיות, הן ניצבות במרכזם של מספר מחקרים אקדמיים (Badea *et al*. 2018; Kestemont ,2014; Koppel *et al*. 2006).

על מנת לספק לסטודנטים את הכלים החישוביים הדרושים להתמודדות עם אתגרים באקדמיה ובתעשייה, תוכניות הכשרה של IS מרחיבות בהדרגה את תוכניות הלימודים שלהן כך שיכללו לא רק תכנים "קלאסיים" של IS, אלא גם תכנים מתקדמים, כגון מדעי הרוח הדיגיטליים, אינפורמטיקה משפטית ותכנים חישוביים

במדעי הנתונים. קורסי מדעי הנתונים, בניגוד לתוכני IS הקלאסיים, קשורים יותר לתחום המכונה חשיבה חישובית (Wing, 2006). בדרך כלל עיסוק ב-DS דורש כתיבת סקריפטים אוטומטיים, יישום של מודלים סטטיסטיים ועיבוד כמויות גדולות של נתונים מובנים ו/או לא מובנים לצורך חילוץ דפוסים ותובנות בעלות משמעות מתמטית. משום כך הכנסת תוכני DS בתוכניות IS הובילה את הצוות האקדמי של המחלקה והסטודנטים הפוטנציאליים להעלות חששות לגבי הרקע המתאים הנדרש לסטודנטים כדי להצליח בחלק זה של הכשרתם. בדרך כלל דאגות אלו נסבות על כמה חששות מרכזיים: רקע אקדמי רלוונטי, הצלחה קודמת במבחני מיון אקדמיים או בתואר אקדמי קודם, גיל (סטודנטים עלולים לחוש מבוגרים מכדי להבין חומר טכנולוגי) ומגדר (הדעה השלטת היא שתחומים הקשורים למחשבים נשלטים על ידי גברים).

הספרות המקצועית בחנה את החששות אשר הוזכרו לעיל בלימודים אקדמיים באופן כללי ובדגש על דיסציפלינות ספציפיות. מחקרים הראו כי יש קשר ישיר בין חששות סטודנטים לבין הצלחתם בלימודים אקדמיים. למשל פרץ ואחרים (Perez *et al*., 2014) הראו כי סטודנטים אשר חששו שאינם בעלי הרקע המתאים לתחום לימודי מסוים, לא הצליחו באותו תחום. חששות אלו יכולים להשפיע גם על מועמדים עוד בטרם התנסו בתחום דעת מסוים. ספציפית בתחום הטכנולוגי, שיעור הגברים אשר מתחילים את לימודי התואר הראשון גדול משמעותית משיעור הנשים, ואחת הסיבות המרכזיות לכך היא הפחד של נשים והחשש שלא יצליחו בתחום זה (Sax *et al*., 2015; Speer, 2021). נוסף לחששות על בסיס רקע אקדמי ומגדר, מחקרי עבר מצביעים על חששות דומים גם בקרב סטודנטים מבוגרים (Yunus *et al*., 2016).

מטרת מחקר זה הינה בחינת החששות שהוזכרו לעיל באמצעות מחקר תצפית גישוש הבוחן את אוכלוסיית הבוגרים לשנת 2020 בהכשרת בוגרי מדעי המידע של אוניברסיטת בר-אילן, תוך התמקדות בקורסים של מדעי הנתונים החישוביים ובפרט למידת מכונה מפוקחת ולא מפוקחת. למיטב ידיעתנו, חששות אלו טרם נבחנו בספרות הקיימת.

במסגרת מחקר זה התמקדנו במספר מדדי מפתח לצורך אפיון הסטודנטים הנבדקים והצלחתם. מדדים אלו כוללים נתונים דמוגרפיים (דוגמת גיל ומגדר), נתוני קבלה (כגון ציוני המבחן הפסיכומטרי וציוני התואר הראשון) וציוני הצלחה בהכשרה (כגון ממוצע ציונים בתואר השני והציונים בקורסי מדעי הנתונים החישוביים). המדדים הדמוגרפיים הנבחנים מבוססים היטב בספרות הקיימת באשר לקשריהם עם הצלחה אקדמית. למשל, הדומיננטיות הגברית במקצועות המדעיים נחקרה רבות לאורך השנים והסברים מגוונים ניתנו למקור ההבדל המגדרי בהצלחה האקדמית (Kanny *et al*., 2014; Pezzoni *et al*., 2016; Sax, 2001; Shapiro & Sax, 2011; Speer, 2021). בדומה לכך נחקרה הצלחתם של סטודנטים מבוגרים יותר בלימודים אקדמיים (Hoskins *et al*., 1997; McNeil *et al*., 2014). נתוני הקבלה ששימשו אותנו במחקר נסמכים גם הם על הספרות הקיימת, שהראתה כי תוצאות המבחן הפסיכומטרי הן המדד העקבי והמבוסס ביותר להערכת הצלחה אקדמית של מועמדים (Burton & Ramist, 2001), וגם ציוני התואר

הראשון מנבאים היטב הצלחה בתארים מתקדמים במגוון תחומי דעת (Braunstein, 2002; Kuncel *et al*., 2007; McKee *et al*., 2001). מדדים אלו טרם נבחנו במסגרת הכשרת בוגרים ב-IS וגם לא ספציפית בהכשרתם החישובית של סטודנטים במדעי המידע.

סקירת ספרות

הכשרת אנשי IS מתבצעת בדרכים שונות, לרבות תוכניות וקורסים מיוחדים באוניברסיטאות, סדנאות וכנסים ייעודיים. מספר תוכניות ה-IS במוסדות אקדמיים גדל באופן אקספוננציאלי בשנים האחרונות (Wang, 2018). הגידול האדיר יכול להעיד על צורך באנשי מקצוע בתחום זה. הסעיפים הבאים מורכבים מסקר קצר על התפקיד של DS בהכשרה של בוגרי תוכניות IS, תיאור וסקירה קצרה של הגישות המרכזיות בתחום מדעי הנתונים וביטוין בתוכנית הלימודים שלנו.

הקניית מיומנויות DS לתלמידי IS

תוכניות הכשרה למדעי המידע ברחבי העולם ממציאות את עצמן מחדש באופן תדיר ומפתחות תוכניות לימודים חדשות כדי להכשיר אנשי מקצוע בתחום המידע עם הידע והכישורים הנכונים שיתאימו לשינויים בצרכים החברתיים ובשוק העבודה (Bronstein, 2015; Johnson, 1999; Juznic & Badovinac, 2005, Hjørland, 2002). היבט מרכזי בתהליך זה הוא השילוב והפיתוח של מיומנויות DS בתוכניות הכשרה של IS (Wang, 2018; Zuo *et al*., 2017). DS הוא תחום דעת רחב אשר הוגדר בצורות שונות במהלך קיומו (VanDyk *et al*., 2015; Garber, 2019). גילו האמיתי, הקשר שלו לתחומים קיימים כמו סטטיסטיקה ומדעי המחשב, ואפילו הפרופיל של העוסקים בו, נדון בהרחבה בספרות המקצועית (Davenport, 2020). באופן מסורתי DS הוגדר כהרחבה של סטטיסטיקה ומתמטיקה (Cleveland, 2001). עם זאת, במהלך השנים התברר כי DS מצריך מיומנויות נרחבות ומקיפות אשר אינן מסתכמות בסטטיסטיקה, כגון מיומנויות אנליטיות, מיומנויות תקשורת, מיומנויות מתמטיות, מיומנויות תכנות ועוד רבות אחרות (Doyle, 2019). המיומנויות המגוונות של מומחי DS באות לידי ביטוי גם ברקע האקדמי וההכשרה המגוונת שלהם (Davenport & Patil, 2012), החל ממדעים מדויקים, כגון פיזיקה ניסויית, ועד למדעי החברה, כגון סוציולוגיה. כתוצאה מכך אין כיום הגדרה מוסכמת של המיומנויות של עוסקים במדעי הנתונים (Fayyad & Hamutcu, 2020), על אף ניסיונות שונים לנסח אותה כראוי בהקשרים שונים (למשל: Agarwal & Dhar, 2014; Cao, 2017; Costa & Santos, 2017; Dhar, 2013; van der Aalst, 2014).

חוקרים ואנשי חינוך ניסו להגדיר כיצד אמורה להיראות תוכנית הלימודים ב-DS (Baumer, 2015; Brunner & Kim, 2016). מאחר שמתודולוגיות ומיומנויות DS נדרשות על ידי מומחים מדיסציפלינות שונות, קורסי DS מתפתחים ומוצעים בדיסציפלינות שונות וזמינים לרוב הסטודנטים ללא קשר לרקע שלהם ולמקצוע הראשי שהם לומדים (Dichev & Dicheva, 2017). מחקרים קודמים מצביעים על כך שהמיקוד העיקרי של קורסי DS צריך להיות בסטטיסטיקה, למידת מכונה, ויזואליזציה, אתיקה וחישוביות (Dichev & Dicheva, 2017). המכון למתמטיקה של פארק סיטי פרסם דו"ח ובו רשימה של קורסים המבטיחה הכשרה של בוגרי DS עם היכולות הנדרשות לתחום (DeVeaux *et al*., 2017). המחברים זיהו שישה תחומי נושא עיקריים של DS: תיאור ואינטגרציית נתונים, יסודות מתמטיים, חשיבה חישובית, חשיבה סטטיסטית, תקשורת בין-אישית ואתיקה בשימוש בנתונים. מחקר עדכני שסקר 69 חברי סגל המלמדים DS במכללות ובאוניברסיטאות הרכיב רשימה של נושאים הנלמדים לרוב בקורסי מבוא ל-DS ומשותפים לכל הדיסציפלינות. ואלה הנושאים: הדמיית נתונים, ניקוי נתונים, אתיקה, ניהול נתונים, שיטות סטטיסטיות, ארכיטקטורת נתונים ולמידת מכונה (Schwab-McCoy *et al*., 2020).

את נושאי ה-DS העולים מרשימות אלו ואחרות ניתן לייחס לארבע מיומנויות היסוד ב-DS הנדרשות כמעט בכל פרויקט מבוסס DS: עיבוד מקדים של נתונים, חקר נתונים, ניתוח נתונים והצגת נתונים (Kang *et al*., 2015). למעשה, כל מערך המיומנויות הזה בא לידי ביטוי בתהליך של פיתוח ויישום אפליקציות של למידת מכונה (Jordan & Mitchell, 2015), והמיומנויות המרכזיות הללו נצרכות בעבודתם היום-יומית של רוב העוסקים במדעי הנתונים. כלומר, היכולת למנף אלגוריתמים של למידת מכונה מפוקחת ולא מפוקחת ולהשתמש בהם כדי להתמודד עם סוגים שונים של נתונים היא יכולת הכרחית עבור העוסקים ב-DS, וכוללת את רוב מיומנויות ה-DS שנדונו לעיל. במחקר זה אנו מתמקדים בשני סוגי המיומנויות הללו תוך התייחסות לשני הקורסים האקדמיים שניתנו בתוכנית ההכשרה שלנו (ראו פירוט בסעיף "תוכנית הכשרת בוגרים במדעי המידע באוניברסיטת בר-אילן"). בהמשך נדון בשני סוגי המיומנויות האלו בהרחבה.

מדעי הנתונים ולמידת מכונה

DS הוא תחום רחב המתכלל ידע מעמיק בסטטיסטיקה, ויזואליזציה של מידע, חישוביות ותכנות (Dichev & Dicheva, 2017). במחקר זה אנו מתמקדים בחוד החנית של התחום שנקרא למידת מכונה (Machine Learning – ML) ועוסק בפיתוח תוכנות מחשב אשר יכולות לבצע משימות מורכבות מבלי להיות מתוכנתות לכך בצורה מפורשת (Shalev-Shwartz & Ben-David, 2014). באופן מסורתי טכניקות ML מחולקות לקטגוריות המשתנות בהנחות הבסיסיות שלהן, בבסיס התיאורטי, במדדי הערכה ובהגדרות היישום שלהן. שתי הקטגוריות הבסיסיות ביותר הן: 1) **למידה מפוקחת**: בלמידה מסוג זה התוכנית מקבלת דוגמאות שעליה ללמוד מהן כקלט (הידועות גם בשם נתוני אימון) יחד עם התוצאות הרצויות שלהן (התיוג שלהן), שניתנו על ידי **המפקח** (בדרך כלל, כותב אנושי). מטרת התוכנית היא ללמוד את הכלל

הממפה קלט חדש (אשר התוכנית לא ראתה בעבר) לפלט הנכון. 2) **למידה לא מפוקחת**: בלמידה מסוג זה התוכנית אינה מקבלת פלטים ועליה לזהות ולפענח לבדה אילו מבנים ודפוסים (ייתכן שאלה יהיו מורכבים) מסתתרים בקלט שלה.

בהמשך פרק זה אנו מגדירים בצורה פורמלית את המונחים למידת מכונה מפוקחת ולמידת מכונה לא מפוקחת ומדגישים את העקרונות והטכניקות שנלמדות בתוכנית הלימודים שלנו.

למידת מכונה מפוקחת

למידת מכונה מפוקחת עוסקת בלימוד מיפוי בין קלט לפלט בהתבסס על סט דוגמאות המורכבות מקלט ופלט שהתוכנית מתאמנת איתן (Russell & Norvig, 2002). באופן פורמלי, נניח שיש לנו מערך נתונים $D = \{(x_1, t_1), \ldots, (x_N, t_N)\}$, שבו x_i הוא ייצוג (בדרך כלל וקטור) של הקלט, ו-t_i הוא הפלט המשויך אליו (בדרך כלל תווית או ערך אמיתי). טכניקות ML מפוקחות עוסקות בלמידת מיפוי טוב לניבוי תוצאות עבור קלט חדש אשר התוכנית לא ראתה בשלב האימון והלמידה. המשמעות של מיפוי טוב היא שילוב של מאפיינים מתמטיים כגון דיוק וחוסן (robustness).

שתי הגדרות החיזוי הבולטות ביותר של תחום למידת מכונה מפוקחת הן סיווג ורגרסיה. בבעיות מסוג סיווג המשימה היא חיזוי תווית או קטגוריה נפרדת (בדרך כלל מקבוצה לא מסודרת של תוויות). אם ניקח את דוגמה מספר 1 מההקדמה, בהינתן כתב יד x אנו עשויים להיות מעוניינים בזיהוי אוטומטי של מחברו מתוך רשימה של מחברים פוטנציאליים $t \in T$. בדוגמה זו נתוני האימון D עשויים להיות מורכבים מקבוצה של כתבי יד (x_i מיוצגים בצורה סטנדרטית כלשהי), שכל אחד מהם משויך לשם המחבר שלו ($t_i \in T$). לכן אלגוריתם הסיווג שנבחר צריך לחזות את המחבר הנכון עבור כתב יד חדש שלא נראה. לעומת זאת, בבעיות מסוג רגרסיה המשימה היא חיזוי מספר שלם או רציף (בדרך כלל ערך אמיתי). אם ניקח את דוגמה מספר 1 מההקדמה, הערך הרציף עשוי להתייחס לחיזוי שנת כתיבה של כתב יד. בדומה לדוגמת החיזוי, נתוני האימון יהיו כנראה מורכבים מכתבי יד (x_i), וכאן כל אחד מהם קשור לשנת הפרסום שלו (t_i). לפיכך אלגוריתם הרגרסיה שנבחר יצטרך לחזות את שנת הפרסום של כתב יד חדש שלא נראה. ההבדלים בין שתי הגדרות הלמידה המפוקחות גוררים אלגוריתמים שונים, קריטריוני בחירה ומדדי הערכה שונים ועוד (אנו מפנים את הקורא המתעניין לקריאה נוספת אצל Shalev-Shwartz & Ben-David, 2014).

בתוכנית ההכשרה שלנו אנו מלמדים את העקרונות הבסיסיים של מזעור סיכונים אמפיריים ואת טכניקות הסיווג הקלאסיות האלה: Support Vector Machines (SVM), Naïve Bayes, עצי החלטה, אלגוריתם K-nearest neighbor ו-Neural Networks. נוסף לכך אנו מלמדים את טכניקות הרגרסיה האלה:

רגרסיה לינארית ורגרסיה לוגיסטית. בתוכנית שלנו יש תשומת לב מיוחדת לבחירת האלגוריתם הנכון, הערכה והשוואה של טכניקות ML מפוקחות, שימוש בשיטות שונות לבחירת מאפיינים ואופטימיזציה של פרמטרים כחלק מתהליך הלמידה.

למידת מכונה לא מפוקחת

למידה לא מפוקחת עוסקת בלמידת דפוסים שלא זוהו בעבר במערך נתונים כאשר לא קיים פלט/תיוג לדוגמאות האימון (Russell & Norvig, 2002). בניגוד ללמידה מפוקחת, העושה שימוש בצמדי קלט-פלט, טכניקות למידה לא מפוקחת מקבלות קלט בלבד ומתמודדות עם האתגר של זיהוי דפוסים וחריגות אך ורק על סמך הקלט. באופן פורמלי, אלגוריתם לא מפוקח מקבל כקלט מערך נתונים $D = \{x_1, \dots, x_N\}$, שבו x_i הוא ייצוג (בדרך כלל וקטור) של הקלט, ומפיק הבחנות של המבנים הנסתרים בתוך הנתונים.

מכיוון שאין אנו יודעים מה המשמעות של כל דוגמת אימון, קשה לנתח את יעילותו של אלגוריתם אשר אומן בשיטת למידה לא מפוקחת. כתוצאה מכך פותח מגוון רחב של אתגרים מסקרנים, מגבלות, שיקולים ושיטות עבודה מומלצות לשימוש בלמידה לא מפוקחת. במסגרת תוכנית ההכשרה שלנו אנו מתמקדים בשלוש הגישות הבולטות ביותר: 1) **אשכולות:** האלגוריתם צריך לזהות מופעי נתונים הדומים זה לזה ולקבץ אותם יחד, בתקווה לחשוף את המבנה הפנימי של הקלט. אם ניקח את דוגמה מספר 2 אשר הופיעה בהקדמה, המשמעות היא חלוקת התמונות לתת-קבוצות נפרדות כך שכל תת-קבוצה תהיה בעלת מאפיינים דומים אך שונה מהאחרות; 2) **זיהוי אנומליות:** האלגוריתם מחפש דפוסים חריגים בקלט. אם ניקח שוב את דוגמה מספר 2, זיהוי אנומליה יכול לבוא לידי ביטוי כזיהוי התמונות שאינן דומות לשום ציור אחר בסט או שאי אפשר לשייך אותן למבנה הכללי בסט. מנקודת מבט יישומית ניתן להשתמש באלגוריתם הלמידה כדי לסמן את התמונות הללו במערך נתונים לצורך שיקול נוסף; 3) **שיוך:** האלגוריתם צריך לאתר תכונות מסוימות של מדגם נתונים המתואמים עם תכונות אחרות של מדגם זה. למשל, יכול להיות שתכונות מפתח של תמונה יהיו משויכות לתכונות אחרות, כגון מקור התמונה עשוי להיות קשור לשימוש בצבעים שונים.

בתוכנית ההכשרה שלנו אנו מלמדים את העקרונות והאתגרים הבסיסיים של למידה לא מפוקחת ואת טכניקות האשכולות הקלאסיות האלה: K-means ו-Hierarchical Clustering; טכניקות זיהוי חריגות: גורם חריג מקומי, זיהוי חריגים מבוסס ניתוח אשכולות וחריגות מחוקי שיוך וערכות פריטים תכופות; לימוד חוקיות אסוציאטיבי: אלגוריתם Aprioiri ו-FP-Growth. בכל שלוש המסגרות מוקדשת תשומת לב מיוחדת לבחירה, הערכה והשוואה של הטכניקות השונות, שימוש בשיטות להפחתת מאפיינים לא משפיעים והדמיית נתונים כחלק מתהליך הלמידה.

תוכנית הכשרת בוגרים במדעי המידע באוניברסיטת בר-אילן

תוכנית ההכשרה שלנו לתואר שני במדעי המידע מוצעת לסטודנטים עם רקע מגוון ואינה דורשת כל רקע טכנולוגי קודם. התוכנית נמשכת שנתיים (חלק מן הסטודנטים פורס אותה לשלוש שנים), ובמהלכן נחשפים הסטודנטים הן לנושאי IS הקלאסיים, כגון מבוא למדעי המידע, חיפוש ואחזור מידע מקוון וארגון מידע, והן לנושאים הקשורים ל-DS. אחת מהמטרות של תוכנית הלימודים שלנו היא לספק לסטודנטים ידע מעמיק ב-DS נוסף לידע רחב ב-IS. קורסי ה-DS העיקריים הנלמדים בתוכנית שלנו הם: יסודות מתמטיים ל-DS, סטטיסטיקה, מבוא לתכנות בפיתון, תכנות מתקדם, הדמיית נתונים, למידה מפוקחת ולמידה לא מפוקחת. קורסים אלו מכוונים במיוחד לפיתוח ארבע מיומנויות ה-DS הבסיסיות אצל התלמידים (Kang *et al*., 2015):

• עיבוד מקדים של נתונים – היכולת לחלץ נתונים שמישים מקבוצה גדולה יותר של נתונים גולמיים (Bartschat *et al*., 2019; Joseph & Thanakumar, 2019).

• חקר נתונים – היכולת לזהות מגמות בנתונים, לבצע ניתוח חקרני כדי להבין את הנתונים ולזהות השערות מעניינות (Russo & Zou, 2019; Simmons *et al*., 2011).

• ניתוח נתונים – היכולת לבנות את המודל הנכון לנתונים, להעביר נתונים לידע קונקרטי ולהעריך את יכולת המודל להתייחס להשערות המחקר (Awan *et al*., 2019; Gibert *et al*., 2010; Raschka, 2018).

• הצגת נתונים ותקשורת – היכולת לתקשר ולהסביר את הנתונים לאנשים מקבוצות מיומנויות שונות (כלומר משכבת הניהול), להסביר את חשיבות הדפוסים בנתונים ולהציע פתרונות (Gilpin *et al*., 2018; Vellido, 2020; Vellido Alcacena *et al*., 2011).

בתחילת הכשרתם הסטודנטים לומדים את קורסי המבוא ל-DS המקנים לתלמידים את המיומנויות המתמטיות, התכנותיות והטכניות הבסיסיות הנדרשות בקורסים המתקדמים ב-DS, שהם למידה מפוקחת ולמידה לא מפוקחת. במחקר זה אנו מתמקדים בשני קורסי הליבה הללו – למידה מפוקחת ולא מפוקחת – שכן הם משקפים בצורה הטובה ביותר את מיומנויות ה-DS אשר הסטודנטים רוכשים במהלך הלימודים לתואר השני.

מתודולוגיה

מחקר זה נעשה בשני שלבים:

1. זיהוי חששות של סטודנטים ושל גורמים במחלקה.
2. בחינת חששות אלו באמצעות נתונים.

על מנת לזהות את חששות הסטודנטים וגורמים במחלקה ראיינו באופן לא רשמי את ראש המחלקה ועוד שני גורמים מנהליים במחלקה האחראים על גיוס סטודנטים ונמצאים בקשר ישיר עם סטודנטים פוטנציאליים. ראיונות קצרים אלו כללו שאלת איתות – מהם החששות העיקריים שהביעו סטודנטים פוטנציאליים?

לצורך בחינת החששות שזוהו במסגרת הראיונות, חקרנו את מחזור הסטודנטים של 2020 ובדקנו את ביצועי התלמידים בקורס למידת מכונה מפוקחת (SML), בקורס למידת מכונה לא מפוקחת (UML) ובתוכנית ההכשרה כולה. תוכני הקורסים SML ו-UML מפורטים בסעיף "מדעי הנתונים ולמידת מכונה". קורס ה-SML ניתן על ידי המחבר השני, שהוא זוכה פרס המרצה המצטיין של אוניברסיטת בר אילן לשנת 2017, והשתתפו בו 26 סטודנטים. קורס UML ניתן על ידי המחבר הראשון, שהוא זוכה פרס המרצה המצטיין של אוניברסיטת בר אילן לשנת 2018, והשתתפו בו 31 סטודנטים. בסך הכול 24 סטודנטים השתתפו בשני הקורסים בכיתה הנבחנת (9 גברים ו-15 נשים, הגיל הממוצע 35 ± 7. נציין שבמהלך המאמר בדרך כלל אנו משתמשים במילה הכללית "סטודנטים" עבור גברים ונשים כאחד, אלא אם כן נושא המגדר הוא שעומד על הפרק). העובדה שהמרצים זכו בפרסי הוראה לפני זמן לא רב מאפשרת לנו להניח כי הקורסים הועברו באופן נאות וברור ושלא היה הבדל מהותי באיכות ההוראה בין הקורסים.

מכיוון ששני הקורסים הללו ניתנו על ידי המחברים, ניתנה לנו ההזדמנות הייחודית לחקור את כישוריהם של התלמידים ממקור ראשון. לשם כך הגדירו המחברים יחדיו שני פרויקטי גמר עבור שני הקורסים: בקורס SML הוטלה על התלמידים המשימה הקלאסית של חיזוי מחיר של חפץ בהינתן סט נתונים מתויג. בקורס UML קיבלו הסטודנטים סט גדול של מסמכים (במקרה שלנו מחקרים אקדמיים), והם התבקשו לחקור את הדפוסים הנסתרים האפשריים בסט. הפרויקטים בוצעו באופן פרטני ונבדקו לאיתור העתקות (לא זוהו כאלו). כל אחד מהכותבים בחן באופן ידני ועצמאי כל עבודה על פי הקריטריונים האלה: 1) עיבוד מוקדם של נתונים; 2) חקר נתונים; 3) ניתוח נתונים; ו-4) הצגת נתונים ותקשורת. נזכיר, קריטריונים אלה תואמים את ארבע מיומנויות ה-DS הבסיסיות כפי שנדונו בסעיף "הקניית מיומנויות DS לתלמידי IS". באופן ספציפי, נוסף למתן ציון בסולם הסטנדרטי – ציון בין 0 ל-100 – המחברים העריכו את רמת המיומנות של כל תלמיד בכל אחד מהקריטריונים שנבדקו בסולם ליקרט של 5 נקודות, הנע בין כשירות נמוכה (1) לכשירות מלאה (5). בסך הכול כל סטודנט קיבל 8 ציונים – 4 ציונים לכל קורס.

נוסף לציונים חילצנו מידע נוסף על התלמידים, התואם את החששות שהועלו:

1. דיסציפלינת התואר ראשון
2. ציון סופי לתואר ראשון
3. ציון מבחן הפסיכומטרי הישראלי (להלן: ציון הפסיכומטרי)

4. גיל

5. מגדר

מידע זה הוצלב עם ציון ממוצע של הסטודנטים בתואר השני, המשמש כאינדיקטור להצלחתם בתוכנית ההכשרה כולה.

ציון הפסיכומטרי הממוצע של התלמידים היה 92±583 (ממוצע ± סטיית תקן). מתוך 24 הסטודנטים, 12 הם בוגרי תואר ראשון במדעי הרוח, 11 בוגרי מדעי החברה ו-1 בוגר מדעים מדויקים. הציון הסופי הממוצע לתואר ראשון של התלמידים הוא 5.4±83.6 והציון הסופי הממוצע לתואר שני הוא 4.2±87.5.

ניתוח ותוצאות

בתום ביצוע הראיונות סיווגנו את חששות הסטודנטים וגורמי המחלקה לסוגיות האלה: חששות לגבי תעסוקה עתידית, חששות לגבי רקע סטודנט מתאים לתוכנית ההכשרה ורקע סטודנט מתאים לחלק ה-DS החישובי של תוכנית ההכשרה. בהתמקדות בחששות האחרונים שהועלו, זיהינו את החששות המרכזיים האלה: 1) בוגרי מדעי הרוח עשויים להיתקל בקשיים בהתמודדות עם תוכני DS בשל הרקע המתמטי המוגבל שלהם; 2) ציון סופי לתואר ראשון (שהוא כלי המיון העיקרי לתוכנית שלנו) עשוי להוות אינדיקטור חלש להצלחת הסטודנטים בתוכנית ההכשרה ובחלק ה-DS שלה בשל השונות ברקע האקדמי של הסטודנטים (כלומר, מוסד, מגמה וכד'); 3) ציוני הפסיכומטרי עשויים לשמש אינדיקטור טוב להצלחה של סטודנטים פוטנציאליים, אך הם אינם מובאים בחשבון בתהליך הגיוס; 4) לתלמידים מבוגרים עשויה להיות אוריינטציה טכנולוגית פחותה ולכן יצליחו פחות בתכנים החישוביים; 5) סטודנטיות פוטנציאליות נרתעות מ- DS מכיוון שהוא נתפס כתחום גברי.

חשוב לציין שהחששות לעיל אינם ייחודיים לתוכנית ההכשרה שלנו ב-IS אלא ניתן למצוא אותם בצורות שונות בתחומים שונים. לדוגמה במחקר שנערך בבריטניה (Tariq & Durrani, 2012) נמצא שסטודנטים גברים, צעירים (18–29) ואלו עם רקע מתמטי אקדמי קודם, נוטים להציג ביטחון רב יותר בכישורי המתמטיקה והחישוב שלהם. בדומה, גאו (Guo, 2017) הראה שסטודנטים מבוגרים יחסית שהשתתפו בקורסי תכנות דיווחו על רמות גבוהות יותר של תסכול, על היעדר הזדמנויות לאינטראקציה עם מורים ועמיתים ועל בעיות בהתמודדות עם טכנולוגיות המשתנות ללא הרף.

על מנת להתייחס כראוי לחששות שהועלו, ראשית היה עלינו לקבוע מה נחשב כהצלחה בחלק ה-DS של תוכנית הלימודים. נזכיר, אנו מתמקדים בשני קורסים נפרדים, SML ו-UML, שבכל אחד מהם ניתנו ציונים לארבע מיומנויות ה-DS (ראו בסקירת הספרות). כדי לענות על השאלה הזאת, ניתן להעלות שאלה מקדימה: האם מיומנויות DS קשורות זו בזו? אנו מחלקים שאלה זו לשני חלקים: 1) האם ציוני הסטודנטים

במיומנויות ה-DS היו תואמים בשני הקורסים?; 2) האם ציוני הסטודנטים במיומנויות ה-DS קשורים זה בזה בכל קורס בנפרד?

ראשית נדון בהשוואה בין שני הקורסים. טבלה מספר 1 מסכמת את התוצאות.

מיומנות	SML	UML
עיבוד מקדים של נתונים	$4.25 \pm 0.6(4)$	$4.20 \pm 0.8(4)$
חקר נתונים	$4.92 \pm 0.4(5)$	$4.33 \pm 0.6(4)$
ניתוח נתונים	$3.37 \pm 1.3(3)$	$3.57 \pm 0.9(3)$
הצגת נתונים ותקשורת	$4.46 \pm 0.8(5)$	$4.04 \pm 0.7(4)$

טבלה מספר 1: ציוני התלמידים בארבע המיומנויות שנבחנו. הציונים הם בסולם של 1–5 (דרגה 5 מייצגת את הציונים הגבוהים ביותר) ומדווחים עם סטיית התקן המעוגלת. החציונים מדווחים בסוגריים.

כפי שניתן לראות מטבלה מספר 1, נראה כי יש הלימה בין הציונים של שני הקורסים. כדי לחקור את המתאם בין שני הקורסים, סיכמנו את הציונים של כל קורס לציון אחד ובדקנו את מתאם דירוג ספירמן (r) ביניהם. התוצאות מצביעות על מתאם חיובי בינוני אך מובהק של $r = 0.47, p < 0.05$. כאשר בוחנים את המתאם בין שני הקורסים עבור כל אחת מארבע המיומנויות בנפרד, נראה כי עבור חלק מהמיומנויות יש קורלציה מובהקת ועבור אחרות אין קורלציה. מתאם ספירמן עבור עיבוד מקדים של נתונים הוא $r = 0.68, p < 0.001$, המתאם עבור חקר נתונים הוא $r = -0.23, p = 0.28$, עבור ניתוח נתונים הוא $r = 0.36, p < 0.1$, ועבור הצגת נתונים ותקשורת הוא $r = 0.04, p = 0.85$. בסך הכול, הקורלציה בין המיומנות עיבוד מקדים של נתונים ושל מיומנות ניתוח נתונים בשני הקורסים נמצאה משמעותית, בעוד שהקורלציה בין מיומנות חקר נתונים ומיומנות הצגת הנתונים לא נמצאה משמעותית.

כעת פנינו לבחון את המתאם בין כל אחת מהמיומנויות בכל אחד משני הקורסים. מטריצת המתאמים מוצגת בטבלה מספר 2. התוצאות מראות כי עבור קורס SML כל המיומנויות נמצאות בקורלציה מובהקת זו עם זו, ואילו עבור קורס UML כל המיומנויות נמצאות בקורלציה מובהקת זו עם זו מלבד מיומנות חקירת נתונים, אשר אינה בקורלציה עם אף מיומנות אחרת.

מיומנות	חקר נתונים	ניתוח נתונים	הצגת נתונים ותקשורת
עיבוד מקדים של נתונים	0.36*,0.31	0.78***,0.5**	0.6***,0.44**
חקר נתונים	-	0.78***,0.17	0.6***,0.29
ניתוח נתונים	-	-	0.58***,0.75 ***

טבלה מספר 2: 0.01>*** ,0.05>** ,0.1>* . מטריצת מתאם בין כל אחת מארבע המיומנויות שנבחנו. כל תא מתאר שני מתאמים: משמאל מוצג המתאם לקורס SML ומימין מוצג המתאם לקורס UML.

לסיכום, בעוד שנמצאו מתאמים מסוימים בין שני הקורסים ובין המיומנויות הנבחנות, התוצאות מצביעות על כך שהמיומנויות הנבדקות אינן יכולות להסתכם במיומנות-על **אחת** של DS וכי הן עשויות לבוא לידי ביטוי שונה מעט בקורס הלמידה המפוקחת ובקורס הלמידה הלא מפוקחת. לכן בניתוח הבא שלנו המשכנו לבחון את ארבע מיומנויות ה-DS בשני הקורסים ולא איחדנו אותן למיומנות אחת או התייחסנו לקורס אחד בלבד.

הצלחה בחלק DS של התוכנית שלנו

תחילה בדקנו אם דיסציפלינת התואר ראשון, הציון סופי של התואר הראשון, ציוני הפסיכומטרי, גיל ומגדר יכולים לספק אינדיקציה להצלחת הסטודנטים בחלק ה-DS של ההכשרה (כפי שנמדד בארבע מיומנויות ה-DS). בדיקה זו בוצעה בשני אופנים: תחילה בעזרת מבחנים השוואתיים המספקים אינדיקציה ראשונית לקשר אפשרי ולאחר מכן בעזרת שני מודלי רגרסיה ליניארית הבוחנים השפעה סיבתית פוטנציאלית, האחד עבור מיומנויות של למידה מפוקחת והשני עבור מיומנויות של למידה לא מפוקחת.

הפרמטר הראשון שבחנו היה דיסציפלינת התואר הראשון של הסטודנטים. השווינו בין בוגרי מדעי החברה לבין בוגרי מדעי הרוח (בניתוח זה הושמט סטודנט בודד שסיים את לימודי התואר הראשון בפקולטה למדעים מדויקים). ראשית, עשינו שימוש במבחן Mann-Whitney U במטרה לבחון את הציונים של הסטודנטים בארבע מיומנויות ה-DS עבור שני הקורסים הנחקרים. מן הממצאים עולה כי קיים הבדל מובהק במיומנות מסוג ניתוח נתונים עבור שני הקורסים – $p < 0.01$ עבור קורס SML ו- $p < 0.1$ עבור קורס UML. בשני הקורסים בוגרי מדעי הרוח קיבלו ציונים גבוהים יותר מאלו של בוגרי מדעי החברה. נוסף לכך, בקורס SML בוגרי מדעי הרוח קיבלו ציונים גבוהים יותר בכל ארבע המיומנויות שנבחנו (בממוצע, אם כי ההבדל אינו מובהק סטטיסטית) בעוד שב-UML הדבר נכון רק עבור שתיים מתוך ארבע המיומנויות. טבלאות 3 ו-4 מסכמות את התוצאות.

מיומנויות	בוגרי מדעי החברה	בוגרי מדעי הרוח
עיבוד מקדים של נתונים	$4 \pm 0.7(4)$	$4.4 \pm 0.5(4)$
חקר נתונים	$4.8 \pm 0.6(5)$	$5(5)$
ניתוח נתונים	$\mathbf{2.7 \pm 1.6(3)}$	**4±0.85** (4)
הצגת נתונים ותקשורת	$4.2 \pm 0.75(4)$	$4.7 \pm 0.64(5)$
סה"כ ציון למידה מפוקחת	$\mathbf{15.6 \pm 3.2(16)}$	$\mathbf{18.2 \pm 1.6(18)}$

טבלה מספר 3: הציונים בארבע המיומנויות שנבדקו בקורס SML. הציונים הם בסולם של 1–5 (5 מייצג את הציונים הגבוהים ביותר) ומדווחים יחד עם סטיית התקן המעוגלת. החציונים מדווחים בסוגריים. התוצאות המובהקות ($p < 0.05$) מודגשות.

מיומנויות	מדעי החברה	מדעי הרוח
עיבוד מקדים של נתונים	$4 \pm 0.77(4)$	$4.45 \pm 0.49(4)$
חקר נתונים	$4.4 \pm 0.66(4)$	$4.2 \pm 0.6(4)$
ניתוח נתונים	$\mathbf{3.14} \pm \mathbf{1.06}(\mathbf{3})$	**3.73±0.64** (4)
הצגת נתונים ותקשורת	$4.2 \pm 0.4(4)$	$4.1 \pm 0.63(4)$
סה"כ ציון למידה לא מפוקחת	$16.1 \pm 2.1(15)$	$16.5 \pm 2(16)$

טבלה מספר 4: הציונים בארבע המיומנויות שנבדקו בקורס UML. הציונים הם בסולם של 1–5 (5 מייצג את הציונים הגבוהים ביותר) ומדווחים יחד עם סטיית התקן המעוגלת. החציונים מדווחים בסוגריים. התוצאה המובהקת ($p < 0.01$) מודגשת.

כאשר בוחנים את הקשר בין הציון הסופי של התואר הראשון לבין הציונים של הסטודנטים בארבע מיומנויות ה-DS הנבחנות, אנו מוצאים כמה מתאמים חיוביים חזקים ביניהם. באופן ספציפי אנו מוצאים מתאמים חיוביים משמעותיים בשני הקורסים בין הציון הסופי של התלמידים בתואר הראשון לבין מיומנויות עיבוד מקדים של נתונים, ניתוח נתונים ואפילו הצגת נתונים ותקשורת. טבלה מספר 5 מסכמת את התוצאות.

מיומנויות	SML	UML
עיבוד מקדים	0.59***	0.43**
חקר נתונים	0.002	0.13
ניתוח נתונים	0.62***	0.44**
הצגת נתונים ותקשורת	0.55***	0.34*

*< 0.1, **< 0.05, ***< 0.01

טבלה 5: קורלציה בין הציון הסופי בתואר הראשון של הסטודנטים לבין המיומנויות הנבחנות בשני הקורסים.

כאשר בחנו את הקשר בין ציוני הפסיכומטרי לבין הציונים של הסטודנטים בארבע מיומנויות ה-DS הנבחנות, לא נצפתה קורלציה ביניהם. כמו כן לא מצאנו קורלציה בין גיל הסטודנטים לבין ציוניהם בארבע מיומנויות ה-DS הנבחנות. באופן ספציפי, באמצעות מבחני קורלציה לא הצלחנו לזהות מתאמים מובהקים סטטיסטית בין ציוני הפסיכומטרי והגיל לבין כל אחת מהמיומנויות שנבדקו במחקר זה. המתאמים נעו בין -0.04 (לגיל) ל-0.2 (עבור ציון פסיכומטרי) וערכי p היו גבוהים יחסית.

כאשר בחנו את הקשר בין המגדר של הסטודנטים לבין ציוניהם בארבע מיומנויות ה-DS הנבחנות באמצעות מבחן Mann-Whitney U, לא הצלחנו לזהות הבדלים מובהקים סטטיסטית בין סטודנטים וסטודנטיות במיומנויות שנבדקו. את חוסר המובהקות הסטטיסטית ניתן לייחס במידה רבה למספר הנמוך יחסית של גברים במאגר הסטודנטים שלנו (9). עם זאת, זוהה הבדל בולט במיומנות ניתוח הנתונים – סטודנטיות קיבלו ציון הגבוה ב-20% משל הסטודנטים (3.6 בממוצע לסטודנטיות בהשוואה ל-3 בממוצע לסטודנטים).

לצורך בחינת סיבתיות אפשרית הותאמו שני מודלי הרגרסיה, האחד עבור קורס למידה מפוקחת והשני עבור קורס למידה לא מפוקחת. בשני המודלים נמצא כי עבור שני הקורסים השפעת ממוצע הציונים בתואר הראשון מובהקת ($p<0.05$) וליתר המשתנים אין השפעה מובהקת. תוצאה זו תואמת את הניתוחים ההשוואתיים לעיל ומעידה על האינדיקטיביות של הציונים לתואר ראשון על הצלחה ב-DS.

מדדי ה-R^2 adjusted עבור שני מודלי הרגרסיה מעידים על שונות רבה שאיננה מוסברת בפרמטרים שנבחנו (R=0.16 עבור קורס למידה מפוקחת ו-R=0.26 עבור קורס למידה לא מפוקחת). תוצאה זו מפיגה במידת מה את החששות המרכזיים שנבחנו במחקר זה.

הצלחה בתוכנית ה-IS בכללותה

בהתאמה לניתוח שנעשה לעיל, כעת פנינו לבחון אם דיסציפלינת התואר ראשון, הציון סופי של התואר הראשון, ציוני הפסיכומטרי, גיל ומגדר יכולים להעיד על הצלחת הסטודנטים בתוכנית שלנו (הצלחה נמדדת בציון הממוצע של התואר השני). לצורך כך השתמשנו גם כאן במבחנים השוואתיים ומודל רגרסיה ליניארית.

הנתונים מצביעים על הבדל מובהק סטטיסטית בין הממוצעים של ציוני התואר השני של סטודנטים מדיסציפלינות שונות: ממוצע הציונים הסופי של בוגרי מדעי הרוח גבוה משמעותית מזה של בוגרי מדעי החברה (89.3 ± 3.04 לעומת 85.6 ± 4.74, $p < 0.05$). עם זאת, חשוב לציין כי לא זוהו הבדלים משמעותיים בעת בחינת ציוני הפסיכומטרי הממוצעים של הסטודנטים הללו (595.3 ± 92 במדעי הרוח לעומת 573.7 ± 100 במדעי החברה, $p = 0.36$).

כמו כן נמצאה התאמה בין ממוצע הציונים הסופי בתואר הראשון לממוצע הציונים הסופי בתואר השני $r = 0.67, p < 0.01$. באופן ספציפי, הציון הממוצע הסופי לתואר שני של סטודנטים עם ציון סופי של תואר ראשון מתחת ל-85 נמוך ב-6 נקודות בממוצע מסטודנטים בעלי ציון ממוצע בתואר הראשון הגבוה מ-85, $p < 0.01$.

למרות המתאם החלש בין ציוני הפסיכומטרי ומיומנויות ה-DS שנבחנו בסעיף הקודם, אנו מוצאים מתאם חזק בין ציוני הפסיכומטרי והצלחת הסטודנטים בתוכנית ההכשרה כולה $r = 0.55$, $p < 0.05$. באופן ספציפי, באמצעות מבחן Mann-Whitney U מצאנו שהציון הסופי הממוצע לתואר שני של סטודנטים עם ציון פסיכומטרי מתחת לממוצע (555>) נמוך ב-4 נקודות בהשוואה לסטודנטים עם ציון פסיכומטרי מעל הממוצע, $p < 0.05$.

כאשר השתמשנו במבחן Mann-Whitney U כדי לבחון את הקשר בין המגדר של הסטודנטים לבין ממוצע הציונים שלהם בתוכנית הלימודים, לא הצלחנו לזהות הבדלים מובהקים סטטיסטית בין ציוני הסטודנטים (86.4 ± 5 וציוני הסטודנטיות 88.2 ± 3.5. הסיבה לכך יכולה להיות מספר הגברים הנמוך יחסית (9) בכיתה שנבחנה.

בדומה לכך, נראה כי הגיל קשור באופן חלש לממוצע ציוני התואר של הסטודנטים בתוכנית שלנו. נמצא מתאם חיובי קל ולא מובהק של $r = 0.16$ בין גיל הסטודנטים לממוצע ציונם בתואר, כלומר סטודנטים מבוגרים יותר הצליחו, בממוצע, יותר.

בבחינת מודל הרגרסיה נמצא כי ההשפעות של ממוצע התואר הראשון ושל ציון הפסיכומטרי מובהקות ($p<0.05$) ואילו יתר המשתנים בעלי השפעה לא מובהקת.

מדד ה-R^2 adjusted עבור מודל הרגרסיה הינו גבוה ומעיד על האינדטיקטיביות הרבה של שני משתנים אלו ($R=0.44$).

מיומנויות DS והצלחה בתוכנית ה-IS שלנו

לפני שסיימנו את ניתוח הנתונים, ביקשנו למקם את חלק ה-DS בתוך תוכנית IS מבחינת הצלחת התלמידים. לשם כך בחנו אם יש קורלציה בין מיומנויות ה-DS שנבדקו לבין הצלחת התלמידים בשאר הכשרת ה-IS שלהם.

כפי שניתן לראות בטבלה מספר 6, עבור קורס SML כל המיומנויות מלבד חקר נתונים נמצאו בקורלציה מובהקת עם הצלחת הסטודנטים בשאר הכשרת ה-IS שלהם, $p < 0.01$. בקורס UML נמצאה קורלציה מובהקת בין מיומנות עיבוד מקדים ומיומנות ניתוח נתונים לבין הצלחה בתואר, בעוד שחקר נתונים והצגת נתונים ותקשורת נמצאו בקורלציה לא מובהקת, אך חיובית, עם הצלחת הסטודנטים בתואר.

מיומנות	SML	UML
עיבוד מקדים	0.66***	0.53***
חקר נתונים	0.05	0.03
ניתוח נתונים	0.77***	0.39*
הצגת נתונים ותקשורת	0.67***	0.25

$*< 0.1, **< 0.05, ***< 0.01$.

טבלה מספר 6: קורלציה בין מיומנויות ה-DS שנבחנו בשני הקורסים לבין הצלחת התלמידים בשאר תוכנית ההכשרה.

דיון

נתחיל בדיון בתוצאות הנוגעות לארבע מיומנויות ה-DS. התוצאות מצביעות על כך שעבור מיומנות עיבוד מקדים של נתונים וניתוח נתונים, יש קורלציה בין הציונים של קורסי SML ו-UML. נתון זה יכול להצביע על כך ששתי מיומנויות DS אלו באות לידי ביטוי בצורה דומה בהקשרי למידה מפוקחים ולא מפוקחים. באופן ספציפי, אנו משערים שעיבוד מקדים וניתוח נתונים הן מיומנויות המושתתות על יכולות תכנות גבוהות ותפיסה טכנולוגית רחבה. ככאלה, סטודנטים נוטים לקבל ציון דומה במיומנויות הללו ללא קשר למסגרת הלמידה. זה איננו המקרה עבור מיומנויות חקר הנתונים והצגת נתונים ותקשורת, שבהן לא נמצאה קורלציה משמעותית בין הקורסים. באשר למיומנות חקר נתונים, נראה כי חוסר הקורלציה יכול להיות מיוחס לציונים הגבוהים שהתקבלו בקריטריונים אלו בקורס SML. חשוב לציין שמיומנות חקר נתונים תופסת תפקיד קטן בתחום הלמידה המפוקחת, במיוחד כאשר ניתנת משימת חיזוי ספציפית מוגדרת במפורש, בהשוואה לאופי הבלתי מובנה יותר של תחום הלמידה הלא מפוקחת. באשר למיומנויות ההצגה ותקשורת של נתונים, אלו משקפות הבנה ופרשנות של התוצאות שהושגו. נראה כי הבנה חזקה של תוצרי למידה מפוקחת אינה מצביעה בהכרח על הבנה חזקה של תוצרי למידה לא מפוקחת ולהפך. לפיכך נראה כי חלוקת התוכנית לשני קורסים נפרדים – למידה מפוקחת ולמידה לא מפוקחת – היא ראויה.

בבואנו לחקור את הקשר האפשרי בין המיומנויות הנבחנות, התגלה שבקורס SML כל המיומנויות שנבחנו נמצאו במתאם, רובן אף במתאם חזק. בקורס UML נמצא מתאם בכל המיומנויות למעט מחקר נתונים, שאינו בקורלציה עם אף אחת מהמיומנויות האחרות. כפי שצוין קודם לכן, האופי הבלתי מובנה של למידה לא מפוקחת מוביל לאפיקים שונים לחקר נתונים. מהבחינה הזאת, הסטודנטים מציגים מגוון רחב של יכולות שאולי אינן קשורות למיומנויות אחרות, כגון תכנות ופרשנות מדעית של התוצאות. העובדה שבדרך כלל ארבע המיומנויות נמצאו בקורלציה חיובית חזקה מעידה שגילוי מוקדם של תלמידים שצפויים להתקשות או להצטיין ב-DS עשוי להיות אפשרי. לדוגמה, הצלחה בקורס בעל מאפיינים תכנותיים עשויה לנבא הצלחה ברוב מיומנויות ה-DS. אנו מתכוונים להמשיך ולחקור נושא זה בעבודה עתידית תוך שימוש בנתונים מפורטים יותר מהקורסים הקודמים.

כעת נפנה להתייחס לדאגות המרכזיות להצלחת התלמידים בחלק ה-DS של הכשרתם. מעניין לציין שבניגוד לחששות שמביעים סטודנטים פוטנציאליים וצוות המחלקה, בוגרי מדעי הרוח וסטודנטיות נוטים לקבל ציון גבוה יותר במיומנות ניתוח נתונים. חשוב לציין שנכון להיום, מקצוע מדעי הנתונים מבוקש ונשלט על ידי גברים שאינם בוגרי מדעי הרוח. התוצאות עשויות להצביע על כך שסטודנטים מרקעים מגוונים יכולים להשתלב בתוכניות IS מובילות ולרכוש את מיומנויות ה-DS הנדרשות מאנשי מקצוע. נוסף לכך, לא זוהה מתאם משמעותי בין הגיל לבין מיומנויות ה-DS שנבדקו. תוצאות אלו עשויות לסייע בהפגת חלק מהדאגות הקשורות לבוגרי מדעי הרוח, סטודנטיות פוטנציאליות וסטודנטים מבוגרים יותר.

התוצאות גם מראות כי ממוצע הציונים הסופי של התואר הראשון נמצא בקורלציה חזקה עם מיומנויות ה-DS שנבחנו, בעוד שציוני הפסיכומטרי של הסטודנטים לא הראו קורלציה כזאת. תוצאות אלו מצביעות על כך שהצלחה קודמת בלימודים אקדמיים (ברמת התואר הראשון) יכולה לנבא הצלחה פוטנציאלית בחלק ה-DS של הכשרת ה-IS שלנו טוב יותר מציוני הפסיכומטרי. מהתוצאות עולה כי ציוני הפסיכומטרי, שגורמי המחלקה האמינו כי יכולים לשמש אינדיקטור להצלחת התלמידים, אינם משקפים את הצלחת הסטודנטים בארבע מיומנויות ה-DS שנבדקו (קיימת קורלציה חלשה בין הפרמטרים). עם זאת, יש לשים לב שגם ממוצע הציונים הסופי של הסטודנטים לתואר ראשון וגם ציוני הפסיכומטרי עומדים בהתאמה עם הצלחת הסטודנטים בתוכנית ההכשרה כולה (ממוצע כללי של התואר השני), כפי שנדון בהמשך. תוצאות אלו עולות בקנה אחד עם הידע הקיים בספרות על הצלחה בתארים מתקדמים באופן כללי (Braunstein, 2002; Kuncel *et al*., 2007; McKee *et al*., 2001).

כעת נפנה להתייחס לחששות בנוגע להצלחת התלמידים בתוכנית ההכשרה כולה. בדומה לניתוח הצלחת התלמידים בחלק ה-DS של הכשרתם, הציון הסופי של בוגרי מדעי הרוח גבוה משמעותית מזה של בוגרי מדעי החברה. נוסף לכך, סטודנטיות מצליחות מעט יותר מסטודנטים – ממוצע ציוני התואר השני של סטודנטיות גבוה בשתי נקודות מממוצע ציוני התואר השני של סטודנטים, אך ההבדל אינו מובהק סטטיסטית. עוד נמצא שיש מתאם חיובי חלש בין גיל להצלחה בתוכנית, כלומר סטודנטים מבוגרים יותר הצליחו מעט יותר מסטודנטים צעירים יותר. תוצאות אלו עקביות עם מחקרי עבר הנוגעים להבדלים על בסיס מגדר וגיל (Kanny *et al*., 2014; Hoskins *et al*., 1997; McNeil *et al*., 2014; Pezzoni *et al*., 2016; Sax, 2001; Shapiro & Sax, 2011; Speer, 2021). בהתאמה למסקנות שהוזכרו בפסקה לעיל, התוצאות הללו יכולות לסייע בהפחתת חלק מהדאגות הנלוות של בוגרי מדעי הרוח, סטודנטיות פוטנציאליות וסטודנטים מבוגרים פוטנציאליים.

באשר לקריטריוני קבלה של סטודנטים לתוכנית הלימודים, ממוצע הציונים הסופי של התואר הראשון נמצא בקורלציה חזקה עם הציון הסופי של הסטודנטים לתואר השני. מעניין שגם ציוני הפסיכומטרי נמצאו בקורלציה חזקה עם הצלחה בתואר השני, וזאת בניגוד לניתוח לעיל שדן בקשר בין ציוני התואר הראשון לבין הצלחה בתוכני ה-DS של התוכנית. תוצאה זו מפתיעה במקצת שכן נמצאו מתאמים חלשים בלבד בין ציוני פסיכומטרי לבין מיומנויות ה-DS. יחד עם זאת, תוצאה זו עקבית עם מחקרם של ברטון ורמיסט (Burton & Ramist, 2001). אנו מאמינים כי זו התוצאה של האופי האינטגרטיבי של ציון הפסיכומטרי וכן הציון הממוצע של הסטודנטים לתואר שני. שני הציונים משקפים מגוון רחב של מיומנויות משנה נפרדות המשתלבות זו בזו באופן מורכב, לעיתים שילוב הדוק, לעיתים פחות ולעתים כלל לא.

טבלה מספר 7 מסכמת את הקריטריונים שנבחנו תוך השוואה של חוזק האינדיקציה שלהם להצלחה בחלק ה-DS של התוכנית וכן בתוכנית כולה.

קריטריון	DS	IS
מגדר	חלש	חלש
גיל	גרוע	חלש
ציון פסיכומטרי	חלש	**חזק**
דיסציפלינת התואר ראשון	**חזק**	**חזק**
ציון סופי לתואר ראשון	**חזק**	**חזק**

טבלה מספר 7: אינדיקציה להצלחת סטודנטים בהכשרה של DS ו-IS. כל תא מייצג את חוזק האינדיקציה. אינדיקציות חזקות מסומנות בהדגשה.

באשר לקשר האפשרי בין ביצועי הסטודנטים בקורסי ה-DS לבין הצלחתם בשאר ההכשרה של התואר השני, אנו מוצאים כי מיומנויות עיבוד מקדים וניתוח נתונים נמצאות בקורלציה עם הצלחת הסטודנטים בתואר. תוצאה זו דומה באופן מפתיע לקורלציה שבין הציון הסופי בתואר הראשון של הסטודנטים לבין מיומנויות ה-DS (טבלאות 5 ו-6). במילים אחרות, ציון סופי לתואר ראשון הוא אינדיקטור טוב להצלחת הסטודנטים במיומנויות עיבוד מקדים וניתוח נתונים, ואלה בתורן מהוות אינדיקציה טובה להצלחה בתוכנית ההכשרה שלנו.

לאורך הניתוח שלנו בלטה ביותר מיומנות ניתוח הנתונים. מיומנות זו מזכה את הסטודנטים בציונים הנמוכים ביותר בממוצע מבין הארבע שנבדקו הן בקורסי SML והן בקורסי UML (טבלה 1), היא בקורלציה עם רוב המיומנויות האחרות (טבלה 2), ציוני בוגרי מדעי הרוח גבוהים משמעותית רק במיומנות זו (רק עבורה יש מובהקות סטטיסטית, ראו טבלאות 3 ו-4), היא נמצאת בקורלציה חזקה עם הציון הסופי של התואר הראשון (טבלה 5) וכן עם הציון הסופי של התואר השני (טבלה 6). תוצאה זו מובילה אותנו להשערה שמיומנות ניתוח נתונים היא המרכזית ביותר בחלק ה-DS של התוכנית שלנו. לאור הנאמר לעיל,

אנו מאמינים שמיומנות זו צריכה לקבל תפקיד גדול יותר בהכשרה ובהערכה של תלמידי IS, אולי בהיקף של קורסים נוספים. דגש רב יותר על מיומנות זו יוביל לשיפור בקרב הסטודנטים, וזה בתורו יכול להוביל להכשרה של מקצועני IS טובים יותר.

מסקנות

אנו מכירים בכך שהמחקר הנוכחי מוגבל מבחינת הכמות והמגוון של הנתונים שבהם נעשה שימוש. המדגם שלנו הוא מחזור בודד (2020) שבו מספר הסטודנטים שנבחנו (24), ובמיוחד סטודנטים גברים (9), היה נמוך יחסית. אך למרות המספר הקטן של הסטודנטים יש לשים לב כי כלל האוכלוסייה – הסטודנטים שלמדו בתוכניתנו בגרסתה הנוכחית – הינו כ-150 סטודנטים. מאחר שקריטריוני הקבלה ונתוני הדמוגרפיה של הסטודנטים שלנו לא השתנו באופן משמעותי במהלך השנים הללו, אנו סבורים כי המדגם מייצג באופן סביר את התוכנית שלנו. עם זאת, אוכלוסיית הסטודנטים שנבחנו במחקר זה הינה מאוניברסיטה אחת (אוניברסיטת בר-אילן) וכולם ישראלים, מה שעלול להגביל את הכללת התוצאות שלנו לתוכניות נוספות. למרבה הצער, הוספת נתונים ומידע על סטודנטים נוספים או מאוניברסיטאות נוספות לא יכולה להתבצע בקלות בשל השוני באוכלוסייה ובשיטות ההוראה. באופן ספציפי, בשל מגיפת COVID19 הרחבת המדגם לשנים נוספות יכולה לייצר ערפלנים אשר יהיה קשה לשלוט בהם. לדוגמה, המחזור הנבחן הוא האחרון שעבר את כל לימודיו במתכונת הלמידה הפרונטלית. מחזורי הלימודים שאחריו חוו את הלימודים בצורה מקוונת או היברידית, סטודנטים לא יכלו להפגש בשל סגרים או הגבלות, סטודנטים הורים נאלצו להישאר בבית עם ילדיהם וכו'. על כן שילוב של נתוני מחזור 2021, למשל, עם נתוני קדם-COVID19 עשויים להכניס הטיה משמעותית לתוצאות.

בהתחשב במגבלות המחקר המוזכרות לעיל, תוצאותיו מעידות כי בוגרי מדעי הרוח מצליחים יותר מבוגרי מדעי החברה הן בחלק ה-DS של ההכשרה והן בתוכנית ההכשרה כולה. עוד מצאנו כי הציון הסופי של הסטודנטים בתואר הראשון הוא המדד העקבי ביותר להצלחת הסטודנטים בתוכנית הלימודים שלנו. יתרה מכך, מצאנו שציון הפסיכומטרי של הסטודנטים אינו מעיד על הצלחה בחלק ה-DS של ההכשרה, אך כן מצביע על הצלחה בתוכנית ההכשרה כולה. סטודנטיות נוטות להצליח יותר מסטודנטים בביצועי המדדים אשר נבחנו, ולא נמצא קשר משמעותי בין גיל הסטודנטים לביצועים הנמדדים. אנו מאמינים שתוצאות אלה יכולות להפיג חלק מהחששות שהביעו סטודנטים פוטנציאליים וצוות המחלקה. באופן ספציפי, איננו מוצאים ראיות התומכות ברוב החששות שהועלו, ובמקרים רבים הצלחנו להציג ראיות הפוכות.

עבודתנו בחנה את חששות הסטודנטים הנוגעים להצלחה אקדמית. כמובן יש משתנים נוספים היכולים להסביר הצלחה, כגון מצב משפחתי, רקע תעסוקתי קודם, מוסד הלימודים בתואר הראשון, עבודה במקביל ללימודים ומצב סוציואקונומי. אנו מתכננים לבחון נתונים אלו במחקר המשך. כמו כן ברצוננו לבחון את

השתלבות הבוגרים במסגרות תעסוקתיות. בחינה זו עשויה להיות מאתגרת מכיוון שעצם ההגדרה של מומחה IS בתעשייה אינה מדויקת, וחלק מהסטודנטים הלומדים בתוכנית שלנו כבר מועסקים במקצועות דומים עוד בטרם התחילו את לימודיהם.

מקורות

Agarwal, R., & Dhar, V. (2014). Big data, data science, and analytics: The opportunity and challenge for IS research. *Information Systems Research, 25(3*), 443-448. https://doi.org/10.1287/isre.2014.0546

Awan, S. E., Bennamoun, M., Sohel, F., Sanfilippo, F. M., & Dwivedi, G. (2019). Machine learning-based prediction of heart failure readmission or death: Implications of choosing the right model and the right metrics. *ESC Heart Failure*, *6*(2), 428–435. https://doi.org/10.1002/ehf2.12419

Badea, M., Florea, C., Florea, L., & Vertan, C. (2018). Can we teach computers to understand art? Domain adaptation for enhancing deep networks capacity to de-abstract art. *Image and Vision Computing*, 77, 21-32. https://doi.org/10.1016/j.imavis.2018.06.009

Bartschat, A., Reischl, M., & Mikut, R. (2019). Data mining tools. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(4), e1309. https://doi.org/10.1002/widm.1309

Baumer, B. (2015). A data science course for undergraduates: Thinking with data. *The American Statistician*, *69*(4), 334–342. https://doi.org/10.1080/00031305.2015.1081105

Belkin, N. J., & Robertson, S. E. (1976). Information science and the phenomenon of information. *J. Am. Soc. Inf. Sci., 27*(4), 197–204. https://doi.org/10.1002/asi.4630270402

Borko, H. (1968). Information science: What is it? *American documentation*, *19*(1), 3–5. https://doi.org/10.1002/asi.5090190103

Braunstein, A. W. (2002). Factors determining success in a graduate business program. *College Student Journal*, *36*(3), 471-478.

Bronstein, J. (2015). An exploration of the library and information science professional skills and personal competencies: An Israeli perspective. *Library Information Science Research*, *37*(2):130–138. https://doi.org/10.1016/j.lisr.2015.02.003

Brunner, R. J. & Kim, E. J. (2016). Teaching data science. *Procedia Computer Science*, 80, 1947–1956. https://doi.org/10.1016/j.procs.2016.05.513

Burgoyne, J. A., Fujinaga, I., & Downie, J. S. (2015). Music information retrieval. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A new companion to digital humanities* (pp. 213–228). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118680605.ch15

Burton, N. W., & Ramist, L. (2001). *Predicting success in college: SAT® studies of classes graduating since 1980. Research Report No. 2001-2*. College Entrance Examination Board.

Cao, L. (2017). Data science: Challenges and directions. *Communications of the ACM*, *60*(8), 59–68. https://doi.org/10.1145/3015456

Cleveland, W. S. (2001). Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review*, *69*(1), 21–26.

Costa, C., & Santos, M. Y. (2017). The data scientist profile and its representativeness in the European e-competence framework and the skills framework for the information age. *International Journal of Information Management*, *37*(6), 726–734. https://doi.org/10.1111/j.1751-5823.2001.tb00477.x

Davenport, T. (2020). Beyond unicorns: Educating, classifying, and certifying business data scientists. *Harvard Data Science Review, 2(2). https://doi.org/10.1162/99608f92.55546b4a*

Davenport, T. H., & Patil, D. (2012). Data scientist. *Harvard Business Review*, *90*(5), 70–76.

De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., Bryant, L., Cheng, L. Z., Francis, A., Gould, R., Kim, A. Y., Kretchmar, M., Lu, Q., Moskol, A., Nolan, D., Pelayo, R., Raleigh, S., Sethi, R. J., Sondjaja, M., ... Ye, P. (2017). Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Application*, *4,* 15–30. https://doi.org/10.1146/annurev-statistics-060116-053930

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, *56*(12), 64–73. https://doi.org/10.1145/2500499

Dichev, C., & Dicheva, D. (2017). Towards data science literacy. *Procedia Computer Science*, *108*, 2151-2160. https://doi.org/10.1016/j.procs.2017.05.240

Doyle, A. (2021, January 30). *Important job skills for data scientists*. TBC. https://www.thebalancecareers.com/list-of-data-scientist-skills-2062381.

Eiteljorg, H., II. (2004). Computing for Archaeologists. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A companion to digital humanities* (pages 20–30). John Wiley & Sons, Ltd.

Ess, C. (2004). Revolution? What revolution? Successes and limits of computing technologies in philosophy and religion. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A companion to digital humanities* (pp. 132-144). John Wiley & Sons, Ltd.

Fayyad, U., & Hamutcu, H. (2020). Toward foundations for data science and analytics: A knowledge framework for professional standards. *Harvard Data Science Review. 2.2*(2). https://doi.org/10.1162/99608f92.1a99e67a

Forte, M. (2015). Cyberarchaeology: A post-virtual perspective. *Humanities and the Digital. A Visioning Statement.* MIT Press.

Garber, A. M. (2019). Data science: What the educated citizen needs to know. *Harvard Data Science Review*, 1(1). https://doi.org/10.1162/99608f92.88ba42cb

Gibert, K., Sanchez-Marre, M., & Codina, V. (2010). Choosing the right data mining technique: Classification of methods and intelligent recommendation. *International Environmental Modelling and Software Society.*

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)* (pp. 80–89). https://doi.org/10.1109/DSAA.2018.00018

Guo, P. J. (2017). Older adults learning computer programming: Motivations, frustrations, and design opportunities. In *Proceedings of the 2017 chi conference on human factors in computing systems* (pp. 7070–7083). https://doi.org/10.1145/3025453.3025945

Hajic, J. (2004). Linguistics meets exact sciences. In *S. Schreibman, R. Siemens, & J. Unsworth (Eds.), A companion to digital humanities* (pp. 79-87). John Wiley & Sons Ltd.

Hjørland, B. (2002). Domain analysis in information science: Eleven approaches – traditional as well as innovative. *Journal of Documentation*, *58*(4), 422-462. https://doi.org/10.1108/00220410210431136

Hoskins, S. L., Newstead, S. E., & Dennis, I. (1997). Degree performance as a function of age, gender, prior qualifications and discipline studied. *Assessment & Evaluation in Higher Education*, *22*(3), 317-328. https://doi.org/10.1080/0260293970220305

Johnson, I. M. (1999). Librarians and the informed user: Reorienting library and information science education for the "information society". *Librarian Career Development*. *7*(4), 29-42. https://doi.org/10.1108/09680819910276941

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Joseph, S. I. T., & Thanakumar, I. (2019). Survey of data mining algorithms for intelligent computing system. *Journal of trends in Computer Science and Smart technology (TCSST)*, *1*(01), 14-24. https://doi.org/10.36548/jtcsst.2019.1.002

Juznic, P., & Badovinac, B. (2005). Toward library and information science education in the European Union: A comparative analysis of library and information science programmes of study for new members and other applicant countries to the European Union. *New Library World*. *106*(3/4), 173-186. https://doi.org/10.1108/03074800510587372

Kang, J. W., Holden, E. P., & Yu, Q. (2015). Pillars of analytics applied in MS degree in information sciences and technologies. In *Proceedings of the 16th Annual*

Conference on Information Technology Education (83–88). https://doi.org/10.1145/2808006.2808028

Kanny, M. A., Sax, L. J., & Riggers-Piehl, T. A. (2014). Investigating forty years of STEM research: How explanations for the gender gap have evolved over time. *Journal of Women and Minorities in Science and Engineering*, *20*(2). https://doi.org/10.1615/JWomenMinorScienEng.2014007246

Kestemont, M. (2014, April). Function words in authorship attribution. From black magic to theory? In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)* (pp. 59-66). http://dx.doi.org/10.3115/v1/W14-0908

Koppel, M., Mughaz, D., & Akiva, N. (2006). New methods for attribution of rabbinic literature. *Hebrew Linguistics: A Journal for Hebrew Descriptive, Computational and Applied Linguistics*, *57*, 5-18.

Kuncel, N. R., Credé, M., & Thomas, L. L. (2007). A meta-analysis of the predictive validity of the Graduate Management Admission Test (GMAT) and undergraduate grade point average (UGPA) for graduate student academic performance. *Academy of Management Learning & Education*, *6*(1), 51-68. https://www.jstor.org/stable/40214516

McKee, A. J., Mallory, S. L., & Campbell, J. (2001). The Graduate Record Examination and undergraduate grade point average: Predicting graduate grade point averages in a criminal justice graduate program. *Journal of Criminal Justice Education*, *12*(2), 311-317. https://doi.org/10.1080/10511250100086141

McNeil, J., Long, R., & Ohland, M. W. (2014). Getting better with age: Older students achieve higher grades and graduation rates. In *2014 IEEE Frontiers in Education Conference (FIE) Proceedings* (pp. 1-5). IEEE. https://doi.org/10.1109/FIE.2014.7044164

Perez, T., Cromley, J. G., & Kaplan, A. (2014). The role of identity development, values, and costs in college STEM retention. *Journal of educational psychology*, *106*(1), 315. https://psycnet.apa.org/doi/10.1037/a0034027

Pezzoni, M., Mairesse, J., Stephan, P., & Lane, J. (2016). Gender and the publication output of graduate students: A case study. *PLoS One*, *11*(1), e0145146. https://doi.org/10.1371/journal.pone.0145146

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv, 1811.12808*. https://doi.org/10.48550/arXiv.1811.12808

Rommel, T. (2004). Literary studies. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A companion to digital humanities* (pp. 88–96). John Wiley & Sons Ltd. https://doi.org/10.1002/9780470999875

Russell, S., & Norvig, P. (2002). *Artificial intelligence: A modern approach*. ISBN 9780134610993

Russo, D., & Zou, J. (2019). How much does your data exploration overfit? Controlling bias via information usage. *IEEE Transactions on Information Theory*, *66*(1), 302–323. https://doi.org/10.1109/TIT.2019.2945779

Sax, L. J. (2001). Undergraduate science majors: Gender differences in who goes to graduate school. *The Review of Higher Education*, *24*(2), 153-172. https://doi.org/10.1353/rhe.2000.0030

Sax, L. J., Kanny, M. A., Riggers-Piehl, T. A., Whang, H., & Paulson, L. N. (2015). "But I'm not good at math": The changing salience of mathematical self-concept in shaping women's and men's STEM aspirations. *Research in Higher Education*, *56*(8), 813-842. https://doi.org/10.1007/s11162-015-9375-x

Schwab-McCoy, A., Baker, C. M., & Gasper, R. E. (2020). Data science in 2020: Computing, curricula, and challenges for the next 10 years. *Journal of Statistics Education*, 1–17. https://doi.org/10.1080/10691898.2020.1851159

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press. https://doi.org/10.1017/CBO9781107298019

Shapiro, C. A., & Sax, L. J. (2011). Major selection and persistence for women in STEM. *New Directions for Institutional Research*, *2011*(152), 5-18. https://doi.org/10.1002/ir.404

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Speer, J. (2021). Bye bye Ms. American Sci: Women and the leaky STEM pipeline.

http://dx.doi.org/10.2139/ssrn.3913037

Tariq, V. N., & Durrani, N. (2012). Factors influencing undergraduates self-evaluation of numerical competence. *International Journal of Mathematical Education in Science and Technology*, *43*(3), 337–356. https://doi.org/10.1080/0020739X.2011.618552

Thomas, W. G., & William, G. (2004). Computing and the historical imagination. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A companion to digital humanities* (56-68). John Wiley & Sons Ltd. https://doi.org/10.1002/9780470999875

Van der Aalst, W. M. (2014). Data scientist: The engineer of the future. In K. Mertins, F. Bénaben, R. Poler, & J.-P. Bourrières (Eds.), *Enterprise interoperability VI* (13–26). Springer. https://doi.org/10.1007/978-3-319-04948-9_2

Van Dyk, D., Fuentes, M., Jordan, M. I., Newton, M., Ray, B. K., Lang, D. T., & Wickham, H. (2015). ASA statement on the role of statistics in data science. *Amstat news*, *460*(9), 24.

Vellido Alcacena, A., Martin, J. D., Rossi, F., & Lisboa, P. J. (2011). Seeing is believing: The importance of visualization in real-world machine learning applications. In *Proceedings: 19th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2011* (pp. 219–226).

Vellido, A. (2020). The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural computing and applications*, *32*(24), 18069-18083. https://doi.org/10.1007/s00521-019-04051-w

Wang, L. (2018). Twinning data science with information science in schools of library and information science. *Journal of Documentation, 74*(6). https://doi.org/10.1108/JD-02-2018-0036

Williams, M. E. (1988). Defining information science and the role of ASIS. *Bulletin of the American Society for Information Science*, *14*(2), 17-19.

Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, *49*(3), 33–35. https://doi.org/10.1145/1118178.1118215

Yunus, K., Wahid, W., Omar, S. S., & Ab Rashid, R. (2016). Computer phobia among adult university students. *International Journal of Applied Linguistics and English Literature*, *5*(6), 209-213. http://dx.doi.org/10.7575/aiac.ijalel.v.5n.6p.209

Zaagsma, G. (2013). On digital history. *BMGN-Low Countries Historical Review*, *128*(4), 3–29. https://doi.org/10.18352/bmgn-lchr.9344

Zuo, Z., Zhao, K., & Eichmann, D. (2017). The state and evolution of US iSchools: From talent acquisitions to research outcome. *Journal of the Association for Information Science and Technology*, *68*(5), 1266-1277. https://doi.org/10.1002/asi.23751