**מחקרים בתחום:**

**מאמר 1: Asif et al. (2017)**

קישור למאמר: https://peerj.com/articles/cs-1294/

1. מה יש בדאטה:

- סטודנטים מתואר ראשון במדעי המחשב באוניברסיטה פרטית בפקיסטן.

- סך הכול 287 תצפיות.

- מידע שנאסף: ציוני קורסים משנים א' ו־ב', CGPA סופי, נתוני נוכחות, מגדר, גיל, מחזור לימודים.

2. שאלת המחקר:

- האם ניתן לחזות את ה-CGPA של סטודנט בסיום התואר בהתבסס על ביצועיו בשנתיים הראשונות?

- אילו משתנים משפיעים הכי הרבה על ההצלחה האקדמית?

3. שיטות:

- חמש שיטות ML: Decision Tree (J48), Random Forest, Naïve Bayes, SVM, k-NN.

- הערכת המודלים בוצעה באמצעות 10-fold cross-validation.

4. מודלים חזקים ומדדי ביצוע:

- J48 ו-Random Forest נתנו את התוצאות הטובות ביותר (דיוק > 80%).

- מדדים: Accuracy, Precision, Recall, F1, AUC.

- ציטוט: “The J48 and Random Forest classifiers achieved the best results with overall classification accuracy above 80%.” (עמ' 10)

5. תובנות:

**- ציוני קורסים ספציפיים (כמו תכנות בסיסי, בסיסי נתונים ומתמטיקה בדידה) היו החשובים ביותר.**

ציטוט: “marks of programming fundamental, database systems and discrete structures ... ranked as top predictors.” (עמ' 11)

- **משתנים דמוגרפיים כמו גיל ומגדר השפיעו מעט מאוד**.

ציטוט: “Demographic attributes have less impact...” (עמ' 12)

- **נוכחות תרמה מעט אך פחות מהציונים.**

ציטוט: “Attendance variables have moderate impact, but not as significant as grades.” (עמ' 11)

- מסקנה כוללת: ניתן לחזות בהצלחה את CGPA ע"ב נתוני השנים הראשונות, והציונים המוקדמים הם המנבא המרכזי.

**מאמר 2: Beaulac & Rosenthal (2019)**

קישור למאמר: https://dl.acm.org/doi/10.1145/3287324.3287510

1. מה יש בדאטה:

- מעל 65,000 סטודנטים מאוניברסיטת טורונטו.

- נתוני עתק של ≈1.6 מיליון ציוני קורסים.

- מאפיינים: ציונים של שני הסמסטרים הראשונים, קושי קורס, תחום לימודים, ממוצע תיכון.

2. שאלת המחקר:

- האם ניתן לחזות הצלחה אקדמית (סיום תואר, ממוצע סיום) ע"ב נתוני שני הסמסטרים הראשונים?

- האם ניתן לחזות גם את מגמת הלימודים שבה הסטודנט יתמיד?

3. שיטות:

- שימוש ב-Random Forest לבעיות סיווג וחיזוי.

- בחינת חשיבות משתנים (Feature importance) באמצעות Gini Impurity וניתוח רגרסיות.

4. מודל חזק ומדדים:

- Random Forest השיג AUC ≈ 0.88 עבור חיזוי סיום תואר.

- דיוק גבוה גם בחיזוי תחום לימודים.

- ציטוט: “Grades in early courses — especially ‘gatekeeper’ courses — dominate prediction of graduation outcomes.” (עמ' 3)

5. תובנות:

- **משתנים אקדמיים מוקדמים הם המנבאים המרכזיים** (ציוני שנה א׳).

- **קורסים עם grading strictness גבוה תורמים יותר לניבוי.**

**- ממוצע תיכון תורם אבל פחות מציוני קורס בפועל.**

**- משתנים דמוגרפיים כמו מגדר וגיל תרמו מעט בלבד.**

ציטוט: “Demographics (e.g., gender) had marginal importance compared to academic indicators.” (עמ' 4)

- המסקנה הברורה: הצלחה או קושי בשנה הראשונה מהווים סמן קריטי להצלחת סיום התואר.

**מאמר 3: Yağcı (2022)**

קישור למאמר: https://doi.org/10.1016/j.measurement.2022.111928

1. מה יש בדאטה:

- 1,854 סטודנטים לתואר ראשון מטורקיה.

- משתנים עיקריים: ציון מבחן אמצע (Mid-term), שם הפקולטה, חוג לימוד, מגדר, גיל.

2. שאלת המחקר:

- האם ניתן לחזות את ציון הסיום בקורס (Final grade) בהתבסס על ציון האמצע, תחום לימוד ומאפיינים דמוגרפיים?

- מהם הגורמים המשפיעים ביותר על ביצועים אקדמיים?

3. שיטות:

- השוואת אלגוריתמים: Random Forest, SVM, k-NN, Decision Trees, Logistic Regression.

- בוצעה חלוקה ל־80/20 train/test עם השוואת דיוק, Precision, Recall ו־F1.

4. מודל חזק ומדדים:

- Random Forest נתן את התוצאה הטובה ביותר.

- Accuracy: ≈ 75% עם Mid-term בלבד.

- ציטוט: “Midterm scores alone explained over 70% of the variation in final scores.” (עמ' 6)

- תוספת של משתני דמוגרפיה כמעט לא שיפרה את המודל.

5. תובנות:

**- משתנה המידטרם היה המנבא המרכזי והעיקרי.**

**- חוג הלימוד ופקולטה תרמו מעט בלבד.**

**- מגדר וגיל לא תרמו משמעותית.**

ציטוט: “Adding gender and age variables did not improve model accuracy.” (עמ' 7)

- ממצא מובהק: אפשר להגיע לחיזוי טוב גם בלי דמוגרפיה – באמצעות ציונים קודמים בלבד.

**מאמר 4: Falat & Piscová (2022)**

קישור למאמר: <https://www.sciencedirect.com/science/article/abs/pii/S0957417422010431>

1. מה יש בדאטה:

- שאלון שנאסף מ-400 סטודנטים באוניברסיטאות בסלובקיה.

- שלושה סוגי משתנים:

- פסיכולוגיים: מוטיבציה, שביעות רצון.

- סוציולוגיים: השכלת הורים, הכנסה.

- לימודיים: שעות למידה, דפוסי למידה, שימוש בטכנולוגיה.

2. שאלת המחקר:

- מהם הגורמים המנבאים בצורה הטובה ביותר את ממוצע הציונים (GPA)?

- האם למוטיבציה והשכלה הורית יש השפעה חזקה יותר משעות למידה?

3. שיטות:

- אלגוריתמים: Random Forest Regression, Linear Regression, Gradient Boosting.

- מדדי הערכה: MAPE (Mean Absolute Percentage Error), RMSE, R².

4. מודל חזק ומדדים:

- Random Forest היה החזק ביותר.

- MAPE של ≈ 11%, RMSE ≈ 0.42.

- ציטוט: “Random Forest outperformed linear models in accuracy and robustness.” (עמ' 8)

5. תובנות:

**- מוטיבציה פנימית הייתה אחד הגורמים החזקים ביותר לחיזוי GPA.**

ציטוט: “Intrinsic motivation showed the strongest correlation with academic performance.” (עמ' 7)

**- השכלה של ההורים תרמה משמעותית גם כן.**

ציטוט: “Parental education had significant predictive value, especially the mother’s education level.” (עמ' 7)

**- שעות למידה לא נמצאו כגורם חשוב במיוחד.**

ציטוט: “Study time showed weak and inconsistent relationship with GPA.” (עמ' 6)

**- שילוב של פסיכולוגי וסוציולוגי תורם יותר מאשר גורמים לימודיים טכניים.**

**מאמר 5: Cruz-Jesus et al. (2020)**

קישור למאמר: https://doi.org/10.1016/j.compedu.2020.103857

1. מה יש בדאטה:

- נתונים מ-1,654 סטודנטים להשכלה גבוהה בפורטוגל.

- משתנים דמוגרפיים: גיל, מגדר, נוכחות בשיעורים, גישה לאינטרנט ומחשב אישי, כמות קורסים, קורסים שנכשלו, ועוד.

2. שאלת המחקר:

- האם ניתן לחזות ביצועים אקדמיים (ציונים סופיים בקורסים) ע"ב משתני רקע בלבד?

- אילו משתנים משפיעים הכי הרבה על הצלחה אקדמית?

3. שיטות:

- שימוש ב-Random Forest, Support Vector Machines, Logistic Regression, Naive Bayes.

- הערכת המודלים לפי Accuracy, ROC-AUC ו-F1-score.

4. מודל חזק ומדדים:

- Random Forest נתן את התוצאות הטובות ביותר.

- Accuracy של ≈ 93% עבור חיזוי ציונים בקורסים.

- ציטוט: “RF classifier clearly outperformed other ML algorithms in predicting academic success.” (עמ' 9)

5. תובנות:

- שלושת הגורמים החשובים ביותר: נוכחות בכיתה, גיל, וגישה לאינטרנט.

ציטוט: “Attendance, age, and internet access consistently emerged as the most influential predictors.” (עמ' 8)

- מגדר תרם מעט מאוד לחיזוי.

ציטוט: “Gender had negligible importance in most models.” (עמ' 9)

- השפעה של גישה לטכנולוגיה (כמו מחשב אישי) נצפתה אך לא תמיד הייתה חזקה.

- משתנים כמו מספר קורסים שכשלו סיפקו גם הם תרומה מסוימת.

**מאמר 6: Khalil & Ahmad (2021)**

קישור למאמר: https://doi.org/10.1016/j.heliyon.2021.e07611

1. מה יש בדאטה:

- מדובר בסקירת ספרות שיטתית (Systematic Literature Review) של מעל 80 מחקרים בין השנים 2010–2020 בתחום חיזוי ביצועים אקדמיים.

- כללו מאמרים מתחום Educational Data Mining ו-Machine Learning על סטודנטים באקדמיה.

2. שאלת המחקר:

- אילו שיטות ML נפוצות ביותר לחיזוי ביצועים אקדמיים?

- מהם המשתנים שנמצאו הכי חשובים שוב ושוב?

- אילו מגמות חוזרות קיימות במודלים ובתכונות?

3. שיטות:

- ניתוח איכותני וכמותי של הספרות לפי קריטריונים ברורים (PRISMA).

- זיהוי שיטות ML, מאפייני דאטה, תוצאות ביצוע, ושכיחות תכונות.

- סיווג מחקרים לפי מאפיינים: סיווג מול רגרסיה, סוגי משתנים, רמת חינוך וכו׳.

4. ממצאים על שיטות ומדדים:

- האלגוריתמים הפופולריים ביותר: Decision Tree, Random Forest, Naïve Bayes, SVM.

- רוב המאמרים מדדו דיוק (Accuracy), ROC-AUC, Precision, Recall.

- ציטוט: “Decision tree-based models were the most frequently used and generally performed well.” (עמ' 6)

5. תובנות:

- המשתנה שחוזר הכי הרבה כמנבא משמעותי: ציונים קודמים.

ציטוט: “Past grades were identified as the strongest predictor of future academic performance in over 60% of studies.” (עמ' 7)

- משתנים התנהגותיים (כמו log-ins למערכות למידה) הופיעו במעמד שני.

לדוגמה:

* מספר הכניסות ל־Moodle / Canvas / Zoom.
* משך זמן צפייה בהרצאות מוקלטות.
* מספר שליחויות של תרגילים.

במילים אחרות – **התנהגות סטודנטלית בפועל**, כפי שנמדדת בפלטפורמות הלמידה, היא מדד טוב יותר ממין/גיל – אך פחות מדויק לעומת ציוני עבר.

🧠 ציטוט מהמאמר:

“Behavioral data (e.g., frequency of LMS log-ins, time spent on tasks) were the second most predictive category after past academic records.” *(Khalil & Ahmad, 2021, p. 7)*

פרויקט שלנו אין את המשתנים ההתנהגותיים הללו – אבל אפשר לחפש מקבילים עקיפים כמו:

* **נוכחות בשיעורים** (כ proxy להתנהגות)
* **מספר שעות לימוד ביום**
* **התייעצות עם מרצים**

- משתנים דמוגרפיים (מין, גיל, הכנסה) הופיעו ברבים אך תרומתם לרוב הייתה מוגבלת.

ציטוט: “Demographic variables were often included but showed limited predictive power.” (עמ' 8)

- ממצא נוסף: מחקרים שמיקדו תחום (כמו מדעי המחשב או רפואה) הצליחו טוב יותר במידול.

**מאמר 7:**

Shaninah M. & Noor M. (2024) – The Impact of Big-Five Personality Traits on Student Academic Performance

PDF: https://files.eric.ed.gov/fulltext/EJ1415377.pdf

1. מה יש בדאטה

- 305 סטודנטים לתואר ראשון (אוניברסיטת אל-זינטאן, לוב).

- שאלון 44 פריטים (Big-Five Personality) + גיל, מגדר, תעסוקה, הכנסה.

(Table 1, p. 4)

2. מה שאלת המחקר

- האם תכונות האישיות והדמוגרפיה מספיקות לחיזוי Student Academic Performance ללא ציוני עבר?

(Objectives, p. 3)

3. באילו שיטות השתמשו

- PLS-SEM בהשוואה ל-CB-SEM.

- בדיקות מהימנות Cronbach α, Composite Reliability, AVE.

(Methods, pp. 5-6)

4. מה המודל שעבד הכי טוב ועל-פי מה נמדד

- מודל PLS-SEM: R² ≈ 0.48 להסבר SAP.

(Results, Table 3, p. 8)

5. מה התובנות

- כל חמש תכונות Big-Five מובהקות; אישיות הקטגוריה הדומיננטית.

- דמוגרפיה תרמה מעט בלבד.

- אין נתוני ציוני עבר → מאפשר לבדוק אצלכם תרומת אישיות מול משתנים לימודיים.

**מאמר 8:**

Liu H. et al. (2023) – Predicting Student Success with LSTM on Clickstream Data (OULAD)

PDF: https://www.mdpi.com/2227-7102/13/1/17/pdf

1. מה יש בדאטה

- 5 341 סטודנטים של Open University-UK.

- 213 000 אירועי קליקים ב-LMS (content, quiz, subpage, homepage).

(Section 3.1 – Data Description)

2. מה שאלת המחקר

- האם קליקים בלבד יכולים לחזות Pass/Fail בקורס וללא ציוני עבר.

(Introduction – p. 2)

3. באילו שיטות השתמשו

- אלגוריתמים: LSTM, 1D-CNN, Random Forest, GBDT, Logistic Regression.

- SMOTE לאיזון, 10-fold CV.

(Section 3.2 – Methods)

4. מה המודל שעבר טוב ועל-פי מה

- LSTM – Accuracy ≈ 90 % (Table 4 – Results).

- מדדים: Accuracy, Precision, Recall, F1.

5. מה התובנות

- אירועי ‘content’ ו-‘quiz’ בעלי חשיבות גבוהה ביותר.

- דמוגרפיה לא שיפרה ביצועים.

- מראה שחיזוי מדויק אפשרי ממידע התנהגותי בלבד – מתאים לבדוק את Attendance/Hours אצלכם.

**מאמר 9:**

Araka E. et al. (2022) – Self-Regulated Learning Profiles and Course Achievement

PDF: https://files.eric.ed.gov/fulltext/EJ1332590.pdf

1. מה יש בדאטה

- Trace-logs + שאלון Self-Regulated Learning (SRL) של Open University UK.

- 4 פרופילי-SRL: Exemplary, Strategic, Struggling, Poor.

(Table 2 – p. 8)

2. מה שאלת המחקר

- כיצד פרופיל-SRL קשור לשיעור הצלחה וציון סופי.

(Research Questions – p. 5)

3. באילו שיטות השתמשו

- Agglomerative Hierarchical Clustering להגדרת פרופילים.

- ANOVA וקורלציה בין פרופיל לציון.

(Methods – p. 7)

4. מה המודל שעבר טוב ועל-פי מה

- לא מודל חיזוי; מתאם ρ ≈ 0.32 בין הפרופיל לציון סופי.

(Results – p. 10)

5. מה התובנות

- פרופיל ‘Exemplary’ מציג שיעור מעבר כפול לעומת ‘Poor’.

- מדגים את חשיבות מיומנויות ניהול-עצמי – ייתכן למנף משתנה Skill-Dev Hours אצלכם.

**מאמר 10:**

Al-Alawi M. et al. (2023) – Predicting Academic Probation with Non-GPA Factors

PDF: https://pmc.ncbi.nlm.nih.gov/articles/PMC9999331/pdf/nihpp-2023.03.22.23287732.pdf

1. מה יש בדאטה

- 6 514 סטודנטים באוניברסיטה בעומאן (שנים 2010-2021).

- פיצ’רים: משך לימודים, מגדר, פקולטה, ציוני תיכון; ללא GPA אוניברסיטאי.

(Dataset – Table 1, p. 4)

2. מה שאלת המחקר

- אילו גורמים מנבאים סבירות לעמוד על-תנאי / השעיה אקדמית.

(Objectives – p. 3)

3. באילו שיטות השתמשו

- Feature Selection InfoGain.

- Ensemble Models: LogitBoost, Bagging, Voting; 10-fold CV.

(Methods – pp. 5-6)

4. מה המודל שעבר טוב ועל-פי מה

- LogitBoost – AUC ≈ 0.79 (Table 3, p. 8).

5. מה התובנות

- משך הלימודים באוניברסיטה וציון תיכון הם המנבאים החזקים ביותר.

- מגדר ופקולטה מוסיפים תרומה בינונית; גיל זניח.

- מדגים פוטנציאל חיזוי ללא שימוש ב-GPA קודם – ניתן להשוות למודל אצלכם אם תסירו SGPA.

**מאמר 11:**

Emhemed A. et al. (2021) – Personality Traits, Study Effort and GPA

PDF: https://www.researchgate.net/profile/Anis-Emhemed/publication/349360269\_Prediction\_of\_Students\_Performance\_Based\_on\_Personal\_Traits\_and\_Study\_Effort/links/60327e2f92851c6de5f4bd5d/Prediction-of-Students-Performance-Based-on-Personal-Traits-and-Study-Effort.pdf

1. מה יש בדאטה

- 437 סטודנטים באוניברסיטה קנדית.

- משתנים: Big-Five Personality, שעות לימוד שבועיות, עבודה, שימוש ברשתות.

(Dataset – Table 1, p. 3)

2. מה שאלת המחקר

- האם שילוב אישיות + מאמץ לימודי חוזה GPA ללא ציונים קודמים.

(Introduction – p. 2)

3. באילו שיטות השתמשו

- Random Forest, SVM, Gradient Boosting; 5-fold CV.

(Methods – p. 4)

4. מה המודל שעבר טוב ועל-פי מה

- Random Forest – RMSE ≈ 0.37 (Results – Table 3, p. 6).

5. מה התובנות

- Conscientiousness ושעות לימוד שבועיות הם המאפיינים החזקים ביותר.

- שימוש ברשתות חברתיות מקושר ל-GPA נמוך.

- אישיות מוסיפה הסבר ניכר מעבר לדמוגרפיה.

**מאמר 12:**

## 2. The Impact of Extracurricular Activities on Students – Western Pennsylvania (2019)

\*קישור ל‑PDF:\* https://files.eric.ed.gov/fulltext/EJ1208711.pdf :contentReference[oaicite:2]{index=2}

1. \*מה יש בדאטה:\*

– N≈148 תלמידי י"א בתיכון כפרי בפנסילבניה

– משתנים: ECA (כן/לא), GPA, תחושת שייכות, נוכחות, התנהגות

2. \*מה שאלת המחקר:\*

– האם משתתפים ב‑ECA מציגים GPA גבוה יותר, תחושת שייכות והתנהגות טובה יותר לעומת לא‑משתתפים?

3. \*שיטות:\*

– סקרים (Likert) + השוואת ממוצעים (t-test, ANOVA)

4. \*מודל שעבד טוב:\*

– t-test/ANOVA על GPA ושייכות: GPA של 3.456 בין משתתפים לעומת 2.578 לא‑משתתפים (p<0.001), תחושת שייכות 56.9% מול 37.5%

5. \*התובנות:\*

– ECA מובילה לציונים גבוהים יותר, שייכות מוגברת והתנהגות טובה

– לא נבדק עומס או סיבתיות; תמיכה משמעותית בפעילות חוץ‑לימודית

**תובנות:**1. במאמרים שיש ציונים של אוניברסיטה זה המנבא החזק.

2. דמוגרפי חלש יחסית.

3. אישיות והתנהגות לימודית משפיעים לפעמים על המודל.

היפוטזות:

1. **איך תחומי עניין משפיעים על החיזוי הציון האם יש קורלציות בין התחומים השונים (האם חלוקת הסטודנטים לתתי קבוצות יכולה להשפיע על feature important).**
2. תחומי עניין **- בר**
3. להפוך את הציון לטווחים **- אורן**
4. Skills **- נעם**
5. דמוגפי - טווחים - income **free style**
6. רמת אנגלית **- דותן**
7. ועוד... **- free style**

**פיצירים - כל הפיצרים שמסומנים בבולד/צהוב.**

1. **איך הרגלי למידה והתנהגות משפיעים על החיזוי של הציון. האם בן אדם שעושה "מעבר" הציונים שלו גבוהים יותר.**

**פיצרים – מה שמסמן בקו תחתון.**

רעיונות מיצוי:

1. Clustering
2. לראות אם יחס אידאלי בין משתנים
3. עד מתי משתנים משפיעים ע"ב הערך.
4. חשיבות features

================

Team name 2

This Markdown file describes the data folder structure and organization of Academic performance dataset of Bangladeshi students:

1. "University Admission year" (chr): the range of enter the university is 2013-2023.

**2. "Gender" (chr): Options- male/female.**

**3. "Age" (chr) the age range is 18-27.**

4. "H.S.C passing year" (dbl): High School Certificate completion year the range is 2012-2022 (with an outlier in 2028).

5. "Program" (chr): all of them in Bachelor of Computer Science and Engineering.

6. "Current Semester" (dbl): values range 1-24.

7. "Do you have meritorious scholarship?" (chr): A merit-based financial award given to academically high-achieving students (Options- No/Yes).

8. "Do you use University transportation?" (chr): A service provided by the university to help students commute to campus (Options- No/Yes).

**9. "How many hour do you study daily?" (chr): values range 0-13 hours.**

**10. "How many times do you seat for study in a day?" (dbl): values range 0-7 times**

**11. "What is your preferable learning mode?" (chr): Options- Offline/Online.**

**12. "Do you use smart phone?"(chr): Options- No/Yes.**

**13. "Do you have personal Computer?" (chr): Options- No/Yes.**

**14. "How many hour do you spent daily in social media?" (chr): values range 0-20 hours.**

**15. "Status of your English language proficiency" (chr): Options - Intermediate/Basic/Advanced.**

**16. "Average attendance on class" (chr): values range 0-100 percentage.**

**17. "Did you ever fall in probation?" (chr): A formal warning status assigned to students who fail to maintain minimum academic standards (Options- No/Yes).**

**18. "Did you ever got suspension?" (chr): A temporary removal from the university due to severe academic underperformance or rule violations (Options- No/Yes).**

**19. "Do you attend in teacher consultancy for any kind of academical problems?" (chr): Options- No/Yes.**

**20. "What are the skills do you have ?" (chr): Options are multiple skills. The most common: Programming (44.6%), Web development (21.0%), Networking (12.8%).**

**21. "How many hour do you spent daily on your skill development?" (chr): values range 0-12 hours**

**22. "What is you interested area?" (chr): Options are multiple area listed. The most common: Software (54.3%), Hardware (13.8%), Data Science (11.1%).**

23. "What is your relationship status?" (chr): Options- Single/Relationship/Married/Engaged.

**24. "Are you engaged with any co curriculum activities?" (chr): Options- No/Yes.**

**25. "With whom you are living with?" (chr): Options - Family/Bachelor.**

26. "Do you have any health issues?" (chr): Options- No/Yes.

27. "What was your previous SGPA?" (chr): A measure of a student's academic performance in their most recent completed semester. Values range - 0-4.

28. "Do you have any physical disabilities?" (chr): Options- No/Yes.

29. "What is your current CGPA?" (chr): The average of all grade points earned across all semesters to date. Values range - 0-4.

**30. "How many Credit did you have completed?" (chr): values range 0-145.**

**31. "What is your monthly family income?" (chr): values range 4000-2,000,000.**