1. לקרוא על סביבת  [Open AI Gym](https://gym.openai.com/) ולהתקין אותה בכדי שתוכלו להשתמש בה. הסביבה נקראת היום: Gymnasium ונמצאת ב- <https://gymnasium.farama.org> יש שם את כל המידע שאתם צריכים כדי להשתמש בסביבה, אבל תוכלו למצא מידע נוסף באתרים שונים, אם תרצו.

2. יש לפתור את בעיית blackjack מתוך סביבת toy text. הפתרון יבוסס על שני שלבים:

א. למידת המודל. אתם יודעים את המצבים האפשריים, ואת הפעולות האפשריות. אתם יכולים להניח שאתם גם יודעים את הפרס. הניחו שאין התייחסות מיוחדת למקרה של natural blackjact. (ראו בדוקומנטציה, כלומר הפרסים הם 1,1-,0). עליכם ללמוד את מטריצת המעברים על ידי דגימה של הרבה ניסיונות. כתבו פונקצייה שנקראת: **transition\_matrix** שמדפיסה את מטריצת מעברים ככה שכל מספר בעל 3 נקודות עשרוניות ( למשל 0.001).

ב. אחרי שיש לכם את מטריצת המעברים, פתרו את הבעיה בעזרת modified policy iteration. הניחו שגאמה=1 (אין בעיה בגלל שכל משחק סופי וחסום, ולכן אין צורך כאן במקדם דעיכה). לצורך חישוב ערך מדיניות, הסתפקו במספר קבוע של צעדים k שיהיה פרמטר שלכם. אתחלו אותו ל-5.

שימו לב, בתחילת המשחק יש לדילר קלף מוסתר. אין צורך למדל אותו. אין הבדל בין קלף מוסתר לבין פשוט לקבל עוד קלף. (תחת הנחות הדגימה מחבילה אינסופית שיש במימוש הנ״ל). ספקו גרף שמראה את ההתכנסות של המדיניות: תבחרו n>1 מצבים, בציר X יש את מספר האיטרציה ובציר Y את הפעולות. בגרף אחד, עבור כל מצב שבחרתם ואיטרציה מסויימת, תראו את הפעולה הנבחרת. יש לכם חופש הצגה על הפעולה בציר Y.

3. הקוד שלכם יהיה מתועד, וצריך להכיל גם את השיטות הבאות:

א. שיטה שבהנתן מדיניות, מסוגלת להפעיל את המדיניות הנ״ל במשחק.

ב. שיטה שבהנתן פונקצית ערך ומצב, יודעת להחזיר את הערך של מצב. כתבו פונקצייה שנקראת: **value\_function\_q3** שבהינתן ערך ומצב, מדפיסה את הערך של מצב.

הגשה:

יש להגיש קובץ שיקרא assignment1\_<id1>\_<id2>.zip, המכיל את קבצי הקוד יחד עם מסמך pdf קצר בו ת.ז, השמות שלכם, הסבר על הקוד בכלליות, והסבר על שיטות מרכזיות.

כמו כן: אנא ספקו

א. גרף המדגים **שיפור ערך המדיניות.** המדיניות הראשונה תיהיה מדיניות שתמיד מבקשת עוד קלף אם הערך נמוך מ-21. בגרף ציר x יהיה מספר האיטרציה וציר y יהיה הערך (המשוערך) הממוצע של המדיניות באיטרציה הנ״ל על כל המצבים ההתחלתיים האפשריים (כלומר, לדילר יש קלף בודד, ולכם יש שני קלפים). ערך ההתחלתי של המדיניות הראשונה יאותחל ל-0. הערך ההתחלתי של מדיניות אחרת יהיה הערך המשוערך של המדיניות הקודמת.

ב. טבלה שמתארת את המדיניות האופטימלית למצבים התחלתיים אשר בהם לאף שחקן אין אס. כלומר, טבלה בגודל 12 על 17 (ערכים 4-20).

בהצלחה