

נושאים נבחרים בסטטיסטיקה

פרופ' ישראל פרמט



מאיה ברסלאור 208997379

איל ברימן 316062702

משה כהן 316161694



תוכן עניינים

3.....	תקציר
3.....	מבוא
3.....	שאלת המחקר
4.....	שיטה
4.....	תיאור הנתונים
5.....	תיאור המשתנים
9.....	תיאור סטטיסטי של המשתנים
16.....	בחירת משתנים (Feature Selection)
17.....	איחוד קטגוריות של המשתנים שנבחרו
20.....	מודל רגרסיה לינארית
23.....	מודלי סיווג
25.....	מודלי למידת מכונה (Machine Learning Classification)
25.....	Random Forest
26.....	Support Vector Machine (SVM)
27.....	Neural Network
28.....	סיכום תוצאות ערכי ההיפר-פרמטרים אשר כיוונו במודלים השונים
29.....	סיכום ומסקנות
30.....	נספחים



תקציר

במסגרת הפרויקט בנינו מודלי חיזוי להצלחתו של תלמיד בבית הספר במהלך לימודיו. ניבוי הצלחתו של תלמיד יכול לסייע בכך שבמקרה ומאפייניו האישיים והדמוגרפיים של תלמיד משייכים אותו לקבוצות חלשות באוכלוסייה, יהיה ניתן לאפשר לו קבלת ליווי מקצועי וחניכה אישית ובכך להגדיל את סיכויי הצלחתו בלימודים. הפרויקט התבסס על הידע שצברנו בקורסי נתונים וסטטיסטיקה כמו אמידה ומבחני השערות, רגרסיה לינארית, למידת מכונה ונושאים נבחרים בסטטיסטיקה תוך שימוש ב-Python ו-R. ראשית, בחנו את סט הנתונים שברשותנו ולמדנו את התפלגויות הנתונים של כל אחד מהמשתנים ואת הקשרים ביניהם. משום שהמשתנה המוסבר (ציון ממוצע) הוא רציף ביצענו מודל רגרסיה לינארית, אך עקב אי קיום הנחות המודל וקבלת התאמה נמוכה בין המודל לנתונים, המרנו את הבעיה לבעיית סיווג בינארית (עבר בהצלחה/נכשל בשנת הלימודים). מצאנו את המשתנים המשפיעים ביותר על המשתנה המוסבר בטכניקות שונות וביצענו מספר מודלים: Logistic Regression, Random Forest, SVM ו-Neural Network. במטרה למצוא ולדייק את המודל הטוב ביותר לבעיה. לאחר בחינת כל המודלים, נמצא כי המודל המתאים עבור הבעיה הוא Random Forest, אשר מנבא בצורה הטובה ביותר האם התלמיד ייכשל/יעבור בהצלחה את שנת הלימודים.

מילות מפתח: מודלים לינאריים, מודלי סיווג, למידת מכונה.

מבוא

היכולת לנבא כיצד ישפיעו מאפיינים אישיים של תלמידים על ביצועיהם הינה בעיה אשר מעסיקה רבות תחום החינוך והתעסוקה, זאת משום שישנה הסתמכות רבה על ציוני התלמידים שאמורים להעיד על יכולותיהם. אמנם שיטה זו נתפשת מיושנת ולעתים לא רלוונטית אך נראה כי מערכות רבות עדין ממשיכות להסתמך עליה. לכן, בחרנו לבחון את הנושא באמצעות מידע על תלמידי תיכון, תוך מציאת התאמה בין ציוני התלמידים למאפייניהם האישיים. ציוני תלמידי התיכון מושפעים מגורמים שונים, שביניהם מדדים דמוגרפיים ואישיים. מדדים אלו, הכוללים הקשרים סוציו-אקונומיים ומשפחתיים לצד הרגלי צריכת אלקוהול ומצב בריאותי, עשויים להשפיע על מסירותם של התלמידים ללימודיהם ועל מידת הפניות ויכולת ההשקעה, שבסופו של דבר כנראה משפיעים על הציונים שלהם. מחקר זה נערך בשני בתי ספר שונים בליסבון שפורטוגל (גבריאל פריירה ומוסיניו דה סילבירה) ומתמקד בהבנת ההשפעה של תכונות דמוגרפיות ספציפיות על ביצועי התלמידים בבית הספר. קבוצת המחקר מכילה גברים ונשים בגילאי 15-22 ואיסוף המידע התבצע באמצעות שאלונים אישיים. הנתונים נאספו בשנת 2008 על ידי פ. קורטז וא. סילבה והוצג בכנס FUBUTEC 2008 בפורטו שפורטוגל.

שאלת המחקר

האם ניתן לחזות את הצלחתו של תלמיד תיכון על בסיס מאפייניו האישיים?



שיטה

תיאור הנתונים

הנתונים איתם עבדנו בפרויקט הינם נתונים דמוגרפיים, מגדריים, אישיים ומשפחתיים אודות תלמידי תיכון. נתונים אלו מוצגים בטבלת Excel המכילה 649 רשומות כך שכל רשומה מציגה נתונים אודות תלמיד תיכון. בנוסף, טבלת הנתונים מכילה 31 משתנים, כך ש-29 מתוכם הם משתנים מסבירים ו-2 מתוכם הם משתנים מוסברים - כל אחד מהם מציין את ציון התלמיד בכל סמסטר, סמסטר א'/ב'. נתייחס בפרויקט זה לממוצע הציונים כמשתנה המוסבר אותו נרצה לחזות. הנתונים נלקחו מתוך אתר Kaggle ונאספו על ידי החוקרים פ. קורטז ו-א. סילבה לצורך מחקר אשר הוצג בכנס FUBUTEC בפורטו אשר בפורטוגל בשנת 2008.



תיאור המשתנים

• משתנים מסבירים:

משתנה	סימון	סוג המשתנה	הסבר
school	X_1	קטגוריאל (נומינלי)	בית הספר בו לומד התלמיד: GP (Gabriel Pereira) , MS (Mousinho da Silveira)
sex	X_2	קטגוריאל (נומינלי)	מגדר התלמיד: F (female) , M (male)
age	X_3	בדיד	גיל התלמיד: 15-22
Housing Type	X_4	קטגוריאל (נומינלי)	סוג אזור המגורים של התלמיד: U (urban) , R (rural)
Family Size	X_5	קטגוריאל (אורדינלי)	גודל משפחת התלמיד: LE3 - משפחה עם עד שלוש נפשות GT3 - משפחה עם יותר משלוש נפשות
Parental Status	X_6	קטגוריאל (נומינלי)	סטטוס הזוגיות של הורי התלמיד: T - ההורים חיים יחד A - ההורים חיים בנפרד
Mothers Education	X_7	קטגוריאל (אורדינלי)	רמת ההשכלה של אם התלמיד: 0 - ללא השכלה 1 - בית ספר יסודי 2 - חטיבת ביניים 3 - בית ספר תיכון 4 - השכלה גבוהה
Fathers Education	X_8	קטגוריאל (אורדינלי)	רמת ההשכלה של אב התלמיד: 0 - ללא השכלה 1 - בית ספר יסודי 2 - חטיבת ביניים 3 - בית ספר תיכון 4 - השכלה גבוהה
Mothers Work	X_9	קטגוריאל (נומינלי)	סוג העבודה של אם התלמיד: מורה, בריאות, שירות, בית (לא עובדת), אחר
Father Work	X_{10}	קטגוריאל (נומינלי)	סוג העבודה של אב התלמיד: מורה, בריאות, שירות, בית (לא עובד), אחר
Reason School Choice	X_{11}	קטגוריאל (נומינלי)	סיבת הבחירה בבית הספר: קרוב לבית, מוניטין, קורס (מועדף המועבר בבית הספר), אחר

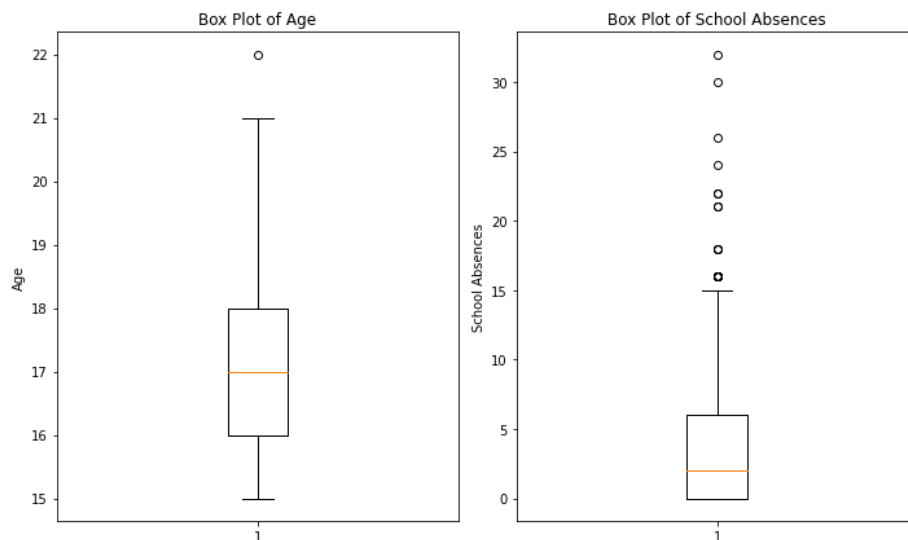


אפוטרופוס התלמיד : אמא, אבא, אחר	קטגוריאל (נומינלי)	X_{12}	Legal Responsibility
זמן הנסיעה לבית הספר : 0 - פחות מ-15 דקות 1 - בין 15 דקות לחצי שעה 2 - בין חצי שעה לשעה 3 - למעלה משעה	קטגוריאל (אורדינלי)	X_{13}	Commute Time
זמן למידה שבועי : 1 - עד לשעתיים 2 - בין שתיים לחמש שעות 3 - בין חמש לעשר שעות 4 - למעלה מעשר שעות	קטגוריאל (אורדינלי)	X_{14}	Weekly Study Time
האם קיימת תמיכה חינוכית נוספת (מעבר לבית הספר) : כן, לא.	קטגוריאל (נומינלי)	X_{15}	Extra Educational Support
האם קיימת תמיכה חינוכית משפחתית : כן, לא.	קטגוריאל (נומינלי)	X_{16}	Parental Educational Support
האם התלמיד לוקח שיעורים פרטיים בתשלום (בקורסים הנלמדים בבית הספר) : כן, לא.	קטגוריאל (נומינלי)	X_{17}	Private Tutoring
האם התלמיד מבצע פעילויות מחוץ לבית הספר : כן, לא.	קטגוריאל (נומינלי)	X_{18}	Extracurricular Activities
האם התלמיד לומד במעון יום : כן, לא.	קטגוריאל (נומינלי)	X_{19}	Attended Daycare
האם לתלמיד יש רצון להמשיך לתואר אקדמאי : כן, לא.	קטגוריאל (נומינלי)	X_{20}	Desire Graduate Education
האם לתלמיד יש גישה בביתו לאינטרנט : כן, לא.	קטגוריאל (נומינלי)	X_{21}	Has Internet
האם התלמיד בקשר רומנטי : כן, לא	קטגוריאל (נומינלי)	X_{22}	Is Dating
מידת היחסים של התלמיד עם משפחתו (1-5) : 1-5 ; 1 - גרוע מאוד , 5 - מצוין	קטגוריאל (אורדינלי)	X_{23}	Good Family Relationship
מידת הזמן הפנוי של התלמיד לאחר הלימודים (1-5) : 1 - נמוך מאוד , 5 - גבוה מאוד	קטגוריאל (אורדינלי)	X_{24}	Free Time After School



מידת הזמן בו התלמיד נפגש עם חברים (1-5): 1 - נמוך מאוד, 5 - גבוה מאוד	קטגוריאל (אורדינלי)	X_{25}	Time with Friends
מידת צריכת האלכוהול של התלמיד בימי לימודים (1-5): 1 - נמוך מאוד, 5 - גבוה מאוד	קטגוריאל (אורדינלי)	X_{26}	Alcohol Weekdays
מידת צריכת האלכוהול של התלמיד בסוף שבוע (1-5): 1 - נמוך מאוד, 5 - גבוה מאוד	קטגוריאל (אורדינלי)	X_{27}	Alcohol Weekends
מצבו הבריאותי של התלמיד (1-5): 1 - רע מאוד, 5 - טוב מאוד	קטגוריאל (אורדינלי)	X_{28}	Health Status
מספר ההיעדרות של התלמיד מבית הספר: 0-32	בדיד	X_{29}	School Absence

בכדי ללמוד על התפלגות הנתונים של המשתנים הבדידים, ייצרנו תרשימי קופסה (Box-Plot) עבור כל אחד מהם. ניתן לראות כי הגיל החציוני של התלמידים שנבדקו הוא 17 והפיזור סביב ערך זה מתנהג באופן אחיד. כלומר, טווח הגילאים של התלמידים שנבדקו הוא מפורז סביב ערך זה באופן דומה. כמו כן, רוב התלמידים נעדרים לא יותר משישה ימים במהלך שנת הלימודים.



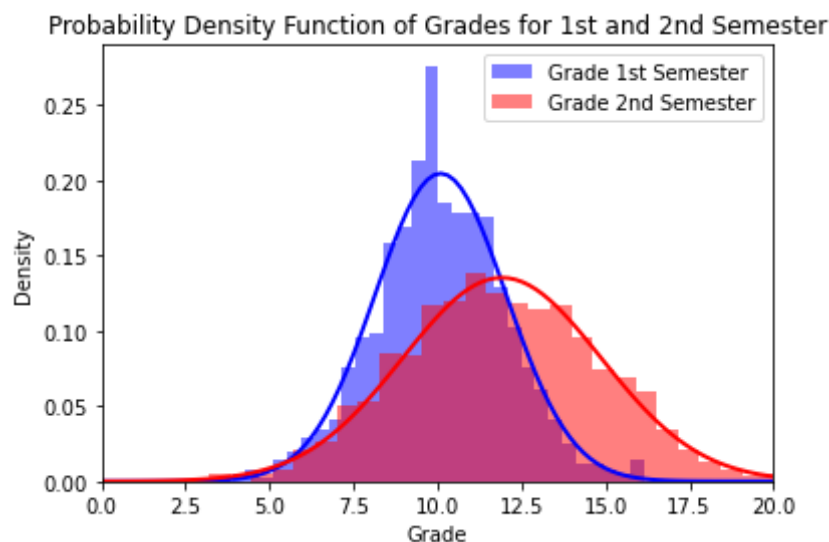
בנוסף, בתרשימים אלו ניתן לזהות תצפיות חריגות ביחס להתפלגות המשתנים. עבור משתנה הגיל (Age), זוהי חריג אחד המייצג תלמיד בודד שגילו גבוה ביחס לשאר התלמידים. עבור משתנה מספר ההיעדרויות (School Absences) זוהי שמונה חריגים המייצגים תלמידים שנעדרו במהלך השנה יותר ביחס לשאר התלמידים. משום שברשותנו מאגר נתונים רחב המכיל מספר רב של תצפיות (649), חריגים אלו זניחים. כמו כן, הם מייצגים מקרים רלוונטיים שנרצה שהמודל שלנו ילמד מהם, ולכן החלטנו לא להסיר אותם בעת אימון כל המודלים.



• משתנים מוסברים:

משתנה	סימון	סוג המשתנה	הסבר
Grade 1st Semester	Y_1	בדיד	ציון הסטודנט בסמסטר הראשון (0-20)
Grade 2nd Semester	Y_2	בדיד	ציון הסטודנט בסמסטר השני (0-20)
Grade	Y	רציף	ציון הממוצע של הסטודנט בשני הסמסטרים (0-100) $\frac{(G1 + G2)}{2}$

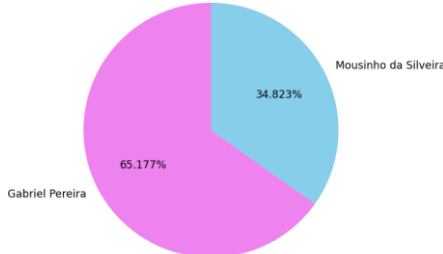
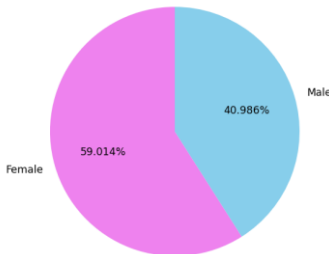
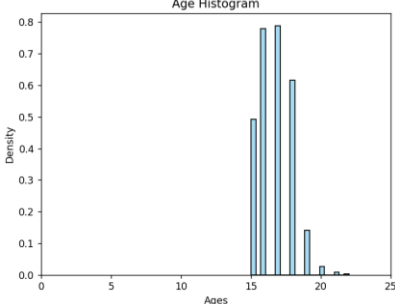
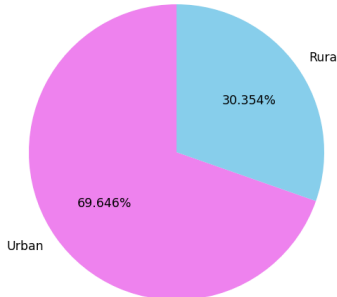
כחלק מבחינת התפלגות הציונים של התלמידים בסמסטרים השונים ביצענו את מבחן טי-מזווג (Paired t-test) שהתקבל כמובחק (ברמת מובהקות 95%). כלומר, ניתן לומר כי יש הבדל בין הציון של התלמיד בסמסטר א' לעומת ציונו בסמסטר ב'. לאחר בחינת הממוצעים בשני הסמסטרים (סמסטר א' - 11.4, סמסטר ב' - 11.57), מצאנו כי ברוב המקרים התלמיד ישפר את הציון שלו בסמסטר השני.



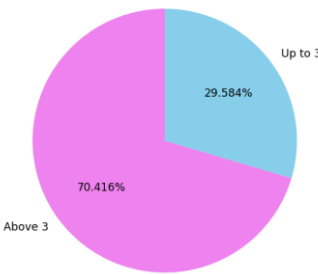
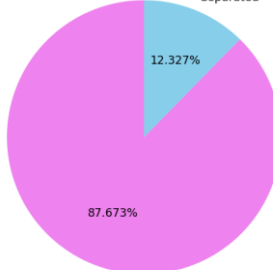
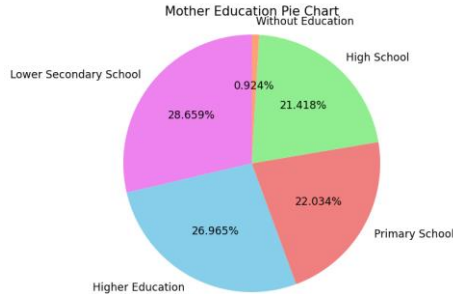
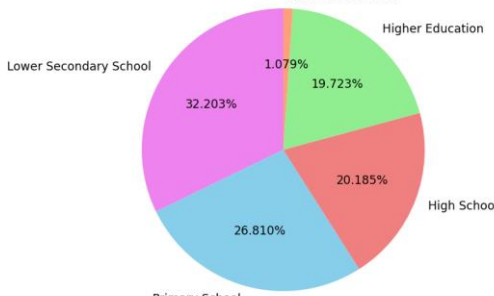
Paired t-test results:
t-statistic: -2.906349849150211
p-value: 0.0037839647405636585
Reject the null hypothesis. There is a significant difference between the two columns.

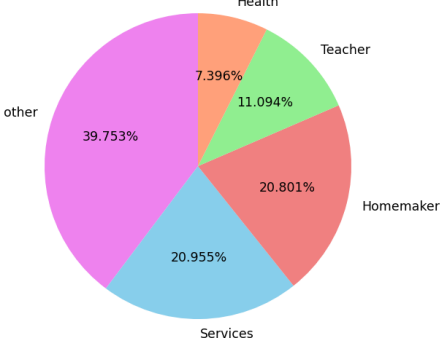
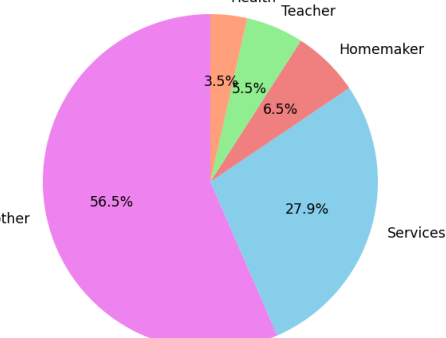
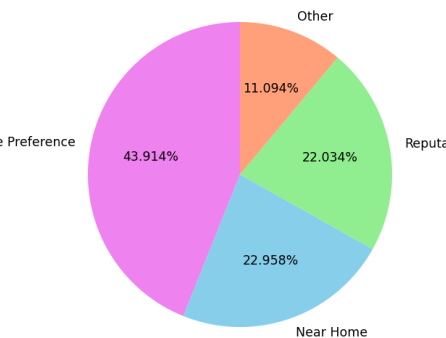
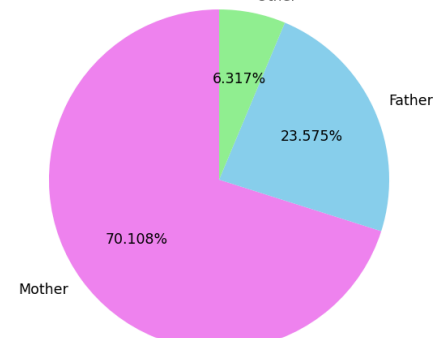
לצורך הפרויקט המנסה לנבא ציון ממוצע כללי של התלמיד במהלך שנת הלימודים, ביצענו ממוצע של ציוני התלמיד בשני הסמסטרים באמצעות הגדרה של משתנה מוסבר חדש (Grade) בשביל לשקף את ביצועי התלמיד בשני הסמסטרים.

תיאור סטטיסטי של המשתנים

המשתנה	תיאור סטטיסטי - התפלגות	הסבר מילולי																
School	<p>School Pie Chart</p>  <table><thead><tr><th>School</th><th>Percentage</th></tr></thead><tbody><tr><td>Gabriel Pereira</td><td>65.177%</td></tr><tr><td>Mousinho da Silveira</td><td>34.823%</td></tr></tbody></table>	School	Percentage	Gabriel Pereira	65.177%	Mousinho da Silveira	34.823%	<p>ניתן לראות כי התלמידים שנבדקו לומדים בשני בתי ספר: Gabriel Pereira, בו לומדים כ-65% מהתלמידים ו-Mousinho da Silveira, בו לומדים כ-35% מהתלמידים.</p> <p><u>מקרא:</u> כחול - Mousinho da Silveira סגול - Gabriel Pereira</p>										
School	Percentage																	
Gabriel Pereira	65.177%																	
Mousinho da Silveira	34.823%																	
Gender	<p>Gender Pie Chart</p>  <table><thead><tr><th>Gender</th><th>Percentage</th></tr></thead><tbody><tr><td>Male</td><td>40.986%</td></tr><tr><td>Female</td><td>59.014%</td></tr></tbody></table>	Gender	Percentage	Male	40.986%	Female	59.014%	<p>ניתן לראות שאוכלוסיית התלמידים הנבדקת מורכבת מכ-59% נשים וכ-41% גברים. זאת בשונה מהתפלגות האוכלוסייה בעולם אשר מורכבת מ-50% נשים ו-50% גברים, עם יתרון קל לנשים.</p> <p><u>מקרא:</u> כחול - Male סגול - Female</p>										
Gender	Percentage																	
Male	40.986%																	
Female	59.014%																	
Age	<p>Age Histogram</p>  <table><thead><tr><th>Ages</th><th>Density</th></tr></thead><tbody><tr><td>15</td><td>0.50</td></tr><tr><td>16</td><td>0.78</td></tr><tr><td>17</td><td>0.78</td></tr><tr><td>18</td><td>0.62</td></tr><tr><td>19</td><td>0.15</td></tr><tr><td>20</td><td>0.05</td></tr><tr><td>21</td><td>0.02</td></tr></tbody></table>	Ages	Density	15	0.50	16	0.78	17	0.78	18	0.62	19	0.15	20	0.05	21	0.02	<p>ניתן לראות כי התלמידים הנבדקים הם בגילאים 15-22. רוב התלמידים נמצאים בגילאי 15-17: ממוצע הגילאים הוא 16.74 והחציון עומד על 17.</p>
Ages	Density																	
15	0.50																	
16	0.78																	
17	0.78																	
18	0.62																	
19	0.15																	
20	0.05																	
21	0.02																	
Housing Type	<p>Housing Type Pie Chart</p>  <table><thead><tr><th>Housing Type</th><th>Percentage</th></tr></thead><tbody><tr><td>Rural</td><td>30.354%</td></tr><tr><td>Urban</td><td>69.646%</td></tr></tbody></table>	Housing Type	Percentage	Rural	30.354%	Urban	69.646%	<p>כ-30% מהתלמידים גרים באזור כפרי וכ-70% גרים באזור עירוני.</p> <p><u>מקרא:</u> כחול - Rural סגול - Urban</p>										
Housing Type	Percentage																	
Rural	30.354%																	
Urban	69.646%																	



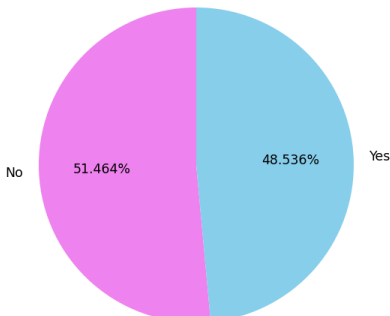
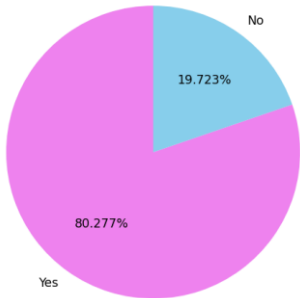
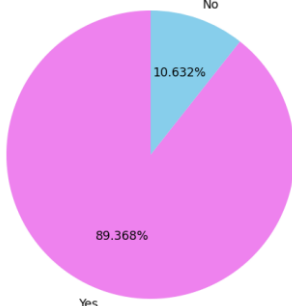
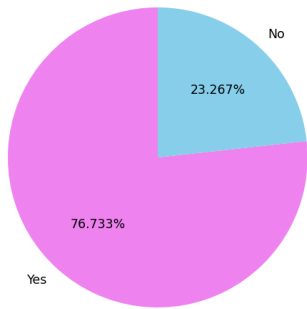
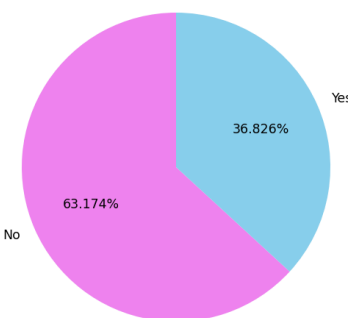
<p>ניתן לראות כי רוב משפחות התלמידים מכילות יותר מ-3 נפשות, כ-70%. 30% הנותרים הם משפחות המכילות עד 3 נפשות.</p> <p><u>מקרא:</u> כחול - Up to 3 סגול - Above 3</p>	<p>Family Size Pie Chart</p>  <table><tr><th>Family Size</th><th>Percentage</th></tr><tr><td>Up to 3</td><td>29.584%</td></tr><tr><td>Above 3</td><td>70.416%</td></tr></table>	Family Size	Percentage	Up to 3	29.584%	Above 3	70.416%	<p>Family Size</p>						
Family Size	Percentage													
Up to 3	29.584%													
Above 3	70.416%													
<p>ניתן לראות כי הרוב המוחלט של הורי התלמידים חיים ביחד, כ-88%. יתר ההורים, כ-12% חיים בנפרד.</p> <p><u>מקרא:</u> כחול - Separated סגול - Living Together</p>	<p>Parental Status Pie Chart</p>  <table><tr><th>Parental Status</th><th>Percentage</th></tr><tr><td>Separated</td><td>12.327%</td></tr><tr><td>Living Together</td><td>87.673%</td></tr></table>	Parental Status	Percentage	Separated	12.327%	Living Together	87.673%	<p>Parental Status</p>						
Parental Status	Percentage													
Separated	12.327%													
Living Together	87.673%													
<p>ניתן לראות שכ-1% מאימהות התלמידים ללא השכלה, כ-21% מהן בעלות השכלה תיכונית, כ-22% מהן בעלות השכלה חטיבתית, כ-27% מהן בעלות השכלה גבוהה וכ-29% מהן בעלות השכלה של בית ספר יסודי.</p> <p><u>מקרא:</u> כחול - High Education סגול - Lower Secondary School כתום - Without Education ירוק - High School אדום - Primary School</p>	<p>Mother Education Pie Chart</p>  <table><tr><th>Mother Education</th><th>Percentage</th></tr><tr><td>Without Education</td><td>0.924%</td></tr><tr><td>High School</td><td>21.418%</td></tr><tr><td>Primary School</td><td>22.034%</td></tr><tr><td>Higher Education</td><td>26.965%</td></tr><tr><td>Lower Secondary School</td><td>28.659%</td></tr></table>	Mother Education	Percentage	Without Education	0.924%	High School	21.418%	Primary School	22.034%	Higher Education	26.965%	Lower Secondary School	28.659%	<p>Mother Education</p>
Mother Education	Percentage													
Without Education	0.924%													
High School	21.418%													
Primary School	22.034%													
Higher Education	26.965%													
Lower Secondary School	28.659%													
<p>ניתן לראות שכ-1% מאבות התלמידים ללא השכלה, כ-20% מהם בעלי השכלה גבוהה, כ-20% מהם בעלי השכלה תיכונית, כ-27% מהם בעלי השכלה חטיבתית וכ-32% מהם בעלי השכלה של בית ספר יסודי.</p> <p><u>מקרא:</u> כחול - Primary School סגול - Lower Secondary School כתום - Without Education ירוק - Higher Education אדום - High School</p>	<p>Father Education Pie Chart</p>  <table><tr><th>Father Education</th><th>Percentage</th></tr><tr><td>Without Education</td><td>1.079%</td></tr><tr><td>Higher Education</td><td>19.723%</td></tr><tr><td>High School</td><td>20.185%</td></tr><tr><td>Primary School</td><td>26.810%</td></tr><tr><td>Lower Secondary School</td><td>32.203%</td></tr></table>	Father Education	Percentage	Without Education	1.079%	Higher Education	19.723%	High School	20.185%	Primary School	26.810%	Lower Secondary School	32.203%	<p>Father Education</p>
Father Education	Percentage													
Without Education	1.079%													
Higher Education	19.723%													
High School	20.185%													
Primary School	26.810%													
Lower Secondary School	32.203%													

<p>ניתן לראות כי ב-7% מהאימהות עובדות במשרות בריאות, כ-11% מורות, כ-21% עקרות בית, כ-21% עוסקות בעבודות שירות ו-40% עוסקות בעבודות אחרות.</p> <p><u>מקרא:</u> כחול - Services סגול - Other כתום - Health ירוק - Teacher אדום - Homemaker</p>	<p>Mother Work Pie Chart</p> 	<p>Mother Work</p>
<p>ניתן לראות כי כ-3.5% מהאבות עובדים במשרות בתחום הבריאות, כ-5.5% מורים, כ-6.5% עקרי בית, כ-30% עובדים בעבודות שירות וכ-56.5% עובדים בעבודות אחרות.</p> <p><u>מקרא:</u> כחול - Services סגול - Other כתום - Health ירוק - Teacher אדום - Homemaker</p>	<p>Father Work Pie Chart</p> 	<p>Father Work</p>
<p>ניתן לראות כ-22% מהתלמידים בחרו בבית הספר בגלל מוניטין, כ-23% בחרו בבית הספר בגלל קרבה לבית, כ-44% בחרו בבית הספר בגלל העדפות קורסים המתקיימים בו וה-11% בחרו בו מסיבות אחרות.</p> <p><u>מקרא:</u> כחול - Near Home סגול - Course Preference כתום - Other ירוק - Reputation</p>	<p>Reason School Choice Pie Chart</p> 	<p>Reason School Choice</p>
<p>ניתן לראות כי האפטרופוס (בעל אחריות משפטית) של כ-70% מהתלמידים הוא האמא, של כ-24% מהתלמידים הוא האבא ושל כ-6% מהתלמידים הוא אדם אחר.</p> <p><u>מקרא:</u> כחול - Father סגול - Mother ירוק - Other</p>	<p>Legal Responsibility Pie Chart</p> 	<p>Legal Responsibility</p>



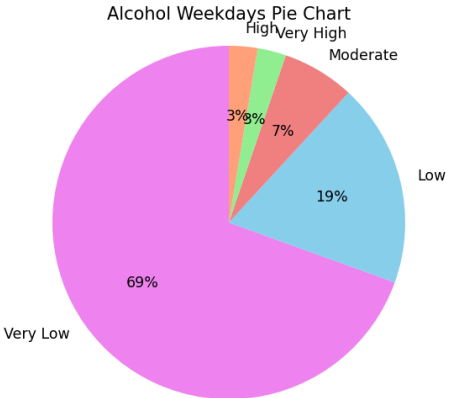
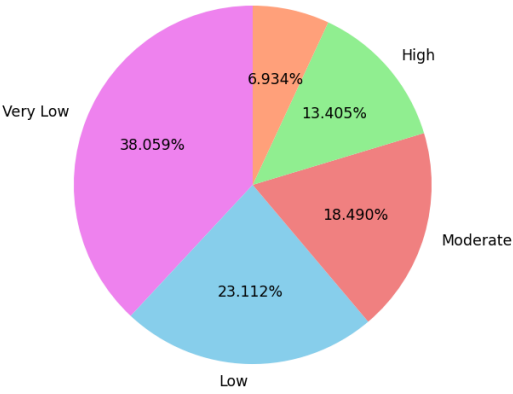
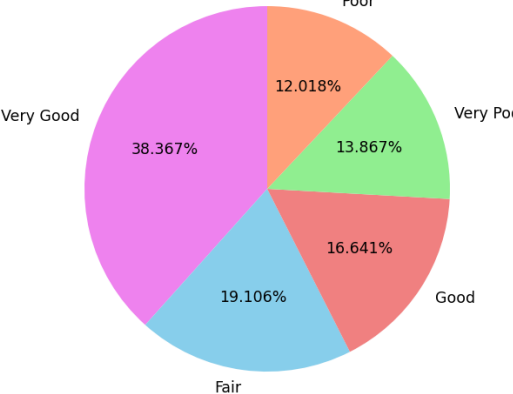
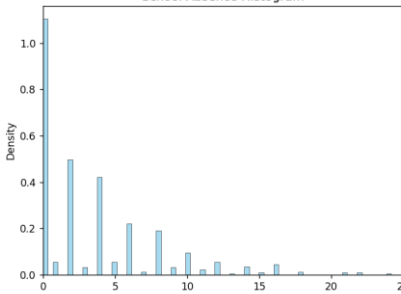
<p>ניתן לראות שלכ-56.4% מהתלמידים לוקח עד 15% דק' לנסוע לבית הספר, ל-32.8% מהתלמידים לוקח בין 15 ל-30 דק', ל-8.3% מהתלמידים לוקח בין 30 דק' לשעה ול-2.5% מהתלמידים לוקח יותר משעה לנסוע לבית הספר.</p> <p><u>מקרא:</u> כחול - 15 to 30 min סגול - Up to 15 min כתום - More than 1h ירוק - 30 min to 1h</p>	<p>Commute Time Pie Chart</p> <table><thead><tr><th>Travel Time</th><th>Percentage</th></tr></thead><tbody><tr><td>Up to 15 min</td><td>56.4%</td></tr><tr><td>15 to 30 min</td><td>32.8%</td></tr><tr><td>30 min to 1h</td><td>8.3%</td></tr><tr><td>More than 1h</td><td>2.5%</td></tr></tbody></table>	Travel Time	Percentage	Up to 15 min	56.4%	15 to 30 min	32.8%	30 min to 1h	8.3%	More than 1h	2.5%	<p>Commute Time</p>
Travel Time	Percentage											
Up to 15 min	56.4%											
15 to 30 min	32.8%											
30 min to 1h	8.3%											
More than 1h	2.5%											
<p>ניתן לראות שכ-90% מהתלמידים בעלי תמיכה חינוכית נוספת ול-10% אין תמיכה כזו.</p> <p><u>מקרא:</u> כחול - Yes סגול - No</p>	<p>Extra Educational Support Pie Chart</p> <table><thead><tr><th>Support</th><th>Percentage</th></tr></thead><tbody><tr><td>Yes</td><td>10.478%</td></tr><tr><td>No</td><td>89.522%</td></tr></tbody></table>	Support	Percentage	Yes	10.478%	No	89.522%	<p>Extra Educational Support</p>				
Support	Percentage											
Yes	10.478%											
No	89.522%											
<p>ניתן לראות כי ל-61% מהתלמידים יש תמיכה חינוכית משפחתית ולכ-39% מהתלמידים אין תמיכה כזו.</p> <p><u>מקרא:</u> כחול - No סגול - Yes</p>	<p>Parental Educational Support Pie Chart</p> <table><thead><tr><th>Support</th><th>Percentage</th></tr></thead><tbody><tr><td>Yes</td><td>61.325%</td></tr><tr><td>No</td><td>38.675%</td></tr></tbody></table>	Support	Percentage	Yes	61.325%	No	38.675%	<p>Parental Educational Support</p>				
Support	Percentage											
Yes	61.325%											
No	38.675%											
<p>ניתן לראות כי הרוב המוחלט של התלמידים (כ-94%) לא לוקח שיעורים פרטיים, לעומת 4% הנותרים של התלמידים אשר נעזרים בשיעורים פרטיים.</p> <p><u>מקרא:</u> כחול - Yes סגול - No</p>	<p>Private Tutoring Pie Chart</p> <table><thead><tr><th>Tutoring</th><th>Percentage</th></tr></thead><tbody><tr><td>Yes</td><td>6.009%</td></tr><tr><td>No</td><td>93.991%</td></tr></tbody></table>	Tutoring	Percentage	Yes	6.009%	No	93.991%	<p>Private Tutoring</p>				
Tutoring	Percentage											
Yes	6.009%											
No	93.991%											



<p>ניתן לראות כי משתנה זה כמעט מאוזן – כ-51% מהתלמידים מבצעים פעילות מחוץ לבית הספר וכ-49% מהתלמידים לא מבצעים פעילות מחוץ לבית הספר.</p> <p><u>מקרא:</u> כחול - Yes סגול - No</p>	<p>Extracurricular Activities Pie Chart</p>  <table><tr><th>Response</th><th>Percentage</th></tr><tr><td>Yes</td><td>48.536%</td></tr><tr><td>No</td><td>51.464%</td></tr></table>	Response	Percentage	Yes	48.536%	No	51.464%	<p>Extracurricular Activities</p>
Response	Percentage							
Yes	48.536%							
No	51.464%							
<p>ניתן לראות כי רוב התלמידים, כ-80% לומדים במעון יום וכ-20% אינם לומדים במעון יום.</p> <p><u>מקרא:</u> כחול - No סגול - Yes</p>	<p>Attended Daycare Pie Chart</p>  <table><tr><th>Response</th><th>Percentage</th></tr><tr><td>Yes</td><td>80.277%</td></tr><tr><td>No</td><td>19.723%</td></tr></table>	Response	Percentage	Yes	80.277%	No	19.723%	<p>Attended Daycare</p>
Response	Percentage							
Yes	80.277%							
No	19.723%							
<p>ניתן לראות מתוך הגרף כי הרוב המוחלט של התלמידים, 89.368%, מעוניינים להמשיך לתואר אקדמאי ו-10.632% אינם מעוניינים.</p> <p><u>מקרא:</u> כחול - No סגול - Yes</p>	<p>Desire Graduate Education Pie Chart</p>  <table><tr><th>Response</th><th>Percentage</th></tr><tr><td>Yes</td><td>89.368%</td></tr><tr><td>No</td><td>10.632%</td></tr></table>	Response	Percentage	Yes	89.368%	No	10.632%	<p>Desire Graduate Education</p>
Response	Percentage							
Yes	89.368%							
No	10.632%							
<p>ניתן לראות מתוך הגרף של-23% מהתלמידים אין גישה לאינטרנט וכ-77% מהתלמידים בעלי גישה לאינטרנט.</p> <p><u>מקרא:</u> כחול - No סגול - Yes</p>	<p>Has Internet Pie Chart</p>  <table><tr><th>Response</th><th>Percentage</th></tr><tr><td>Yes</td><td>76.733%</td></tr><tr><td>No</td><td>23.267%</td></tr></table>	Response	Percentage	Yes	76.733%	No	23.267%	<p>Has Internet</p>
Response	Percentage							
Yes	76.733%							
No	23.267%							
<p>ניתן לראות כי כ-63% מהתלמידים אינם נמצאים בקשר רומנטי וכ-37% נמצאים בקשר רומנטי.</p> <p><u>מקרא:</u> כחול - Yes סגול - No</p>	<p>Is Dating Pie Chart</p>  <table><tr><th>Response</th><th>Percentage</th></tr><tr><td>Yes</td><td>36.826%</td></tr><tr><td>No</td><td>63.174%</td></tr></table>	Response	Percentage	Yes	36.826%	No	63.174%	<p>Is Dating</p>
Response	Percentage							
Yes	36.826%							
No	63.174%							



<p>ניתן לראות כי כ-3.4% מהתלמידים ביחסים גרועים מאוד עם משפחותיהם, כ-4.5% ביחסים גרועים עם משפחותיהם, כ-15.6% ביחסים בינוניים עם משפחותיהם, כ-48.8% ביחסים טובים עם משפחותיהם וכ-27.7% ביחסים מצוינים עם משפחותיהם.</p> <p><u>מקרא:</u> כחול - Excellent סגול - Good כתום - Very Poor ירוק - Poor אדום - Fair</p>	<p>Good Family Relationship Pie Chart</p> <table><tr><th>Category</th><th>Percentage</th></tr><tr><td>Excellent</td><td>27.7%</td></tr><tr><td>Good</td><td>48.8%</td></tr><tr><td>Fair</td><td>15.6%</td></tr><tr><td>Poor</td><td>4.5%</td></tr><tr><td>Very Poor</td><td>3.4%</td></tr></table>	Category	Percentage	Excellent	27.7%	Good	48.8%	Fair	15.6%	Poor	4.5%	Very Poor	3.4%	<p>Good Family Relationship</p>
Category	Percentage													
Excellent	27.7%													
Good	48.8%													
Fair	15.6%													
Poor	4.5%													
Very Poor	3.4%													
<p>ניתן לראות מתוך הגרף כי ל-7% מהתלמידים יש מעט מאוד זמן פנוי, ל-11% יש הרבה מאוד זמן פנוי, ל-16% יש מעט זמן פנוי, ל-27% יש הרבה זמן פנוי ול-39% יש זמן פנוי במידה בינונית.</p> <p><u>מקרא:</u> כחול - High סגול - Moderate כתום - Very Poor ירוק - Very High אדום - Low</p>	<p>Free Time After School Pie Chart</p> <table><tr><th>Category</th><th>Percentage</th></tr><tr><td>High</td><td>27.427%</td></tr><tr><td>Moderate</td><td>38.675%</td></tr><tr><td>Low</td><td>16.487%</td></tr><tr><td>Very High</td><td>10.478%</td></tr><tr><td>Very Low</td><td>6.934%</td></tr></table>	Category	Percentage	High	27.427%	Moderate	38.675%	Low	16.487%	Very High	10.478%	Very Low	6.934%	<p>Free Time After School</p>
Category	Percentage													
High	27.427%													
Moderate	38.675%													
Low	16.487%													
Very High	10.478%													
Very Low	6.934%													
<p>ניתן לראות שכ-17% מהתלמידים מבליים במידה רבה מאוד עם חברים, כ-22% מבליים במידה רבה עם חברים, כ-32% במידה בינונית, כ-22% מבליים מעט וב-7% מבליים מעט מאוד עם חברים.</p> <p><u>מקרא:</u> כחול - Low סגול - Moderate כתום - Very Low ירוק - Very High אדום - High</p>	<p>Time with Friends Pie Chart</p> <table><tr><th>Category</th><th>Percentage</th></tr><tr><td>High</td><td>21.726%</td></tr><tr><td>Moderate</td><td>31.587%</td></tr><tr><td>Low</td><td>22.342%</td></tr><tr><td>Very High</td><td>16.949%</td></tr><tr><td>Very Low</td><td>7.396%</td></tr></table>	Category	Percentage	High	21.726%	Moderate	31.587%	Low	22.342%	Very High	16.949%	Very Low	7.396%	<p>Time With Friends</p>
Category	Percentage													
High	21.726%													
Moderate	31.587%													
Low	22.342%													
Very High	16.949%													
Very Low	7.396%													

<p>ניתן לראות כי כ-3% מהתלמידים שותים אלכוהול במהלך השבוע במידה רבה מאוד, כ-3% שותים במידה רבה, כ-7% שותים בכמות בינונית, כ-19% שותים במידה מעטה, והרוב המוחלט, כ-69% שותים אלכוהול במידה מעטה מאוד במהלך השבוע.</p> <p><u>מקרא:</u> כחול - Low סגול - Very Low כתום - High ירוק - Very High אדום - Moderate</p>	<p>Alcohol Weekdays Pie Chart</p> 	<p>Alcohol Weekdays</p>
<p>ניתן לראות שכ-7% מהתלמידים צורכים אלכוהול בסופי שבוע במידה רבה מאוד, כ-13% במידה רבה, כ-19% במידה בינונית, כ-23% צורכים מעט אלכוהול בסופי שבוע וכ-38% צורכים מעט מאוד.</p> <p><u>מקרא:</u> כחול - Low סגול - Very Low כתום - Very High ירוק - High אדום - Moderate</p>	<p>Alcohol Weekends Pie Chart</p> 	<p>Alcohol Weekends</p>
<p>ניתן לראות כי כ-38% מהתלמידים במצב בריאותי טוב מאוד, כ-17% במצב בריאותי טוב, כ-19% במצב בריאותי בינוני, כ-12% במצב בריאותי לא טוב וכ-14% במצב בריאותי מאוד לא טוב.</p> <p><u>מקרא:</u> כחול - Fair סגול - Very Good כתום - Poor ירוק - Very Poor אדום - Good</p>	<p>Health Status Pie Chart</p> 	<p>Health Status</p>
<p>ניתן לראות לפי התפלגות היעדרויות התלמידים כי רוב התלמידים נוכחים באופן מלא, וכי מעט תלמידים נעדרים באופן תדיר.</p>	<p>School Absence Histogram</p> 	<p>School Absence</p>



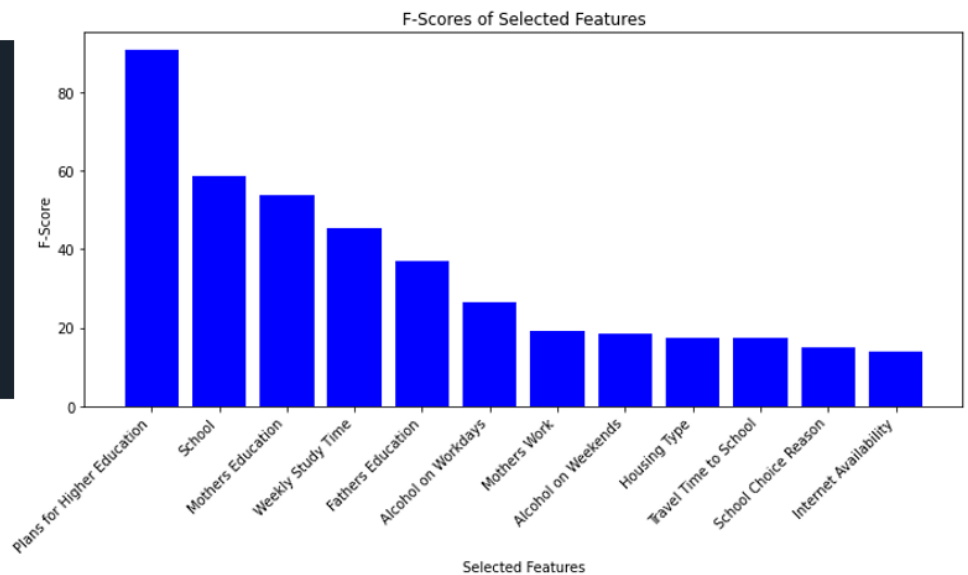
בחירת משתנים (Feature Selection)

סט הנתונים שברשותנו מכיל מספר רב של משתנים מסבירים (סה"כ 29 משתנים). כידוע, מימד גבוה עשוי להוות אתגר משמעותי בחיזוי המשתנה המוסבר. לכן, כשלב מקדים לבניית המודל נרצה לבחור את המשתנים שדרכם ניתן ללמוד בצורה היעילה ביותר על המשתנה המוסבר. כלומר, נבחר את המשתנים שהשונויות המוסברת שלהם היא גבוהה ביותר ביחס לשאר המשתנים, ובכך מספקים מידע טוב יותר על הציון הממוצע של כל תלמיד. בחרנו לבצע את המבחן הסטטיסטי F בכדי לבחון האם יש הבדל בין הקבוצות של המשתנה המסביר ביחס למוסבר. במבחן זה במידה ונקבל ערך F סטטיסטי גבוה ו-P value נמוך נוכל לקבוע ברמת אמינות מוגדרת (0.95) כי יש הבדל בין הקבוצות השונות, ובסבירות גבוהה המשתנה המסביר מספק הפרדה טובה בין הערכים של המשתנה המוסבר. ביצענו את המבחן עבור כל אחד מהמשתנים המסבירים, ולאחר מכן בחרנו את שניים-עשר המשתנים המסבירים בעלי ערכי F סטטיסטי הגבוהים ביותר. בחרנו לצמצם ל-12 משתנים בשביל להוריד את מימד הבעיה משמעותית תוך מתן אפשרות בחירה נוספת של מספר משתנים בהמשך.

להלן רשימת המשתנים שנבחרו :

Plans for Higher Education, School, Mothers Education, Weekly Study Time, Fathers Education, Alcohol on Workdays, Mothers Work, Alcohol on Weekends, Housing Type, Travel Time to School, School Choice Reason, Internet Availability.

Feature	Score
Plans for Higher Education	90.727338
School	58.551567
Mothers Education	53.640535
Weekly Study Time	45.407437
Fathers Education	36.888045
Alcohol on Workdays	26.489317
Mothers Work	19.097469
Alcohol on Weekends	18.391224
Housing Type	17.489191
Travel Time to School	17.401939
School Choice Reason	14.857098
Internet Availability	13.966213





איחוד קטגוריות של המשתנים שנבחרו

בכדי להכליל את בעיית החיזוי, נרצה לבחון האם ניתן לאחד קטגוריות של המשתנים המסבירים שנבחרו. לצורך כך ביצענו מבחן Tukey HSD הבודק את השוני בין כל הזוגות האפשריים של הקטגוריות. במידה ונקבל תוצאה מובהקת עבור זוג מסוים נאמר כי יש הבדל בין הקבוצות, אך אם אין הבדל ביניהם נשקול האם יש לאחד את הקטגוריות. בנוסף, איחדנו קטגוריות בהתאם למספר התצפיות מכל קטגוריה, כך שקטגוריה עם מספר תצפיות מועט אוחדה עם קטגוריה מתאימה בהתאם לסדר האורדינלי שלה. כמו כן, איחדנו על בסיס הקשר לוגי, וכאשר החציון ומידת הפיזור של טווח ערכי הקטגוריות דומים. יש לציין כי מרבית המשתנים הם קטגוריאליים אורדינליים, ולכן מרבית מאיחוד הקטגוריות נבע מהמיקום של הקטגוריה ביחס לסקאלת הערכים של המשתנה, כך שלרוב קטגוריה אוחדה עם קטגוריה הסמוכה לה.

להלן טבלה המסכמת את איחוד הקטגוריות של המשתנים המסבירים שנבחרו:

משתנה	קטגוריות לפני איחוד	קטגוריות אחרי איחוד	סיבת האיחוד
Plans for Higher Education	2 קטגוריות: כן, לא	ללא שינוי	-
School	2 קטגוריות: GP (Gabriel Pereira), MS (Mousinho da Silveira)	ללא שינוי	-
Mothers Education	5 קטגוריות: 0 - ללא השכלה 1 - בית ספר יסודי 2 - חטיבת ביניים 3 - בית ספר תיכון 4 - השכלה גבוהה	3 קטגוריות: 0 - השכלה נמוכה 1 - השכלת ביניים 2 - השכלה גבוהה	כמות תצפיות קטנה לקטגוריית המקור 0 (6 תצפיות בלבד), ולכן הוחלט לאחד אותה עם קטגוריה 1. כמו כן, הדימיון הרב של הפיזור והחציון של קטגוריות המקור 3 ו-4 גרם לאיחוד קטגוריות אלה.
Weekly Study Time	4 קטגוריות: 1 - עד שעתיים 2 - בין שעתיים לחמש שעות 3 - בין חמש לעשר שעות 4 - למעלה מעשר שעות	2 קטגוריות: 1 - עד שעתיים 2 - יותר משעתיים	מבחן Tukey HSD קבע באופן מובהק כי אין קשר בין קטגוריה 1 ליתר הקבוצות, ומנגד כי ישנו קשר בין יתר הקטגוריות. כמו כן, מידת הפיזור והחציון של יתר הקטגוריות דומה.
Fathers Education	5 קטגוריות: 0 - ללא השכלה 1 - בית ספר יסודי 2 - חטיבת ביניים 3 - בית ספר תיכון 4 - השכלה גבוהה	2 קטגוריות: 0 - השכלה נמוכה 1 - השכלה גבוהה	כמות תצפיות קטנה לקטגוריית המקור 0 (7 תצפיות בלבד), והדימיון בפיזור ובחציון בקטגוריות המקור 1 ו-2 הובילו לאיחוד שלוש קטגוריות אלה. בנוסף, הדימיון בפיזור ובחציון של קטגוריות המקור 3 ו-4 הוביל לאיחוד שתי קטגוריות אלה.



Alcohol on Workdays	5 קטגוריות : 0 - נמוך מאוד 1 - נמוך 2 - בינוני 3 - גבוה 4 - גבוה מאוד	2 קטגוריות : 0 - נמוך 1 - גבוה	הקטגוריות אוחדה על בסיס מידת הדימיון של הפיזור והחציון של הערכים. ניתן לחלק את הקטגוריות לשתי קטגוריות חדשות כך שכל אחת מהן מאוד דומה במדדים אלה. לכן, הוחלט לחלק לשתי קטגוריות מרכזיות : נמוך (המכיל את קטגוריות המקור 0 ו-1), וגבוה (המכיל את קטגוריות המקור 2,3,4).
Mothers Work	5 קטגוריות : 0 - מורה 1 - בריאות 2 - שירות 3 - בית (לא עובדת) 4 - אחר	2 קטגוריות : 0 - עובדת 1 - לא עובדת	מבחן Tukey HSD קבע באופן מובהק כי אין קשר בין קטגוריה 3 ליתר הקבוצות, ומנגד כי ישנו קשר בין יתר הקטגוריות. בנוסף, החציון והפיזור של יתר הקטגוריות מאוד דומה ביחס לקטגוריית המקור 3. גם בהקשר הלוגי, מצאנו לנכון לאחד לשתי קטגוריות חדשות המגדירות האם האמא העובדת או לא.
Alcohol on Weekends	5 קטגוריות : 0 - נמוך מאוד 1 - נמוך 2 - בינוני 3 - גבוה 4 - גבוה מאוד	2 קטגוריות : 0 - נמוך 1 - גבוה	הקטגוריות אוחדה על בסיס מידת הדימיון של הפיזור והחציון של הערכים. ניתן לחלק את הקטגוריות לשתי קטגוריות חדשות כך שכל אחת מהן מאוד דומה במדדים אלה. לכן, הוחלט לחלק לשתי קטגוריות מרכזיות : נמוך (המכיל את קטגוריות המקור 0,1,2), וגבוה (המכיל את קטגוריות המקור 3,4).
Housing Type	2 קטגוריות : U (urban) , R (rural)	ללא שינוי	-
Travel Time to School	4 קטגוריות : 0 - פחות מ-15 דקות 1 - בין 15 דקות לחצי שעה 2 - בין חצי שעה לשעה 3 - למעלה משעה	2 קטגוריות : 0 - פחות מחצי שעה 1 - יותר מחצי שעה	עקב מספר התצפיות המועט של קטגוריית המקור 3, ותרשים הפיזור של הנתונים הוחלט לחלק את המשתנה לשתי קטגוריות מרכזיות : קטגוריה ראשונה מאחדת את קטגוריות 0 ו-1, קטגוריה שנייה מאחדת את קטגוריות 2 ו-3.
School Choice Reason	4 קטגוריות : 0 - קרוב לבית 1 - מוניתין	3 קטגוריות : 0 - קרוב לבית ואחר 1 - מוניתין	ניתן לראות בתרשימי ההתפלגות כי החציון והפיזור של קטגוריית המקור 0 ו-3 הן מאוד דומות ולכן הוחלט לאחד



אותם. בנוסף, יש היגיון לוגי בהשארת קטגוריות המקור 1 ו-2 כפי שהן משום שהן אכן מייצגות סיבות שונות במהותן בבחירת בית הספר.	2 - קורס מועדף	2 - קורס מועדף 3 - אחר	
-	ללא שינוי	2 קטגוריות : 0 - אין 1 - יש	Internet Availability



מודל רגרסיה לינארית

לאחר בחירת המשתנים, ביצענו מודל רגרסיה לינארית לצורך חיזוי הציון של התלמידים. על מנת לעשות זאת, ביצענו רגרסיה לפני, לאחר ובצעדים על בסיס מדד AIC. הרצת אלגוריתמים אלו הובילה לתוצאות הבאות:

רגרסיה לאחר/בצעדים:

```
Call:
lm(formula = Grade ~ Weekly_Study_Time + Alcohol_Weekends + School +
    Housing_Type + Desire_Graduate_Education + Has_Internet +
    Housing_Type:Has_Internet, data = dataset)

Residuals:
    Min       1Q   Median       3Q      Max
-8.5024 -1.6167 -0.2431  1.4976  6.7053

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      9.0729     0.4834  18.768 < 2e-16 ***
Weekly_Study_TimeUp to 2h -0.5518     0.2129  -2.591 0.009775 **
Alcohol_WeekendsLow      0.9001     0.2408   3.738 0.000202 ***
SchoolMousinho da Silveira -1.1144     0.2194  -5.078 5.01e-07 ***
Housing_TypeUrban      -0.7289     0.3955  -1.843 0.065770 .
Desire_Graduate_EducationYes 2.5134     0.3172   7.924 1.02e-14 ***
Has_InternetYes      -0.3697     0.3613  -1.023 0.306495
Housing_TypeUrban:Has_InternetYes 1.4213     0.4677   3.039 0.002473 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.407 on 641 degrees of freedom
Multiple R-squared:  0.2326,    Adjusted R-squared:  0.2242
F-statistic: 27.76 on 7 and 641 DF,  p-value: < 2.2e-16
```

AIC: 2991.666

רגרסיה לפני: ערכי ה-AIC ו- R^2 adjusted:

Adjusted R-squared: 0.2406649 AIC: 3074.016

❖ לאחר התבוננות בתוצאות – נבחר את המודל לפי מדד BIC מינימאלי. לכן, נבחר את המודל שהתקבל ברגרסיה לאחר/בצעדים בעל ערך BIC של 3031.945.

המודל המתקבל:

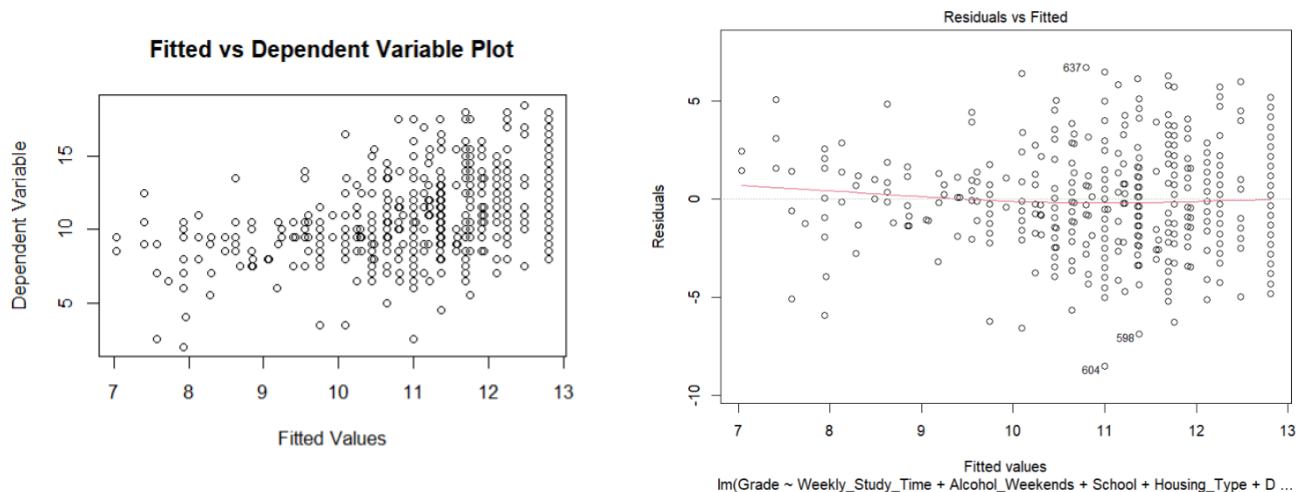
$$Y = 9.0729 - 0.5518 \cdot X_{14Up\ to\ 2h} + 0.9001 \cdot X_{26Low} - 1.1144 \cdot X_{1Mousinho\ da\ Silveira} \\ - 0.7289 \cdot X_{4Urban} + 2.5134 \cdot X_{20Yes} - 0.3697 \cdot X_{21Yes} \\ + 1.4213 \cdot X_{Urban} \cdot X_{21Yes}$$



כעת, נבחן האם הנחות המודל מתקיימות :

אי תלות בין תצפיות : סט הנתונים הנחקרים, מכיל רשומות שונות כך שכל אחת מייצגת תלמיד שונה בבית ספר תיכון, ולכן אין תלות בין התצפיות השונות.

הנחת הלינאריות : מתרשים השאריות ניתן לראות כי קו הרגרסיה קרוב יחסית לקו האפס. ניתן לראות זאת גם בפלט מבחן Chow כי הפונקציה מונוטונית. כמו כן, מתרשים הערכים החזויים וערכי המשתנה התלוי (ציונים) נראה דווקא כי הנחת הלינאריות לא מתקיימת, ככל הנראה בגלל קרבתו של ערך ה-pvalue ל-0.05.



M-fluctuation test

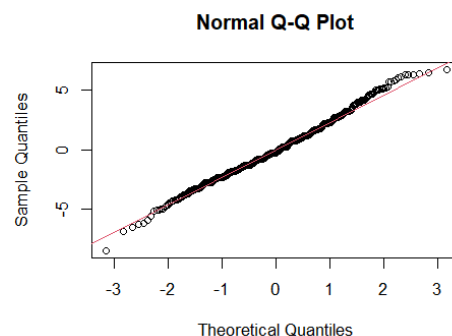
data: backward_model
f(efp) = 1.6047, p-value = 0.08912

הנחת שוויון השונות : על מנת לבדוק את קיום הנחה זו נבצע Breusch-Pagan test אשר בודק הומוסקדסטיות. מהפלט שהתקבל ניתן לראות כי התוצאה מאוד מובהקת, ולכן ניתן לומר כי הנחת שוויון השונות לא מתקיימת במודל זה (ניתן לראות זאת גם בתרשים השאריות).

studentized Breusch-Pagan test

data: backward_model
BP = 38.955, df = 7, p-value = 1.994e-06

הנחת הנורמליות : ניתן לראות מתוך התרשים כי הנחת הנורמליות מתקיימת. עם זאת, נבדוק את קיומה של הנחה זו גם באמצעות מבחן שפירו-וילקס. מפלט המבחן נראה כי ערכו של p-value קטן מ-0.05, ולכן ניתן לומר כי הנחת הנורמליות אינה מתקיימת במודל זה.



Shapiro-wilk normality test

data: residuals
W = 0.99392, p-value = 0.0104



עד כה, ניתן לראות כי המודל שנבחן אינו מסביר את המשתנה המוסבר בצורה טובה (ערך ה- $R^2 adjusted$ נמוך) וכן ההנחות מתקיימות, גם הנחת הלינאריות שקשה לקבוע את קיומה מהפליטים, מה שאף מחזק את ההבנה כי מודל הרגרסיה הלינארית איננו מתאים לבעיה הנבדקת. ניתן גם לחשוב על כך כי הניסיון לבצע חיזוי למשתנה רציף על סמך מספר מועט של משתנים קטגוריאליים (לא רציפים) אשר ברובם בינאריים (בעלי שתי קטגוריות) מקשה על ביצועי המודל.

על כן, הגענו למסקנה כי על המשתנה המוסבר לעבור דיסקרטיזציה אשר לא תפגע ביכולות שלנו לענות על שאלת המחקר. לכן, החלטנו להמיר את המשתנה המוסבר, ציון התלמיד, לבינארי כך שהקטגוריות שלו מוגדרות כ-*"Fail"* ו-*"Pass"*, והחלטנו לבצע בחירת משתנים מחדש עקב שינוי המשתנה המוסבר תוך יישום של מודל רגרסיה לוגיסטית ומודלי למידת מכונה נוספים.



מודלי סיווג

כאמור, בסיס הנתונים מכיל משתנה מוסבר רציף המייצג את הציון הממוצע השנתי של התלמיד. ניתן ללמוד מציון זה האם התלמיד עבר בהצלחה את שנת הלימודים, כך שניתן לחלק את רצף הערכים של המשתנה לשתי מחלקות עיקריות בהתאם לציון "עובר" הנהוג בפורטוגל (ציון 10): 0 - ציון ממוצע נכשל (עד ציון ממוצע 10), 1 - ציון ממוצע עובר (החל מציון ממוצע 10 עד 20). החלטנו לבצע שלושה מודלים של למידת מכונה לצורך הסיווג למחלקות, כך שבהינתן המאפיינים של התלמיד נוכל לחזות מראש האם יעבור או לא יעבור את שנת הלימודים. יש לציין כי הבעיה אינה מאוזנת – מעל 70% מהתצפיות מסווגות למחלקה 1 (ציון עובר). לכן, בעת בחינת ההצלחה של המודל על סט הבחינה נשתמש במדד ה-f1 המתחשב במדד בדיוק (precision) ובמדד ההיזכרות (recall).

גם בעיית סיווג זו בעלת מימד גבוה (מספר גבוה של משתנים מסבירים), ולכן בחרנו לבצע בחירת משתנים משמעותיים (Feature Selection) בכדי לצמצם את המימד ולפשט את הבעיה. כשלב מקדים, המרנו את המשתנים הקטגוריאליים למשתנה דמה (dummy variables) באמצעות הוספת משתנים עבור כל אחד מערכי הקטגוריות. בשיטה זו, המשתנה החדש קיבל את הערך 1 במידה ולתצפיות יש את ערך הקטגוריה המסוימת, ואת הערך 0 במידה ולא (One-Hot Vector). המרה זו הינה המרה מחייבת בטרם ביצוע המודלים עליהם נפרט בהמשך.

לאחר מכן, בחנו את מידת המידע ההדדי (mutual information) בין כל משתנה דמה למשתנה המוסבר. המידע ההדדי מודד את התלות בין כל מאפיין (feature) למשתנה המטרה, כך שמידע הדדי גבוה מצביע על קשר חזק ביניהם דרך ההסתברות המשותפת שלהם, ולכן ניתן להסיק כי יש הפרדה טובה יותר באמצעות משתנה זה ביחס לקטגוריות של המשתנה המוסבר.

המידע ההדדי MI בין שני משתנים X, Y מחושב לפי הנוסחה הבאה:

$$MI(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \left(\frac{p(x, y)}{p(x) \cdot p(y)} \right)$$

בתהליך זה נמצאו 12 המשתנים המשמעותיים ביותר (מסיבה דומה לבחירת המשתנים ברגרסיה לינארית), אותם נספק כפיצ'רים למודלים הבאים לצורך סיווג התלמידים ל"נכשל" ול"עובר":

Desire Graduate Education,
Time with Friends, Weekly
Study Time, Weekly Study
Time, School, Private
Tutoring, Commute Time,
Fathers Education, Mothers
Education, Free Time After
School, Desire Graduate
Education, Free Time After
School.

```
Sorted features by importance:
Desire Graduate Education_No: 0.05906658526156994
Time with Friends_High: 0.042371287499056276
Weekly Study Time_5 to 10h: 0.04161315012915612
School_Gabriel Pereira: 0.04150125557830342
Weekly Study Time_2 to 5h: 0.04104404025313002
School_Mousinho da Silveira: 0.0384645038747049
Private Tutoring_Yes: 0.03692160530380284
Commute Time_15 to 30 min: 0.03599535294273215
Fathers Education_Primary School: 0.03450379080177379
Mothers Education_Higher Education: 0.03195350175964329
Free Time After School_Low: 0.03057459330775414
Desire Graduate Education_Yes: 0.030106831532266343
Free Time After School_Very High: 0.029880317087207287
```



מודל רגרסיה לוגיסטית

משום שהחלטנו לבצע דיסקרטיזציה של שתי קטגוריות למשתנה המוסבר נרצה כעת למצוא את המשתנים אשר מסבירים אותו בצורה הטובה ביותר מתוך 12 המשתנים אשר נבחרו בשלב ה-feature selection הקודם. משום שמדובר כעת במשימת סיווג בחרנו לבצע רגרסיה לוגיסטית – נשתמש ברגרסיה לפנים, לאחר ובצעדים כדי למצוא את המודל הטוב ביותר. בחרנו את המודל הלוגיסטי לפי ערך ה-AIC הנמוך ביותר – בלפנים/לאחור/בצעדים התקבל אותו ערך AIC. פלט ה-summary שהתקבל מהמודל:

```
Call:
glm(formula = Grade ~ Desire_Graduate_Education + School, family = "binomial",
     data = train_data)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.2141     0.3471  -0.617   0.537
Desire_Graduate_EducationYes  2.1509     0.3616   5.947 2.72e-09 ***
SchoolMousinho da Silveira  -1.4797     0.2386  -6.202 5.59e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 525.01  on 454  degrees of freedom
Residual deviance: 439.99  on 452  degrees of freedom
AIC: 445.99

Number of Fisher Scoring iterations: 4
```

המודל המתקבל:

$$P = \frac{1}{1 + e^{-(0.2141 + 2.1509X_{20Y_{es}} - 1.4797X_{1Mousinho da Silveira})}}$$

$$Y = \frac{1}{1 + e^{-\beta x}}$$

משום שהבעיה אינה מאוזנת (שכן רק 30% נכשלו מתוך כלל הסטודנטים), בחרנו למדוד את טיב המודל באמצעות מדד ה-F1 עם 0.8 threshold (סף זה נבחר בצורה היוריסטית על ידי מקסום ערך ה-recall, בחרנו ערכי threshold בטווח של 0.6-0.9 בקפיצות של 0.01 ובחרנו את ה-threshold הנותן את המדד הגבוהה ביותר), בכדי להגביר את הרגישות של המודל (recall) תוך צמצום תוצאות ה-false negative ("מענישים" את העוברים). לאחר אימון המודל בדקנו את ערך מדד זה על סט הבחינה והערך שהתקבל הוא 0.5873. משום שמדובר בבעיית סיווג בינארית וערך המדד קרוב ל-0.5, ניתן לומר כי מודל זה אינו מוסיף מידע מהותי לצורך סיווג התלמידים (שקול להטלת מטבע).



מודלי למידת מכונה (Machine Learning Classification)

כעת נרצה לבחון מודלים אחרים לבעיות סיווג על סט הנתונים שלנו ולבחור בסופו של דבר את המודל המתאים ביותר (זה אשר ייתן את ה-F1 score הגבוה ביותר).

Random Forest

בשונה מעץ החלטה פשוט (Decision Tree) מודל זה מייצר יותר מעץ אחד בודד. כאשר כל אחד מהעצים מחושב באופן בלתי תלוי ליתר העצים. בכדי לכוון את ההיפר-פרמטרים של המודל השתמשנו ב-GridSearch עם הפרמטרים המצורפים מטה:

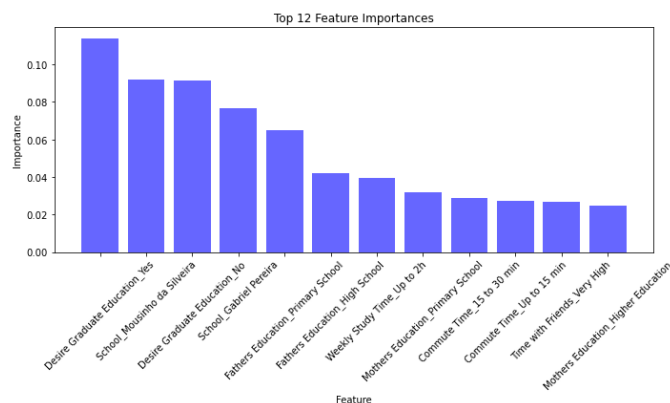
```
param_grid = {
    'n_estimators': [50, 100, 200],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [2, 4, 6]
}
```

- n_estimators : מספר העצים ביער (במודל).
- max_depth : העומק המקסימלי של כל אחד מהעצים.
- min_samples_split : מספר מינימלי של תצפיות בכדי לבצע פיצול בעץ (יצירה של צומת חדש).
- min_samples_leaf : מספר מינימלי של תצפיות בכל עלה בעץ.

העצים המרכיבים את היער עצמאיים לחלוטין ואינם תלויים זה בזה, כך שהנתונים המגדירים את הצמתים וההסתעפויות בעצים הינם אקראיים בכל אחד מהעצים. בנוסף, מספר התצפיות ששימשו לבניית המודל מספק, וללא תצפיות חריגות כך שבעזרת המודל ניתן יהיה לבצע תחזיות של תלמידים נוספים. בנוסף, בסיס הנתונים אינו מכיל נתונים חסרים ולכן ניתן להריץ את המודל ללא שגיאות. לאחר בניית המודל ובחינתו על סט בחינה בלתי תלוי לסט האימון התקבלו התוצאות הבאות:

```
Best Parameters: {'max_depth': None, 'min_samples_leaf': 2,
'min_samples_split': 10, 'n_estimators': 200}
F1: 0.864516129032258
```

הגרף הבא מציג את חשיבות המשתנים כפי שנקבע עם שימוש במודל זה. חשיבות המשתנים נקבעה לפי מדד mean decrease impurity:





Support Vector Machine (SVM)

מודל ה-SVM ידוע כמודל שיכול להתמודד עם בעיות סיווג כאשר מימד הבעיה הוא גבוה (מספר רב של משתנים מסבירים), ולכן בחרנו לבחון את טיב המודל בבעיה שלנו. בכדי לכוון את ההיפר-פרמטרים של המודל השתמשנו ב-GridSearch.

```
param_grid = {
    'C': [0.1, 1], # Regularization parameter
    'gamma': [0.1, 1], # Kernel coefficient
    'kernel': ['linear', 'poly'] # Kernel type
}
```

- C: שולט על הטרייד-אוף בין מיקסוס המרווח בין המחלקות לבין מזעור שגיאת הסיווג.
- gamma: קובע את השפעת המרחק של כל תצפית בסט האימון על עקומת ההפרדה בין המחלקות.
- kernel: סוג פונקציית הליבה המבצעת טרנספורמציה למרחב עם מימד גבוה יותר.

יש לציין כי בטרם בניית במודל בדקנו שאכן כל הנחות המודל מתקיימות. מימד הבעיה שלנו הוא גבוה (12 משתנים מסבירים) כך מודל זה מתאים כך שיכול למקסם את המרווח בין גבול ההפרדה לכל אחת מהמחלקות ובכך ליצור גבולות החלטה מוכללים לסט הבחינה. אנו יישמנו טכניקות עיבוד מקדים להפחתת רעשים וחריגים, ובכך מבטיחים את היציבות והאמינות של המודל. בנוסף, אנו מיישמים פונקציות שאינן ליניאריות בכדי לשפר את הגמישות והביצועים של המודל על פני מערכי נתונים מגוונים.

לאחר בניית המודל ובחינתו על סט בחינה בלתי תלוי לסט האימון התקבלו התוצאות הבאות:

```
Best Parameters: {'C': 0.1, 'gamma': 0.1, 'kernel': 'linear'}
f1: 0.8
```



Neural Network

מודל זה בנוי בהשראת המבנה של הנוירונים במוח האנושי, ומטרתו לקשור בין מספר רב של משתנים לטובת ביצוע משימות סיווג מורכבות על בסיס למידת ערכי המשקלים בקשתות השונות ברשת. בכדי לכוון את ההיפר-פרמטרים של המודל השתמשנו ב-GridSearch.

```
param_grid = {
    'hidden_layer_sizes': [(50,50),(100,100),(150,150)], # Number of neurons in hidden layers
    'alpha': [0.0001, 0.001, 0.01], # Regularization parameter
    'learning_rate_init': [0.001, 0.01, 0.1] # Initial learning rate
}
```

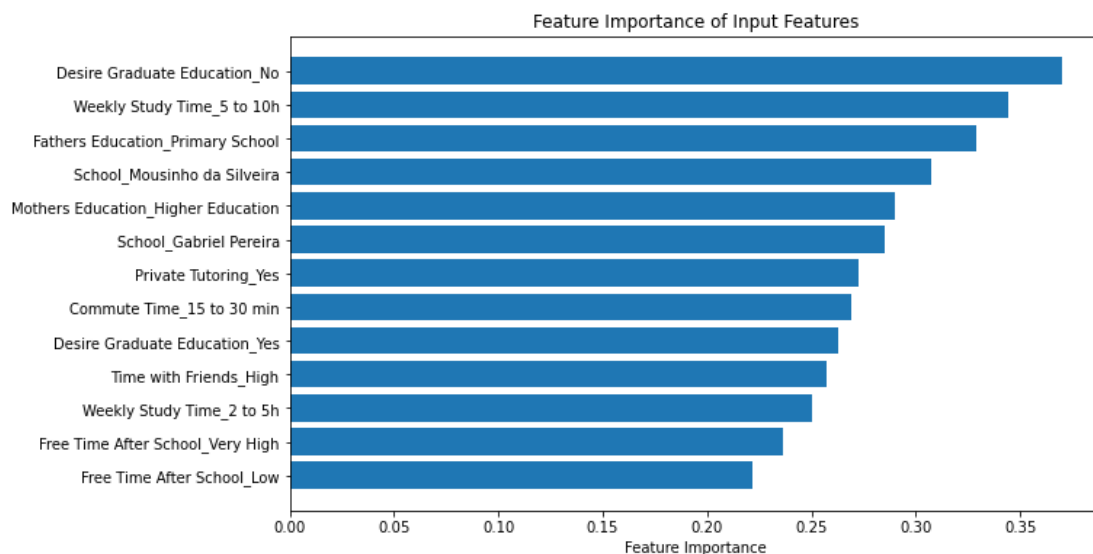
- hidden_Layer_sizes : מספר וגודל השכבות החבויות.
- alpha : מקדם הלמידה של הרשת (בין שתי איטרציות של למידת המשקלים).
- Learning_rate_init : מקדם הלמידה ההתחלתי של הרשת.

יש לציין כי בטרם בניית במודל בדקנו שאכן כל הנחות המודל מתקיימות. רשת הנוירונים דורשת כמות גדולה של נתונים לצורך ביצוע הלמידה המשקפים את הבעיה. במקרה שלנו יש מספר גבוה של תצפיות המספק את בניית הרשת. כמו כן, יש לבצע נורמליזציה של ערכי המשתנים לסולם אחיד בכדי לשמור על סטנדרטיזציה. משום שכל המשתנים שלנו הם משתני דמה (one-hot vector) הסולם הוא אחיד וללא הטעיה. כמו כן, אין בבסיס הנתונים נתונים חסרים או ערכים חריגים, כך שהאלגוריתם יכול לבצע ללא שגיאות.

לאחר בניית המודל ובחינתו על סט בחינה בלתי תלוי לסט האימון התקבלו התוצאות הבאות:

```
Best Parameters: {'alpha': 0.0001, 'hidden_layer_sizes': (100,),
'learning_rate_init': 0.1}
Accuracy: 0.7948717948717948
```

הגרף הבא מציג את חשיבות המשתנים כפי שנקבע עם שימוש במודל זה. חשיבות המשתנים נקבעה לפי מדד mean weight magnitudes:





סיכום תוצאות ערכי ההיפר-פרמטרים אשר כיוונו במודלים השונים

מודל	ערכי הפרמטרים של המודל הנבחר
Random Forest	max_depth: None min_samples_leaf: 2 min_samples_split: 10 n_estimators: 200
SVM	c: 0.1 gamma: 0.1 kernel: linear
Neural Network	alpha: 0.0001 hidden_layer_sizes: (100,100) learning_rate_init: 0.1
Logistic Regression	Threshold: 0.8



סיכום ומסקנות

בעת ביצוע הפרויקט ביצענו מודלים שונים בכדי לחזות את הציון השנתי הממוצע של תלמיד על בסיס מאפייניו האישיים. למרות שמשנתה המוסבר הוא רציף, ההנחות של רגרסיה לינארית לא התקיימו ומדד ה- R_{adj} שהתקבל הוא נמוך אשר העיד על חוסר התאמה של המודל לבעיה. עקב כך, החלטנו להמיר את הבעיה לבעיית סיווג בינארית (שתי מחלקות; 0 - ציון לא עובר, 1 - ציון עובר). לאחר מכן, ביצענו מודל רגרסיה לוגיסטית, שגם קיבל ערך מדד F1 נמוך.

בעקבות תוצאות נמוכות של מודלים אלו, החלטנו להשתמש במודלים נוספים המסוגלים להתמודד עם מספר רב של משתנים, תוך כיוונון מיטבי של ההיפר-פרמטרים המגדירים את המודל. החלטנו לבצע שלושה מודלי סיווג ובכל אחד מהם בדקנו את תוצאותיו על סט בחינה שאינו תלוי בסט אימון המודל. משום שהבעיה שלנו אינה מאוזנת (חלוקה שאינה שווה בין שתי המחלקות) החלטנו להשתמש במדד ה-f1 המתאים לבעיות מסוג זה.

להלן טבלה המסכמת את התוצאות של חמשת המודלים:

Results	Model
$R_{Adj}^2 = 0.24$	Linear regression
f1 = 0.587	Logistic regression
f1 = 0.864	Random Forest
f1 = 0.8	SVM
f1 = 0.794	Neural Network

התוצאות של שלושת המודלים הנוספים היו טובות משמעותית מתוצאות מודלי הרגרסיה. המודל שקיבל את התוצאה הטובה ביותר הוא Random forest (0.864). מודל זה יוצר עצי החלטה שאינם תלויים אחד באחד בשני ובכך מגביר את האקראיות של בניית המודל. כמו כן, מודל זה מתאים למשתנים קטגוריאליים ויחסית אינטואיטיבי להבנה.

בנוסף, המודלים הצביעו על המשתנים המשמעותיים ביותר מבין המשתנים הקיימים. המשתנה שנמצא כמשמעותי במודל הנבחר הוא Desire Graduate Education המייצג את מידת הרצון התלמיד בלימודים אקדמאיים, כך שרוב התלמידים המעידים כמעוניינים בלימודים גבוהים יקבלו ציונים טובים יותר מאשר יתר התלמידים. לכן, משתנה זה מספק הפרדה טובה ביחס למשתנה המוסבר.

לסיכום, המודלים שיצרנו מצליחים לנבא במידה מסוימת את הצלחתו של תלמיד תיכון על בסיס מאפייניו האישיים. המודל שאנו ממליצים להשתמש בו לצורך חיזוי הצלחת התלמיד הוא Random forest, כך שניתן לסווג את התלמיד כתלמיד שיסיים את לימודיו בהצלחה או בכישלון. כמו כן, ייתכן כי במידה והיו ברשותנו משתנים נוספים ומדגם גדול יותר היינו מצליחים להמשיך ולשפר את המודל.



נספחים

פלטי מבחן Tukey HSD, box-plot, Bar-Chart לצורך איחוד קטגוריות:
Mother Education

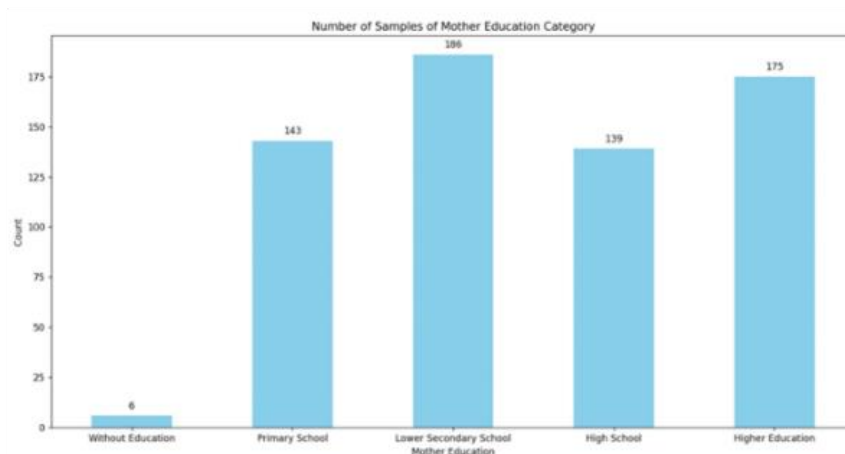
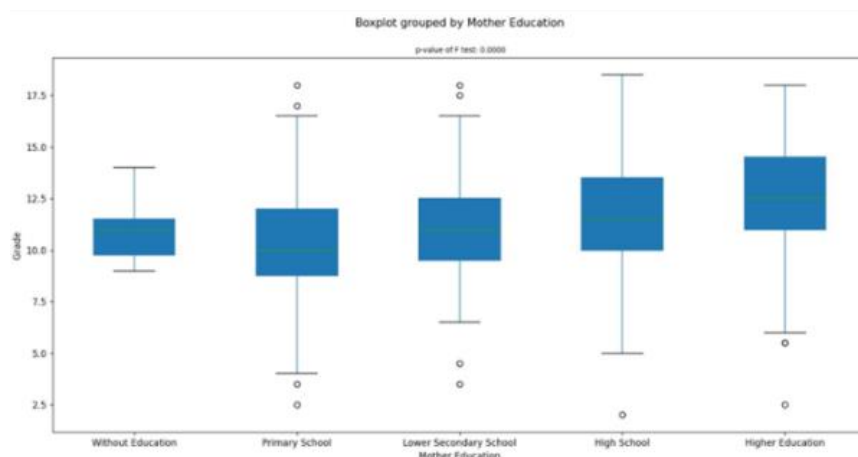
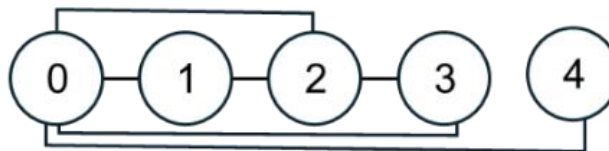
Mother Education

Tukey's HSD results for Mother Education:
Multiple Comparison of Means - Tukey HSD, FWER=0.05

```
=====
```

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	-0.5385	0.9881	-3.5344	2.4574	False
0	2	0.2446	0.9994	-2.7373	3.2265	False
0	3	0.4424	0.9944	-2.5552	3.4401	False
0	4	1.6257	0.5694	-1.3591	4.6106	False
1	2	0.7831	0.0582	-0.0165	1.5826	False
1	3	0.9809	0.0155	0.1246	1.8372	True
1	4	2.1642	0.0	1.3538	2.9746	True
2	3	0.1978	0.9625	-0.6082	1.0039	False
2	4	1.3811	0.0	0.624	2.1382	True
3	4	1.1833	0.0008	0.3665	2.0001	True

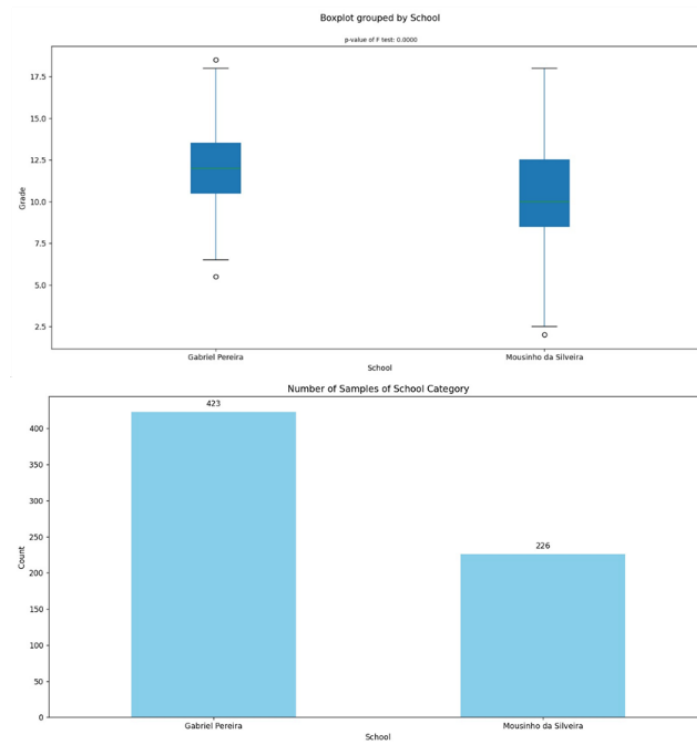
=====





: School

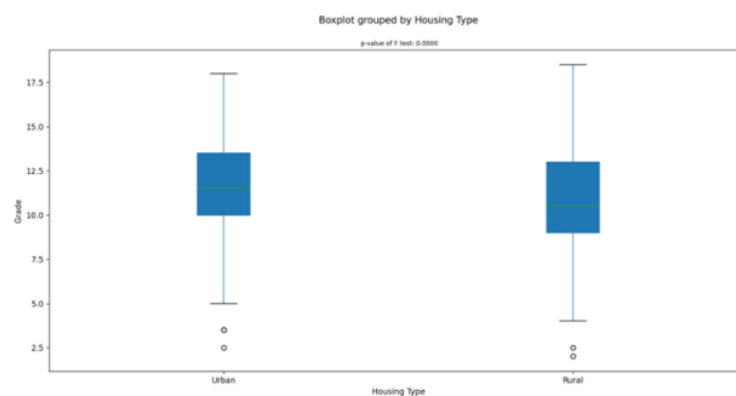
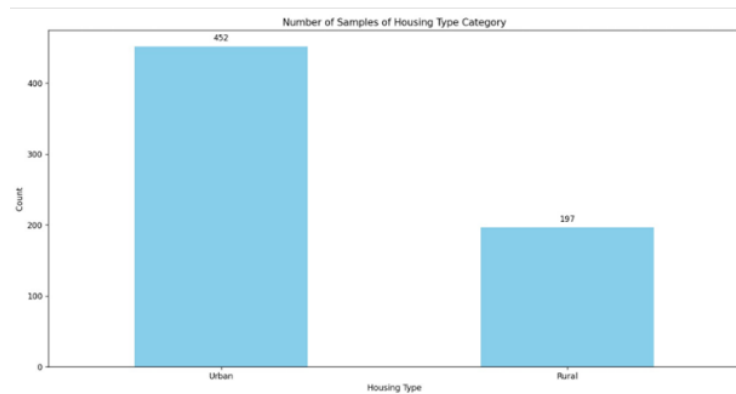
```
School
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====
group1 group2 meandiff p-adj lower upper reject
-----
0      1      -1.675   0.0  -2.1049 -1.2452  True
```





:Housing Type

```
Housing Type
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====
group1 group2 meandiff p-adj lower upper reject
-----
0      1      -0.9751  0.0  -1.433 -0.5173  True
```





: Father Education

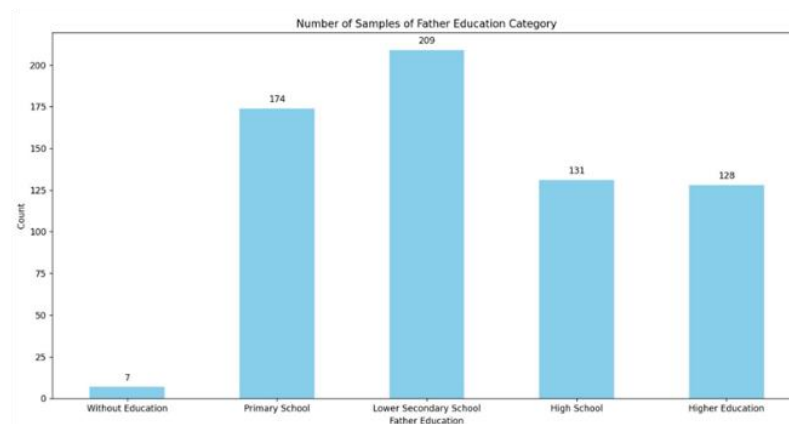
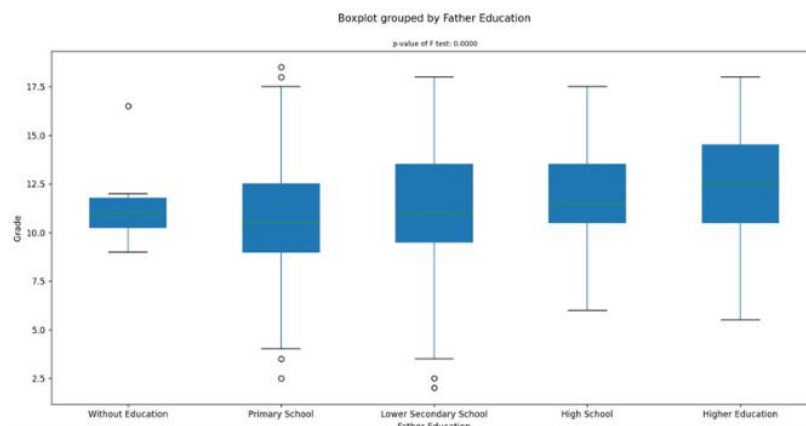
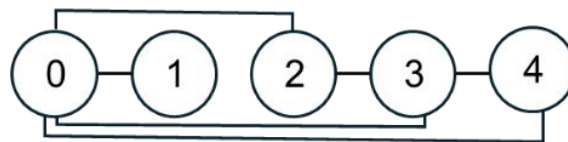
Father Education

Tukey's HSD results for Father Education:
Multiple Comparison of Means - Tukey HSD, FWER=0.05

```
=====
```

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	-0.8908	0.9086	-3.6989	1.9173	False
0	2	-0.11	1.0	-2.9091	2.689	False
0	3	0.3626	0.9967	-2.4633	3.1885	False
0	4	0.9414	0.8928	-1.8862	3.769	False
1	2	0.7808	0.0356	0.0332	1.5283	True
1	3	1.2534	0.0005	0.4108	2.096	True
1	4	1.8322	0.0	0.984	2.6805	True
2	3	0.4726	0.5026	-0.3391	1.2844	False
2	4	1.0515	0.0042	0.2339	1.869	True
3	4	0.5788	0.4047	-0.3265	1.4841	False

```
=====
```

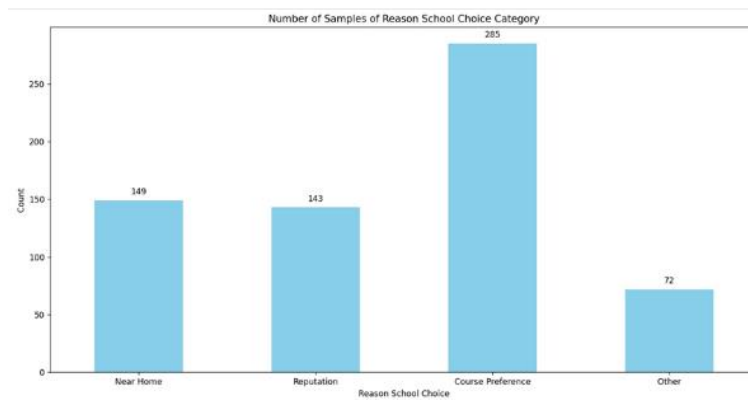
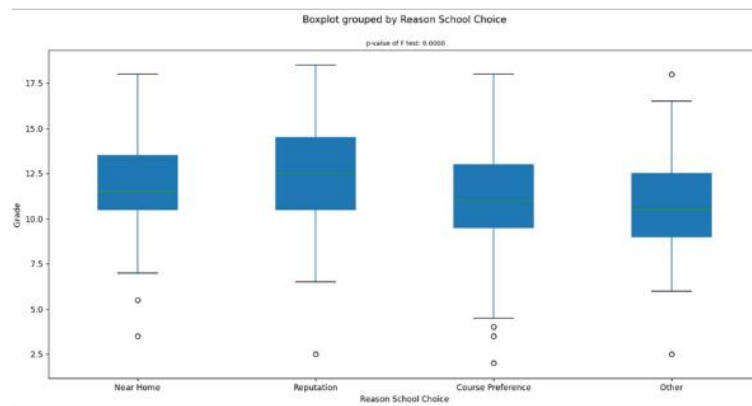
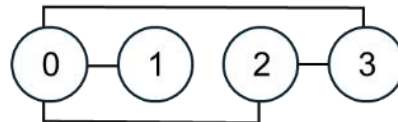




: School Choice Reason

School Choice Reason
Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	0.7359	0.0933	-0.0792	1.551	False
0	2	-0.6359	0.0933	-1.3402	0.0684	False
0	3	-0.9862	0.0582	-1.9953	0.0228	False
1	2	-1.3718	0.0	-2.0859	-0.6578	True
1	3	-1.7222	0.0001	-2.7381	-0.7063	True
2	3	-0.3504	0.766	-1.2797	0.579	False

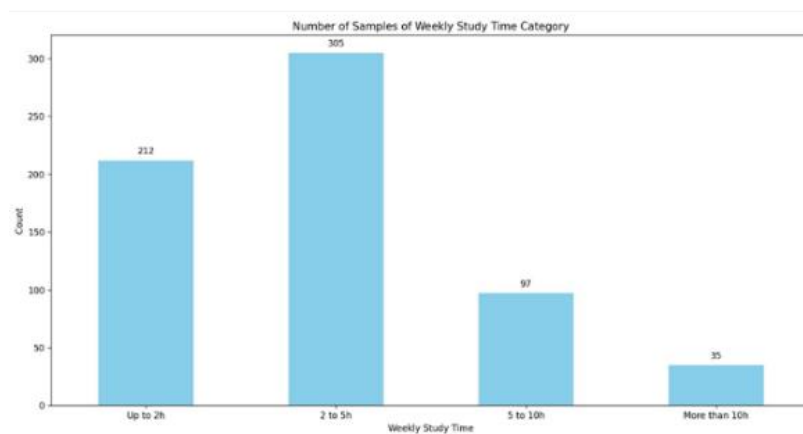
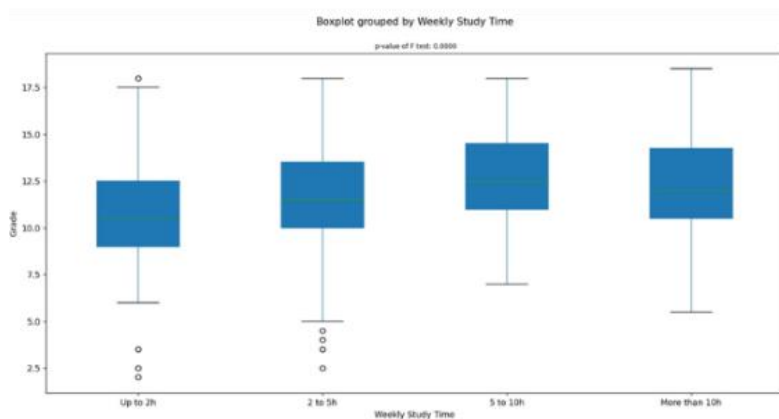
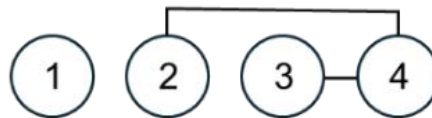




: Weekly Study Time

Weekly Study Time
Tukey's HSD results for Weekly Study Time:
Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
1	2	0.9936	0.0002	0.3754	1.6118	True
1	3	1.9913	0.0	1.1465	2.8361	True
1	4	2.1329	0.0001	0.8679	3.3978	True
2	3	0.9977	0.0076	0.1973	1.7981	True
2	4	1.1393	0.083	-0.0964	2.375	False
3	4	0.1415	0.9933	-1.2217	1.5048	False

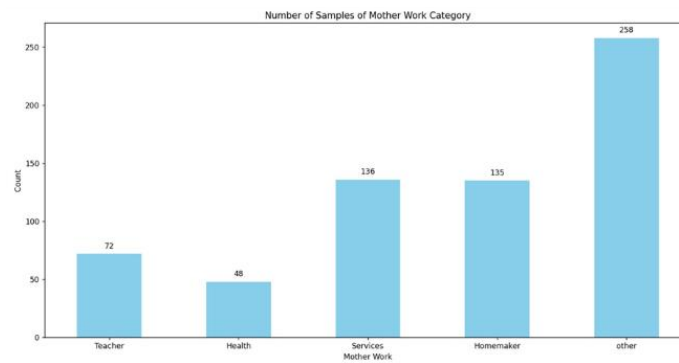
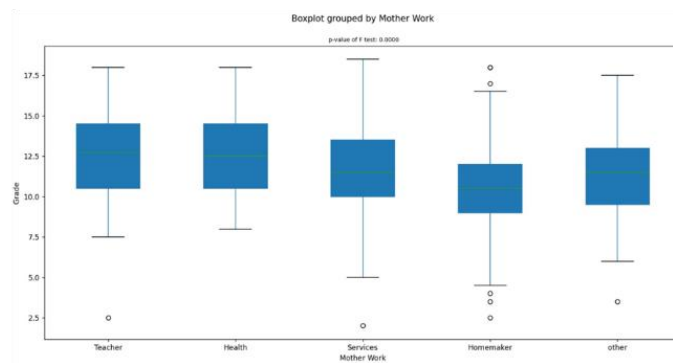
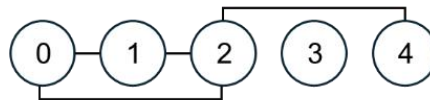




: Mothers Work

Mothers Work
Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	-0.12	0.9993	-1.4868	1.2467	False
0	2	-0.9754	0.0938	-2.0463	0.0955	False
0	3	-2.0874	0.0	-3.1653	-1.0095	True
0	4	-1.2863	0.0034	-2.2694	-0.3031	True
1	2	-0.8554	0.3153	-2.0833	0.3726	False
1	3	-1.9674	0.0001	-3.2014	-0.7333	True
1	4	-1.1663	0.0456	-2.3185	-0.014	True
2	3	-1.112	0.0065	-2.0074	-0.2166	True
2	4	-0.3109	0.8107	-1.0896	0.4679	False
3	4	0.8011	0.0443	0.0128	1.5895	True

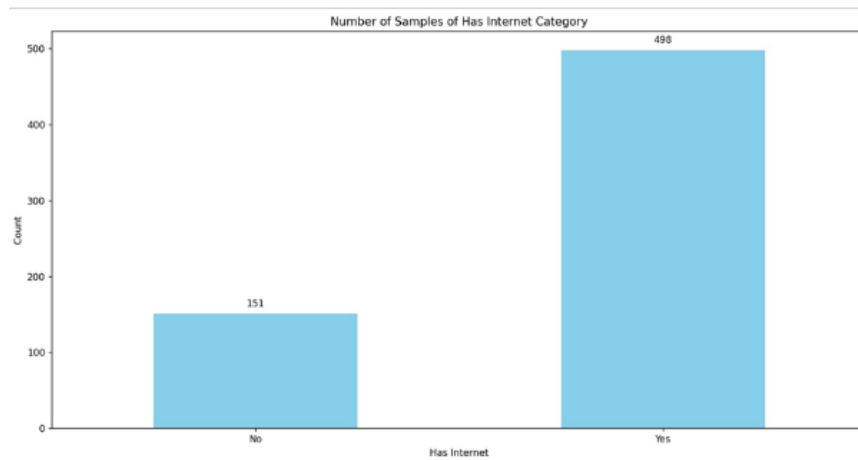
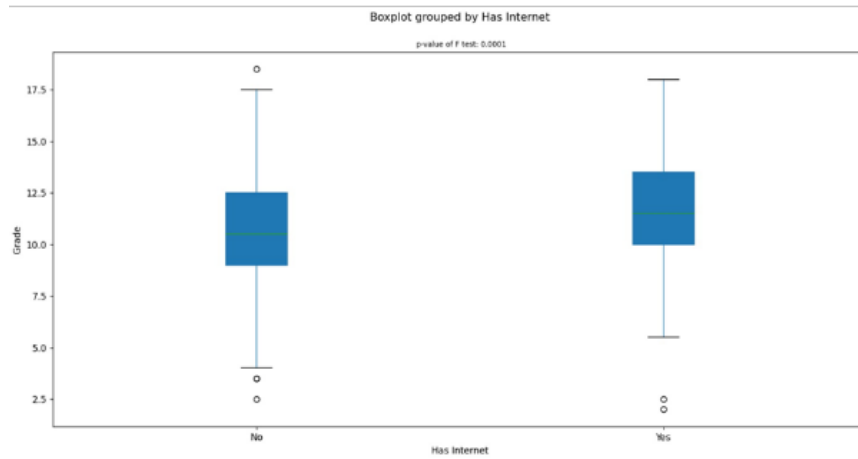




: Internet Availability

```
Internet Availability
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====
group1 group2 meandiff p-adj lower upper reject
-----
0      1      0.9567 0.0002 0.454 1.4594  True
```

0 1

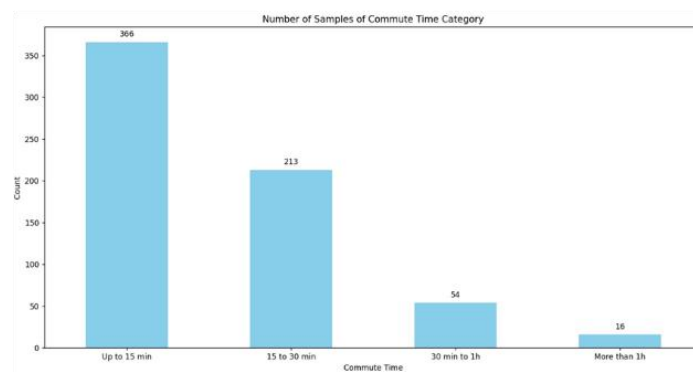
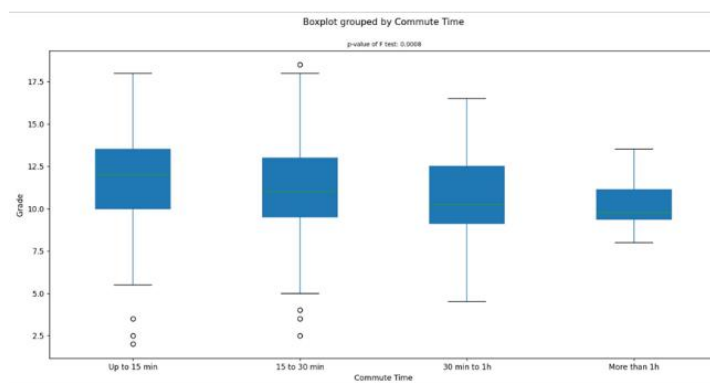
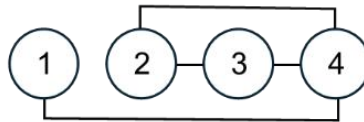




: Travel Time to School

Travel Time to School
Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
1	2	-0.6771	0.0228	-1.2876	-0.0667	True
1	3	-1.1354	0.0236	-2.1628	-0.108	True
1	4	-1.718	0.0641	-3.5028	0.0669	False
2	3	-0.4583	0.6918	-1.535	0.6184	False
2	4	-1.0409	0.4513	-2.8545	0.7728	False
3	4	-0.5825	0.8754	-2.5756	1.4105	False





: Plans for Higher Education

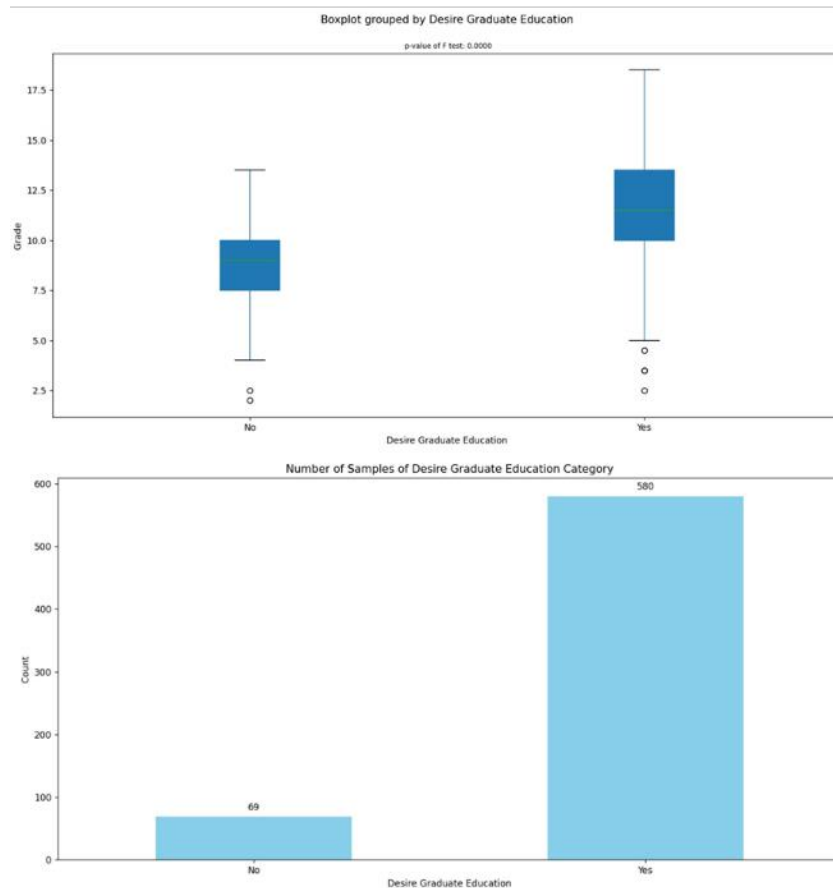
Plans for Higher Education
Multiple Comparison of Means - Tukey HSD, FWER=0.05

```
=====
```

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	3.1387	0.0	2.4916	3.7857	True

```
=====
```

0 1

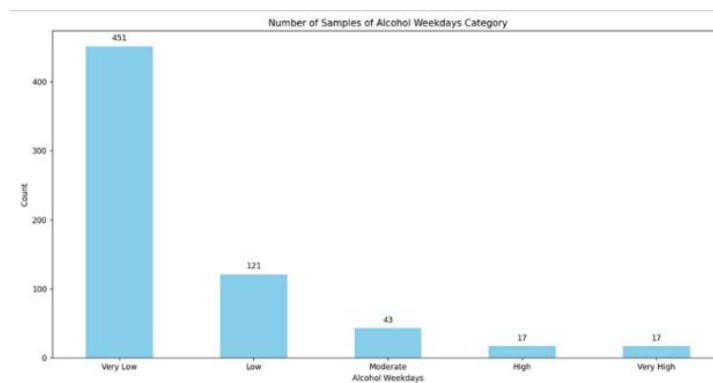
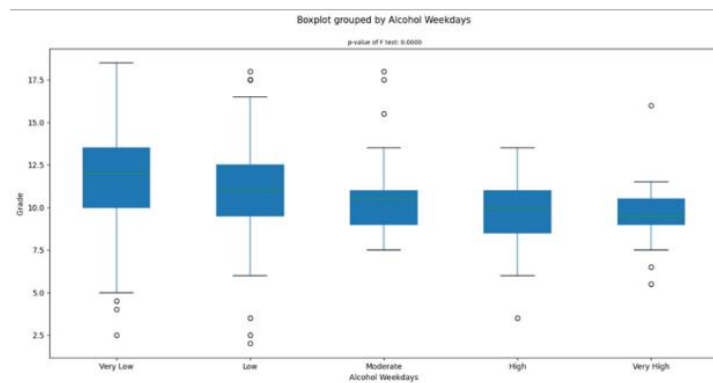
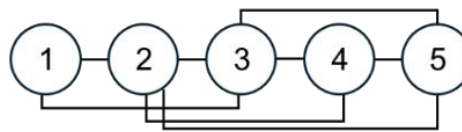




: Alcohol on Workdays

Alcohol on Workdays
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====

group1	group2	meandiff	p-adj	lower	upper	reject
1	2	-0.7111	0.0802	-1.4724	0.0503	False
1	3	-1.1767	0.0544	-2.3671	0.0136	False
1	4	-2.1018	0.0144	-3.9238	-0.2798	True
1	5	-2.0136	0.0218	-3.8356	-0.1916	True
2	3	-0.4657	0.8717	-1.7888	0.8574	False
2	4	-1.3908	0.2718	-3.3022	0.5207	False
2	5	-1.3025	0.3379	-3.2139	0.6089	False
3	4	-0.9251	0.7548	-3.0442	1.1941	False
3	5	-0.8368	0.8167	-2.956	1.2823	False
4	5	0.0882	1.0	-2.4403	2.6168	False





: Alcohol on Weekends

Alcohol on Weekends
Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
1	2	-0.0537	0.9997	-0.8272	0.7198	False
1	3	-0.4566	0.5613	-1.2879	0.3748	False
1	4	-1.1529	0.0068	-2.0847	-0.2211	True
1	5	-1.4185	0.0128	-2.6341	-0.2029	True
2	3	-0.4029	0.7457	-1.3139	0.5081	False
2	4	-1.0992	0.0237	-2.1027	-0.0957	True
2	5	-1.3648	0.0283	-2.6362	-0.0934	True
3	4	-0.6963	0.365	-1.7451	0.3525	False
3	5	-0.9619	0.261	-2.2693	0.3455	False
4	5	-0.2656	0.9844	-1.6391	1.1079	False

