|  |
| --- |
| Photo displaying partial image of two pie charts on a canvas-textured page |
| **تخمین وضعیت تحصیلی دانش آموزان**  **رامتین صمدی - سید محمد رضا موسوی تبار**  پروژه درس مبانی داده کاوی |
| |  |  |  | | --- | --- | --- | | مهر 402 |  | استاد الهام محمدی | |

**فهرست مطالب**

[**طرح پروژه 2**](#_Toc156775276)

[**مدل مفهومی 3**](#_Toc156775277)

[**شناخت داده 3**](#_Toc156775278)

[**گزارش اجرایی 7**](#_Toc156775279)

[**تقسیم بندی داده ها 9**](#_Toc156775279)

[**مدل سازی 10**](#_Toc156775280)

[**آنالیز و آزمون مدل ها 15**](#_Toc156775282)

[**پیشنهادات 17**](#_Toc156775283)

# طرح پروژه

**شناخت کسب و کار:** آموزش در مقاطع اولیه و سنین پایین از اهمیت ویژه ای برخوردار است. این دوران، دوران شکل گیری شخصیت و هویت فرد است و تجربیات این دوران تأثیر عمیقی بر آینده فرد خواهد داشت. کیفیت آموزش در این مقاطع می تواند تأثیر مستقیمی بر عملکرد و کیفیت زندگی دانش آموزان نوجوان و در مراتب بعدی جامعه داشته باشد.

**عوامل تأثیرگذار** بر عملکرد و کیفیت زندگی دانش آموزان نوجوان:

* **کیفیت آموزش**:  تأثیر مستقیمی بر عملکرد تحصیلی و یادگیری دانش آموزان دارد. آموزش باکیفیت می تواند به دانش آموزان کمک کند تا مهارت های لازم برای موفقیت در تحصیل و زندگی را کسب کنند.
* **محیط خانوادگی**:  تأثیر زیادی بر عملکرد و کیفیت زندگی دانش آموزان نوجوان دارد. خانواده هایی که از نظر عاطفی و حمایتی قوی هستند، می توانند به دانش آموزان کمک کنند تا در مدرسه و زندگی موفق شوند.
* **عوامل فردی**: عوامل فردی نیز می توانند بر عملکرد و کیفیت زندگی دانش آموزان نوجوان تأثیر بگذارند. عواملی مانند هوش، انگیزه، و سلامت جسمی و روانی می توانند بر عملکرد تحصیلی و کیفیت زندگی دانش آموزان تأثیر بگذارند.

**اهداف پروژه:** در این پروژه میخواهیم **عوامل موثر بر پیشرفت تحصیلی** دانش آموزان را شناسایی کنیم و به بهبود آنها کمک کنیم و همچنین عواملی که باعث افت تحصیلی دانش آموزان **و هدر رفت وقت، انرژی و منابع آموزشی و شخصی** میشوند را شناسایی کنیم و به طور کلی به فهم و ارتباطی منطقی میان متغییر ها و فاکتور های متفاوت اندازه گیری شده تا کنون برسیم.

که این یافته ها به سیستم آموزشی مورد نظر کمک خواهد کرد تا فرصت ها و گزینه های اصلاح و بهبود دانش آموزان و خدمات خود را پیش بینی و شناسایی کند و بهبود آن ها سهل تر خواهد شد.

و هم چنین یافته های ما میتواند کمک کند تا معایب و مشکلات سیستم و اجزا مختلف هم سریع تر و راحت تر یافته و اصلاح گردد.

هدف نهایی ما پیش بینی مقدار معدل نهایی هر دانش آموز است که مقداری عددی از 0 تا 20 میباشد و برای دقیق تر بودن برسی هایمان نمیخواهیم نمرات را دسته بندی کنیم. پس از یک مدل پیش بینی **رگرسیونی** استفاده کنیم و مقدار کمی را پیش بینی کنیم.

# مدل مفهومی

فرآیند داده‌کاوی شامل گام‌های زیر است:

* پاک‌سازی داد (Data Cleaning)
* یکپارچه‌سازی داد (Data Integration)
* انتخاب داده (Data Selection)
* تبدیل داده (Data Transformation)
* کاوش داده (Data Mining)
* ارزیابی الگو (Pattern Evaluation)
* ارائه دانش (Knowledge Representation)

1. منطبق با این استاندارد میبایست ابتدا داده های موجود را به خوبی برسی و تحلیل کنیم و ایرادات خطا ها مقادیر گم شده و یا حذف شده و غیره را با روش های متناسب اصلاح کنیم.
2. سپس باید مقادیر را به گونه ای تغییر دهیم که مدل سازی سریع تر و آسان تر و کم هزینه تر انجام شود و درصد خطای نهایی را کاهش دهد. تکنیک هایی مانند گسسته سازی و نرمال سازی کاهش نویز و ...
3. داده ها را به 2 بخش جهت آموزش مدل و سپس تست و ارزیابی مدل تقسیم بندی میکنیم.
4. با استفاده از الگوریتم های درخت CHAID و CART مدل رگرسیون را تشکیل میدهیم.(با مرحله 2 در تعامل است و ممکن است نیاز باشد پاکسازی و یکپارچه سازی و تبدیلات متفاوتی برای مدل انجام گیرد)
5. الگو و مدل های بدست آمده را با روش های آماری راستی آزمایی و صحت سنجی میکنیم. (با مرحله 4 در تعامل است و باید تا رسیدن به نرخ های اطمینان و دقت مورد نظر ارتربیوت های مدل بهینه سازی شوند)
6. در نهایت در گزارشات نهایی نتایج قوانین درخت ها و روش پیش بینی را به کار فرما ارایه میدهیم.

# شناخت داده

دیتاست بدست آمده از 2 مدرسه متفاوت در 3 سال تحصیلی متوالی گرد آوری شده و شامل **33 فیچر** یا ویژگی و **396 رکورد** است و داده میسینگ یا گم شده نداریم. usability دیتاست 10 از 10 ارزیابی شده است و اصطلاحا clean data نیز میباشد.

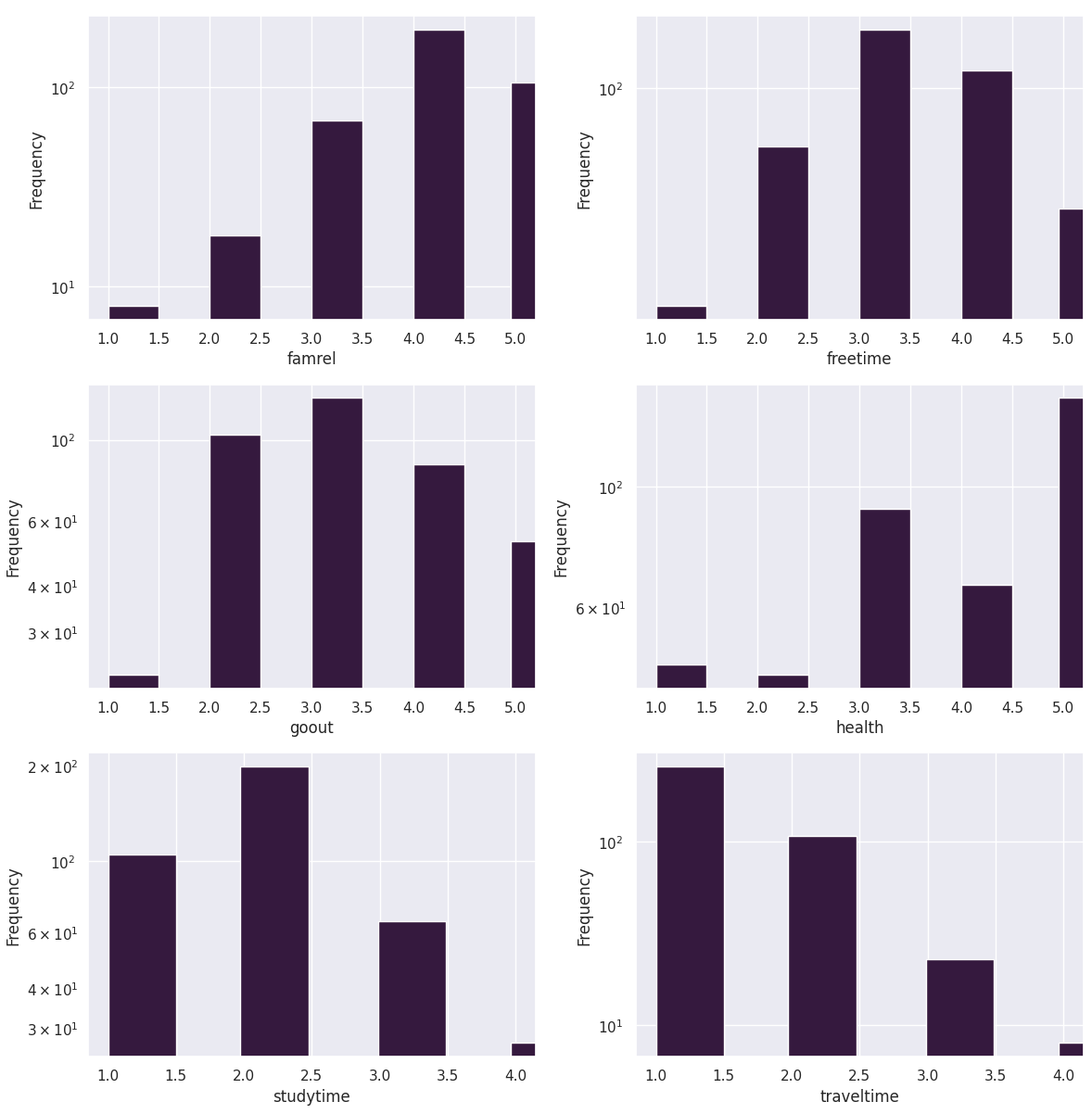
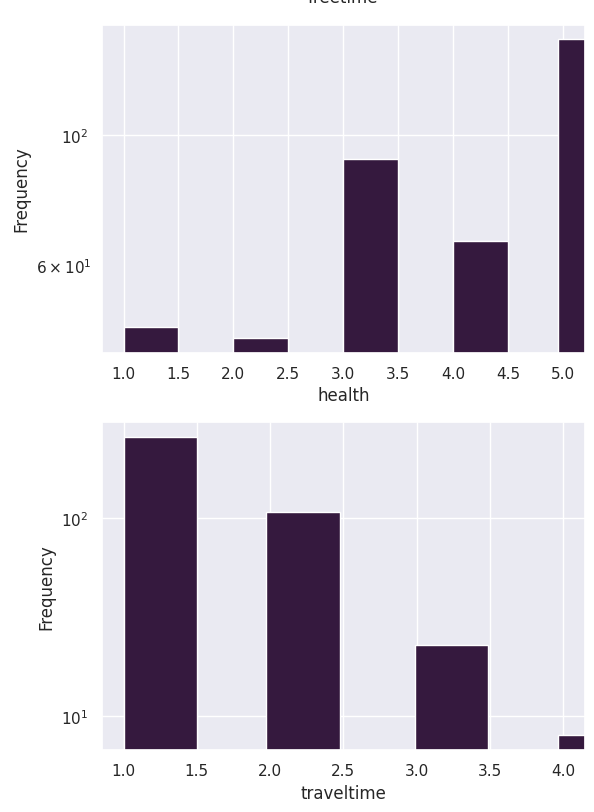
توصیف فیچر های موجود:

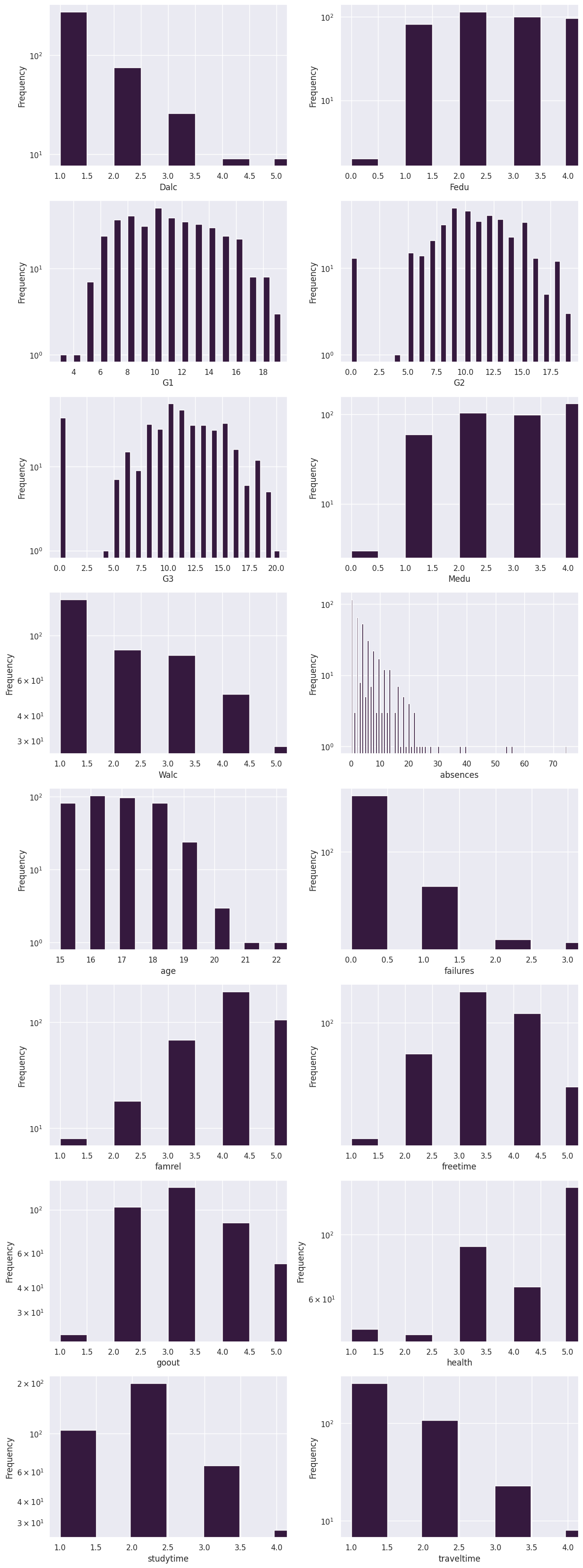
|  |  |  |
| --- | --- | --- |
| نام ویژگی | توضیح | بازه مقدار دهی شده |
| School | نام مدرسه | 'GP' یا 'MS' |
| Sex | جنسیت | F – M |
| Age | سن | 15 تا 22 |
| Address | نوع آدرس (داخل یا بیرون از شهر) | U = شهر و R = بیرون از شهر |
| Famsize | اندازه خانواده (کم یا بیشتر از 3 نفر) | LE3 , GT3 |
| Pstatus | وضعیت تاهل والدین | T = باهم و A = جدا از هم |
| Medu | تحصیلات مادر (صفر - هیچ، ۱ - آموزش ابتدایی (پایه چهارم)، ۲ - پایه پنجم تا نهم، ۳ - متوسطه یا ۴ - آموزش عالی) | 0 - 4 |
| Fadu | تحصیلات پدر (0 به تدریج تا 4 آموزش عالی) | 0 - 4 |
| Mjob | شغل مادر |  |
| Fjob | شغل پدر | Teacher, health, services, at\_home, other |
| Reason | دلیل انتخاب این مدرسه (نزدیکی به خانه، شهرت مدرسه، ترجیح درس یا دلایل دیگر) | home', 'reputation', 'course' , 'other' |
| Guardian | سرپرست دانش آموز | Mother, father, other |
| Traveltime | زمان رفت آمد مدرسه تا خانه | 1-4 |
| Studytime | مقدار مطالعه هفتگی | 1-4 |
| Failures | تعداد دروس رد شده | 1-4 |
| famsup | حمایت تحصیلی خانواده | Yes - no |
| Paid | کلاس های تقویتی ثبت نام شده یا خیر | Yes - no |
| Activities | فعالیت ها فوق برنامه حضور دارد یا خیر | Yes – no |
| Nursery | آیا مهد کودک رفته است یا خیر | Yes - no |
| Higher | میخواهد تحصیلات عالیه داشته باشد یا خیر | Yes - no |
| Internet | دسترسی به اینترنت در خانه | Yes - no |
| Romantic | رابطه عاطفی دارد یا نه | Yes - no |
| Famrel | کیفیت روابط خانوادگی 1خیلی بد تا 5 عالی | 1-5 |
| Freetime | زمان آزاد بعد از مدرسه | 1-5 |
| Gout | بیرون رفتن با دوستان | 1-5 |
| Dalc | مصرف الکل در روز های هفته | 1-5 |
| Walc | مصرف الکل آخر هفته ها | 1-5 |
| Health | وضعیت سلامت عمومی | 1-5 |
| Absences | تعداد غیبت ها | 0-93 |
| G1 | نمره دوره اول | 0-20 |
| G2 | نمره دوره دوم | 0-20 |
| G3 | نمره دوره سوم | 0-20 |

دسته بندی و طبقه بندی همگون فیلد های گسسته مختلف انجام و شده لیبل خورده است.

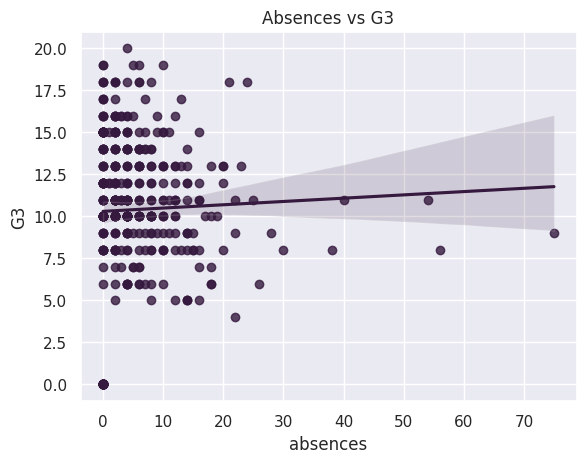
برای **بصری سازی داده** ها اطلاعات را با کمک چند نمودار نمایش میدهیم:

**نمودار های فراوانی هر فیلد:**

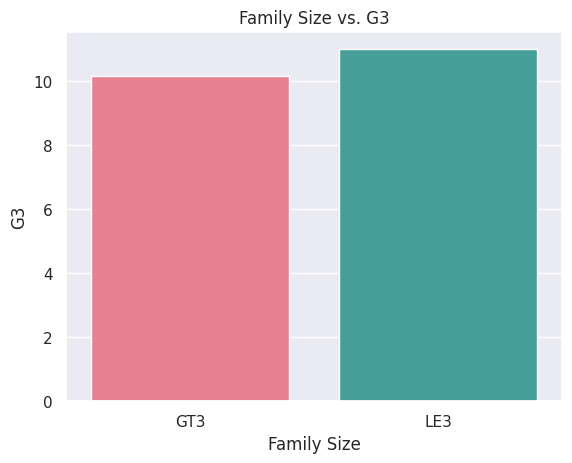




**برسی رابطه میان فیلد های مختلف:**

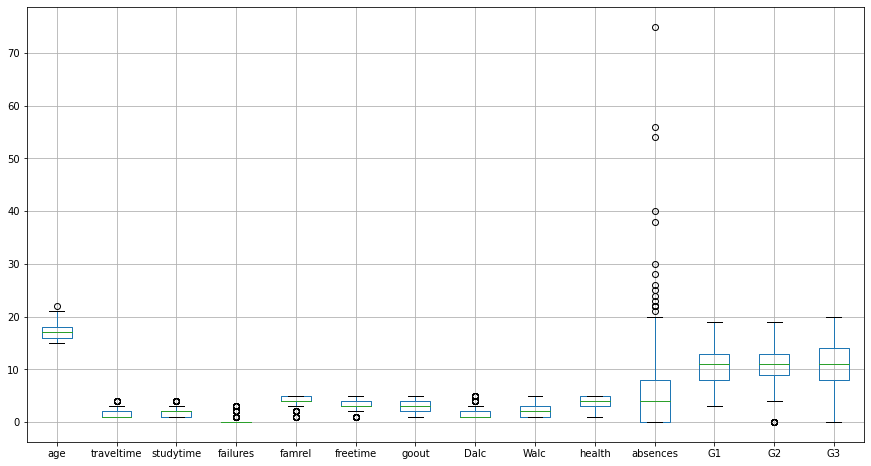


نمودار نقطه ای تعداد غیبت به نسبت نمره نهایی: حضور نداشتن در چند جلسه از کلاس فهم مطالب آموزشی را سخت تر میکند و دانش آموزان باید تلاش بیشتری بکند. و پس از پاکسازی داده ها این مولفه برسی خواهد شد.



نمودار میله ای تعداد اعضا خانواده و نمره نهایی G3: رابطه معنا داری را نشان نمیدهد با این که افرادی که خواهر و برادر ندارند اندکی عملکرد بهتری داشته اند اما قابل چشم پوشی است.

**نمودار جعبه ای تمام فیلد ها:** این نموار همچنین داده های پرت و outlier را میتواند به ما نشان دهد.



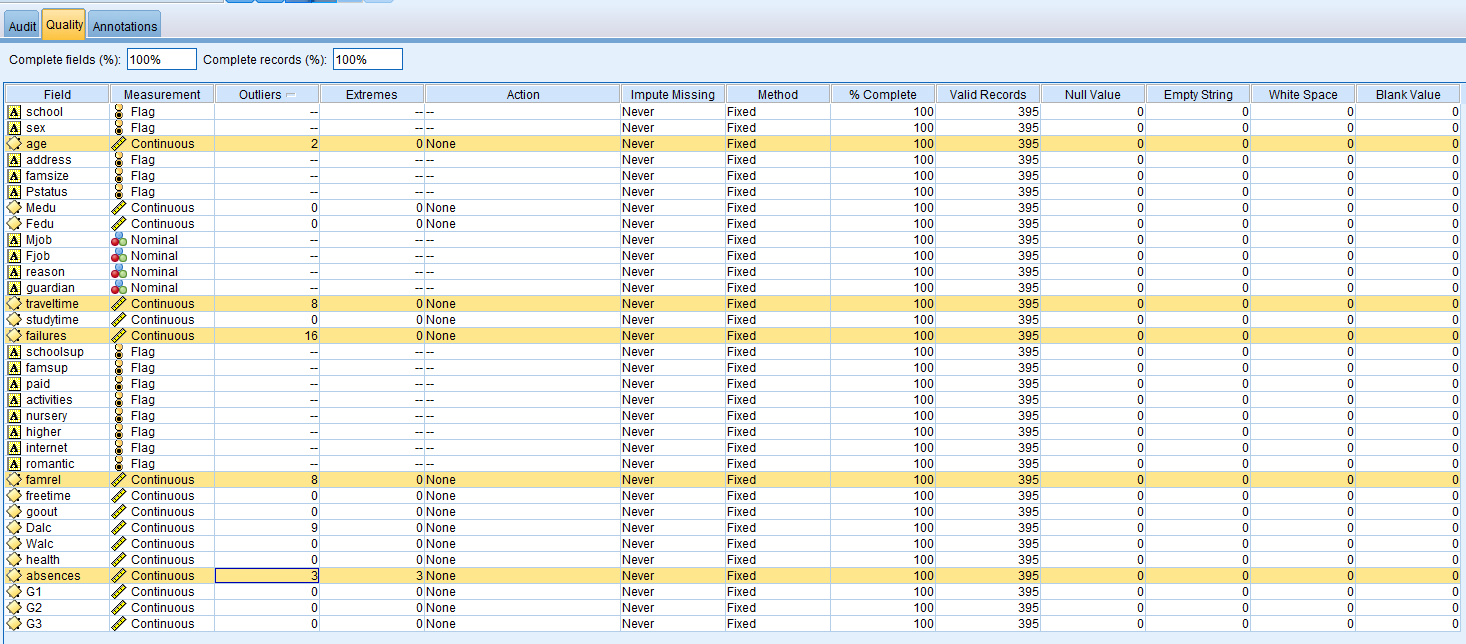
# گزارش اجرایی

ابتدا فایل دیتاست "student\_data.csv" را وارد نرم افزار spss modeler میکنیم .

با افزودن نود data audit از بخش Output وضعیت داده ها و کیفیت و یک پارچگی آنها را برسی میکنیم.

گزینه های graphs – basic statistics – advanced stat را فعال میکنیم و Run میکنیم.

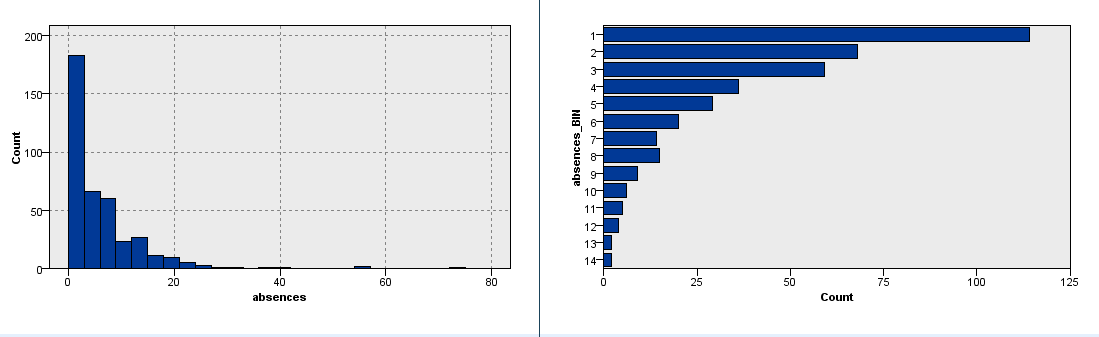
در بخش quality جدول فوق را بدست آورده ایم:



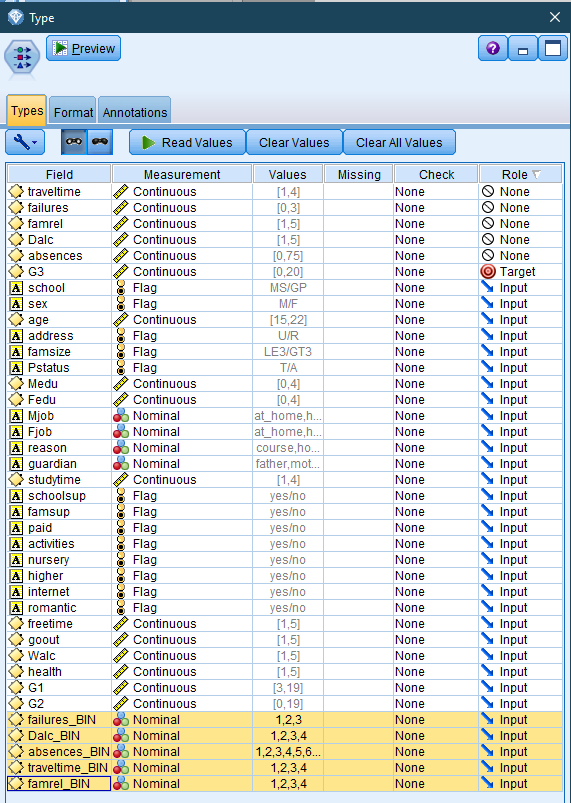
در فیلد های- Dalc age – traveltime – failures – famrel – absences همان طور که مشخص شده outlier هایی داریم که در مدل سازی خلل ایجاد میکنند و باید آنها را پاک سازی کنیم که در 5 مرحله و با استفاده از نود های binning و select این کار را کرده ایم:

* + 1. بایندینگ: فیلد failures که 16 outlier داشت را به 3 دسته همگون تقسیم کرده ایم که در یک ستون به نامfailures\_BIN اضافه میشود و داده های پرت از بین میروند.
    2. شرط انتخاب: رکورد هایی با فیلدage سن های بالای 20 سال را حذف کرده ایم و چون تعدادشان هم زیاد نبوده مشکلی ایجاد نمیکند و داده های پرت فیلد سن را به صفر میرساند.
    3. بایندینگ: فیلد Dalc را به 4 دسته با فواصل یکسان تقسیم بندی و گسسته سازی کرده ایم. که فیلد Dalc\_BIN را به جدول اضافه میکند.
    4. رکورد ها با غیبت های بالای 28 حذف شده + و سپس در نود بایندینگ گسسته سازی 2 واحدی برای کاهش نویز انجام شده است. با این کار نمودار توزیع absences را هموار و بدون نویز کرده ایم.
    5. فیچر های famrel و traveltime را به 4 دسته با فاصله یکسان گسسته سازی کرده ایم که فیلد های famrel\_BIN و traveltime\_BIN را اضافه میکند و بدون هیچ داده پرت ای.

نمودار توزیع فراوانی فیلد absences و absences\_BIN بعد از پاکسازی و آماده سازی این فیلد:

 بعد از اعمال تغییرات قبل از آماده سازی

در برسی مجدد پس از انجام پاکسازی های گفته شده **تمامی آوتلایر ها به صفر رسیدند** و توزیع ها نرمال شد و نویز فیلد غیبت ها از بین رفت هم چنین تمامی داده های پرت از بین رفتند و داده ها برای مدل سازی آماده شده اند.



بعد از پاک سازی داده ها با استفاده از نود TYPE **فیلد هدف** را مشخص میکنیم و فیلد هایی که نمیخواهیم در مدل سازی شرکت بدیم را با لیبل NONE مشخص میکنیم.

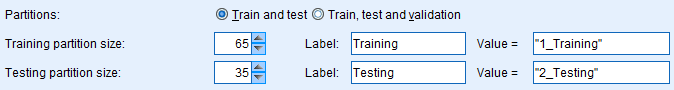
5 فیلد traveltime – failures famrel – absences و Dalcکه گسسته سازی کرده بودیم و هر کدام یک ما به ازای BIN دارند باید از مدل سازی کنار گذاشته می شوند.

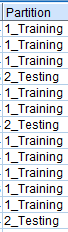
نمره نهایی سال پایانی G3 را فیلد هدف یا Target قرار میدهیم.

در گام بعدی مدلسازی به طراحی آزمون میپردازیم: (**تقسیم بندی داده ها**)

از تب field ops یک نود partition به خروجی تایپ وصل میکنیم و چون تعداد نمونه ها 390 رکورد است داده ها را به 2 بخش 65% جهت آموزش مدل (train) و 35% برای ارزیابی مدل (test) تقسیم میکنیم با این کار نرم افزار به هر رکورد یک لیبل اضافه میکند.

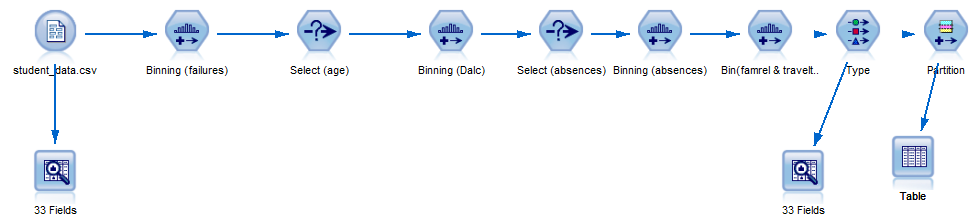
(اگر تعداد رکورد ها به اندازه کافی زیاد بود بهتر بود 3 بخش تقسیم میکیردیم و برای بهبود مدل از آن استفاده میکردیم)



همان طور که در عکس سمت راست مشاهده میشود ستون پارتیشن اضافه شده و مقادیر 1\_training و 2\_testing به آن نسبت داده شده است.

(برای مشاهده جدول جدید از بخش Output یک نودtable وارد کنید و به گره پارتیشن وصل کنید سپس وارد آن شوید و روی RUN کلیک کنید تا جدول پارتیشن بندی شده را به شما نمایش دهد.)

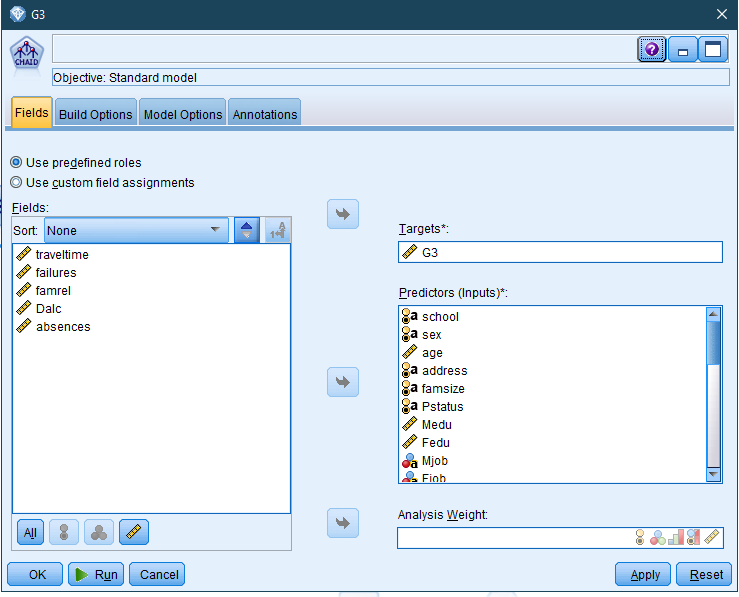
استریم پروژه تا این مرحله به این شکل میباشد:



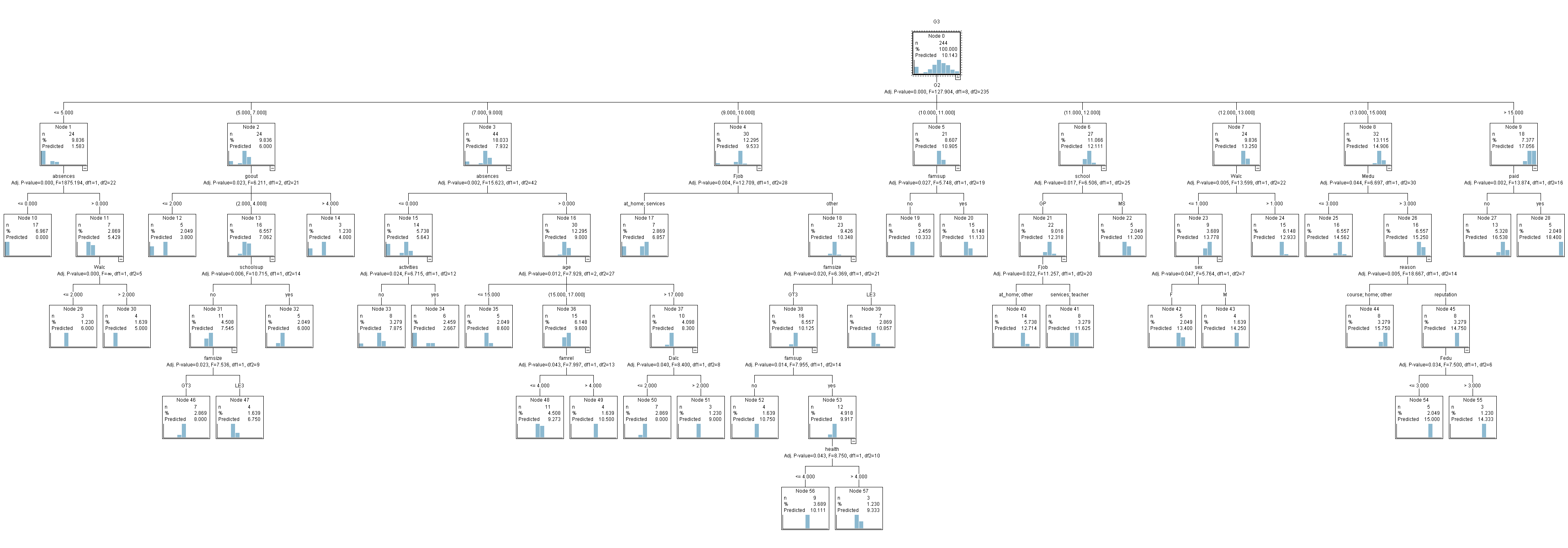
# مدل سازی

برای پیش بینی نمره نهایی دانش آموزان از 2 درخت رگرسیون CHAID و CART استفاده کرده ایم.

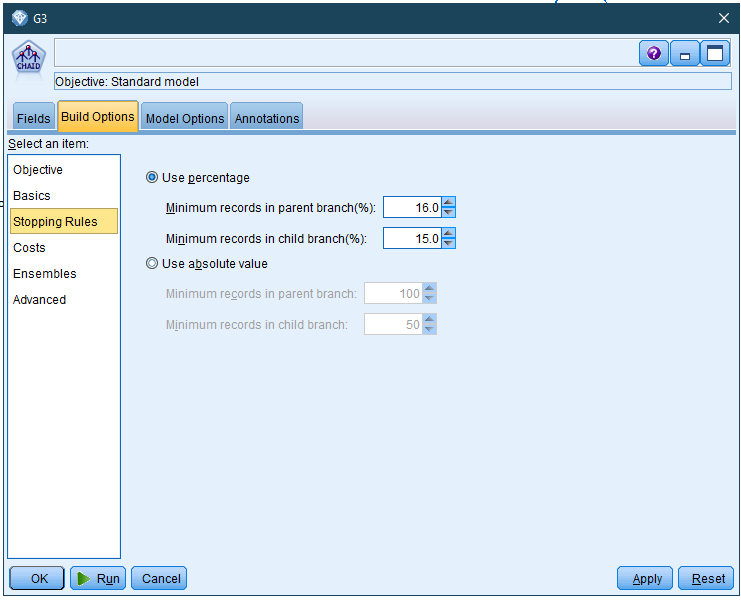
در قسمت مدلینگ درخت CHAID را اضافه میکنیم و خروجی پارتیشن را به آن وصل میکنیم.



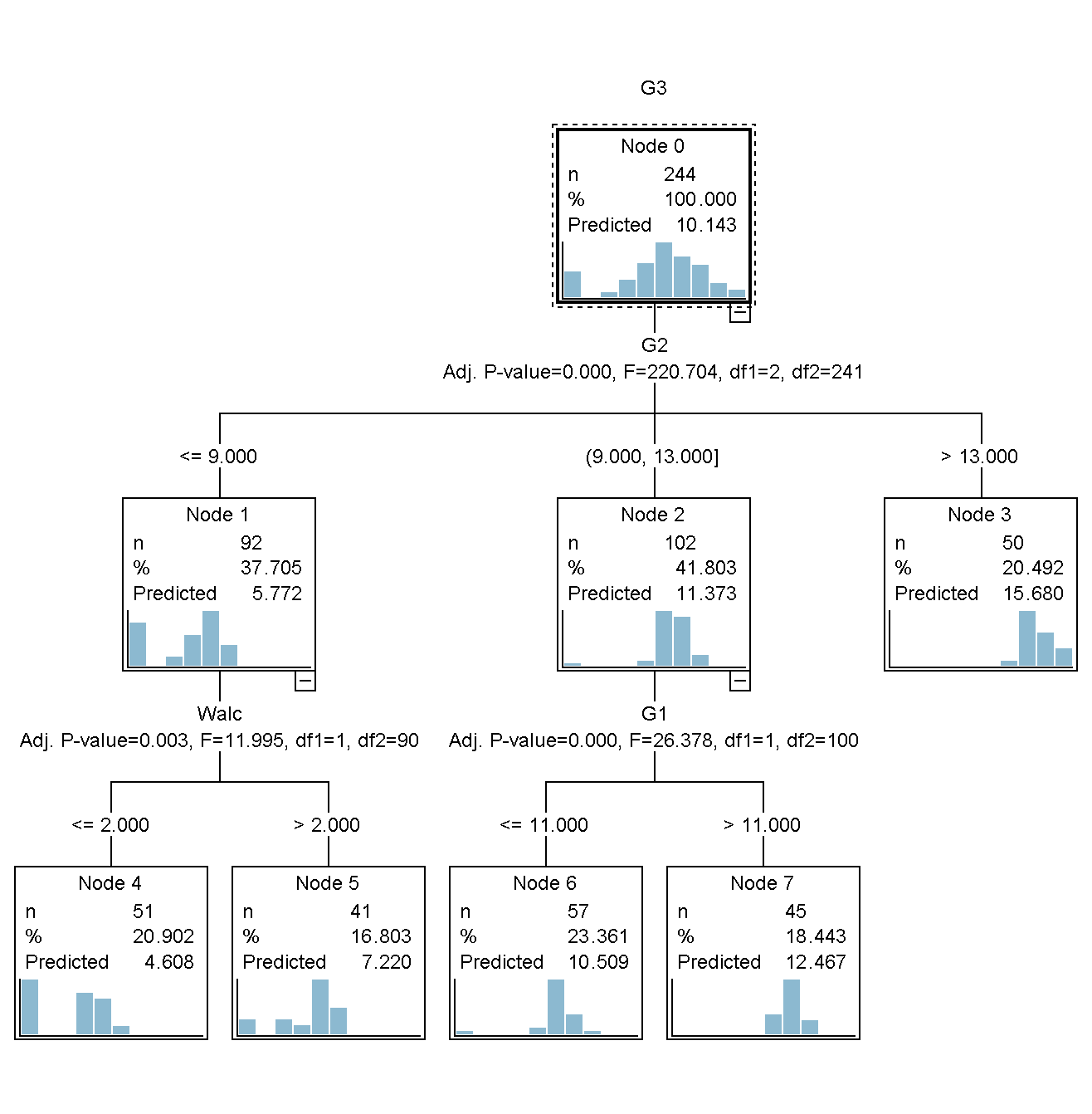
ابتدا با تنظیمات پیش فرض درخت را RUN میکنیم و میبینیم که شکل درخت زیر را میسازد، این درخت روی دیتای ترینینگ تنها 4.4% خطا دارد اما اختلاف دقت train و test در آن 18 درصد است و جزعیات عمق و برگ ها و تعداد قوانین استخراج شده از آن بسیار بالاست که همگی نشانه هایOverFit بودن مدل ما هستند.



پس باید تنظیمات مدل سازی درخت را به گونه ای تغییر دهیم تا مدل منعطف تر و فیت تری بسازد.

پس در تنظیمات درخت در تب Build Options در تنظیمات Stopping Rules حداقل رکورد در شاخه والد را 16درصد و حداقل رکورد شاخه فرزند 15درصد را وارد میکنیم تا جلوی پیش روی و پیچیدگی اضافی درخت را بگیرد و درختی ساده تر و حرص شده به ما بدهد: 

و در نهایت خروجی درخت CHAID اصلاح شده به این صورت خواهد بود:



5 قانون بدست آمده از درخت CHAID:

G2 <= 9 [ Ave: 5.772, Effect: -4.372 ]

Walc <= 2 [ Ave: 4.608, Effect: -1.164 ] => 4.608

Walc > 2 [ Ave: 7.22, Effect: 1.448 ] => 7.22

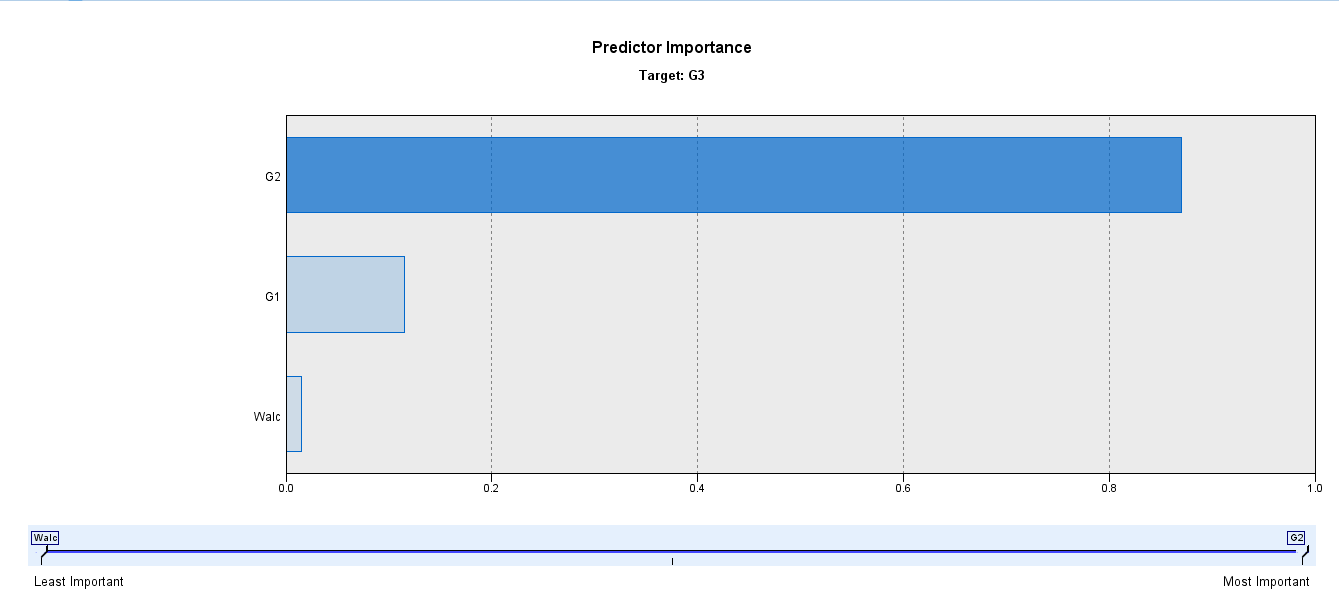
G2 > 9 and G2 <= 13 [ Ave: 11.373, Effect: 1.229 ]

G1 <= 11 [ Ave: 10.509, Effect: -0.864 ] => 10.509

G1 > 11 [ Ave: 12.467, Effect: 1.094 ] => 12.467

G2 > 13 [ Ave: 15.68, Effect: 5.537 ] => 15.68

طبق تحلیل این درخت معدل سال های گذشته و **مصرف الکل در طی هفته** بیشترین تاثیر را در پیش بینی داشته اند:

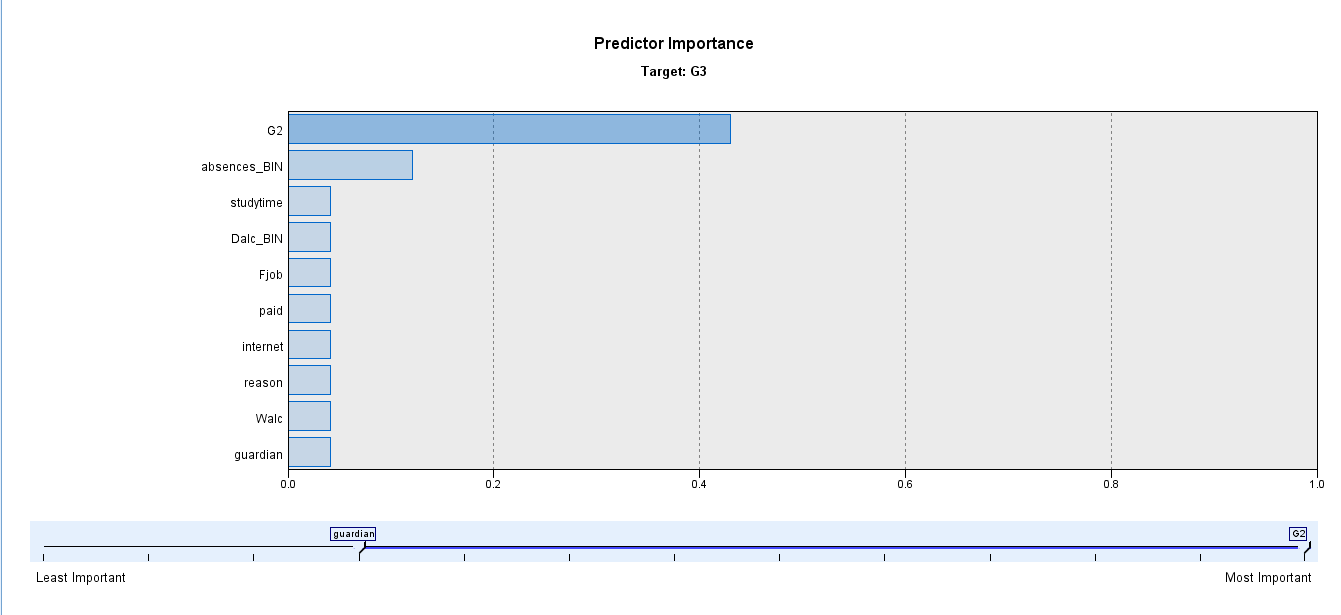


و موالفه مصرف الکل در طی هفته در محاسبه نمره های 4.6 و 7.2 هم بکار رفته است. هم چنین ارتباط داشته است با معدل ترم قبلی پایین تر از 10 پس این مولفه باید مورد توجه مسولان و خانواده ها باشد.

دومین مدل سازی ای که انجام میدهیم استفاده از الگوریتم درخت رگرسیون **CARD** است و مشابه درخت CHAID از بخش مدلینگ یک درخت CRT اضافه میکنیم و به نود پارتیشن متصل میکنیم و برای کاهش اور فیتینگ مقادیرStopping rules را 16 و 15 درصد قرار میدهیم.

حاصل مدل سازی را در ادامه اورده ایم:

مدل عوامل معدل ترم قبل – تعداد غیبت ها – زمان مطالعه – مصرف الکل در آخر هفته ها – شغل پدر – کلاس های فوق برنامه – دسترسی به اینترنت در منزل – دلیل دانش آموز برای انتخاب این مدرسه – مصرف الکل در طی هفته و نوع سرپرستی دانش آموزان را به ترتیب پر اهمیت و تاثیر گذار در نمره نهایی تشخیص داده است:



4 قانون بدست آمده:

G2 <= 9.500 [ Ave: 5.672, Effect: -4.317 ]

absences\_BIN in [ 1 ] [ Ave: 2.5, Effect: -3.172 ] => 2.5

absences\_BIN in [ 2 3 4 5 6 7 8 9 10 11 12 14 ] [ Ave: 7.949, Effect: 2.277 ] => 7.949

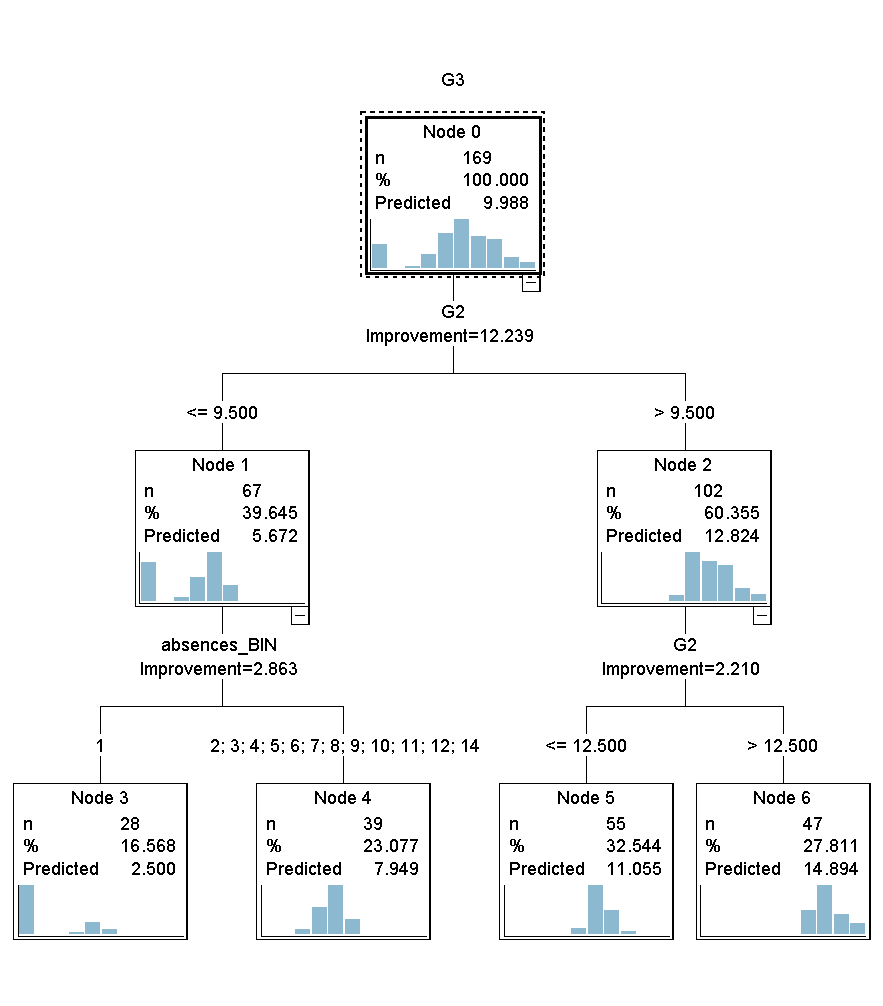
G2 > 9.500 [ Ave: 12.824, Effect: 2.835 ]

G2 <= 12.500 [ Ave: 11.055, Effect: -1.769 ] => 11.055

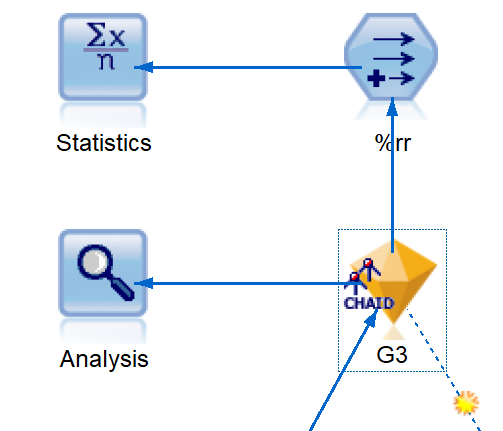
G2 > 12.500 [ Ave: 14.894, Effect: 2.07 ] => 14.894

الگوریتم chaid در نهایت قوانین خود را از روی معدل ترم قبل و **تعداد غیبت های دانش آموز** استخراج کرده است.

گراف درخت **CARD به این شکل خواهد بود:**



عمق آن برابر 2 است و در مجموع 4 قانون دارد و در ابتدا (ریشه) G2 را مورد برسی قرار میدهد -> سپس غیبت ها و G2



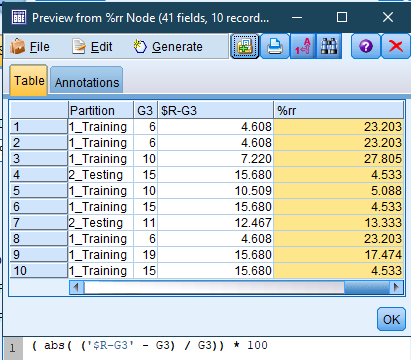
# آنالیز و آزمون مدل ها

برای محاسبه خطای مقادیر پیش بینی شده از بخش Field Ops نود **Derive** را

به خروجی مدل وصل میکنیم و فرمول (abs( ('$R-G3' - G3) / G3)) \* 100

را وارد میکنیم ('$R-G3' مقدار پیش بینی شده توسط مدل است و G3 مقدار واقعی)

این نود برای هر فیلد مقدار خطای مدل را محاسبه میکند و به جدول اضافه میکند:

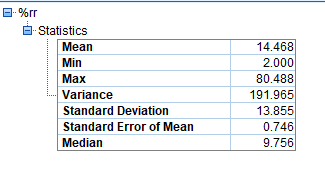
حال برای محاسبه میانگین و میانه و تحلیل آماری خطا های مدل از گره **Statistics** که در بخش output است استفاده میکنیم.

مشخص میکنیم که مولفه %rr را مورد برسی قرار دهد و min max mean median و واریانس و انحرافات معیار را تیک میزنیم تا محاسبه کند.

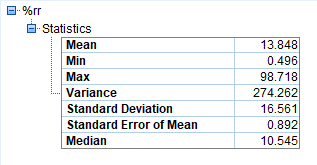
با RUN کردن پنجره استتیستیک محاسبات انجام شده را در جدولی به ما نشان خواهد داد.

تمام این مراحل را برای هر 2 درخت انجام میدهیم.

**نتیجه محاسبه خطای درخت CHAID :**

میانگین خطا 14% در بازه 2 تا 80 که واریانس بالایی دارد و نشان میدهد تعداد کمی خطای 80 درصد توزیع را بهم زده است.

میانه که توصیف بهتری برای خطای کل است = 9.7%

**محاسبه خطای درخت CART :**

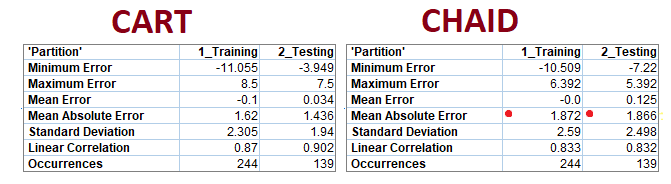
حداقل خطای آن کمتر از درخت چید است.

میانگین خطا تقریبا برابر درخت چید و 13.8% است.

حداکثر خطا 98.7 درصد نزدیک به 100 و واریانس آن بیشتر از درخت قبلی است.

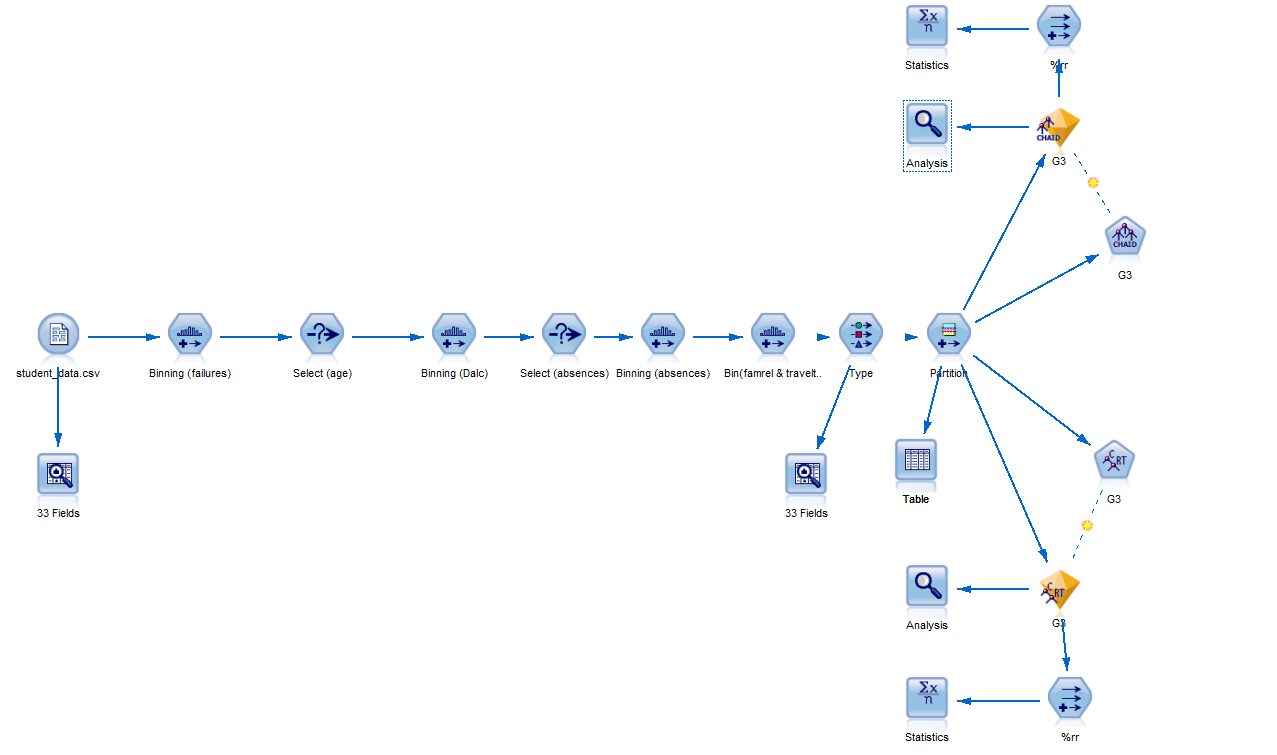
میانه که توصیف بهتری برای خطای کل است = 10.5%

**نتیجه آنالیز های 2 درخت:**

****

درخت CHAID تفاوت خطای کمتری بین داده ترین و تست دارد که نشان میدهد منعطف تر و فیت تر است و میانگین مقدار عددی خطای هر 2 مدل نزدیک به 0.01 یک صدم است که در مقیاس 0 تا 20 عالی است.

در نهایت شکل استریم پروژه به این صورت خواهد بود:



# پیشنهادات

پیشرفت تحصیلی دانش آموزان و ارتقا کیفیت محیط آموزشی از جمله اهداف مهم آموزش و پرورش است. در این راستا، استفاده از تکنولوژی می تواند نقش مهمی ایفا کند. در ادامه، برخی از پیشنهادات در این زمینه ارائه می شود:

* آموزش و توانمندسازی معلمان در زمینه استفاده از تکنولوژی آموزشی: معلمان نقش کلیدی در فرایند یاددهی-یادگیری دارند. بنابراین، آموزش و توانمندسازی آنها در زمینه استفاده از تکنولوژی آموزشی ضروری است. این امر می تواند به معلمان کمک کند تا با استفاده از ابزارها و فناوری های آموزشی، فرایند یادگیری را برای دانش آموزان جذاب تر و اثربخش تر کنند. **مثلا در جریان گزارشات آنلاین هفتگی و ماهیانه از عملکرد یک دیگر و دانش آموزان قرار بگیرند.**
* تجهیز مدارس به امکانات و تجهیزات آموزشی: یکی از پیش نیازهای استفاده از تکنولوژی آموزشی، تجهیز مدارس به امکانات و تجهیزات آموزشی است. این امر شامل تامین رایانه، تبلت، لپ تاپ، پروژکتور، و سایر تجهیزات آموزشی می شود.
* توسعه محتوای آموزشی الکترونیکی: محتوای آموزشی الکترونیکی می تواند به دانش آموزان کمک کند تا در هر زمان و مکانی به آموزش دسترسی داشته باشند. بنابراین، توسعه محتوای آموزشی الکترونیکی با کیفیت و متناسب با نیازهای دانش آموزان ضروری است. **محتوا به گونه ای که بتواند توانایی رقابت با محتوای عادی شبکه های مجازی را داشته باشد.**
* ارائه آموزش های آنلاین و غیرحضوری: آموزش های آنلاین و غیرحضوری می تواند به دانش آموزانی که به دلایل مختلف امکان حضور در کلاس های حضوری را ندارند، کمک کند. بنابراین، ارائه آموزش های آنلاین و غیرحضوری می تواند فرصت های یادگیری را برای همه دانش آموزان فراهم کند.

همچنین میتوان از ابزار ها و روش های نوین برای **جمع آوری بیشتر اطلاعات جهت تحلیل استفاده نمود مثل داده های دوربین های مدار بسته موجود در مدارس مختلف.**

**در بستر یک شبکه تمامی مدارس کشور و محل های مخلف و موسسه های آموزشی را یک پارچه کرد یا بر بستر ابری داده های خود را جهت استفاده به اشتراک بگذارند.**  و میتوان برای سنگین نشدن باز سیستم و کاهش هزینه ها برخی پردازش ها را در محل مدرسه انجام داد.

هوش مصنوعی می تواند برای شخصی سازی آموزش و یادگیری دانش آموزان استفاده شود. به عنوان مثال، **هوش مصنوعی می تواند بر اساس نیازهای فردی دانش آموزان، محتوای آموزشی را به آنها ارائه دهد.**