یک روش جدید استخراج داده ها برای پیشگیری از مصرف الکل دانشجویان دوره متوسطه

چکیده

در این مقاله ما مجموعه داده ای در مورد دانش آموزان پرتغالی در دو دوره متوسطه را بررسی کردیم که توسط پائولو کورتز و آلیس سیلوا،دانشگاه مینو پرتغال ارائه شده. کار ما در نظر دارد با اعتیاد دانشجویی به الکل در سطح ثانویه با استفاده از تکنیک های کسب و کار (BI) و داده های معدن (DM) مقابله کند. نتیجه نشان می دهد که دقت پیش بینی خوبی می تواند به دست آید با توجه به اینکه اعتیاد الکل بر عملکرد دانشجویان اثر می گذارد،به عبارت دیگر، همچنین این نتیجه ارتباط بین مصرف الکل و اجتماعی، جنسیت و ویژگی های زمان مطالعه برای هر دانش آموز را نیز نشان می دهد.به عنوان یک نتیجه مستقیم از این پروژه، ما توانسته ایم از ابزارهای موجود برای کاهش اثرات مضر الکل در زندگی دانشجویان استفاده کنیم.در این مقاله از درخت تصمیم برای داده کاوی و از جنگل تصادفی برای بهبود دقت در روش قبلی استفاده میکنیم.

کلید واژه: دقت ،کارایی ،attributes ، business intelligence

مقدمه

الکل تاثیر زیادی در زندگی ما داشته است،نوشیدن بیش از حد در یک بار یا در طول زمان می تواند عوارض جدی برای سلامت ما داشته باشد.هر کسی که الکل مصرف می کند احتمال دارد که حداقل دست کم برخی از اثرات کوتاه مدت آن را داشته باشد. مثل کم خوابی است.الکل دارای تأثیرات کوتاه مدت و طولانی مدت در سلامت است. مصرف الکل در سن نوجوانی،توانایی های ذهنی و جسمی کودک را کاهش دهید،تاثیر قضاوت و هماهنگی که می تواند منجر به مشکل شود. سطح الکل چنان زیاد است که عملکرد های حیاتی مغز، شامل کنترل تنفس،مسدود شده اند.الکلی ها احتمال بیشتری داشتند که مجروح شوند و یا حوادثی برایشان اتفاق بیفتد و بیشتر نگران کننده است. آنها بیشتر در مسافرت مشغول نوشیدن هستن وبه یک حادثه رانندگی منجر می شود، وقتی بچه ها نوشیدن،مهارت های تصمیم گیری آنها تحت تاثیر قرار می گیرد و احتمال بیشتری دارند که خطرات بزرگی مانند داشتن رابطه جنسی محافظت نشده داشته باشند،این می تواند به بیماری های منتقله از راه جنسی و حاملگی ناخواسته منجر شود. در حالی که نوشیدن بیش از حد توسط نوجوانان مشکل خاص خودش را دارد. افراد نابالغ احتمالا از طیف وسیعی از مسائل مربوط به سلامت رنج می برند از جمله افزایش وزن یا کاهش وزن،پوست بد،خواب مضطرب،سردرد و غیره... در دوران کودکی و نوجوانی،مغز هنوز در حال توسعه است.افزودن الکل به این فرایند، مشکل ایجاد میکند. و می تواند عملکرد حافظه را تحت تاثیر قرار دهد.

واکنش ها،توانایی یادگیری و توجه به ویژه در طول سال های مدرسه بسیار مهم است. نوشیدن می تواند عملکرد کودک در مدرسه را تحت تأثیر قرار دهد و مانع از دستیابی به پتانسیل کامل آنها شود. جوانان که بیش از حد الکل می نوشند احتمال بیشتری نیز دارند که سلامت روان را مختل کنند. هر والدین می خواهند فرزند خود در مدرسه خوب عمل کنند،آمار نشان می دهد نوشیدن افراد زیر سن قانونی باعث می شود بچه هایی که شروع به نوشیدن می کنند در سن ۱۳ سالگی نمرات بدتری داشته باشند. و برای رفتن به مدرسه به مشکل می خورند و در بدترین حالت،از مدرسه حذف می شوند.آنها کنترل کمتری دارند و مغز آنها برای رسمیت شناختن علائم هشدار دهنده تلاش می کند،این می تواند منجر به تجاوز و دعوا شودخطر ابتلا به آنها در خشونت و خرابکاری های جدی به طور مستقیم با توجه به مصرف الکل افزایش می یابد،که می تواند منجر به دستگیری و سابقه کیفری شود.گرایش طبیعی آنها به آزمایش و ریسک افزایش می یابد.اضافه کردن الکل به ترکیب ایده خوبی نیست؛می تواند آنها را در شرایط آسیب پذیر یا خطرناک قرار دهد، [۱] [۲] [۳]

ما مدل های قبلی را که در این مجموعه داده ها کار می کرد، بررسی کردیم. اولین مقاله از رگرسیون و طبقه بندی استفاده می کند و با استفاده از نرم افزار Rminer انجام می شود،در طبقه بندی،مدلها اغلب با استفاده از درصد کسرهای صحیح (PCC) مورد ارزیابی قرار می گیرند در حالی که در رگرسیون، ریشه (RMSE) یک معیار مردمی است (ویتن و فرانک ۲۰۰۵)PCC بالا (i، e، نزدیک به ۱۰۰٪) یک طبقه بندی خوب را نشان می دهد،[۴]

در مقاله دوم،خوشه بندی و نرم افزار WEKA استفاده شده است.خوشه بندی یکی از محبوب ترین روش های یادگیری بی نظیر است. شامل ساختن یک مجموعه از اشیاء فیزیکی یا انتزاعی در کلاس هایی با اشیاء مشابه (Han and Kamber، 2001). به عنوان یک ابزار مستقل برای درک بهتر توزیع داده یا به عنوان مرحله پیش پردازش برای کارهای دیگر استفاده می شود (Hastie .et al، 2001)، ابزار دندانه ای برای درک بهتر،بعد از تغییرات به داده ها و تعدادی از صفات انتخاب شد. انتخاب روش خوشه بندی که مورد استفاده قرار می گرفت، توسعه یافت،با توجه به اینکه کاربران باید به دو دسته تقسیم شوند: انهایی که می نوشند و انهایی که نمی نوشند،برای این فرآیند،روش K MEAN مورد استفاده قرارگرفت.

پس از خوشه بندی داده ها،انتخاب روشهای طبقه بندی مقایسه می شود و آن ها عبارت بودند از:درخت تصمیم و SVM و NAIVE BAYES,LAZY IBK [۵].

با نتیجه این مقاله [۶] و مدل مورد استفاده در آن،ما نتیجه گرفتیم که با استفاده از سایر روش های داده کاوی،ما می توانیم پیشرفت های عملکرد این مقاله را ببینیم و پاسخ های بهتر برای به اشتراک گذاشتن برای دانش آموزان داشته باشیم.،در این کار،ما داده های واقعی دنیای واقعی را از دو مدارس متوسطه پرتغالی بررسی خواهیم کرد . دو منبع مختلف استفاده شد: علامت گزارش ها و پرسشنامه ها.این نرم افزار برای پیاده سازی Rapid Miner استفاده کرده است. و دو الگوریتم (تصمیم گیری درختان، جنگل تصادفی) در ان آزمایش و مدل سازی خواهد شد.به عبارت دیگر،ما جنگل تصادفی را به عنوان بهترین مدل تحلیل خواهیم کرد.

مواد و روش ها

مجموعه داده. ما از مجموعه داده ای در مورد دانش آموز پرتغالی که توسط پائولو کورتز و آلیس سیلوا در دانشگاه مینو.پرتغال ساخته شده است استفاده می کنیم [۷]، در پرتغال: آموزش متوسطه شامل ۳ سال تحصیل،قبل از ۹ سال آموزش ابتدایی و به

دنبال آن آموزش عالی،اکثر دانش آموزان به نظام آموزشی عمومی و آزاد می پیوندند.این مطالعه داده های جمع آوری شده در سال تحصیلی ۲۰۰۶ از دو دانشکده عمومی،از منطقه آلندژو پرتغال به دست امده . از این رو،پایگاه داده از دو منبع ساخته شده است: گزارش های مدرسه بر اساس ورق کاغذ که شامل چند ویژگی می باشد و پرسشنامه ها برای تکمیل اطلاعات قبلی تنظیم شده است.

نرم افزار. RAPID MINER

مدل داده کاوی

ما از مدل طبقه بندی(تصمیم گیری) استفاده می کنیم طبقه بندی یک شکل از تجزیه و تحلیل داده ها است که الگوهای استخراج مدل های مهم داده ها را توصیف می کند،اغلب مدل ها طبقه بندی کننده نامیده می شودکه برچسب های طبقه بندی شده (گسسته، بدون نظم) را پیش بینی می کند.

طبقه بندی داده ها یک فرایند دو مرحله ای است متشکل از یک گام یادگیری (که در آن یک مدل طبقه بندی ساخته شده است) و یک گام طبقه بندی (که در آن مدل برای پیش بینی برچسب های کلاس برای داده های داده شده استفاده می شود).که توصیف یک مجموعه از پیش تعیین شده از کلاس داده ها و یا مفاهیم را به همراه دارد،این گام یادگیری (یا مرحله آموزش) است.جایی که یک الگوریتم طبقه بندی کننده را با تجزیه و تحلیل یا (به دست آوردن) یک مجموعه آموزشی ایجاد شده از تپ های پایگاه داده و برچسب کلاس های مرتبط آنها به دست می اورد،چندین الگوریتم DMهر یک با اهداف و قابلیت های خود،برای وظایف طبقه بندی پیشنهاد شده است. درخت تصمیم (DT) یک ساختار شاخه ای است که مجموعه ای از قوانین را نشان می دهدو ارزش ها را در شکل سلسله مراتبی متمایز می کند [۸] جنگل تصادفی [RF] مجموعه ای از T TN نامحلول است.[8]

درخت تصمیم گیری در نرم افزار Rapid Miner به شرح زیر است: پس از انتخاب اطلاعات مورد نظر و قرار دادن درخت تصمیم حالا ما باید یک برچسب از صفات در مجموعه داده انتخاب کنیم که ارزش عددی نیستند. و مقایسه آنها با یک پاسخ منطقی و قابل فهم و با توجه به برچسب ردیف خط داده ما بررسی می شود و پاسخ نهایی ما با مقادیر برچسب متناسب است. پس از به دست آوردن نتایج آماری این روند،ما از روش دوم استفاده می کنیم یعنی جنگل تصادفی (که همانند درخت تصمیم گیری است)

با تفاوت این که چندین درخت را از مدل موجود ایجاد می کند و ما بهترین را انتخاب می کنیم) برای مقایسه نتایج دو الگوریتم،ما می توانیم بهبود عملکرد روش دوم را در مقایسه با روش اول ببینیم. هنگامی که ما جعبه های مورد نیاز را در Rapid Miner قرار می دهیم، باید پارامترها را تغییر دهیم.به طور مثال هنگام استفاده از جعبه درخت تصمیم گیری یا کادر جنگل تصادفی علامت هشدار به ما می گوید که برچسب مورد نظر خود را انتخاب کنیدو برای تعیین برچسب به یک جعبه نقش مجموعه نیاز داریم.

پیش پردازش

هدف ما یافتن مصرف الکل توسط دانش آموزان دبیرستان است در این مجموعه داده دو ویژگی متفاوت در مورد الکل وجود دارد. اولین مصرف الکل در روز کاری است (Dalc) و دوم مصرف الکل در آخر هفته (واکر) است. فقط یک شاخص را می توان پیش بینی کرد بنابراین ما یک ویژگی را ایجاد کردیم که کل مصرف الکل توسط یک دانش آموز خاص را در یک هفته کامل نشان می دهد. بنابراین این دو ویژگی را به صورت زیر در شکل ۱ ترکیب می کنیم:

$$Alc = \frac{Walc \times 2 + Dalc \times 5}{7}$$

شکل ۱، ترکیب دو صفت

در این بخش ما باید قبل از شروع فرآیند اصلی یک سری از مراحل را انجام دهیم تا پاسخ درستتری را به صورت موثرتری به دست آوریم.

مراحل به صورت زیر می باشد :

مرحله ۱ : از آنجاییکه ما دو مجموعه داده متفاوت با همان زمینه ها داریم (STUDENT MAT و STUDENT POR)، ما باید با هم به یک مجموعه داده جدید (بنام ادغام) برسیم تا مراحل بعدی را دنبال کنیم.

مرحله ۲: همانطور که قبلا گفتیم ما باید یک برچسب برای کشیدن درخت داشته باشیم ، در این پروژه ما باید یک برچسب با استفاده از ویژگی های داده ها که به شیوه ای ریاضی تعیین شده است، بدست آوریم با توجه به محاسبات در شکل بالا برچسب ما Alc خواهد بود و ویژگی جدید نیز بین یک تا پنج تغییر می کند، برای بررسی اینکه آیا دانش آموز یک نوشیدنی است یا خیر، Alc مقدار باینری می شود اگر کمتر از ۳ باشد ۰ است این بدان معنی است که شما یک مصرف کتنده نیستید، در غیر این صورت ۱.

مرحله ۳ : بر اساس اسناد مقاله ما و عدم وجود زمینه ما باید یک فیلد جدید به نام Fabs دریافت کنیم و به مجموعه داده جدیدمان اضافه کنیم اگر یک دانشجو غالبا در مدرسه غیبت کند، او بیشتر از سایرین الکل مصرف می کند بنابراین این ویژگی به مقدار باینری تبدیل می شود، اگر او غالبا غیبت (بیش از ۱۰ روز) داشته باشد مقدار ۰، در غیر این صورت مقدار ۱.

مرحله ۴ : تبدیل مقدار باینری Alc و Fabs به ارزش اسمی برای انجام همبستگی.

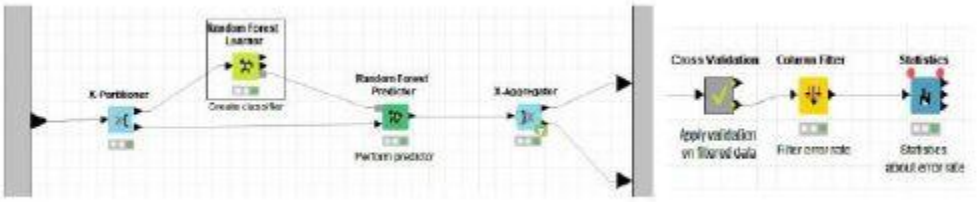
یادگیری و پیش بینی

درخت تصمیم گیری قلب این کار است، آنها برای پیش بینی خوب مورد استفاده قرار گرفتند پیدا کردن همبستگی بین ویژگی ها و همانطور که قبلا مشاهده کردیم برای پیش پردازش مجموعه داده ها استفاده می شوند، هنگامی که الگوریتم درخت تصمیم گیری داریم، باید تصمیم بگیرد که کدام ویژگی در مرحله تقسیم وجود داشته باشد، برای این منظور یک شاخص معادل split وجود دارد در مقاله منبع روش GINI INDEX استفاده شده است. اما در این مقاله ما از gain ratio در روش جنگل تصادفی استفاده می کنیم این پارامتر در درخت تصمیم گیری نحوه انتخاب سطوح ریشه و گره را با توجه به مستندات سند تعیین می کند.

تست

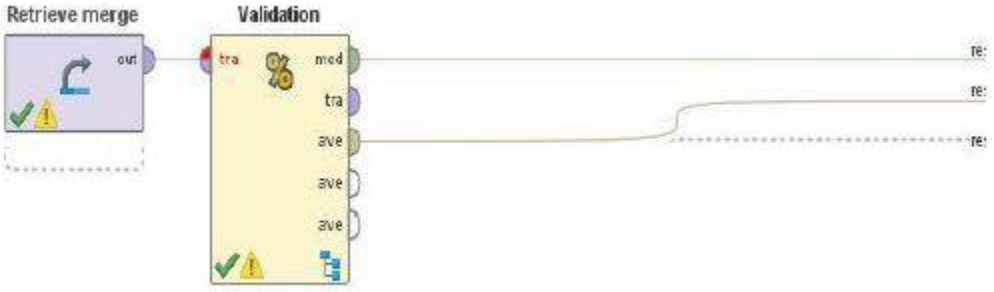
برای تست نتیجه ما ازCROSS VALIDATION متقابل استفاده می کنیم، در این روش، داده ها به صورت تصادفی به زیرمجموعه یا برابر می شوند.و این آموزش و آزمایش چندین بار انجام می شود.

یک پارتیشن به عنوان مجموعه تست ذخیره می شود و پارتیشن باقی مانده در مجموع برای آموزش شکل ۲ استفاده می شود.

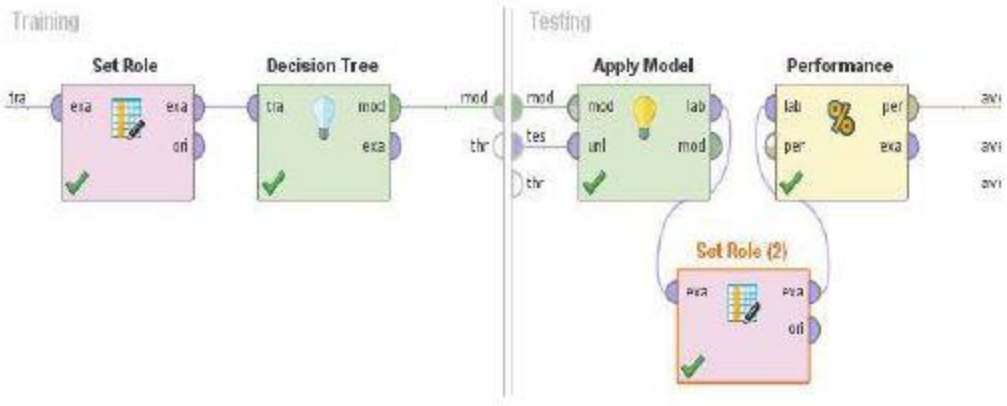


شکل ۲، CROSS VALIDATION و تست با استفاده از KNIME

جنگل تصادفی برای طبقه بندی داده ها استفاده می شود برای ساخت یک طبقه بندی درخت تصمیم، به طور تصادفی هر گره را انتخاب کنید. پیش بینی ویژگی جدیدی از پیش بینی را با ویژگی هدف ما فراهم می کند، با استفاده از X Aggregator، جدول پیش بینی و نرخ خطا را از اعتبارسنجی متقابل پیدا می کنیم، ما از آمار برای پیدا کردن و مشاهده ارزش آماری استفاده می کنیم. شکل ۳ نشان می دهد که بخشی از مدل ما را آزمایش می کند.

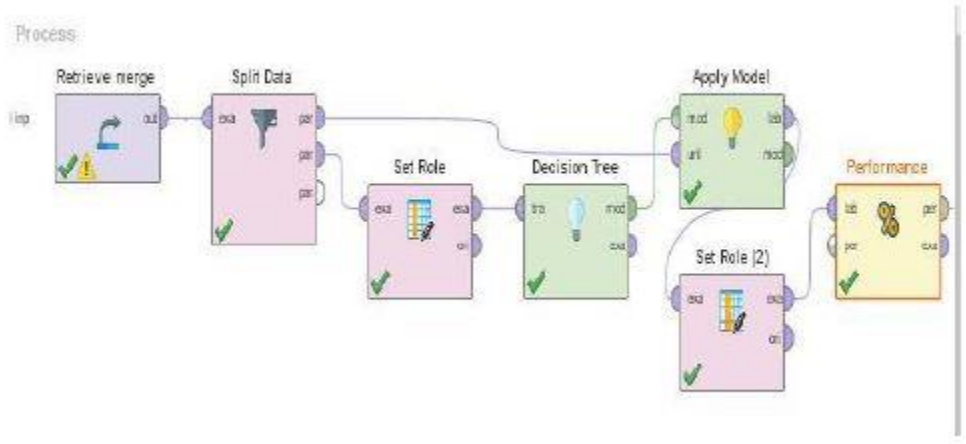


شکل ۳، X-VALIDATION در درخت تصمیم

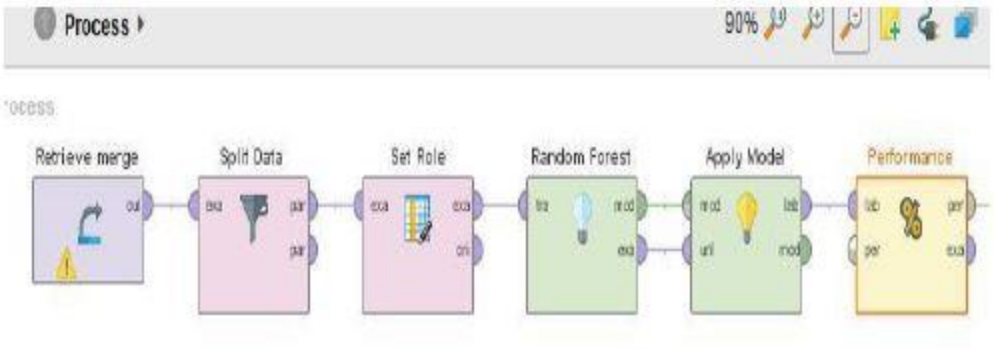


شکل ۴، اعتبار سنجی در درخت تصمیم گیری

در این مقاله برای آزمایش مدل (جنگل تصادفی، درخت تصمیم گیری) و یافتن یک روش موثر بر اساس آمار، ما از دادههای تقسیم شده استفاده کردیم، به این ترتیب ما ۷۰٪ از داده ها را به عنوان اموزش و ۳۰٪ از داده ها به عنوان یک آزمون محاسبه می کنیم،برای این کار روی جعبه داده تقسیم کلیک کنید و مقادیر ۰،۷ و ۰،۳ را به بخش پارتیشن اضافه کنید سپس درخت تصمیم گیری و جنگل تصادفی قرار داده و برچسب آنها را Alc قرار دهید. ما اکنون باید این پروژه را مدل کنیم بنابراین جعبه مدل را اورده و جعبه کارایی را در آن قرار دهید و در قسمت پارامترها ما مشخص میکنیم که نتیجه موضوع را چگونه نمایش دهیم و در نهایت فرم نهایی به شرح زیر است:

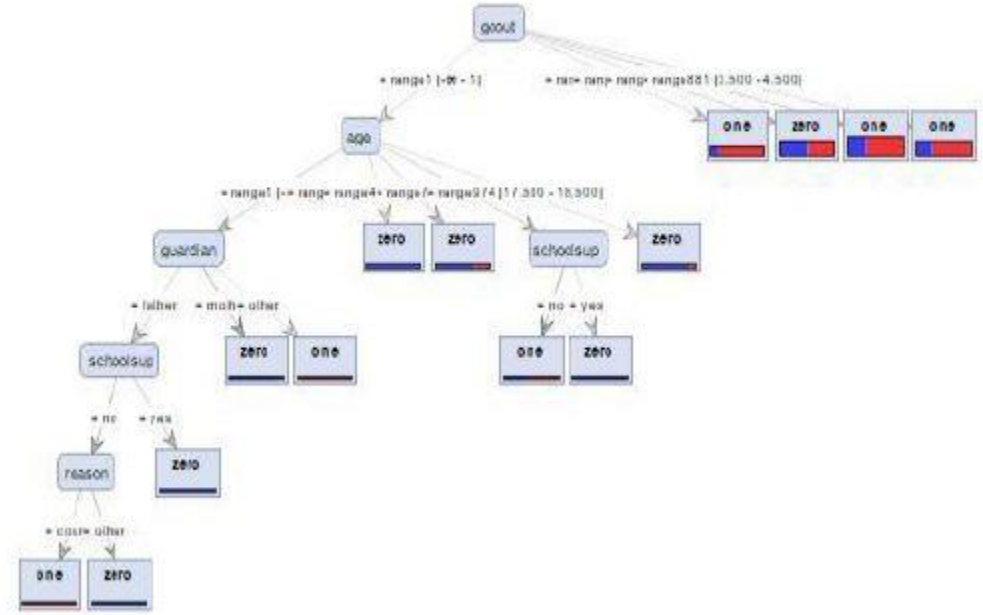


شکل ۵، تصویر نهایی از درخت تصمیم با روش SPLIT DATA

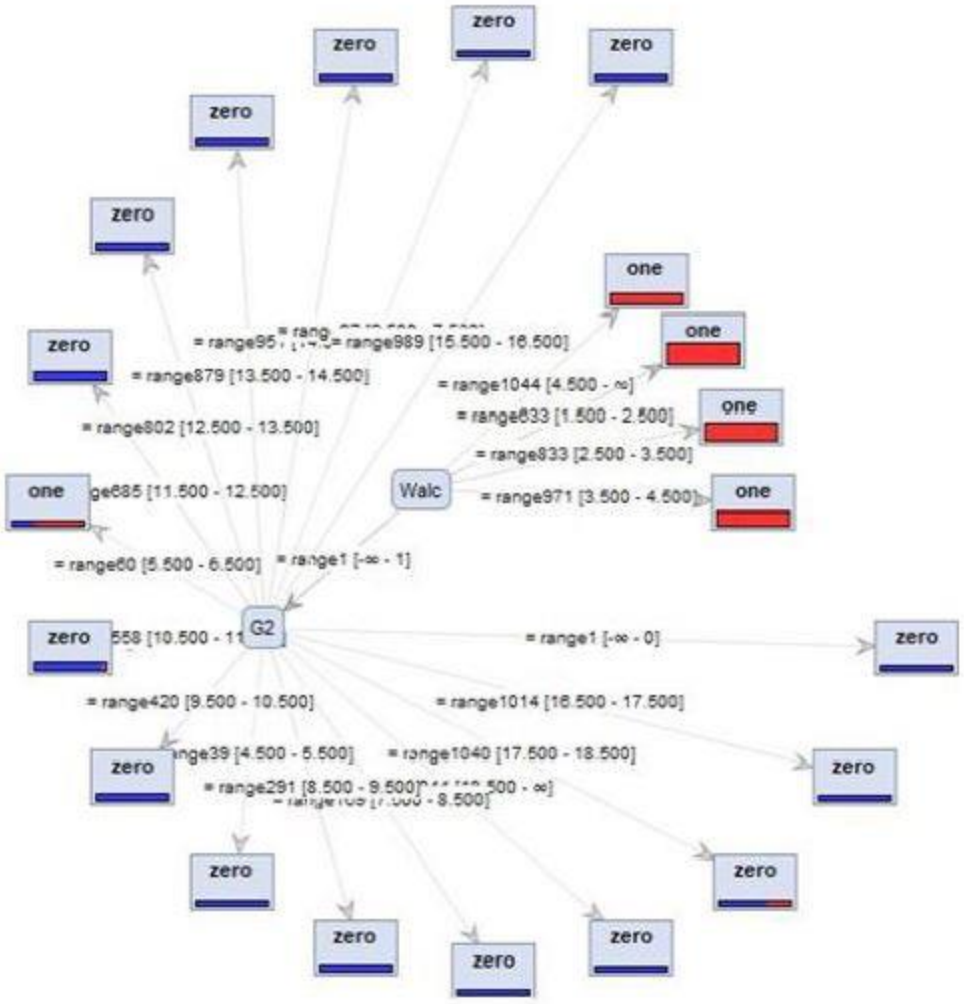


شکل ۶، تصویر نهایی از جنگل تصادفی با روش SPLIT DATA

نمونه درخت ها



شکل ۷.نمونه جنگل تصادفی



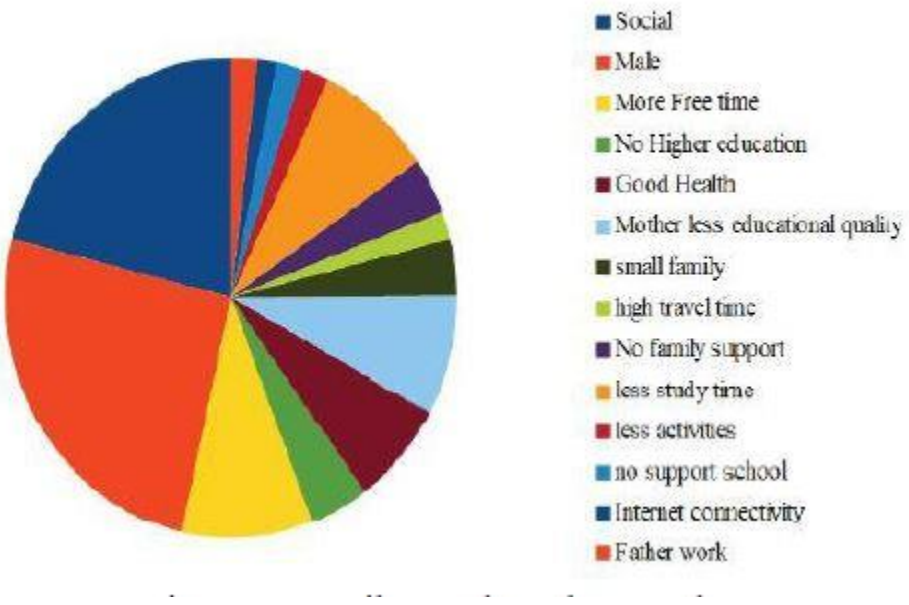
شکل ۸، نمونه جنگل تصادفی

نتایج و بحث

در مطالعه حاضر، نقطه شروع ما، آماده سازی داده ها بود فرایندی که در آن زمینه های جدید مطابق با داده های قبلی ایجاد شد و برای فرآیند آموزش و طبقه بندی مدل استفاده شد. پس از آماده سازی، داده ها کامل شد و ما جنگل تصادفی را برای به دست آوردن بهترین دقت و نتایج استفاده کردیم. فقط ما تنها آن درختی را برمیداریم که دقت خوبی در رابطه با پیش بینی ما دارد.

از آن درختان ما چندین الگوی مشترک را یافتیم گروهی از ویژگی ها را امتحان کردیم و دیدیم که آنها چگونه تاثیر می گذارند (البته بعضی از درخت ها را برای مقدار قابل توجهی نادیده می گیریم).

درصد تاثیر دقت براساس نمودار در شکل ۹ نشان داده شده است و همچنین با جدول در شکل ۱۰،



شکل ۹ ، صفات نشان داده شده با استفاده از گراف

Most impacted attribute

Attribute	Percentage
Male	25.35%
Social	21.13%
More Free time	9.39%
Less study time	8.45%
Mother less educational quality	7.98%
Good Health	7.04%
No Higher education	4.23%
No family support	3.76%
Small family	3.76%
High travel time	1.88%
Less activities	1.88%
No support school	1.88%
Father work	1.88%
Internet connectivity	1.41%

شکل ۱۰، بیشترین ویژگی تاثیر گذار

بر اساس این نتایج مردان بیشتر درگیر الکل هستند و برای تأیید این موضوع ، ما در وب سایت جستجو می کنیم و برخی از کارهای مرتبط با سازمان بهداشت جهانی را پیدا می کنیم که نوشیدن بیش از اندازه زن در گزارش الکل سال ۲۰۱۴ را نشان می دهد ، با پشتیبانی از این مطالعه، بزرگ بودن خانواده و یک کار خوب برای پدر برای رشد کودک اهمیت دارد و همچنین آموزش مادری یک اصل اساسی در زندگی است.

مطالعه ما نشان می دهد که، فقدان برخی خصوصیات شانس بیشتر برای معتاد شدن با الکل برای کودک را بهمراه دارد، بعدا شما می توانید نتایج حاصل از فرایندهای ذکر شده را در روشهای زیر را مشاهده کنید . در شکل ۳و۴، نتایج بدست آمده توسط X-VALIDATION در الگوریتم درخت تصمیم با درصد زیر برای هر پارامتر اورده شده است:

%۱۰۰(class recall(true zero)) ,

100% (class recall(true one)) ,

100% (class precision(pred. zero)) ,

100% (class precision(pred.one)) ,

100% (accuracy)

در شکل ۵ نتایج حاصل از split data در الگوریتم درخت تصمیم با درصد زیر برای هر پارامتراورده شده است:

100% (class recall(true zero)) ,

98.69% (class recall(true one)) ,

97.85% (class precision(pred. zero)) ,

100% (class precision(pred.one)) ,

99.18% (accuracy)

در شکل ۶ نتایج حاصل از split data در الگوریتم جنگل تصادفی با درصد زیر برای هر پارامتراورده شده است:

100% (class recall(true zero)) ,

100% (class recall(true one)) ,

100% (class precision(pred.zero)) ,

100% (class precision(pred.one)) ,

100% (accuracy)

بعد از این تست ما با بازدهی ۱۰۰٪ در روش جنگل تصادفی مواجه شدیم، و این یک نوآوری کارآمد در پروژه ما است.

منابع

[1] Drinkaware.co.uk. Why underage drinking is a risky business.

[2] Fabio Pagnotta, Mohammad Amran Hossain, USING DATA MINING TO PREDICT SECONDARY SCHOOL STUDENT ALCOHOL CONSUMPTION,2017

[3] National Institute on Alcohol Abuse and Alcohlism(NIH). Alcohol's effects on the body.

[4] Cortez, Paulo Silva, Alice Maria Gonçalves , Using data mining to predict secondary school student performance, Apr-2008, https://repositorium.sdum.uminho.pt/handle/1822/8024

[5] Fabio Mendoza, Roberto Morales, Ubaldo Martinez, Alexis Kevin De la Hoz Manotas, Designing A Method for Alcohol Consumption Prediction Based on Clustering and Support Vector Machines, April 2017

https://www.researchgate.net/publication/317348706_Designing_A_Method_for_Alcohol_Consumption_Prediction_Based_on_Clustering_and_Support_Vector_Machines

[6] Fabio Pagnotta, Hossain Amran, using data mining to predict secondary school student alcohol consumption, February 2016, https://www.researchgate.net/publication/296695247_USING_DATA_MINING_TO_PREDICT_SECONDARY_SCHOOL_STUDENT_ALCOHOL_CONSUMPTION

[7] P. Cortez and A. Silva. Using data mining to predict secondary school student performance.[in a.brito and j. teixeira eds. proceedings of 5th future business technology conference (f ubutec2008) pp:5-12 porto portugal]. 2008.

[8] Sepideh Hassankhani Dolatabadi, Farshid Keynia, Designing of Customer and Employee Churn Prediction Model Based on DataMining Method and Neural Predictor,2017

[9] Clustering Algorithms and Bayesian Networks for Distributed Geospatial Data Mining and Knowledge Discovery, 2018; 1(1): 1-8