**به نام خدا**

**پروژه داده کاوی**

بررسی دیتاست پیشگیری از بارداری

1. **مقدمه**

در این تحقیق به دنبال اعمال فرآیندهای داده کاوی بر روی مجموعه ای از داده ها هستیم و سعی می کنیم تا با اعمال روش های مختلف داده کاوی، نتایج قابل قبولی را ارائه دهیم. در ادامه به معرفی مجموعه داده انتخابی می پردازیم و بعد از آن، فرآیندهای تصویرسازی و مدل سازی برای استخراج اطلاعات مفید از مجموعه داده را در پیش می گیریم.

1. **معرفی مجموعه داده**

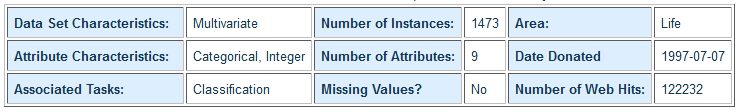
مجموعه داده مورد استفاده ما در این تحقیق، با عنوان مجموعه داده پیشگیری از بارداری شناخته می شود که این مجموعه داده در کشور اندونزی جمع آوری گردیده است. در این مجموعه داده سعی شده است که با پرسش از زوج های مختلف مشخص کنند که چه ویژگی هایی در استفاده و یا عدم استفاده از روش های پیشگیری از بارداری موثر هستند.

این مجموعه داده از آدرس زیر قابل دسترسی می باشد.

<https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice>

سایت UCI خلاصه ی کوتاهی از اطلاعات این مجموعه داده را ارائه داده که در جدول 1 مشاهده می کنیم.

جدول1: اطلاعات مجموعه داده



این مجموعه داده شامل، 1473 رکورد و 9 ویژگی می باشد. همچنین این مجموعه داده فاقد مقادیر گم شده است که این موضوع فرآیندهای داده کاوی را تسهیل می نماید.

ویژگی های مطرح شده در این مجموعه داده در ادامه توضیح داده می شود.

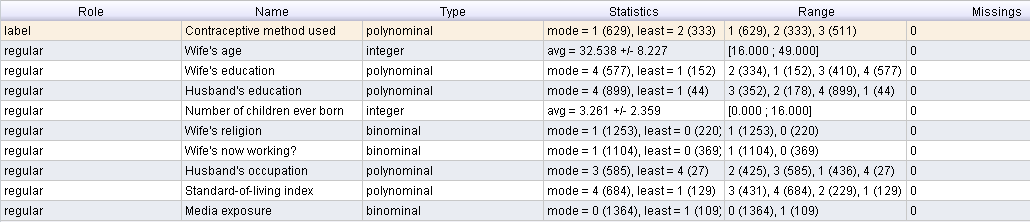
* سن زن (Wife’s age): یک ویژگی عددی است که سن زن را نشان می دهد.
* تحصیلات زن (Wife’s education): این ویژگی تصحیلات زن را از شماره یک تا چهار (کم تا زیاد) نشان می دهد.
* تحصیلات شوهر (Husband’s education): این ویژگی تصحیلات شوهر را از شماره یک تا چهار (کم تا زیاد) نشان می دهد.
* تعداد بچه هایی که تا به حال به دنیا آورده اند (Number of children ever born)
* مذهب زن (Wife’s religion): این ویژگی به دو دسته مسلمان (1) و غیر مسلمان (0) تقسیم شده است.
* وضعیت کاری زن (Wife’s now working?): برای حالتی که زن کار می کند این ویژگی صفر و در غیر این صورت یک می باشد.
* وضعیت شغلی شوهر (Husband’s occupation): این ویژگی از یک تا چهار شماره گذاری شده است، اما توضیحی برای آن قید نشده است.
* شاخص کیفیت زندگی (Standard-of-living index): این ویژگی از یک تا چهار به معنی کم تا زیاد دسته بندی شده است.
* دسترسی به رسانه ها (Media exposure): در صورتی که دسترسی مناسبی وجود داشته باشد، این ویژگی صفر و در غیر این صورت یک می باشد.
* روش پیشگیری از بارداری (Contraceptive method used): این ویژگی، ویژگی برچسب یا Label مجموعه داده می باشد که سه مقدار یک (به معنی عدم استفاده)، دو (استفاده طولانی مدت) و سه (استفاده کوتاه مدت) را می پذیرد.

1. **پاکسازی داده ها**

فرآیند پاکسازی داده ها یکی از مهم ترین قسمت های داده کاوی می باشد، به طوری که بدون اعمال آنها نتایج مناسبی را نمی توان بدست آورد. در پاکسازی داده ها کارهای مختلفی صورت می گیرد. از جمله ی این فعالیت ها، حذف مقادیر گم شده می باشد، ار آنجا که مجموعه داده ما فاقد مقادیر گم شده می باشد نیازی به اعمال این فرآیند نیست. اما از آنجایی که مجموعه داده ی ما مقادیر عددی نظیر سن زن و تعداد فرزندان را دارد باید با سیاست مناسبی آنها را تبدیل به دسته های مختلفی نمود (گسسته سازی) تا بتوان هنگام تصویرسازی داده ها نتایج قابل تفسیری بدست آورد.

جدول 2 فراداده ی مجموعه داده ی اولیه ما نشان می دهد. فرا داده حاوی اطلاعات مفیدی نظیر ویژگی های مجموعه داده ها ، نوع آنها، میانه آنها، تعداد مقادیر گم شده و .... می باشد.

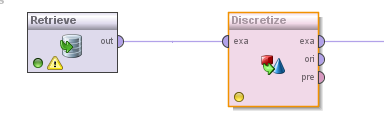
جدول2: فراداده



در ادامه به روش گسسته سازی پرداخته می شود.

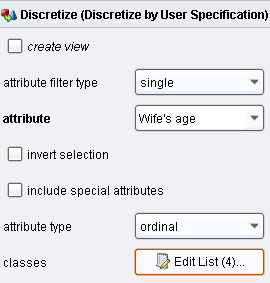
* 1. گسسته سازی سن زن

همان طور که بیان شد، ویژگی سن زن یک ویژگی عددی می باشد که برای تفسیر بهتر نیاز است آن را به ویژگی چندمقداری تبدیل کرد. برای این منظور از عملگر گسسته سازی استفاده می توان نمود. روش استفاده از این عملگر در شکل 1 نشان داده شده است.



شکل1: روش گسسته سازی سن زن

پارامترهای لازم برای عملگر گسسته سازی در شکل 2 و جدول 3 تنظیم شده است.



شکل2: پارامترهای لازم برای گسسته سازی سن زن

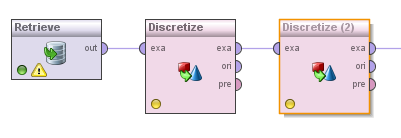
در جدول 3، چهار گروه سنی نوجوان(تا 18 سال)، جوان(تا 35 سال)، میان سال(تا 50 سال) و پیر(تا 100 سال) تعریف شده است.

جدول3: پارامترهای لازم برای گسسته سازی سن زن



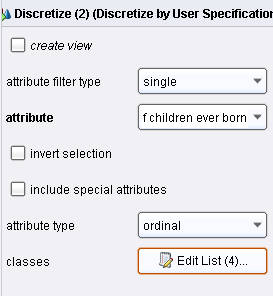
* 1. گسسته سازی تعداد فرزندان متولد شده

شکل 3 روش گسسته سازی تعداد فرزندان متولد شده را نشان می دهد.



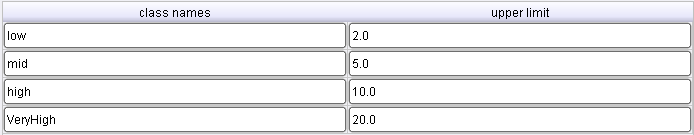
شکل3: روش گسسته سازی تعداد فرزندان متولد شده

پارامترهای لازم برای گسسته سازی تعداد فرزندان متولد شده در شکل 4 و جدول 4 بیان شده است.



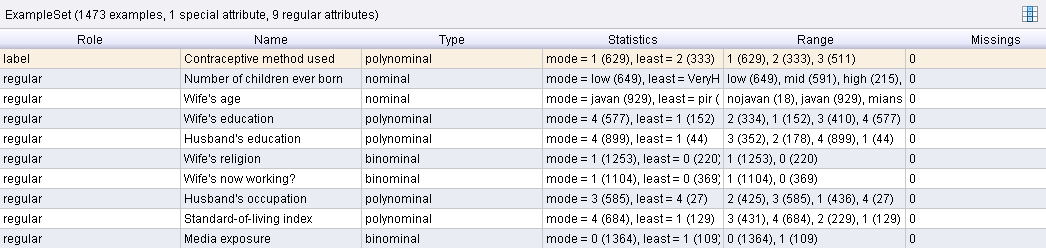
شکل4: پارامترهای لازم برای گسسته سازی تعداد فرزندان متولد شده

جدول4: پارامترهای لازم برای گسسته سازی تعداد فرزندان متولد شده



فراداده ی مجموعه داده جدید بعد از اعمال گسسته سازی در جدول 5 نشان داده شده است. همان طور که دیده می شود این فراداده فاقد ویژگی های عددی می باشد و همه ی ویژگی ها، تبدیل به ویژگی های اسمی و چند مقداری شده اند.

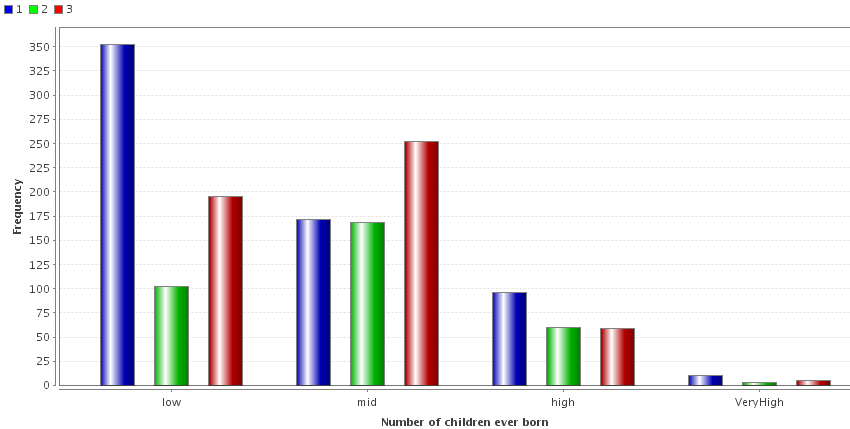
جدول5: فرداده بعد از پاکسازی داده ها



1. **تصویرسازی**

فرآیند تصویرسازی کمک می کند که درک بهتری نسبت به مجموعه داده ی مورد بررسی بدست آوریم. برای همین منظور ما اقدام به تصویرسازی بر اساس ویژگی برچسب نمودیم که نتایج آن در ادامه قید می شود.

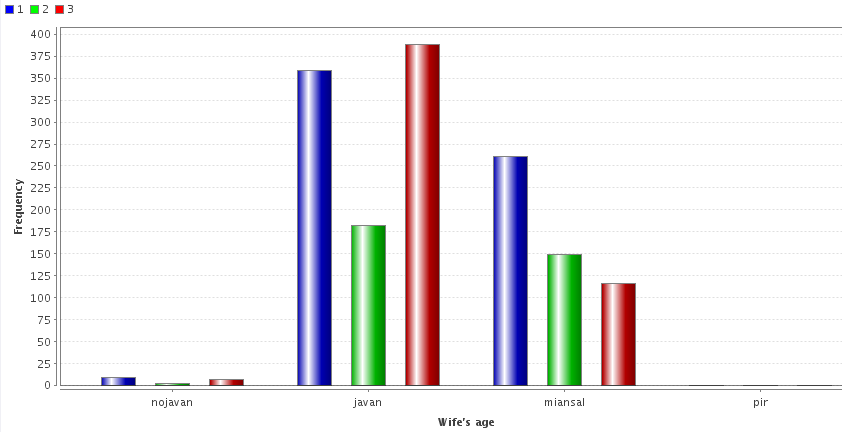
**شکل 5** رابطه ی بین استفاده از روش های پیشگیری از بارداری و تعداد فرزندان متولد شده را نشان می دهد.



شکل5: رابطه ی بین استفاده از روش های پیشگیری از بارداری و تعداد فرزندان متولد شده در نمودار histogram color

این نمودار نظم چندان قابل توجه ای را ارائه نمی دهد.

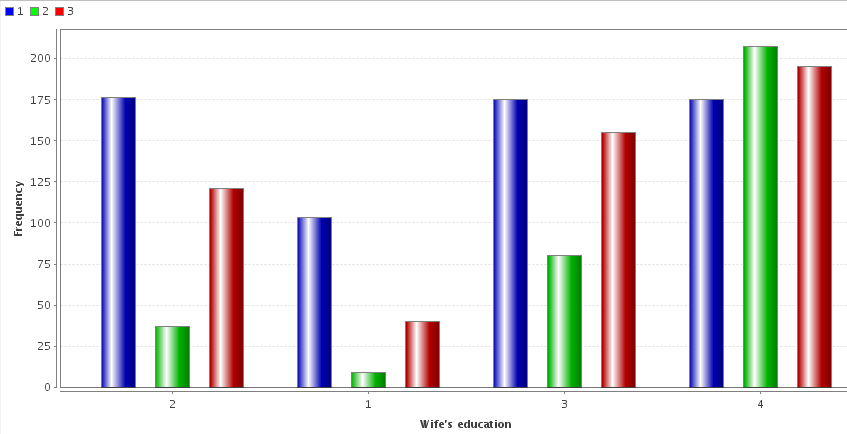
**شکل 6** رابطه ی بین استفاده از روشهای پیشگیری از بارداری و سن زن را نشان می دهد.



شکل6: رابطه ی بین استفاده از روشهای پیشگیری از بارداری و سن زن در نمودار histogram color

این نمودار حاکی از آن است که در گروه سنی جوان، افراد یا از روش های پیشگیری از بارداری استفاده نمی کنند (میله آبی) یا از روش های کوتاه مدت بهره می برند (میله قرمز). اطلاعات دیگری که از این نمودار قابل استخراج است این می باشد، بیشترین روش پیشگیری از بارداری در مجموع روش های کوتاه مدت می باشد، هر چند در افراد میانسال این نسبت به نفع روش های بلند مدت تغییر می کند.

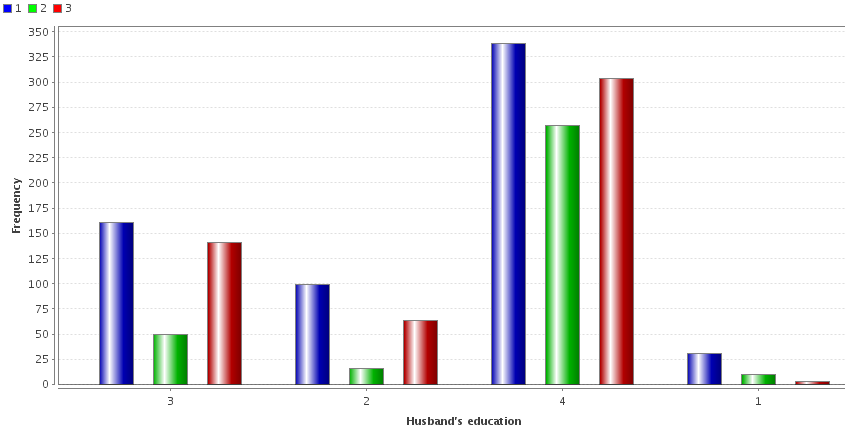
**شکل 7** رابطه ی بین استفاده از روشهای پیشگیری از بارداری و تحصیلات زن را نشان می دهد.



شکل7: رابطه ی بین استفاده از روشهای پیشگیری از بارداری و تحصیلات زن در نمودار histogram color

همان طور که دیده می شود در گروه چهار، (گروهی با تحصیلات بالا) فراوانی افراد تقریبا یکسان است. در گروه یک (گروه با تحصیلات پایین) تعداد افرادی که از روش های بلند مدت پیشگیری از بارداری استفاده می نمایند بسیار کم است که شاید به علت عدم آشنایی یا عدم اعتماد آنها به این روش ها باشد. هر چه میزان تحصیلات بیشتر می شود، میزان استفاده از روش های بلند مدت پیشگیری از بارداری افزایش قابل توجهی می نماید.

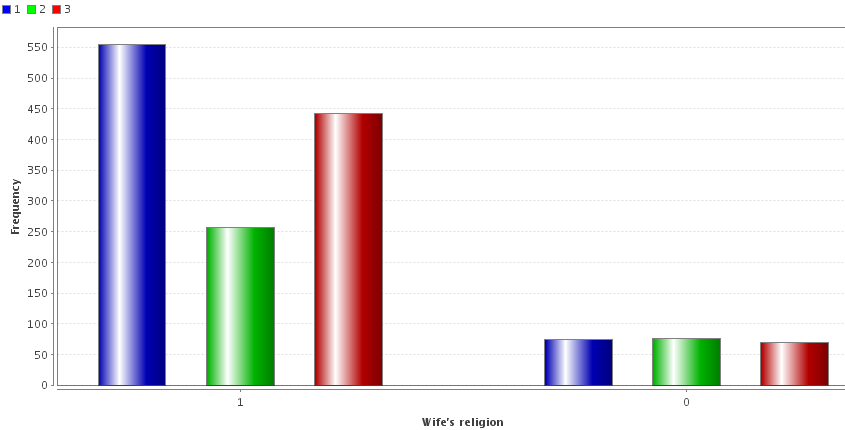
**شکل 8** رابطه ی بین استفاده از روشهای پیشگیری از بارداری و تحصیلات شوهر را نشان می دهد.



شکل8: رابطه ی بین استفادهاز روشهای پیشگیری از بارداری و تحصیلات شوهر در نمودار histogram color

نکته ی جالب در مورد این نمودار، شباهت زیاد آن با نمودار قبلی است. ظاهرا میزان تحصیلات روی زن و شوهر تاثیرات تقریبا یکسانی بر روی نگاه آنها به روش های پیشگیری از بارداری می گذارد.

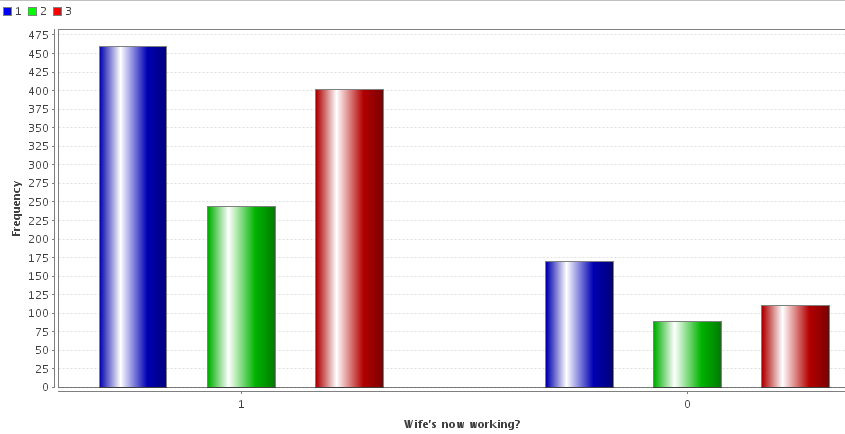
**شکل 9** رابطه ی بین استفاده از روشهای پیشگیری از بارداری و مذهب زن را نشان می دهد.



شکل9: رابطه ی بین استفاده از روشهای پیشگیری از بارداری و مذهب زن در نمودار histogram color

گروه یک، مربوط به مسلمانان می شود و گروه صفر مربوط به غیر مسلمانان. در نگاه اول این نمودار نشان می دهد که بیشتر افراد مورد پرسش قرار گرفته از مسلمانان هستند که با توجه بافت جمعیتی اندونزی دور از ذهن نمی باشد. اطلاعات دیگری که از این نمودار قابل تفسیر می باشد این است که بین غیر مسلمانان نسبت استفاده و عدم استفاده از روش های پیشگیری از بارداری تقریبا یکسان است و غیر مسلمانان بیشتر از مسلمانان از روش های پیشگیری از بارداری استفاده می کنند.

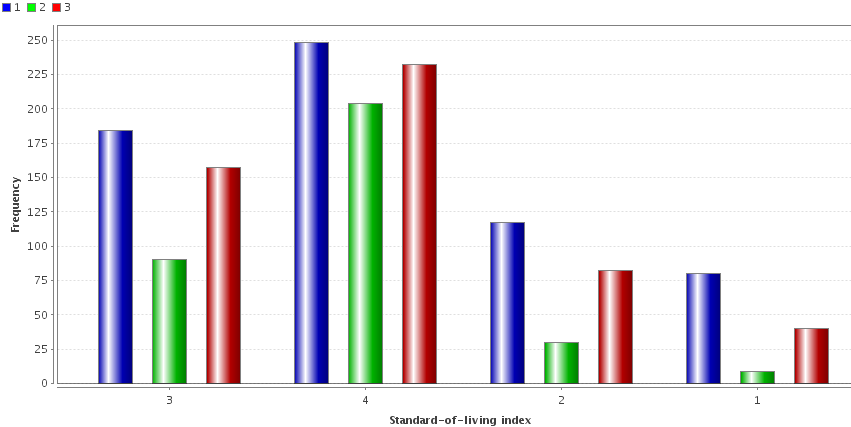
**شکل 10** رابطه ی بین استفاده از روشهای پیشگیری از بارداری و وضعیت شغلی زن را نشان می دهد.



شکل10: رابطه ی بین استفاده از روشهای پیشگیری از بارداری و وضعیت شغلی زن در نمودار histogram color

مشابه سایر نمودارها، این نمودار نیز حاکی از آن است که اکثر افراد تمایل چندانی به روش های بلند مدت پیشگیری ندارند.

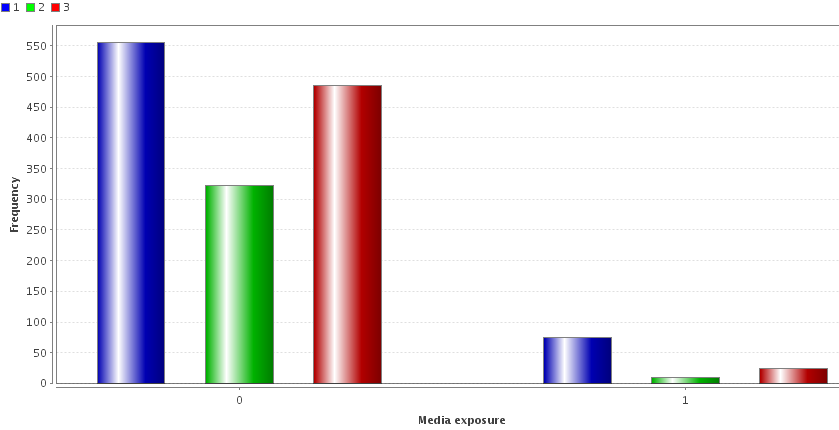
**شکل 11** رابطه ی بین استفاده از روشهای پیشگیری از بارداری و شاخص کیفیت زندگی را نشان می دهد.



شکل11: رابطه ی بین استفاده از روشهای پیشگیری از بارداری و شاخص کیفیت زندگی در نمودار histogram color

این نمودار نشان می دهد که بیشتر افراد مورد بررسی قرار گرفته از شاخص زندگی مناسبی برخوردار بوده اند. نسبت عدم استفاده از روش های پیشگیری از بارداری و روش های کوتاه مدت و بلند مدت در همه ی گروه ها تقریبا یکسان است به جز اینکه هر چه کیفیت زندگی بالاتر می رود استفاده از روش های بلند مدت نیز بیشتر مورد توجه افراد قرار می گیرد.

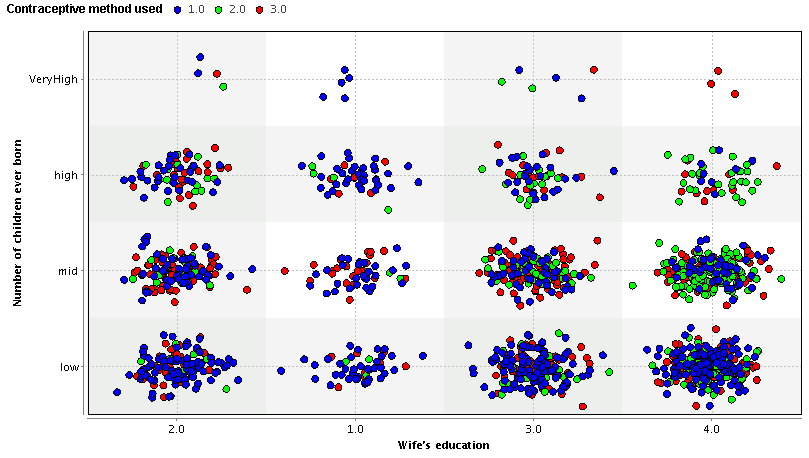
**شکل 12** رابطه ی بین استفاده از روشهای پیشگیری از بارداری و میزان دسترسی به رسانه ها را نشان می دهد.



شکل12: رابطه ی بین استفاده از روشهای پیشگیری از بارداری و میزان دسترسی به رسانه ها در نمودار histogram color

این نمودار نشان می دهد که بیشتر افراد مورد بررسی دسترسی خوبی به رسانه ها داشته اند که گروه صفر دسته بندی شده اند.

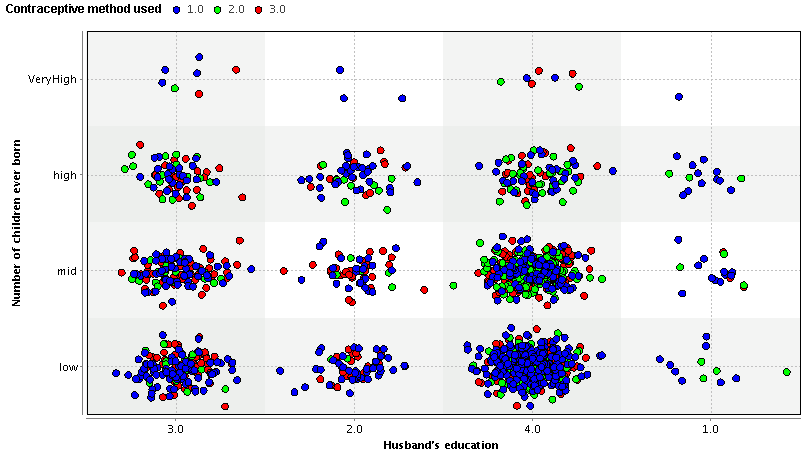
**شکل 13** رابطه ی بین استفاده از روشهای پیشگیری از بارداری و آموزش همسر و تعداد فرزندان را نشان می دهد.



شکل13: رابطه ی بین استفاده از روشهای پیشگیری از بارداری و آموزش همسر و تعداد فرزندان در نمودار scatter

این نمودار نشان می دهد که متناسب با افزایش تعداد فرزندان روش های دائمی پیشگیری از بارداری افزایش پیدا میکند که در ردیف high قابل تشخیص می باشد. نکته قابل توجه در این نمودار این است که تعداد زنان با آموزش بالا وتعداد فرزندان متوسط که روش دائمی پیشگیری را انتخاب نمودن زیاد می باشد که تاثیر آموزش را در این انتخاب را نشان میدهد.

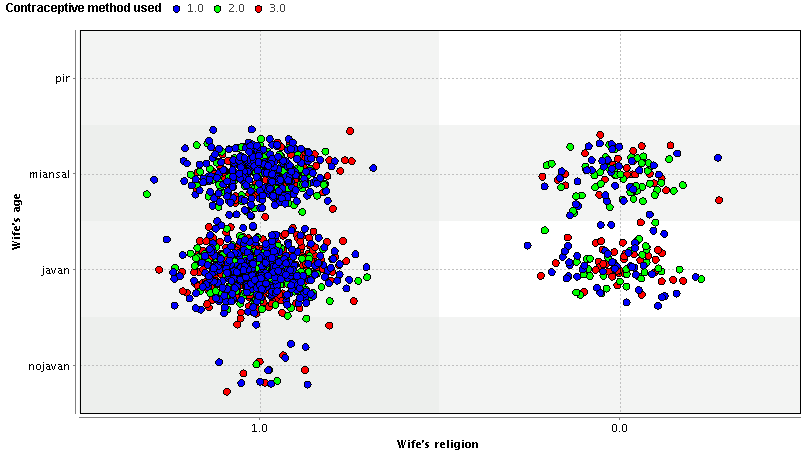
**شکل 14** رابطه ی بین استفاده از روشهای پیشگیری از بارداری و آموزش شوهر و تعداد فرزندان را نشان می دهد.



شکل14: رابطه ی بین استفاده از روشهای پیشگیری از بارداری و آموزش شوهر و تعداد فرزندان در نمودار scatter

این نمودار با توجه به مشابهت آن با نمودار بالایی تاثیرات یکسان آموزش را در زن و مرد نشان میدهد.

**شکل 15** رابطه ی بین استفاده از روشهای پیشگیری از بارداری و مذهب و سن زن را نشان می دهد.

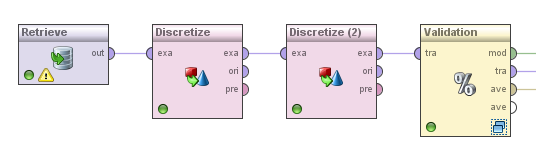


شکل15: رابطه ی بین استفاده از روشهای پیشگیری از بارداری مذهب و سن زن در نمودار scatter

این نمودار نشان میدهد در قشر مسلمان توجه به مسائل پیشگیری نسبت به غیر مسلمان کمتر می باشد.

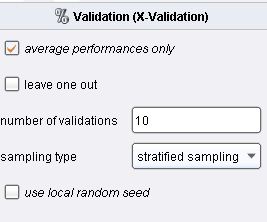
1. **مدل سازی**

در این بخش، به مدل سازی از مجموعه داده مورد نظر می پردازیم. برای مدل سازی نیاز است که از بخشی از مجموعه داده برای آموزش و ساخت مدل بهره برد و از بخش دیگر برای تست مدل ساخته شده استفاده کرد. برای این منظور می توان از عملگر Cross Validation بهره برد. نحوه ی استفاده از این عملگر در شکل 16 نشان داده شده است.



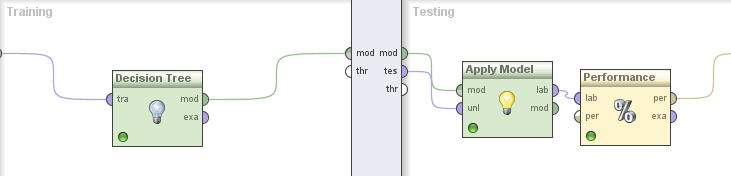
شکل16: روش مدل سازی

پارامترهای لازم برای عملگر Cross Validation در شکل 17 نشان داده شده است.



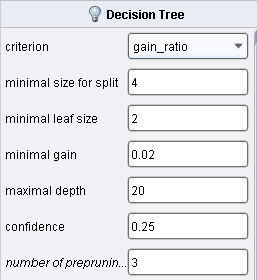
شکل17: پارامترهای لازم برای عملگر Cross Validation

داخل عملگر Cross Validation باید مجموعه ای از عملگرها برای ساخت مدل و تست آن قرار گیرد که ما در نخستین گام از عملگر درخت تصمیم برای این منظور استفاده نمودیم.



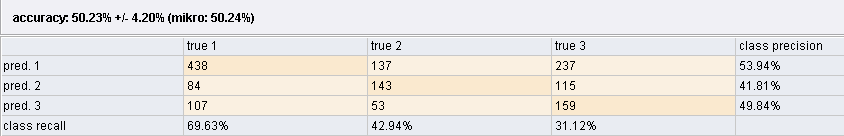
شکل18: عملگرهای داخلی Cross Validation

پارامترهای لازم برای عملگر درخت تصمیم در شکل 19 نشان داده شده است.



شکل19: تنظیمات لازم برای درخت تصمیم

نتیجه ی استفاده از عملگر درخت تصمیم در شکل 20 نشان داده شده است. مدل ساخته شده دارای دقت تقریبا 50 درصد می باشد که دقت چندان بالایی نمی باشد.

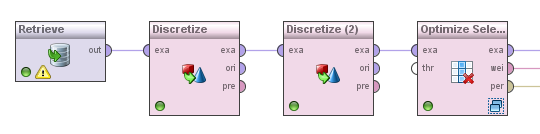


شکل20: دقت مدل ساخته شده با درخت تصمیم

برای اینکه نتایج بهتری بدست آید، می توان تعداد ویژگی مورد بررسی را کاهش داد، برای این منظور باید بهترین زیر مجموعه از ویژگی ها یا به عبارتی تاثیرگذارترین آنها را انتخاب نماییم.

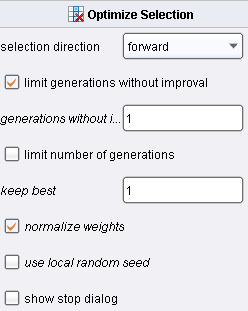
* 1. انتخاب بهترین زیرمجموعه از ویژگی ها:

برای انتخاب بهترین زیرمجموعه از ویژگی ها می توان به صورت مطرح شده در شکل 21 عمل نمود. عملگر Optimize Selection برای این منظور طراحی شده است. در داخل این عملگر باید یک عملگر ساخت مدل و ارزیابی قرار گیرد که می توان از Cross Validation به همراه درخت تصمیم بهره برد.



شکل21: روش اعمال انتخاب بهترین زیرمجموعه از ویژگی ها

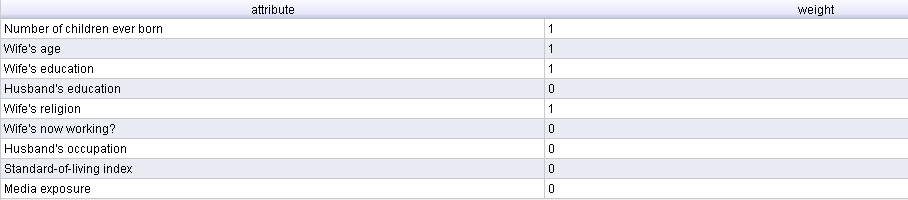
پارامترهای لازم برای عملگر Optimize Selection در شکل 22 نشان داده شده است.



شکل22: تنظیمات لازم برای عملگر Optimize Selection

نتیجه ی استفاده از عملگر Optimize Selection در شکل زیر نشان داده شده است. از بین تمام ویژگی های مورد بررسی در این مجموعه داده تنها 4 ویژگی مورد انتخاب قرار گرفته اند که در جدول 6، با وزن یک نشان داده شده اند.

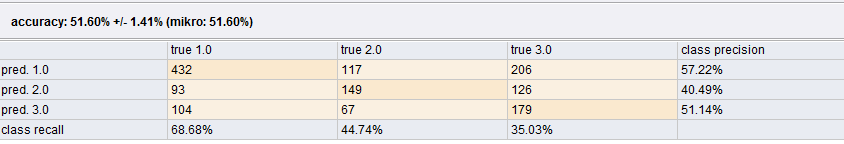
جدول6: بهترین زیرمجموعه از ویژگی ها



* 1. مدل سازی مجدد با درخت تصمیم

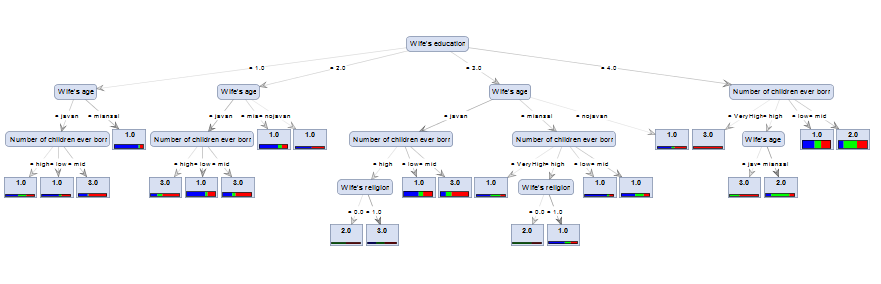
بعد از انتخاب بهترین زیرمجموعه از ویژگی ها، مجددا اقدام به ساخت مدل با درخت تصمیم نمودیم که نتیجه ی آن در جدول 7 نشان داده شده است.

جدول7: دقت مدل ساخته شده با درخت تصمیم



در مدل جدید، مقداری دقت حاصل از کار افزایش یافته و به بیش از 51 درصد رسیده است.

درخت ساخته شده با استفاده از مدل ذکر شده در شکل 23 نشان داده شده است. این درخت نشان می دهد که تحصیلات زن که در ریشه ی درخت قرار گرفته موثرترین عامل برای تفکیک افراد از نظر استفاده از روش های پیشگیری از بارداری می باشد.

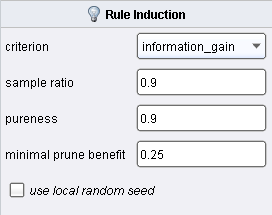


شکل23: درخت تصمیم

* 1. مدل سازی با استنتاج قوانین

همان طور که بیان شد، مدل های مختلفی می توان برای یادگیری و ارزیابی ساخت. یکی از مدل ها که نتایج جالبی ارائه می دهد مدل Rule Induction می باشد. این عملگر، مجموعه قوانین متعددی را می سازد و بر اساس آن ها اقدام به ارزیابی مجموعه داده می نماید.

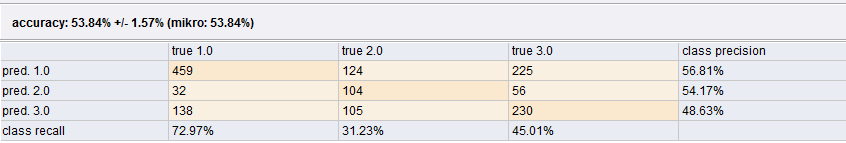
پارامترهای لازم برای این عملگر در شکل 24 نشان داده شده است.



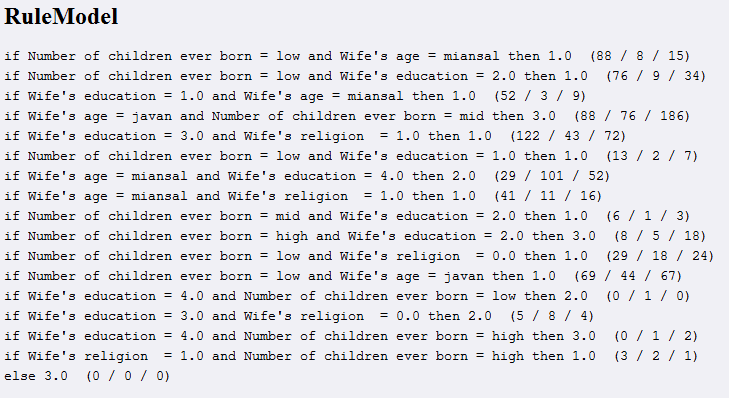
شکل24: تنظیمات لازم برای عملگر Rule Induction

نتیجه ی استفاده عملگر Rule Induction در جدول 8 نشان داده شده است که دقت 53 درصدی را نشان می دهد.

جدول8: دقت مدل ساخته شده با عملگر Rule Induction



شکل 25 مجموعه قوانین حاصل از اعمال عملگر Rule Induction را نشان می دهد.

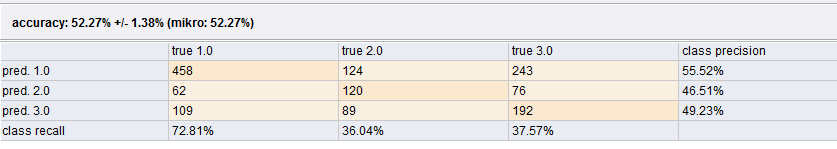


شکل25: مجموعه قوانین استخراج شده

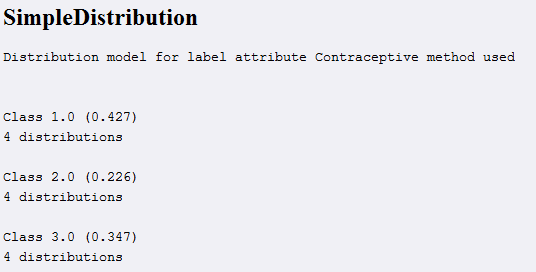
* 1. مدل سازی با الگوریتم Naive Bayes

نتیجه ی استفاده عملگر Naive Bayes در جدول 9 نشان داده شده است که دقت 52 درصدی را نشان می دهد.

جدول9: دقت مدل ساخته شده با عملگر Naive Bayes



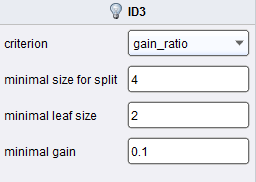
شکل 26 مجموعه توزیع ها ی حاصل از اعمال عملگر Naive Bayes را نشان می دهد.



شکل26: مجموعه توزیع ها ی حاصل

* 1. مدل سازی با الگوریتم ID3

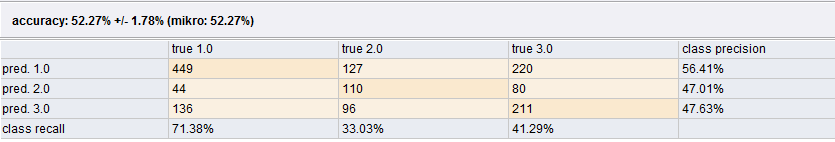
پارامترهای لازم برای این عملگر در شکل 27 نشان داده شده است.



شکل27 : تنظیمات لازم برای عملگر ID3

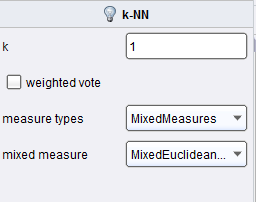
نتیجه ی استفاده عملگر ID3 در جدول 10 نشان داده شده است که دقت 52 درصدی را نشان می دهد.

جدول10: دقت مدل ساخته شده با عملگر ID3



* 1. مدل سازی با الگوریتم K-NN

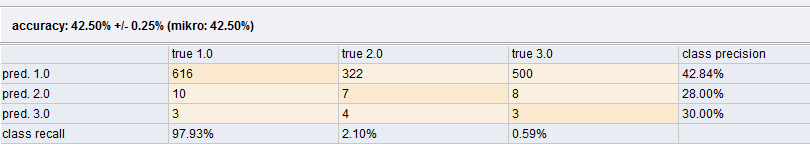
پارامترهای لازم برای این عملگر در شکل 28 نشان داده شده است.



شکل28 : تنظیمات لازم برای عملگر K-NN

نتیجه ی استفاده عملگر K-NN در جدول 11 نشان داده شده است که دقت 42 درصدی را نشان می دهد که دقت مناسبی نمی باشد.

جدول11: دقت مدل ساخته شده با عملگر K-NN



1. **جمع بندی**

در این تحقیق به بررسی مجموعه داده پیشگیری از بارداری پرداخته شده است. فرآیندهای مختلف داده کاوی روی این مجموعه داده اعمال شد. بعد از فرآیند پاکسازی داده ها که در آن بیشتر به گسسته سازی مقادیری عددی توجه شد، اقدام به تصویرسازی داده نمودیم. نتایج جالبی در بخش تصویرسازی حاصل شد و بعد از آن به مدل سازی با درخت تصمیم و عملگر استنتاج قانون و مدل های دیگر پرداخته شده که نتیجه ی آن ساخت مدل هایی با دقت نسبتا خوب بود.