بسمه تعالی

Data mining: Case Presentation

Dr O. Pournik ***MD, MPH, MSc, PhD*** **pournik@gmail.com**

Case Presentation
Screening Models

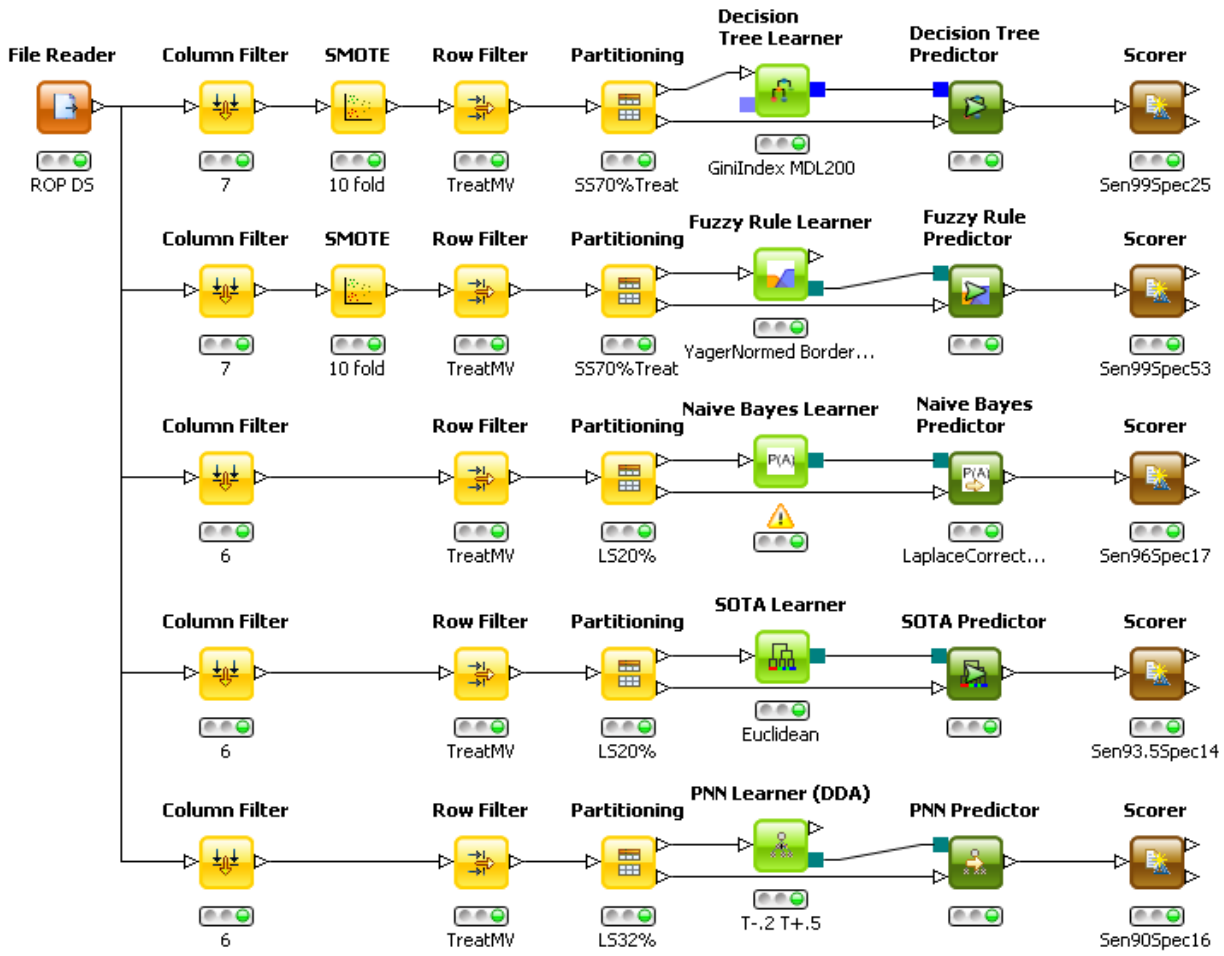
استفاده از روش‌های داده‌کاوی
برای ساخت مدل غربال‌گری
رتینوپاتی نارسی

مدل‌سازی

- پس از تایید کیفیت داده‌ها، مدل سازی با توجه به ساختار categorical متغیر هدف انجام شد.
- مجموعه داده‌ها به صورت خوشه‌ای (stratified) به دو گروه داده های آموزشی (۷۰٪) و ارزیابی (۳۰٪) بر اساس متغیر هدف تقسیم شدند.
- برای مدل سازی از نزم افزار KNIME (نسخه ۲.۳.۱) و Clementine ۱۲ استفاده شد.
- مدل سازی با ۵ متد Decision tree، Fuzzy rule، Naïve Bayes، SOTA و PNN انجام شد.

مدل‌سازی

فرآیند و جریان ساخت مدل‌ها



مدل‌سازی

متغیرهای مورد استفاده در هر یک از مدل ها

مدل	متغیرهای استفاده شده در مدل
درخت تصمیم (Decision tree)	سن بارداری، وزن تولد
قاعده فازی (Fuzzy Rule)	سن بارداری، سابقه سقط، آپگار بدو تولد، باروری با روشهای کمکی
Naïve Bayes	سن بارداری، وزن تولد، جنس، آپگار بدو تولد، آپگار دقیقه ۵
الگوریتم درختی خود سازمان یافته (SOTA)	سن بارداری، وزن تولد، جنس، آپگار بدو تولد، آپگار دقیقه ۵
شبکه عصبی احتمالی (PNN)	سن بارداری، وزن تولد، جنس، آپگار بدو تولد، آپگار دقیقه ۵

ارزیابی

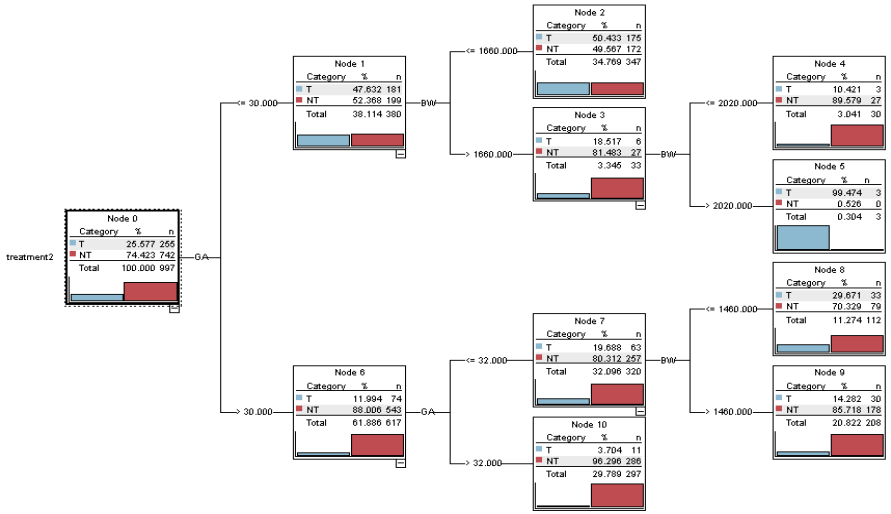
- مجموعه داده‌ها به صورت stratified به دو گروه داده های آموزشی (۷۰٪) و ارزیابی (۳۰٪) بر اساس متغیر هدف تقسیم شدند.

مدل	حساسیت	ویژگی	تعداد فیلدها در مدل	سطح زیر نمودار AUC
درخت تصمیم (Decision tree)	۰,۹۹۱	۰,۲۴۴	۲	۰,۸۷۵
قاعده فازی (Fuzzy Rule)	۰,۹۸۷	۰,۵۳۲	۴	۰,۹۰۶
Naïve Bayes	۰,۹۵۶	۰,۱۷۲	۵	۰,۸۱۵
الگوریتم درختی خود سازمان یافته (SOTA)	۰,۹۳۵	۰,۱۳۲	۵	۰,۸۱۴
شبکه عصبی احتمالی (PNN)	۰,۸۳۲	۰,۱۶۱	۵	۰,۷۶۵

مدل‌سازی-۲

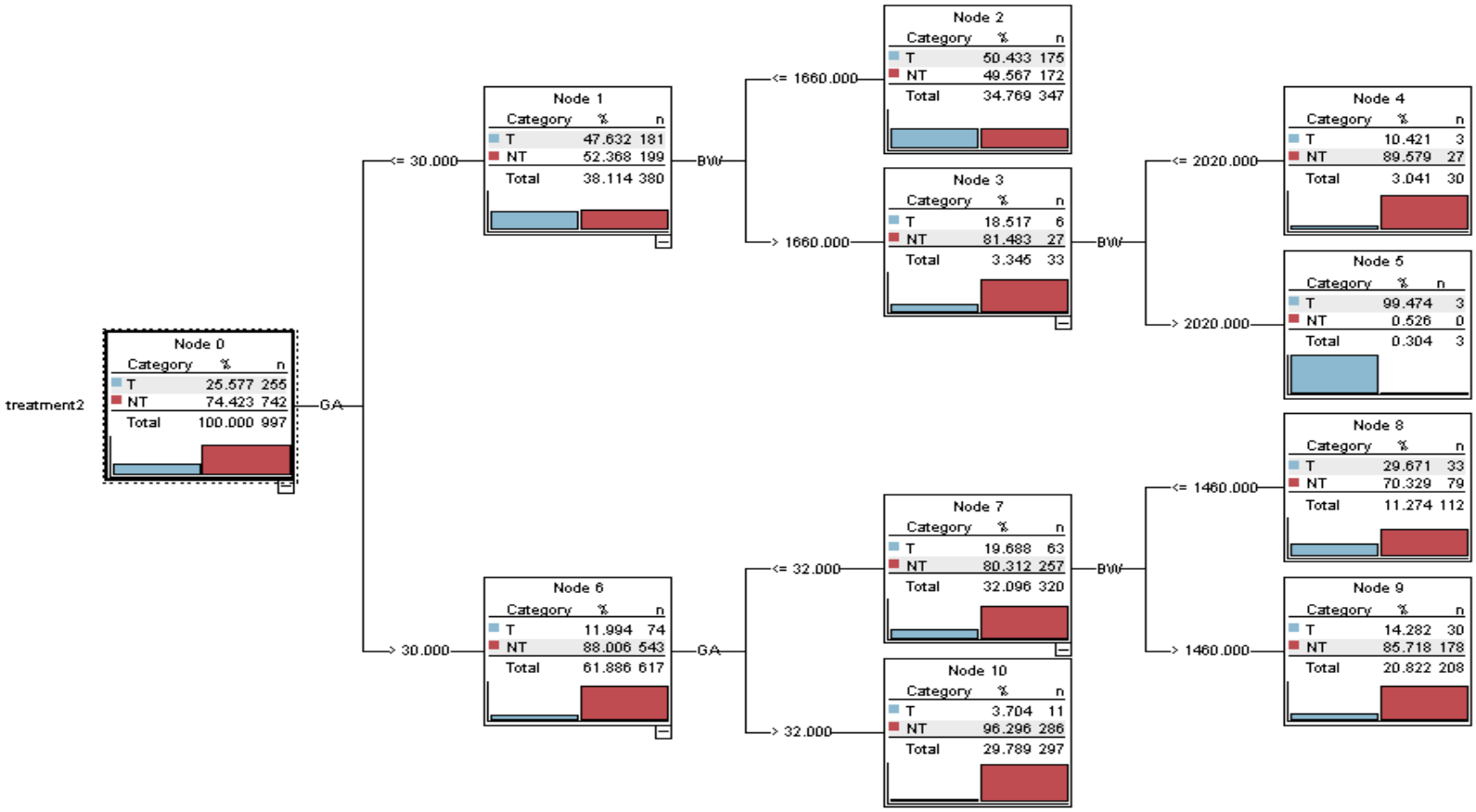
ارتقای مدل کاربردی

- با توجه به امکان شناخت ساده نحوه عمل‌کرد مدل درخت تصمیم و امکان به کارگیری آن به وسیله افراد بدون نیاز به رایانه، این مدل مورد بررسی بیشتری قرار گرفت.
- مدل اولیه درخت تصمیم‌گیری، مدل پیچیده‌ای بود که با بهره‌گیری از شاخص کیفیت GINI و بعد از حرس‌کردن (Pruning) با متد MDL (Minimum description Length) به شکل زیر تبدیل شد. در این مدل حساسیت به ۹۴٪ کاهش پیدا کرد و ویژگی به ۴۸.۷٪ افزایش یافت.



مدل‌سازی-۲

نمونه مدل کاربردی

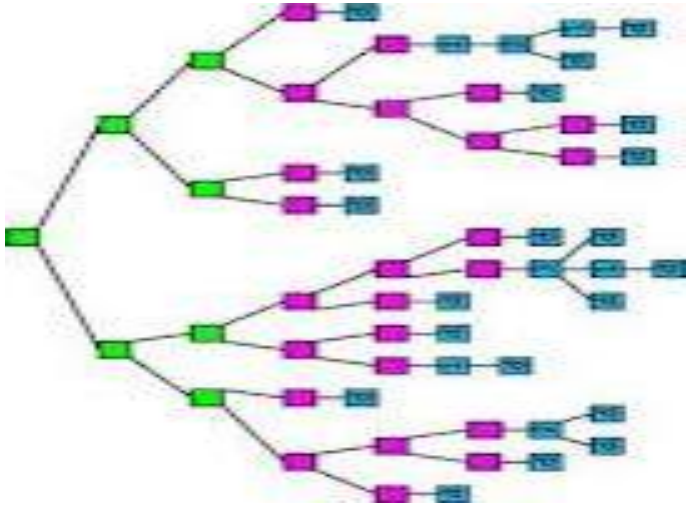


استقرار – نتیجه گیری

- تکنیک‌های داده کاوی می توانند در طراحی مدل‌های بهینه برای غربالگری نوزادان و کشف موارد ROP موثر باشد.
- در بین مدل های طراحی شده در این مطالعه، عملکرد مدل Fuzzy Rule بهتر از سایر مدل ها است و بهره گیری از آن می‌تواند علاوه بر کمک به متخصصین در کشف موارد نیازمند مداخله، به هزینه اثربخش‌تر شدن غربالگری ROP منجر شود.
- هم‌چنین مدل درختی کاربردی شده نیز می‌تواند با توجه به حساسیت قابل قبول و عدم نیاز به رایانه در صورت نیاز به وسیله سیستم بهداشتی–درمانی مورد استفاده قرار گیرد.

Case Presentation
Diagnostic Models

طراحی مدل تشخیصی HTLV1
با استفاده از شمارش کامل سلول های خونی



روش کار

- برای انجام این مطالعه اطلاعات مربوط بـه نتـایج ۵۹۹ مـورد نمونـه CBC شامل ۲٤٦ نمونه از بیماران مبتلا به HTLV1، ۱٤۲ نمونه از بیماران مبتلا به ALL، ۱۱۰ نمونه از بیماران مبتلا به ATL و ۱۰۱ نمونه از افراد سالم جمـع آوری گردید.
- نتایج آزمایش‌های افراد از سیستم اطلاعات بیمارستانی بیمارستان‌های قائم و امام رضا (ع) مشهد استخراج شد.

روش کار

- مدل سازی بر روی این داده‌ها با استفاده از متد CHAID و به کمک نرم‌افزار Clementine نسخه ۱۲ انجام شد. روش CHAID که مختصر شده عبارت Chi-squared Automatic Interaction Detector می‌باشد یکی از انواع درخت تصمیم است.

نتایج

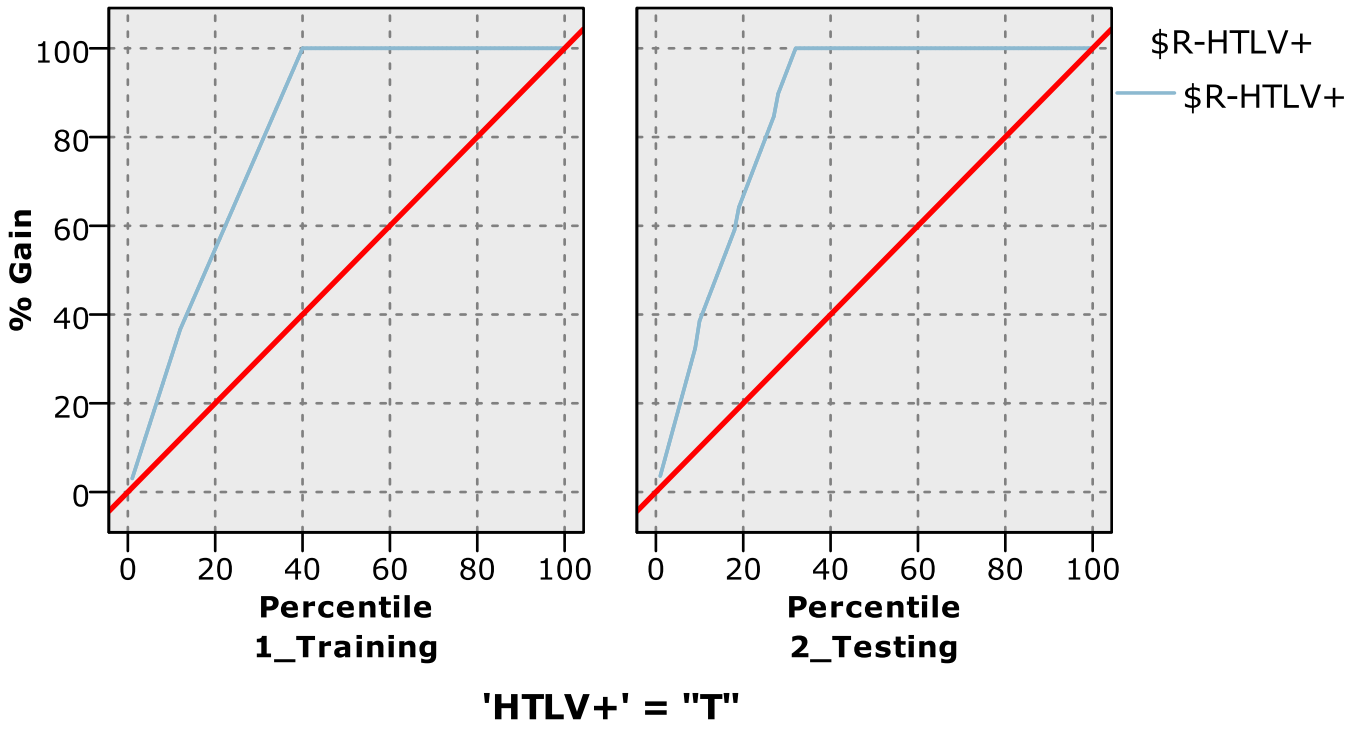
- بانک داده‌های مورد استفاده در این مطالعه محتوی نتایج آزمایش CBC ۳۰۲ نفر شامل ۹٤ بیمار مبتلا به لوکمی، ۱۰۱ فرد نرمال و ۱۰۷ بیمار مبتلا به HTLV1 بود. متغیرهای مورد بررسی شامل تعداد گلبول‌های سفید، تعداد گلبول‌های قرمز، هموگلوبین، هماتوکریت، MCV، MCH، MCHC، پلاکت، درصد پلی مورفونوکلئرها، درصد لنفوسیت‌ها، درصد مونوسیت‌ها و درصد ائوزینوفیل‌ها بود.

نتایج

	HTLV1	Leukemia	Healthy	P value
WBC(10000)	13.7	66.2	8.7	0.698
RBC	4.0	3.6	4.6	<0.001
Hgb	11.6	9.8	13.2	<0.001
MCV	88.7	89.7	85.1	0.002
MCH	29.2	42.3	28.9	0.366
MCHC	34.0	32.9	33.9	0.797
Platelet	178	1009	292	0.286
Lymphocyte	29.3	27.8	32.4	0.142
Monocyte	4.7	5.4	6.2	0.504
Eosinophil	3.6	4.8	2.7	0.074

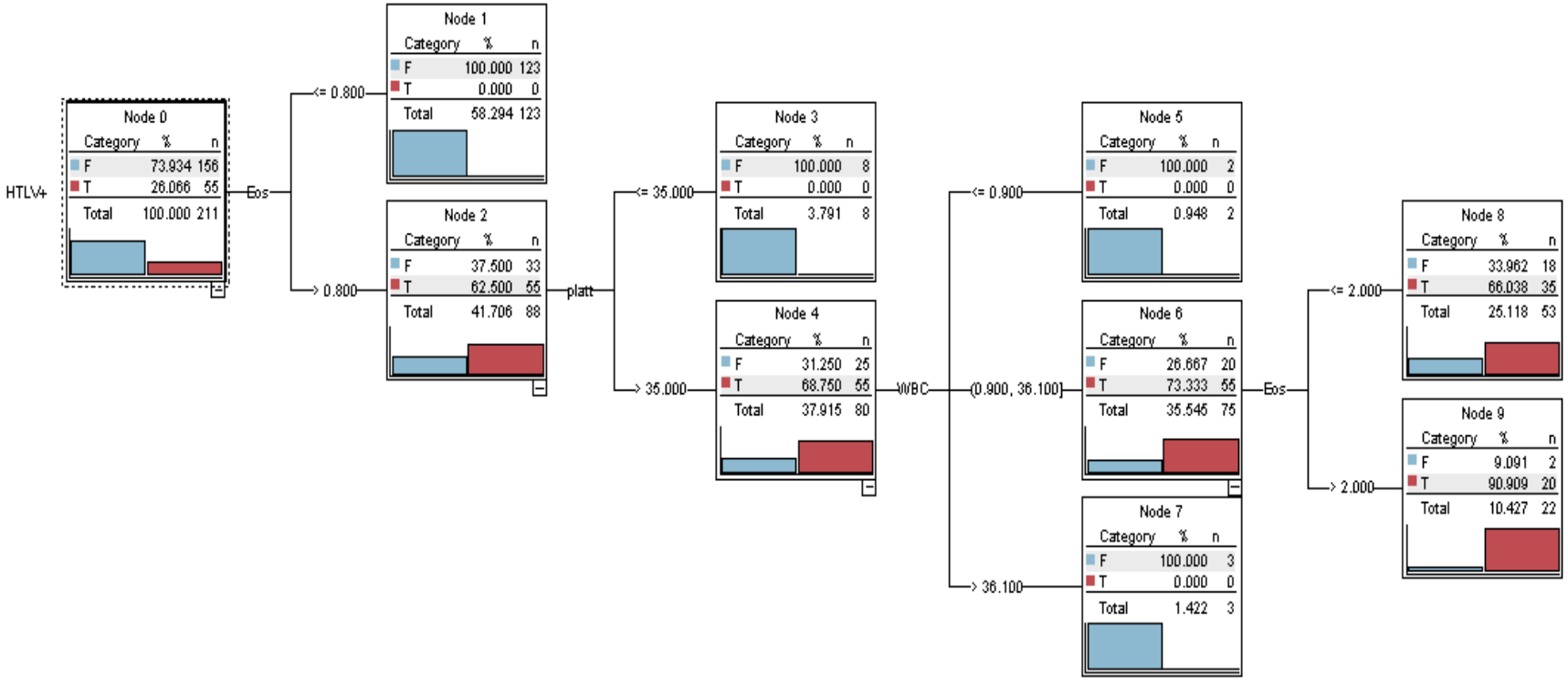
- میانگین متغیرها در گروه های مختلف و سطح معنی داری تفاوت بین گروه های مختلف بر اساس آنالیز واریانس

نتایج



ارزیابی عملکرد مدل(**Gain Chart**)

نتایج



- مدل نهایی به دست آمده در قالب درخت تصمیم

بحث

- این مدل بر اساس روش CHAID ساخته شده که یکی از انواع روش های مبتنی بر درخت تصمیم می باشد. مزیت کلی این روش در سادگی نمایش کلیه مراحل طبقه بندی است. همین امر بهره گیری از مدل به دست آمده را ممکن می‌سازد.

بحث

- با اجرای این مدل بر روی نتایج CBC افراد بدون علامت یا بیماران متفرقه در مناطق اندمیک کشور می توان افراد ناقل احتمالی را کشف و برای آزمایش‌های تکمیلی معرفی کرد. با توجه به احتمال بروز انواع بدخیمی ها و اختلالات سیستم اعصاب در افراد مبتلا، کشف این موارد می تواند به کاهش بار بیماری های مذکور و ارتقای سطح سلامت افراد به ویژه در مناطق اندمیک منجر شود. همچنین پیشنهاد می شود با استفاده از داده های بیشتر امکان توسعه روش و کیفیت مدل فراهم شود.

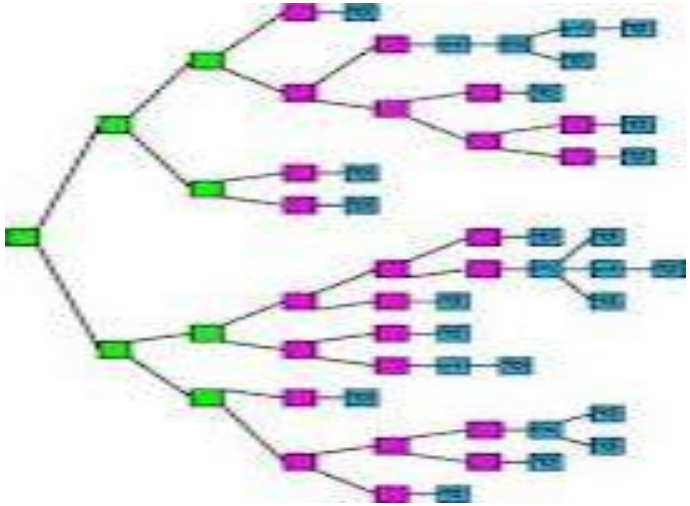
نتیجه گیری

- مدل های تصمیم گیری مبتنی بر داده کاوی و به طور خاص تکنیک به کار گرفته شده در این مقاله و درخت تصمیم به دست آمده می توانند در شناسایی بیماران مبتلا به بیماری HTLV1 کمک کننده بوده و به پزشکان و متخصصین در شناسایی سریع تر و آسان تر این افراد کمک نماید.

Case Presentation
Diagnostic Models

کاربرد داده کاوی پیش بینی کننده در پزشکی:

مقایسه الگوهای منتخب در شناسایی زودهنگام بیماری منژیت باکتریایی در ایران

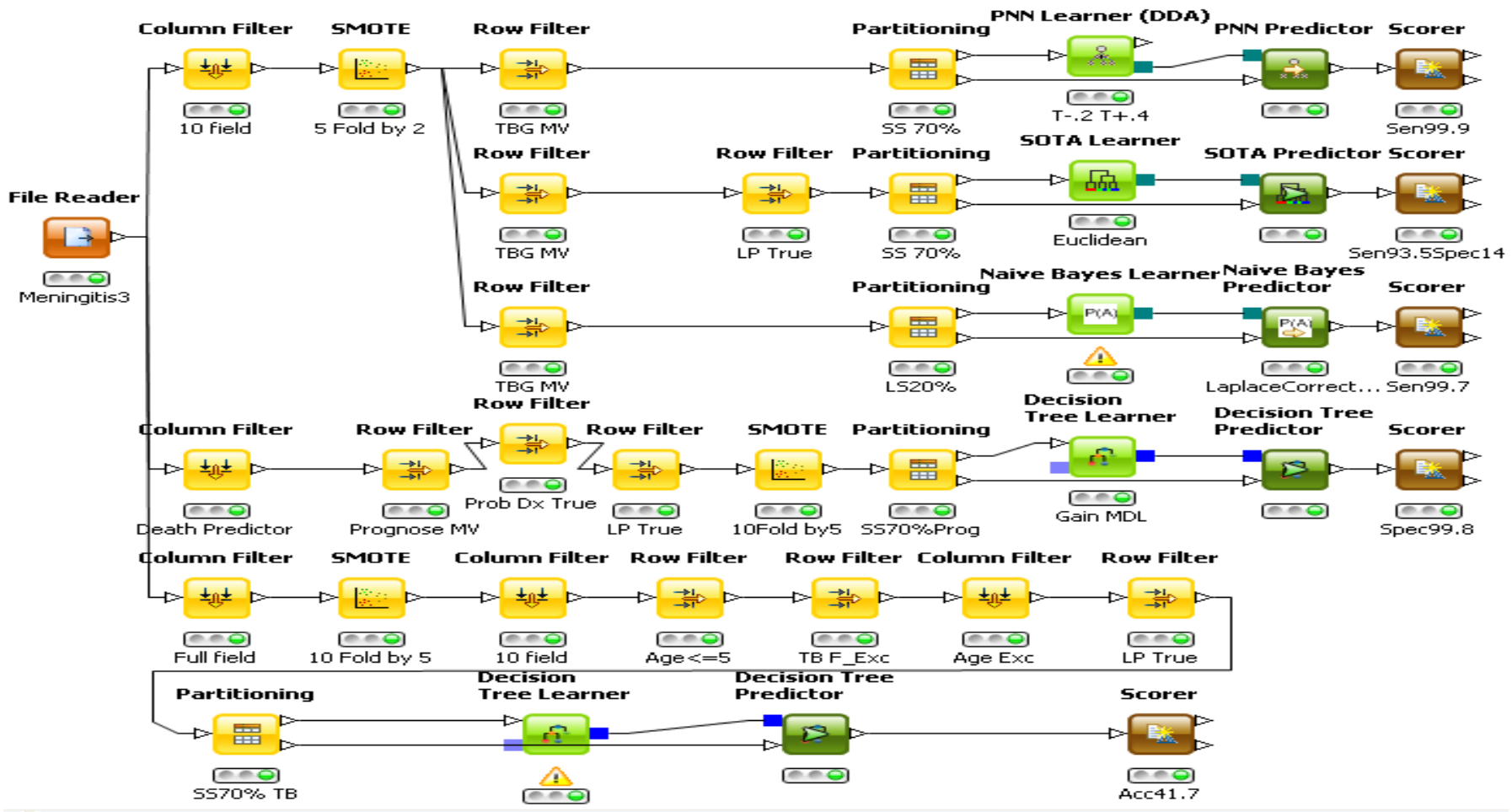


ادراک داده ها

- بانک اطلاعاتی بیماران منتژیتی مرکز مدیریت بیماری های واگیر وزارت بهداشت در سال های ۱۳۸۸ و ۱۳۸۹ از کل کشور و مشتمل بر ۷۵۷۴ مورد

نام متغیر	نام متغیر
سن (سال)	مقدار قند
روز	کشت CSF
جنس	کشت خون
شغل	لاتکس
استان	رنگ آمیزی گرم
دانشگاه	تشخیص محتمل
شهر	تشخیص قطعی
منطقه	تشخیص نهایی
انجام LP	آنتی بیوتیک قبل
ظاهر نمونه	آنتی بیوتیک بعد
تعداد سلول	نتیجه درمان
درصد PMN	دریافت واکسن Hib
درصد لنفوسیت	دریافت واکسن منگوکوکی
مقدار پروتئین	

مدل‌سازی پیش گویی کننده بیماری مننژیت باکتریایی



بحث و نتیجه گیری

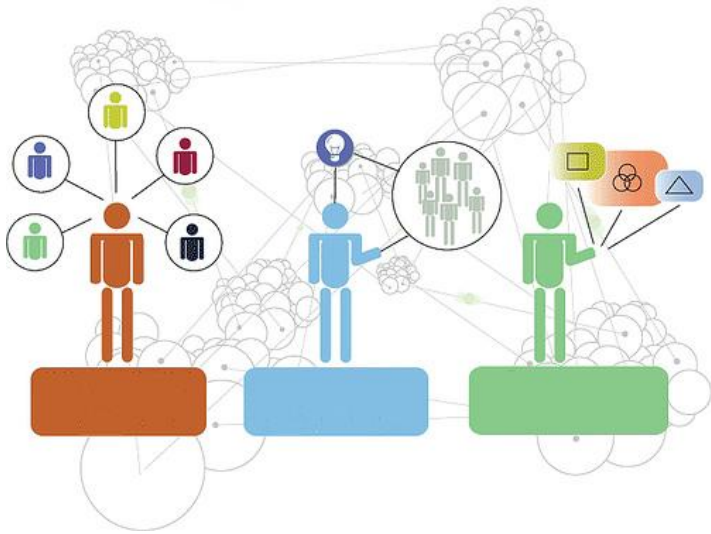
Sensitivity	Accuracy	نام تکنیک
٪۹۴/۷	٪۸۸/۸	SOTA
٪۹۹/۷	٪۹۳/۱	Naïve Bayes

پیشنهادات در خصوص استفاده از نتایج این پژوهش

- نتایج حاصل از این پژوهش کاربرد موفق مدل‌های **Predictive** موثر را در تشخیص و کنترل بیماری‌ها نشان داد.
- این مدل‌ها با داده‌های بیشتر می‌توانند بهینه شده و در سیستم‌های بهداشت و درمان مانند سیستم‌های پایش و نظارتی و **HIS**ها و **DSS**ها مورد استفاده قرار گیرند.

Case Presentation
Fraud or Anomaly Detection

Hepatitis B familial transmission analysis
A Data mining approach



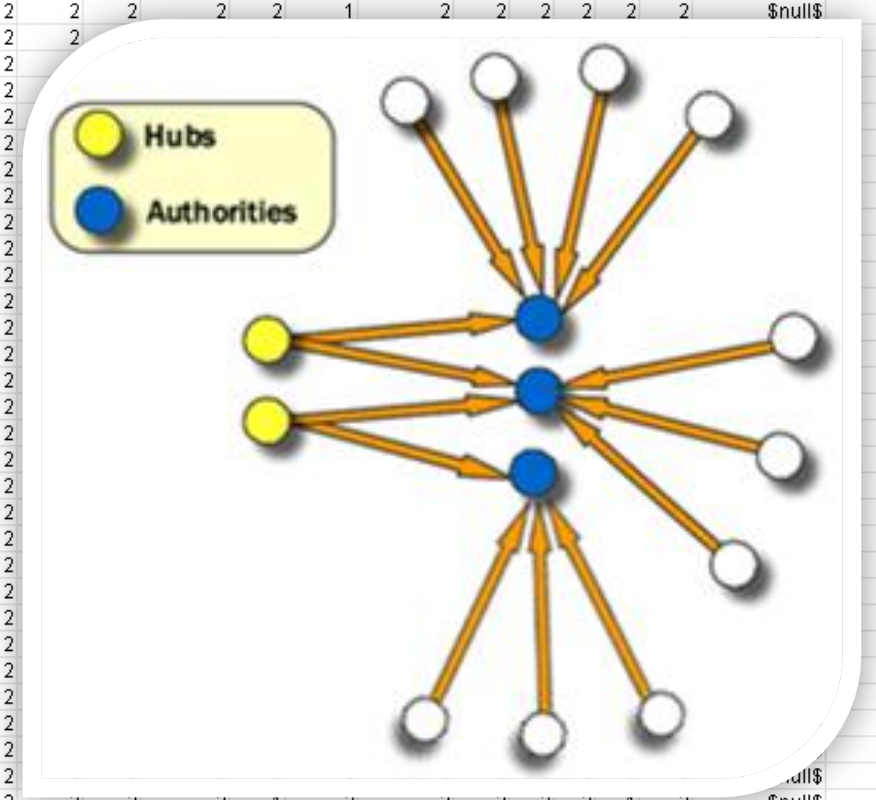
Data Set Specification

- 330 hepatitis B patients
  - Demographic data
  - Viral markers
  - Stage of disease
  - Family history of liver disease mortality
  - Morbidity
- Demographic data and viral markers of their offspring were also collected.

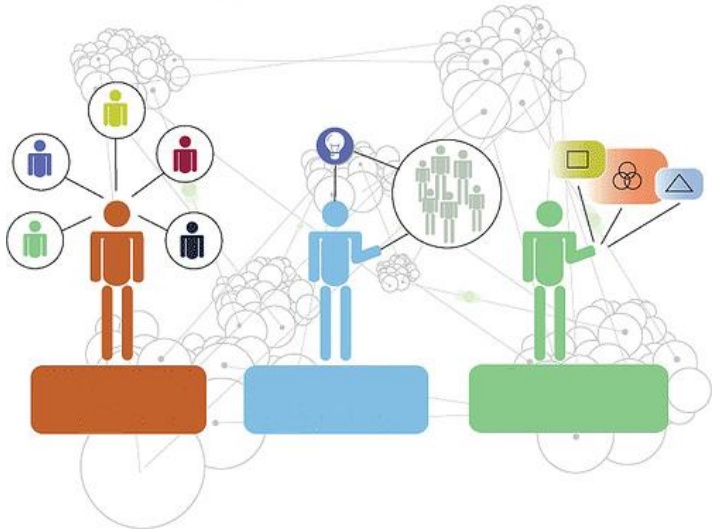
Traditional Data Analysis View

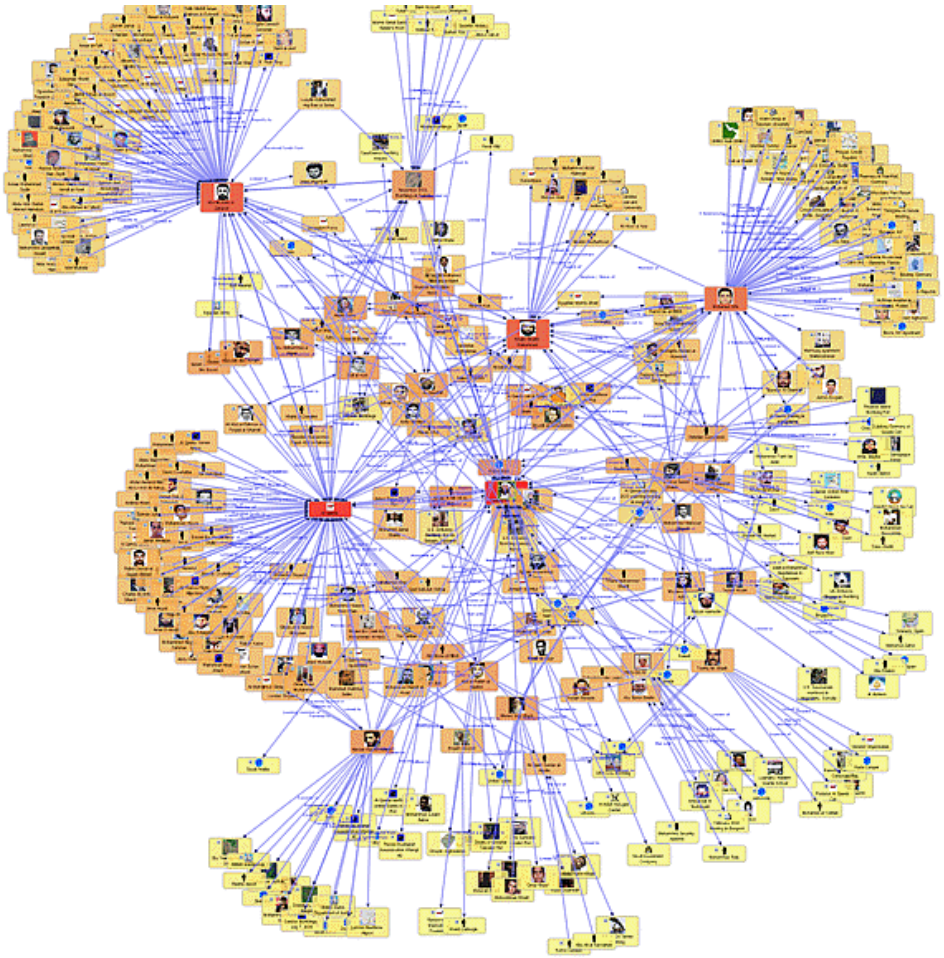
File Edit Generate

	Age	Sex	Diag	Stage	Child	IDUs	Non	Trans	Extra	Hospital	Surg	Dental	Ingection	Liver	Hej	Tat	Oth	HBV	HBV_P_m_M	HBV_P_F_Fa	Mor_L_D_F	HBeAg	HDVAb
1	23	2	4	1	$null$	2	2	2	2	2	2	1	2	2	2	2	2	2	$null$	$null$	2	1	$null$
2	27	2	2	1	$null$	2	2	2												$null$	1	2	2
3	44	1	2	2	$null$	2	2													$null$	2	2	$null$
4	41	1	4	2	$null$	2	2													$null$	2	2	2
5	26	2	4	1	$null$	2	2													$null$	2	2	$null$
6	32	1	4	1	$null$	2	2													$null$	2	2	$null$
7	65	1	4	1	$null$	2	2													$null$	2	2	2
8	25	1	3	2	$null$	2	2													$null$	2	$null$	$null$
9	37	1	3	1	$null$	2	2													$null$	2	2	2
10	29	1	2	2	$null$	2	2													$null$	2	2	2
11	48	2	2	$null$	$null$	2	2													$null$	1	$null$	2
12	8	1	2	2	$null$	2	2													$null$	2	1	2
13	30	1	1	2	$null$	2	2													$null$	2	2	2
14	23	1	2	1	$null$	2	2													1	1	2	$null$
15	37	2	3	1	$null$	2	2													$null$	2	2	2
16	32	1	1	2	$null$	2	2													$null$	2	2	2
17	24	1	2	2	$null$	2	2													$null$	2	1	2
18	46	1	3	2	$null$	2	2													$null$	2	1	$null$
19	13	2	2	1	$null$	2	2													$null$	2	1	$null$
20	21	2	4	2	$null$	2	2													$null$	2	2	2
21	68	1	3	2	$null$	2	2													$null$	2	1	2
22	31	2	2	2	$null$	2	2													$null$	2	2	2
23	29	2	4	2	$null$	2	2													$null$	2	$null$	2
24	22	2	2	2	$null$	2	2													$null$	2	1	$null$
25	14	2	2	2	$null$	2	2													$null$	2	2	2
26	46	1	3	2	$null$	2	2													$null$	2	2	2
27	38	1	4	1	$null$	2	2													$null$	2	$null$	$null$
28	54	1	1	3	$null$	2	2													$null$	2	2	2
29	35	2	4	2	$null$	2	2													$null$	2	$null$	2
30	16	1	2	2	$null$	2	2													$null$	2	$null$	2
31	24	1	3	2	$null$	2	2	2	2	2	1	2	2	2	2	2	1	2	$null$	$null$	2	2	1

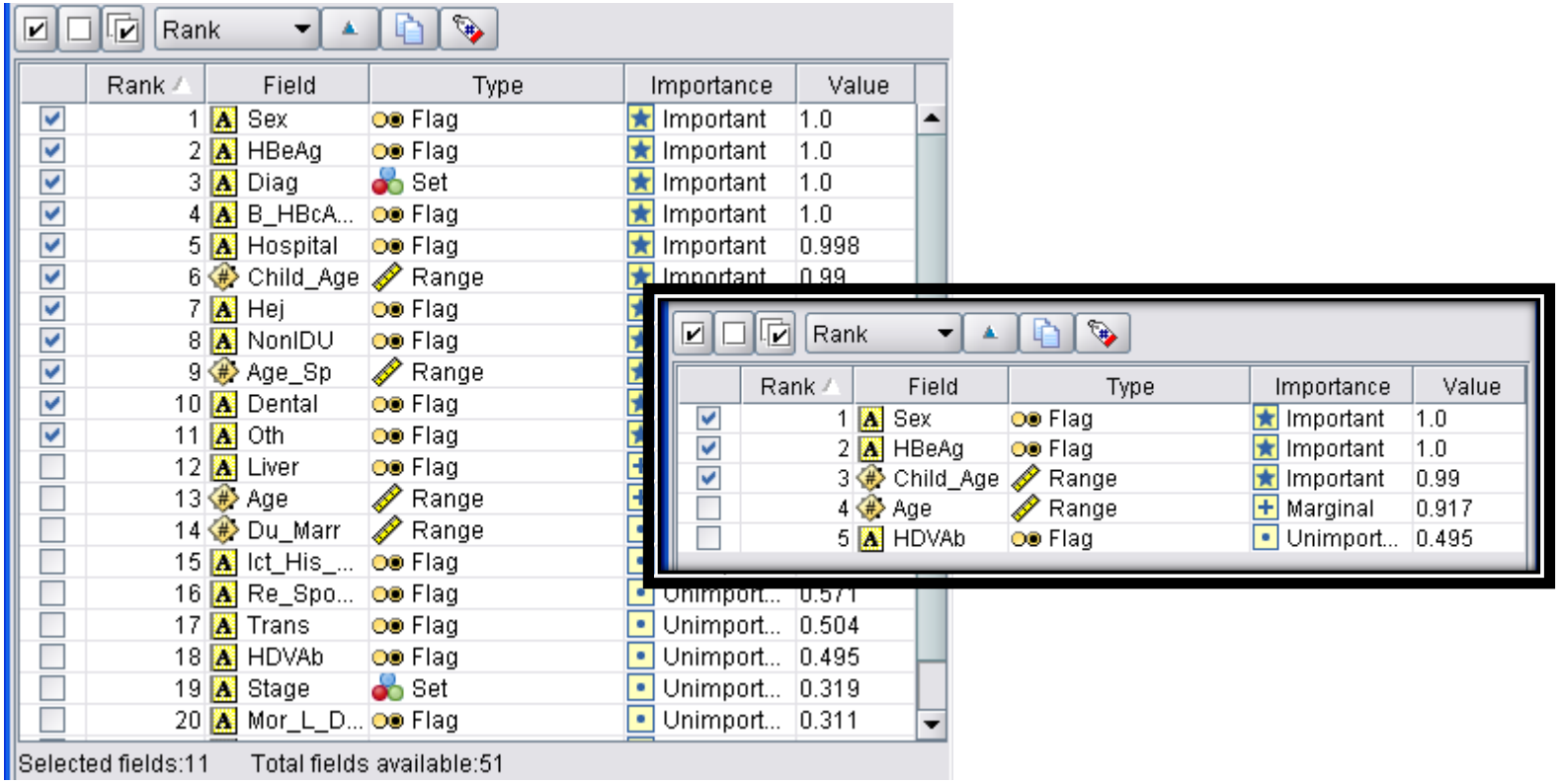


Network Analysis

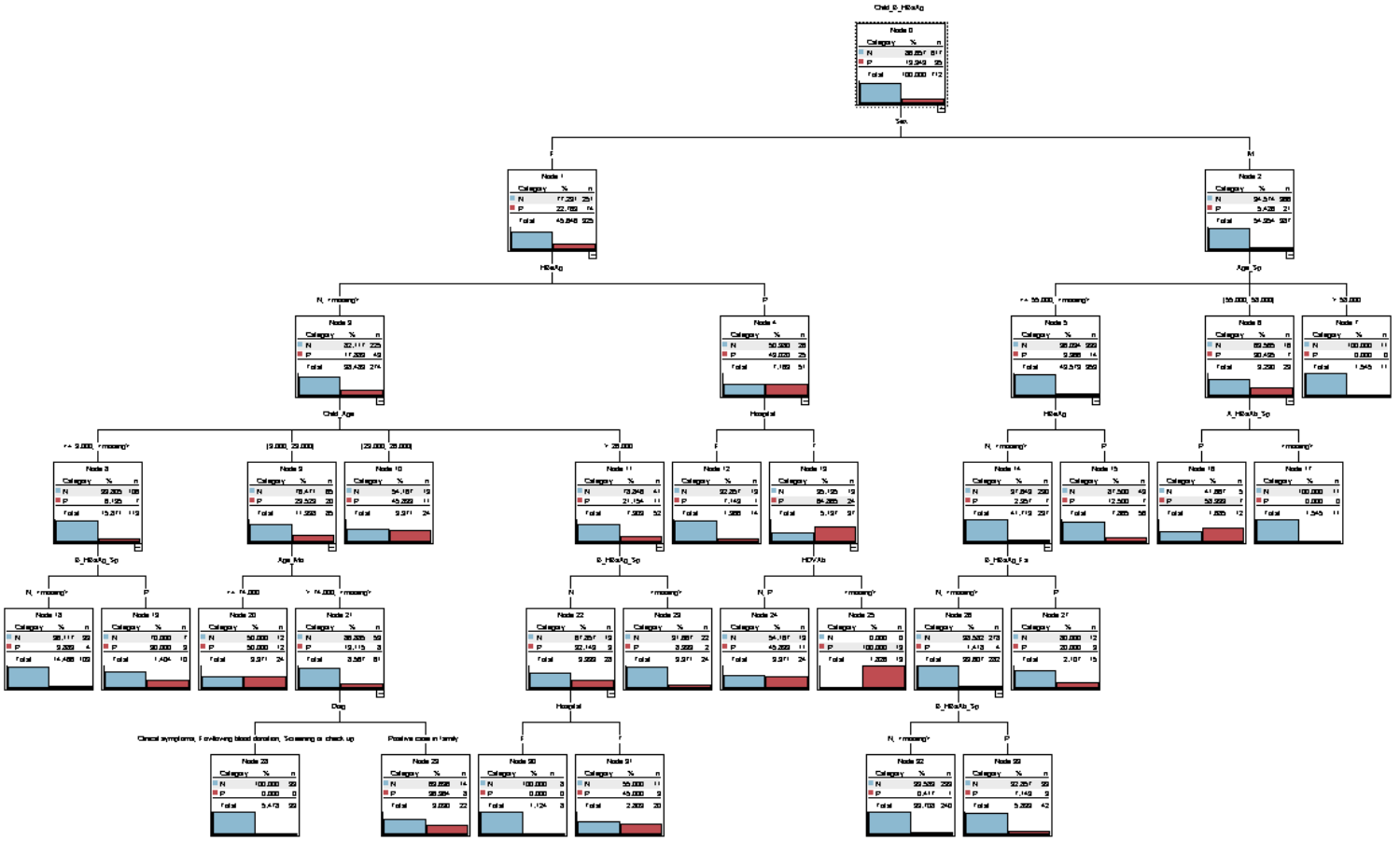




Feature Selection Analysis



Findings (Tree View)



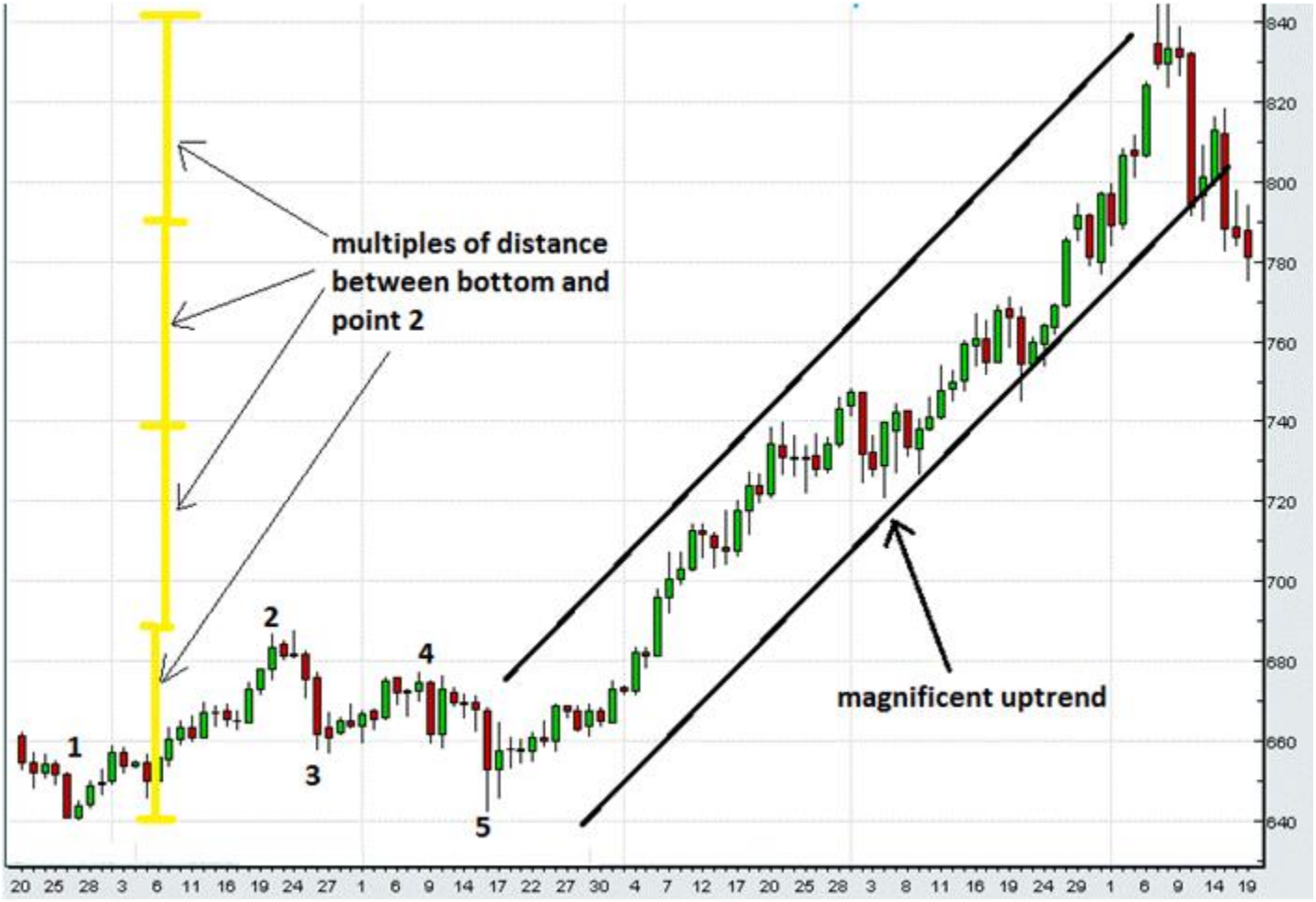
Findings

The most dominant finding in the extracted patterns was the difference in transmission pattern of male and female index cases.

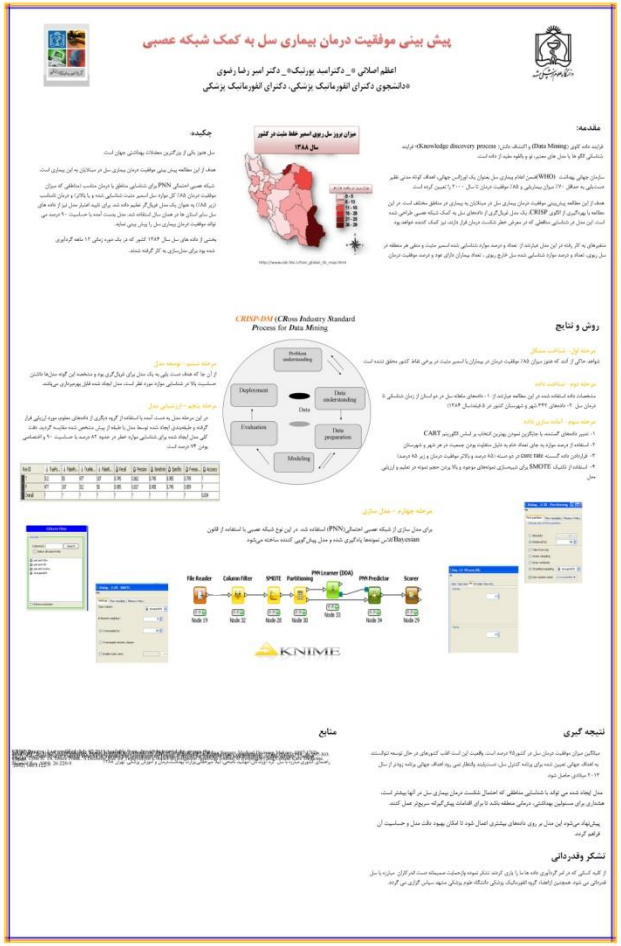
- In male index cases,
  - Family history of liver disease mortality and morbidity was the main predictive factor of disease transmission to offspring.
  - In the next step, proxies of exposure duration such as offspring age were important.

- In female index cases,
  - Rate of transmission was significantly higher in female index cases and the main determining factor was HBeAg.
  - In HBeAg negative patients, the pattern of transmission was different in patients based on their HDVAb. Transmission rate was 25% in HDVAb negative and 5% in HDVAb positive cases.

Anomaly Evaluation Approach



Case Presentation
Health Policy Management



پیش بینی موفقیت درمان بیماری سل به کمک شبکه عصبی

اعظم اصلانی ۱*_ دکتر احمد پورنیک ۲*_ دکتر امیر رضا رضوی ۳

۱*دانشجوی دکترای اپیدمیولوژی پزشکی، دکترای اپیدمیولوژی پزشکی

مقدمه:

چکیده:

روش و نتایج

CRISP-DM (CRoss Industry Standard Process for Data Mining)

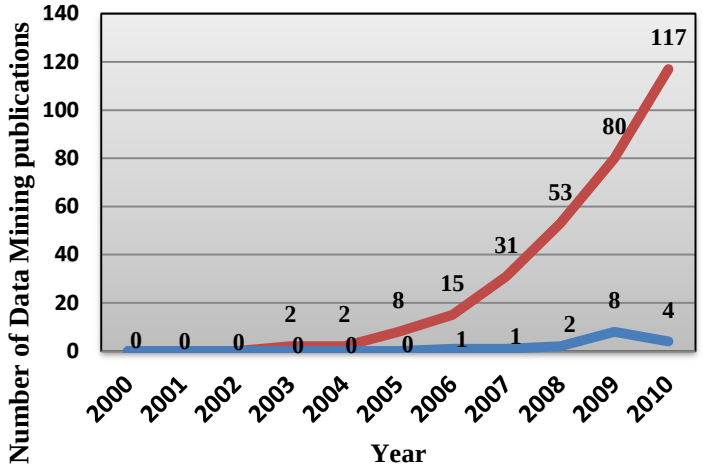
نتیجه گیری

منابع

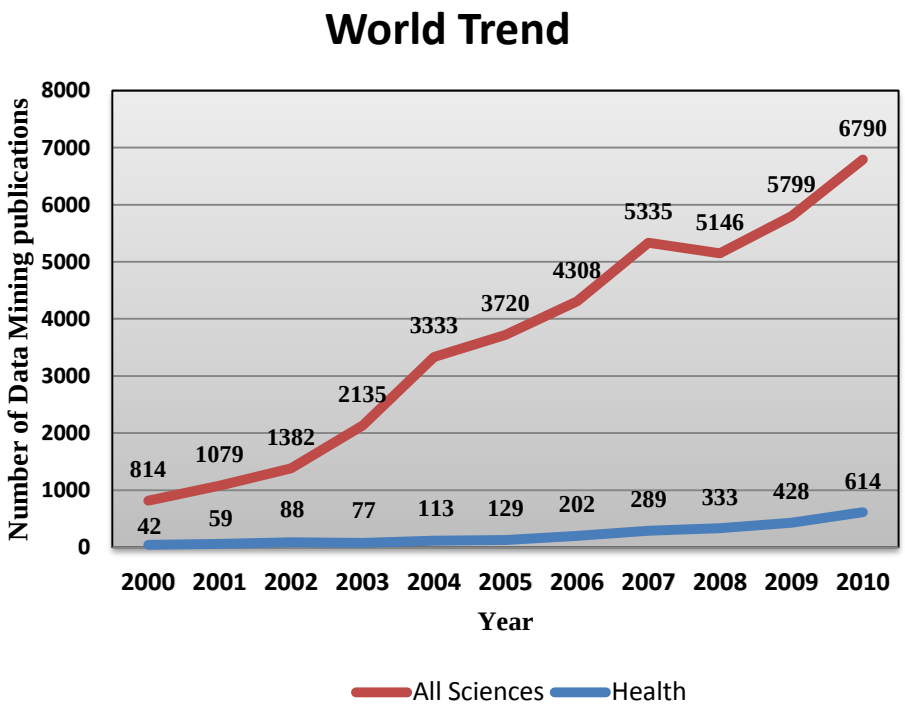
تشکر و قدردانی

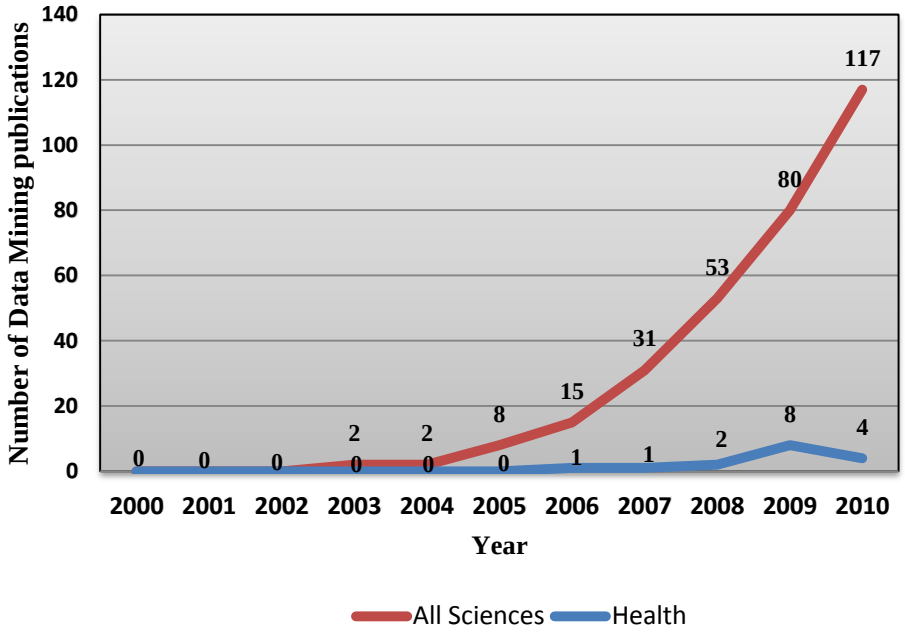
Case Presentation
Scientometric Approach

Data Mining in Medicine: A Scientometrics Perspective

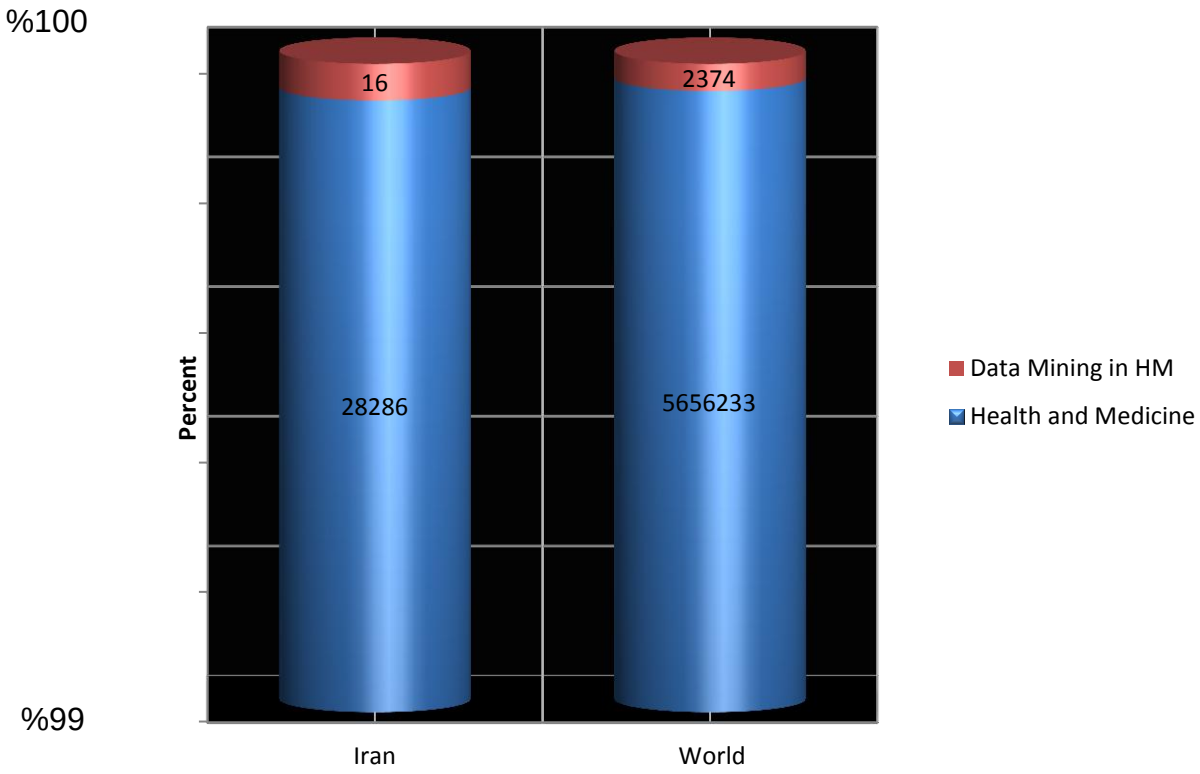


Comparing growth rate of Data Mining publications in health and all sciences in Iran





Share of Data Mining publications in total number of articles in Medicine



Findings and Suggestions of Scientometric Study of DataMining in Medicine

- Publication Trends in Iran is somehow in accordance with world growth trend
- But the volume of publication can be much higher Specially in health and medicine domain.

Any Questions?