بسمه تعالی

گزارش پروژه ی داده کاوی در RapidMiner

مهدی جواهری صابر

۹۲۴۳۰۸۸۰۱۷

داده ی انتخابی : Indian Liver Patient Dataset

بیماری کبد در هندی ها

http://archive.ics.uci.edu/ml/datasets/ILPD+%28Indian+Liver+Patient+Dataset%29

Data Set Characteristics:	Multivariate	**Number of Instances:**	583	**Area:**	Life
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	10	**Date Donated**	2012-05-21
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	36311

در این داده اطلاعات ۵۸۳ فرد هندی مورد بررسی قرار گرفته و بیماری کبد در آنها تشخیص داده شده است

این دیتا حاوی ۴۱۶ نفر بیمار کبد و ۱۶۷ نفر سالم می باشد

۴۴۱ نفر از کل داده مرد و ۱۴۲ نفر زن هستند.

هر فردی که سنش از ۸۹ سال بالاتر بوده گرد شده و با عدد ۹۰ نشان داده شده.

ویژگی ها

- Age
- Gender
- Total Bilirubin (TB)
- Direct Bilirubin (DB)
- Alkphos Alkaline Phosphotase (AAP)
- Sgpt Alamine Aminotransferase (SAlamine)
- Sgot Aspartate Aminotransferase (SAspartate)
- Total Protiens (TP)
- ALB Albumin (ALB)
- A/G Ratio Albumin and Globulin Ratio (AG)
- Selector field used to split the data into two sets (status)

۱- رسم درخت تصمیم با استفاده از الگوریتم Decision Tree

در این الگوریتم داده های ورودی را به فرمت مناسب تبدیل می کنیم (تعیین لیبل و نوع لیبل و ...) و خروجی را به صورت درخت دریافت میکنیم

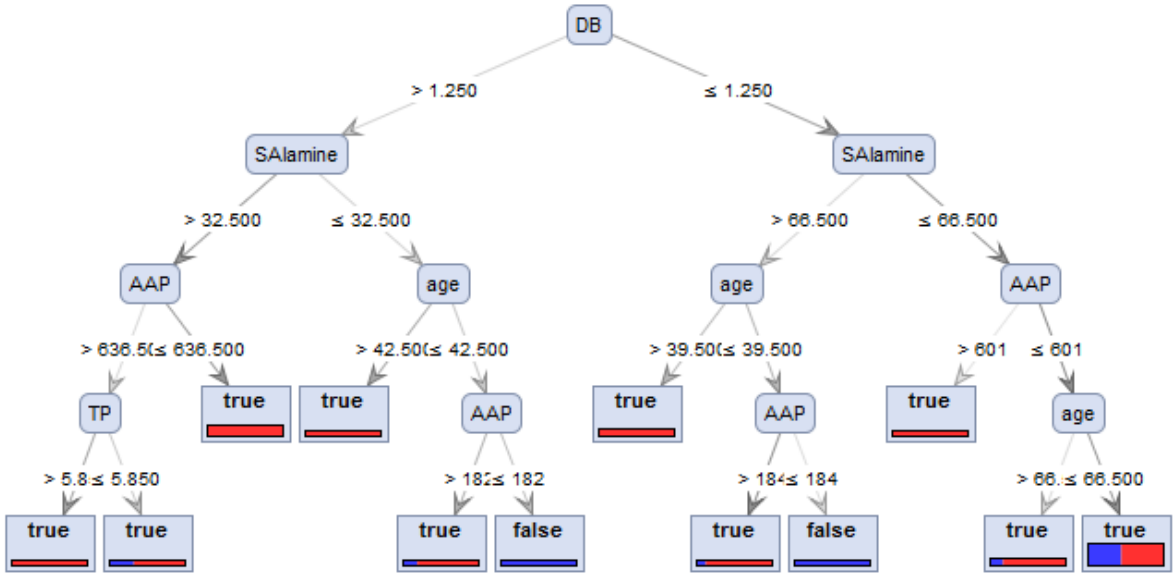
مراحل طی شده :

- ورود داده با اپراتور Read CSV
  - انتخاب داده ی ورودی (csv file)
  - تعیین جدا کننده (column separator)
  - تعیین ستون ویژگی ها (first row as names)
- تعیین لیبل با اپراتور Set Role
  - انتخاب ویژگی (attribute name=status)
  - انتساب خصوصیت لیبل به آن (target role=label)
- تبدیل نوع داده ی عددی به بولین با اپراتور Numerical to Binomial : درخت تصمیم روی داده هایی کار میکند که لیبل آنها از نوع بولین باشد ، برای همین منظور بایستی لیبل را به بولین تبدیل کنیم
  - نوع ویژگی (attribute filter type = single)
  - انتخاب ویژگی (attribute = status)
  - Include special characters

- Min = 1.5
- Max = 2.0

- رسم درخت تصمیم با اپراتور decision tree
  - Criterion = information_gain
  - Maximal depth = 5

خروجی این الگوریتم به صورت زیر میباشد



۲- مقایسه ی ارزیابی در الگوریتم های Decision Tree , Knn , Naïve Bayes

- ارزیابی با اپراتور X-validation
  - در قسمت Training نوع الگوریتم خود را انتخاب میکنیم
    - Decision Tree
      - Criterion = information_gain
      - Maximal depth = 5
    - K-nn
      - K = 5
    - Naïve Bayes
  - در قسمت Testing نوع خروجی ارزیابی را معین میکنیم
    - Apply Model
    - Performance

- Main Criterion = accuracy
  - نمایش خروجی الگوریتم های مذکور

خروجی ارزیابی در الگوریتم Decision Tree

accuracy: 70.14% +/- 3.45% (mikro: 70.15%)

	true false	true true	class precision
pred. false	21	28	42.86%
pred. true	146	388	72.66%
class recall	12.57%	93.27%	

خروجی ارزیابی در الگوریتم K-nn با پارامتر K=5

accuracy: 67.06% +/- 3.73% (mikro: 67.07%)

	true false	true true	class precision
pred. false	54	79	40.60%
pred. true	113	337	74.89%
class recall	32.34%	81.01%	

خروجی ارزیابی در الگوریتم Naïve Bayes

accuracy: 55.90% +/- 3.61% (mikro: 55.92%)

	true false	true true	class precision
pred. false	160	250	39.02%
pred. true	7	166	95.95%
class recall	95.81%	39.90%	

با توجه به خروجی های بدست آمده الگوریتم Decision Tree از الگوریتم های دیگر صحت بالاتری دارد