گزارش ۳:

یادگیری ماشین

درخت تصمیم

Decision Tree

سید علی مجتبوی

گزارش ۲ - بررسی و تحویل انواع روش های Decision Tree

دراین گزارش با انواع روش های Decision Tree آشنا خواهیم شد و آنها را با یکدیگر مقایسه می کنیم

دیتاست Pima Indians Diabetes

دیتاست:

این مجموعه داده در اصل از موسسه ملی دیابت و بیماری های گوارشی و کلیوی تهیه شده است. هدف مجموعه داده این است که براساس اندازه‌گیری‌های تشخیصی خاص موجود در مجموعه داده، پیش بینی کنیم که آیا بیمار به دیابت مبتلا است یا خیر. چندین محدودیت برای انتخاب این نمونه ها از یک پایگاه داده بزرگتر قرار داده شد. به طور خاص، همه بیماران در اینجا زنان حداقل ۲۱ ساله از میراث هندی پیما هستند.

۱. لینک دیتاست: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

۲. مشخصات دیتاست:

مجموعه داده ها شامل چندین متغیر پیش بینی کننده پزشکی و یک متغیر هدف، نتیجه است. متغیرهای پیش بینی کننده شامل تعداد بارداری های بیمار، BMI، سطح انسولین، سن و غیره است.

۳. فیچرهای دیتاست:

# Pregnancies	# Glucose	# BloodPres...	# SkinThick...	# Insulin	# BMI	# DiabetesP...	# Age	# Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0

مقایسه انواع الگوریتم های درخت تصمیم:

- ID3 (Iterative Dichotomiser 3): .1

یک الگوریتم ساخت درخت تصمیم است که توسط Quinlan Ross ارائه شد. این الگوریتم برای ساخت درخت تصمیمی از جنس دودویی) tree binary) استفاده می کندباید به این نکته توجه داشت که این الگوریتم برای مقیاس پذیری به مشکل می خورد و نمی تواند با داده های پیچیده یا مقیاس پذیر به خوبی کار کند .

C4.5: .2

یک نسخه بهبود یافته از ID ۳است و توسط Quinlan Ross نیز ارائه شد ۴.۵ . Cدرخت های تصمیم چندجمله ای می سازد و قابلیت کار با ویژگی های عددی را نیز دارد و دارای قابلیت انتخاب ویژگی ها با اســـتفاده از معیارهای مانند Entropy یا Index Gini CART (Classification. می باشد.

نتایج اجرای ۲ نوع الگوریتم برای این دیتاست:

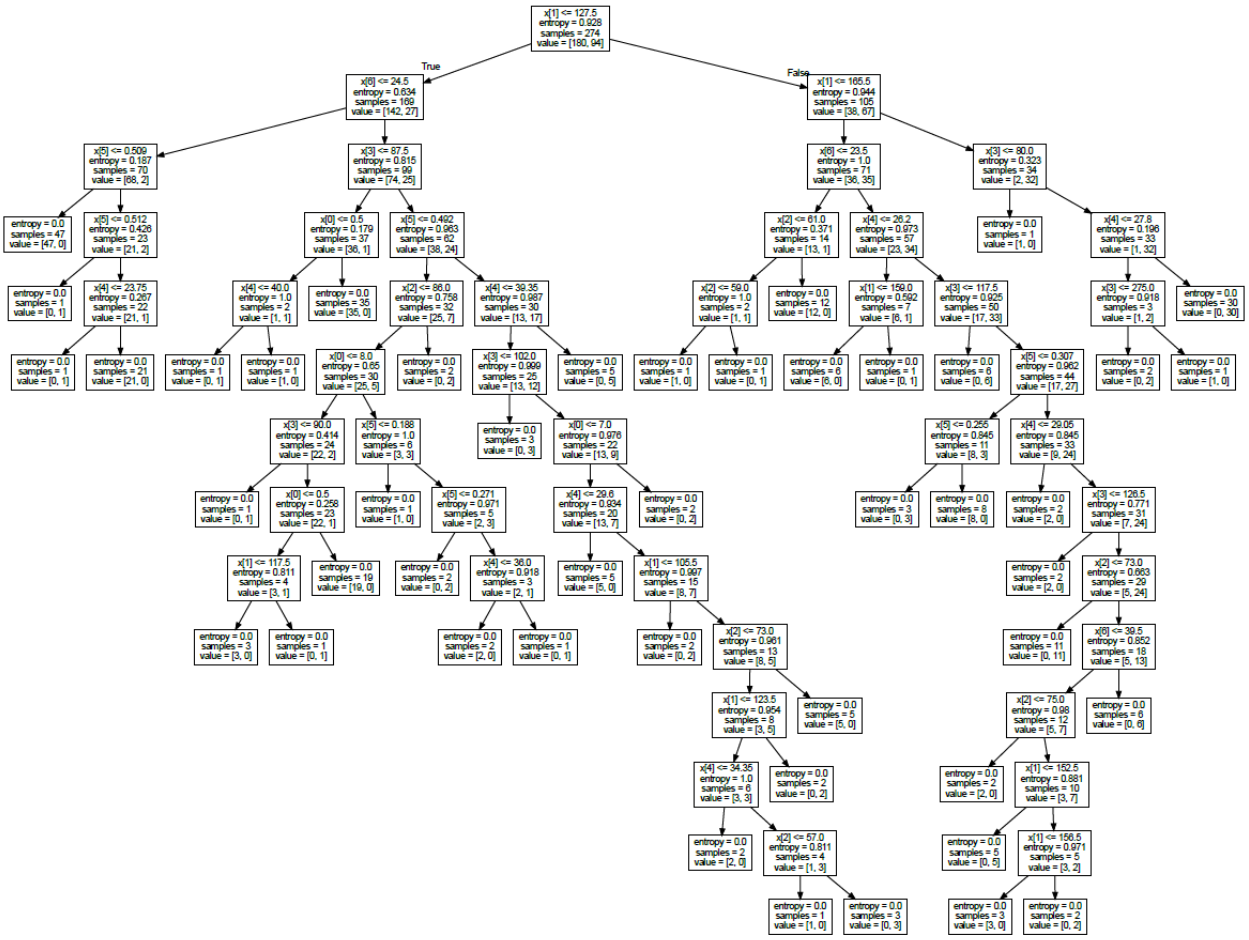
```
TESTING DATA
Confusion Matrix with Testing Data with C4.5 Algorithm:
[[183   0]
 [  0  91]]

Classification Report C4.5 Algorithm:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       183
           1       1.00      1.00      1.00        91

    accuracy                           1.00       274
   macro avg       1.00      1.00      1.00       274
weighted avg       1.00      1.00      1.00       274



TRAINING DATA
Confusion Matrix with Training Data with C4.5 Algorithm:
[[183   0]
 [  0  91]]

Classification Report C4.5 Algorithm:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       183
           1       1.00      1.00      1.00        91

    accuracy                           1.00       274
   macro avg       1.00      1.00      1.00       274
weighted avg       1.00      1.00      1.00       274
```

شکل ۱: نتایج الگوریتم ID3

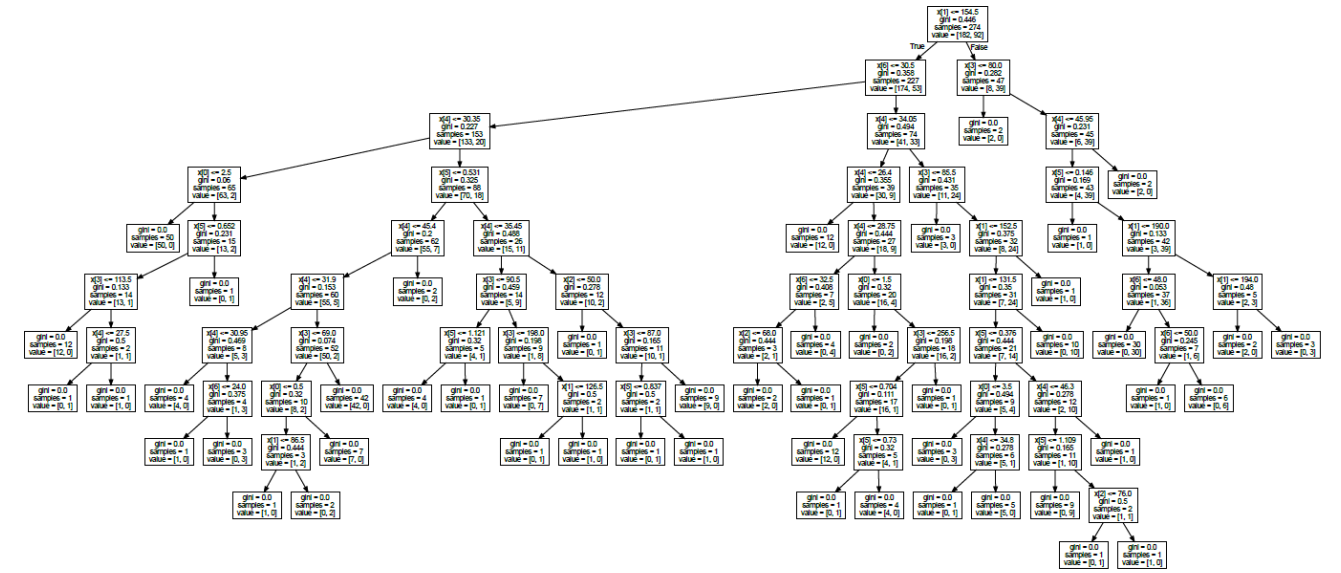
```
TESTING DATA
Confusion Matrix with Testing Data with ID3 Algorithm:
[[66 13]
 [15 24]]

Classification Report ID3 Algorithm:
              precision    recall  f1-score   support

           0       0.81      0.84      0.82        79
           1       0.65      0.62      0.63        39

    accuracy                           0.76       118
   macro avg       0.73      0.73      0.73       118
weighted avg       0.76      0.76      0.76       118



TRAINING DATA
Confusion Matrix with Training Data with ID3 Algorithm:
[[183   0]
 [  0  91]]

Classification Report ID3 Algorithm:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       183
           1       1.00      1.00      1.00        91

    accuracy                           1.00       274
   macro avg       1.00      1.00      1.00       274
weighted avg       1.00      1.00      1.00       274
```

شکل ۲: نتایج الگوریتم C4.5

گراف الگوریتم حاصل از این دو الگوریتم



شکل ۳: گراف الگوریتم ID3



شکل ۴: گراف الگوریتم C4.5

الگوریتم Random Forest

الگوریتم جنگل تصادفی Random Forest یک الگوریتم محبوب یادگیری ماشین از زیرمجموعه هوش مصنوعی است که به تکنیک یادگیری نظارت شده تعلق دارد. می‌تواند برای مشکلات طبقه بندی و رگرسیون (پیش‌بینی و بیان تغییرات یک متغییر بر اساس اطلاعات متغییر دیگر) در یادگیری ماشین استفاده شود. این مبتنی بر مفهوم یادگیری گروه است، که یک فرآیند ترکیب چندین طبقه بندی کننده برای حل یک مسئله پیچیده و بهبود عملکرد مدل است.

همانطور که از نام این الگوریتم پیداست، الگوریتم جنگل تصادفی Random Forest یک طبقه بندی است که شامل تعدادی درخت تصمیم در زیرمجموعه های مختلف مجموعه داده قرار دارد و برای بهبود دقت پیش‌بینی آن مجموعه داده، میانگین می‌گیرد. جنگل تصادفی به جای تکیه بر یک درخت تصمیم، پیش‌بینی را از هر درخت و براساس اکثریت آرا پیش بینی می‌کند و نتیجه نهایی را به عنوان خروجی در نظر می‌گیرد. تعداد بیشتر درختان در جنگل منجر به دقت بالاتری می‌شود و از بروز مشکل Overfitting جلوگیری می‌کند.

نتایج اجرای Random Forest

	precision	recall	f1-score	support
0	0.93	0.81	0.87	147
1	0.84	0.94	0.89	153
accuracy			0.88	300
macro avg	0.88	0.88	0.88	300
weighted avg	0.88	0.88	0.88	300

```
[[119  28]
 [  9 144]]
```

شکل۵: گراف الگوریتم Random Forest