گزارش 3:

یادگیری ماشین

درخت تصمیم

Decision Tree

سید علی مجتبوی

گزارش 2 - 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

دیتاستPima Indians Diabetes

**دیتاست:**

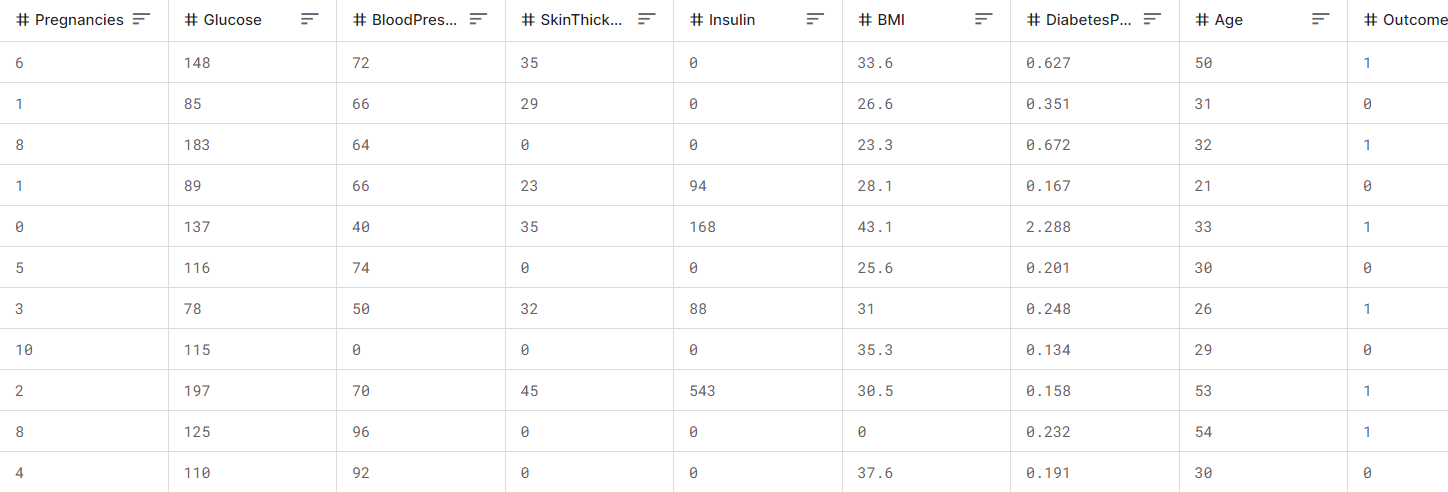
این مجموعه داده در اصل از موسسه ملی دیابت و بیماری های گوارشی و کلیوی تهیه شده است. هدف مجموعه داده این است که براساس اندازه‌گیری‌های تشخیصی خاص موجود در مجموعه داده، پیش بینی کنیم که آیا بیمار به دیابت مبتلا است یا خیر. چندین محدودیت برای انتخاب این نمونه ها از یک پایگاه داده بزرگتر قرار داده شد. به طور خاص، همه بیماران در اینجا زنان حداقل 21 ساله از میراث هندی پیما هستند.

**1. لینک دیتاست:** <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>

**2.** **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.

**3. فیچرهای دیتاست:**



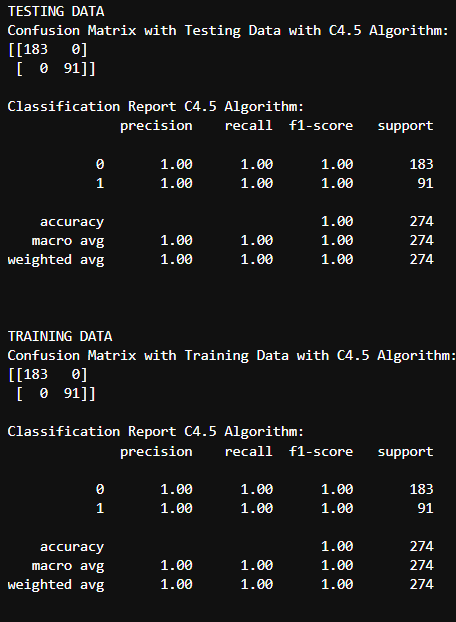
**مقایسه انواع الگوریتم های درخت تصمیم:**

ID3 (Iterative Dichotomiser 3): .1 •

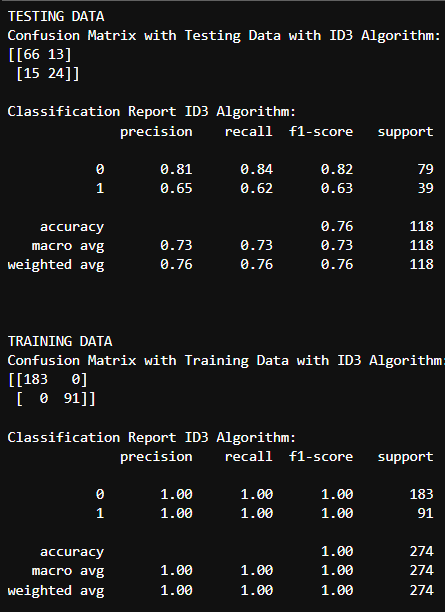
یک الگوریتم ساخت درخت تصمیم است که توسط Quinlan Ross ارائه شد. این الگوریتم برای ساخت درخت تصمیمی از جنس دودویی (tree binary (استفاده می کندباید به این نکته توجه داشت که این الگوریتم برای مقیاس پذیری به مشکل می خورد و نمی تواند با داده های پیچیده یا مقیاس پذیر به خوبی کار کند .

C4.5: .2

یک نسخه بهبود یافته از 3ID است و توسط Quinlan Ross نیز ارائه شد 4.5C . درخت های تصمیم چندجمله ای می سازد و قابلیت کار با ویژگی های عددی را نیز دارد و دارای قابلیت انتخاب ویژگی ها با استفاده از معیارهای مانند Entropy یا.Index Gini CART (Classification می باشد.

****

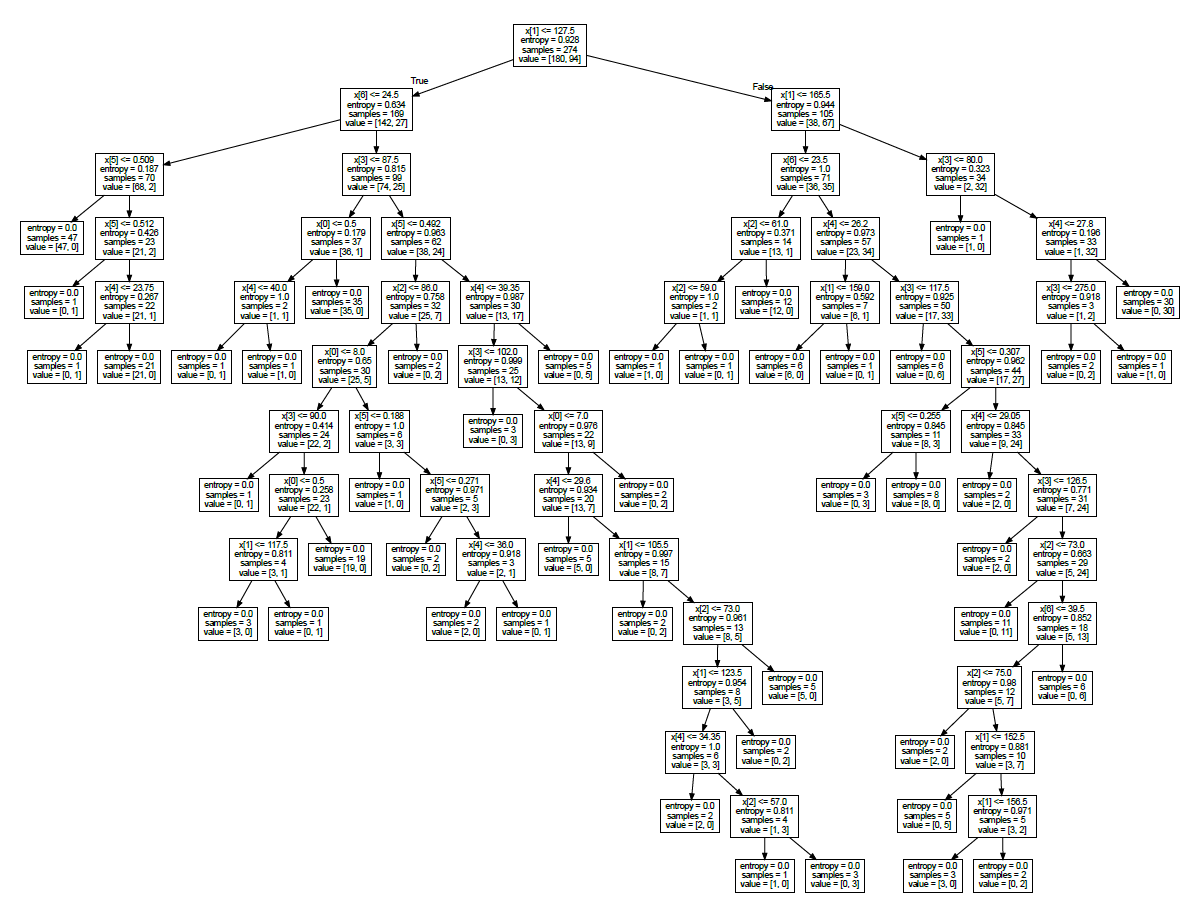
**نتایج اجرای 2 نوع الگوریتم برای این دیتاست:**



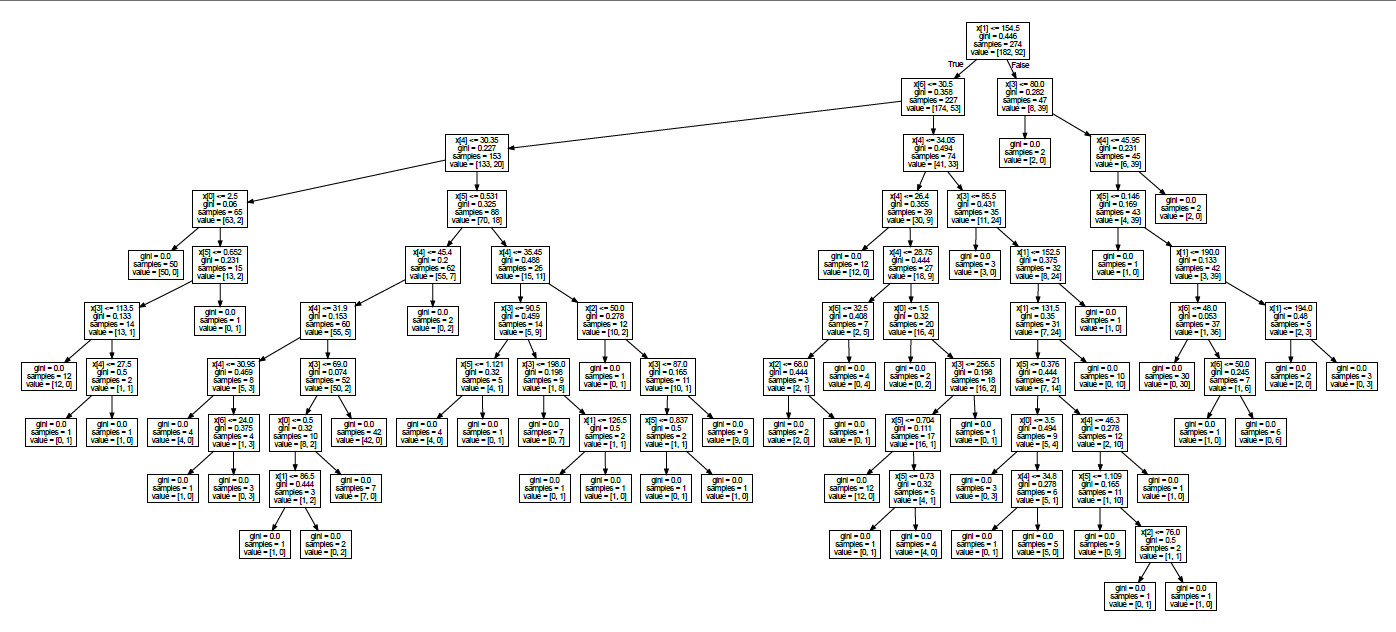
شکل2: نتایج الگوریتم C4.5

شکل1: نتایج الگوریتم ID3

گراف الگوریتم حاصل از این دو الگوریتم



شکل3: گراف الگوریتم ID3



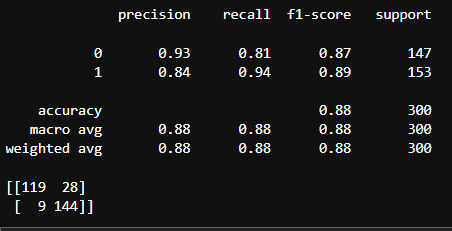
شکل4: گراف الگوریتم C4.5

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 اجرایRandom Forest**



شکل5: گراف الگوریتم Random Forest