با نام خدا

محمدرضا رضایی

گرایش شبکه های کامپیوتری

شماره دانشجویی 9952133007

درس یادگیری ماشین

توضیحات پروژه الگوریتم های یادگیری

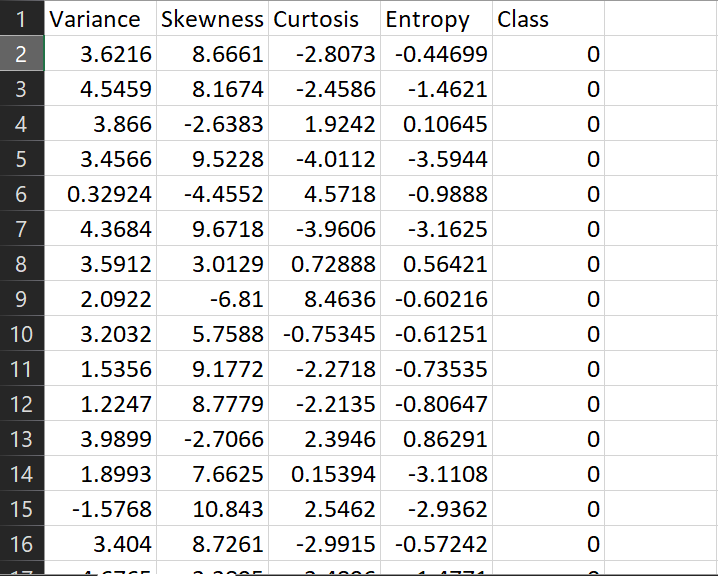
الگوریتم XGboost

الگوریتم Decision Tree

الگوریتم Random forest

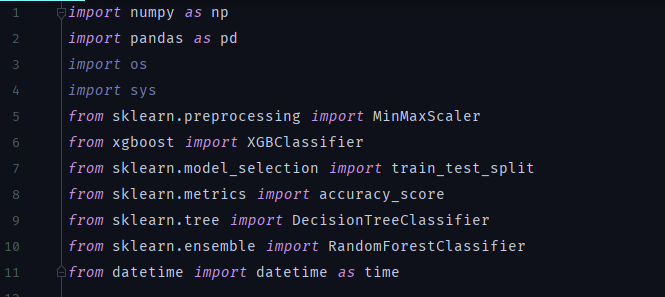
و تفاوت نوع ساختار داده

**نگاهی به دیتاست bill Authentication**

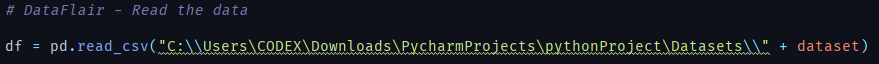
****

**توضیحات الگوریتم Decision Tree برای دیتاست bill Authentication**

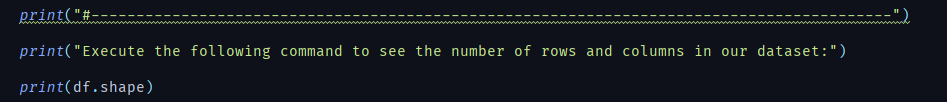
در ابتدا کتاب خانه های مورد نیاز رو فراخوانی می کنیم.



با استفاده از تابع read\_csv() دیتا ست مورد نظر و ادرس دیتا ست در سیستم را به این تابع پاس میکنیم و خروجی را در یک ابجکت میریزیم به نام data flair



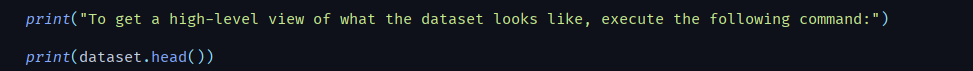
سپس برای مشاهده تعداد سطر و ستون دیتا ست کد زیر را اجرا میکنیم.

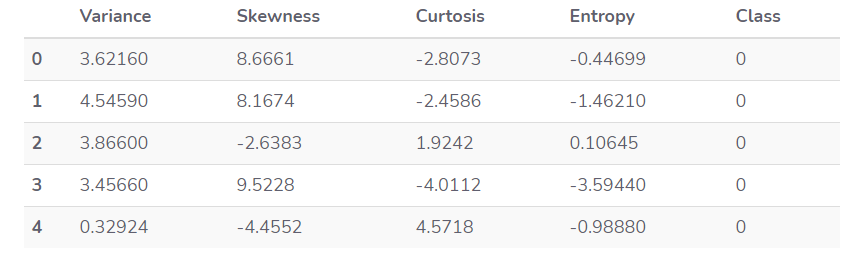


خروجی به شکل زیر است:

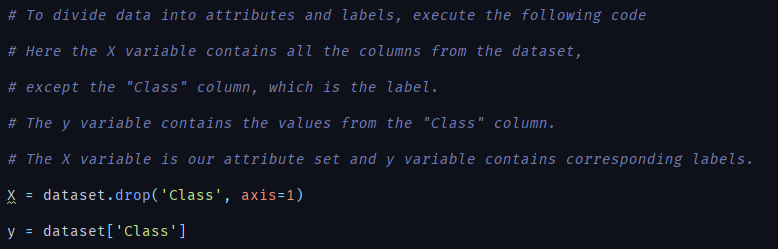


سپس برای مشاهده پیش نمایشی از دیتا ست شامل پنج سطر اول است کد زیر را اجرا میکنیم:





در این مرحله فیچرها و لیبل رو در دیتاست با استفاده از کد زیر جدا میکنیم :



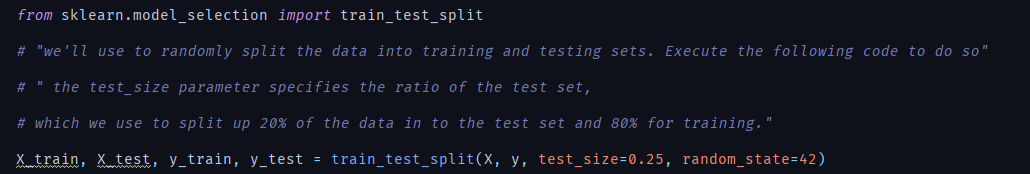
یا میتوان از کد زیر نیز استفاده کرد :



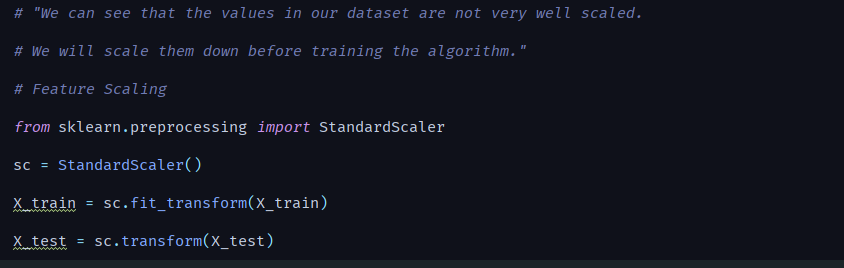
X همون فیچرها هستن که برای جداسازی تمام سطر ها و تمام ستون ها به جز ستون اخر که اسمش class هست رو انتخاب می کنیم.

Y نیز لیبل هست و تنها شامل ستون اخر می شود

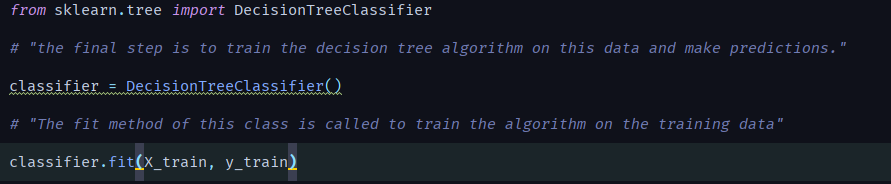
بعد از تابع train\_test\_split() که به عنوان ورودی فیچرها و لیبل همون x , y و اندازه داده های تست که 25 درصد کل داده ها هستند و پارامتر random\_state = 42 را به این تابع پاس میدهیم که کنترل شافلینگ داده ها قبل از جدا سازی را مشخص میکند. خروجی که به ما میدهد شامل جداسازی فیچرها و لیبل به داده های تست و ترین هست. کد به شکل زیر است:



حال باید داده های فیچر را برای دیتای تست و ترین مقیاس بندی کنیم سپس روی داده ها فیت و به این مقیاس منتقل شوند :



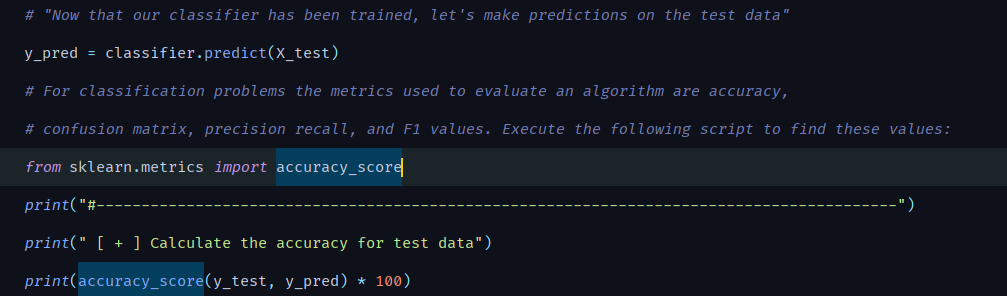
سپس کلسیفایر رو مشخص میکنیم و روی داده های ترین فیت میکنیم تا اموزش ببینه و یادبگیره:



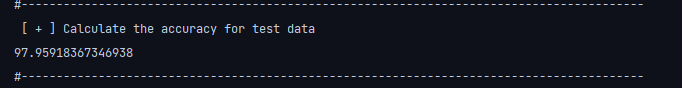
پارامترهای پیش فرض مربوط به درخت تصمیم به شکل زیر هستند:



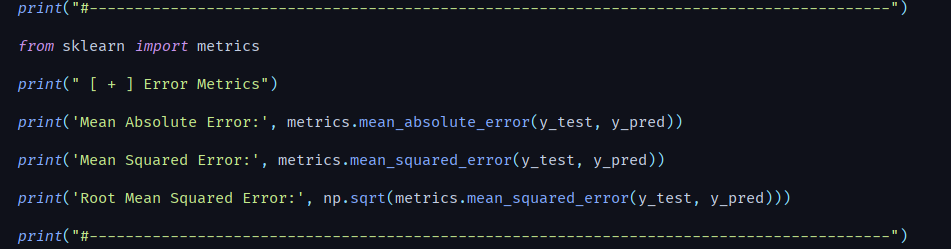
وقتی داده های ترین رو یاد گرفت روی فیچرهای داده های تست عمل پیش بینی انجام میدهیم تا ببینیم به چه میزان دقت داشتیم این کار را با استفاده از تابع accuracy\_score() انجام میدهیم لیبل داده های تست و لیبلی که پیش بینی کردیم رو بهش میدیم و دقت رو برای ما نمایش میدهد :



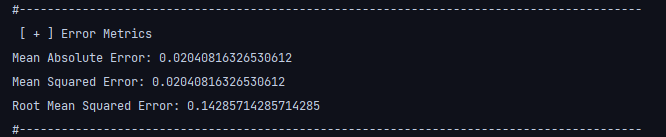
خروجی به شکل زیر است :



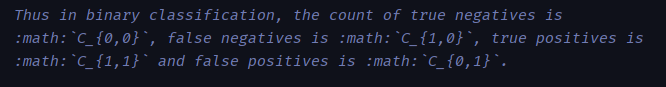
برای محاسبه ی خطا ها نیز کد های زیر را وارد میکنیم :

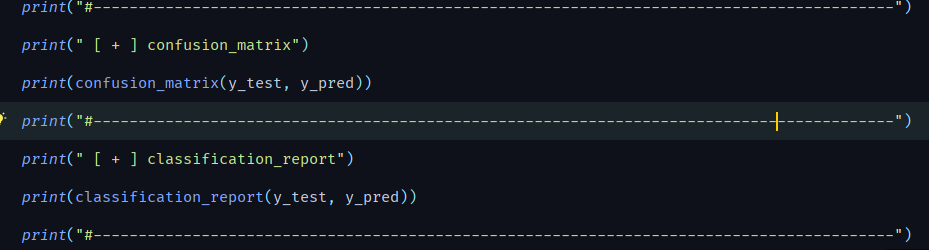


خروجی به شکل زیر است :

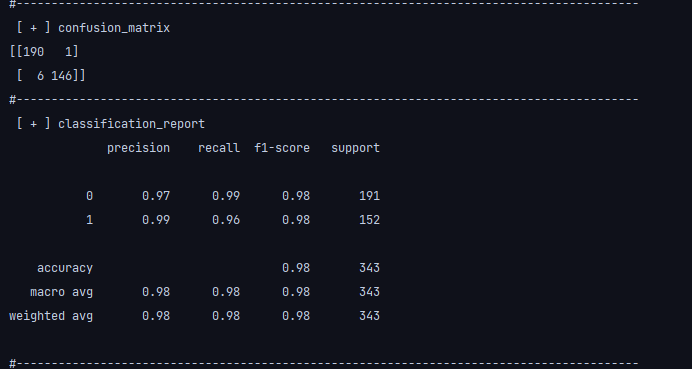


مرحله اخر نیز محاسبه ی confusion matrix , classification report هست که ستون اول شامل negative ها و ستون دوم شامل positive ها هست سطر اول ture و سطر دوم false هست پس هر چقد قطر اصلی بیشتر باشد بهتر است و خطای کمتری داریم :





خروجی به شکل زیر است :



مشاهده میکنیم که طبق ماتریس 7 خطا داریم.

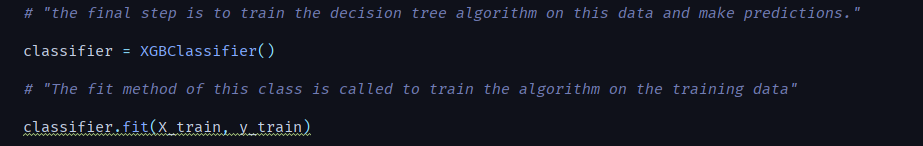
و زمان سپری شده برای الگوریتم درخت تصمیم تا تمام مراحل بالا را انجام دهد :



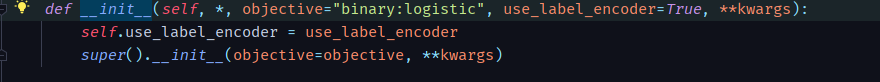
39 میلی ثانیه است.

**توضیحات الگوریتم xgboost برای دیتاست bill Authentication**

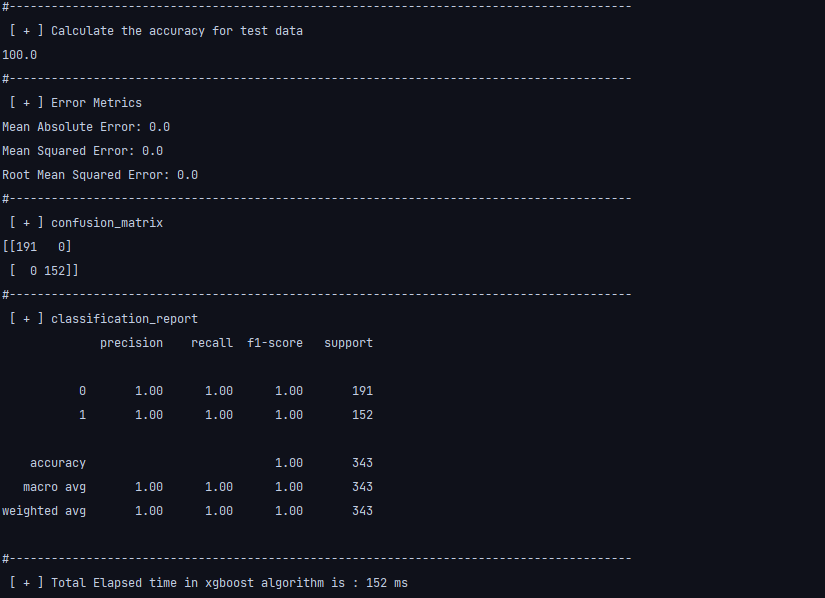
برای این الگوریتم نیز کل کد یکسان است فقط باید کتاب خانه مربوطه ایمپورت شده و به عنوان کلسیفایر انتخاب شود تا داده های ترین را اموزش ببیند :



پارامترهای مربوط به الگوریتم xgboost به شکل زیر است:

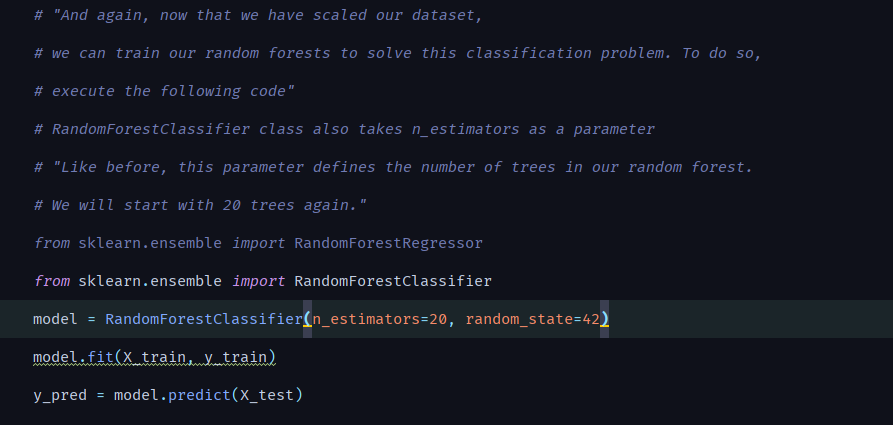


تمام خروجی ها برای الگوریتم xgboost :



**توضیحات الگوریتم random forest برای دیتاست bill Authentication**

برای این الگوریتم نیز کل کد یکسان است فقط باید کتاب خانه مربوطه ایمپورت شده و به عنوان کلسیفایر انتخاب شود تا داده های ترین را اموزش ببیند :

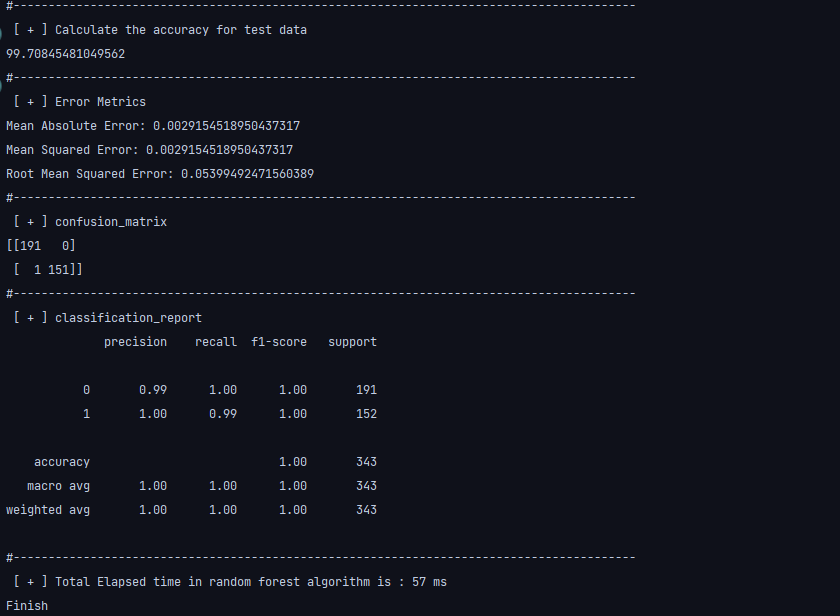


به عنوان پارامتر ها تعداد درختان جنگل رو 20 و random state رو 42 در نظر میگیریم.

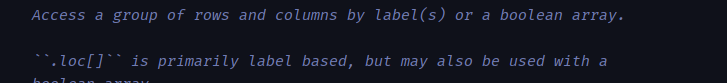
پارامترهای پیش فرض به شکل زیر است:



تمام خروجی ها برای الگوریتم random forest :

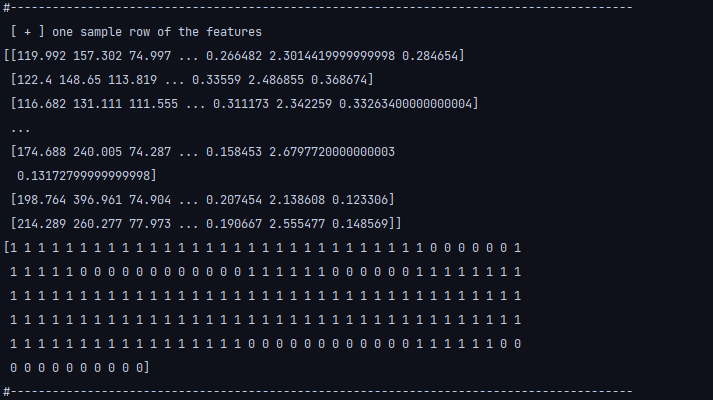


تابع loc که برای دست یابی به گروهی از سطر یا ستونی از داده هاست که لیبلشون مشخصه



حال برای مقایسه نوع داده ها که در کلاس بحث شد:

فیچر ها یک ارایه یا ماتریس دوبعدی است و لیبل یه ارایه ی تک بعدی

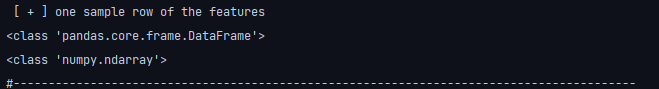


همونطور که خود پایتون هم نشون داده این دوتا از نوع ارایه n بعدی هستند :



ولی خود دیتاست که داخل ابجکت دیتافریم قرار داده شد از کتاب خانه pandas هست و ارایه نیست.

که داخل دیتافریم علاوه بر مقادیر یا value شامل لیبلی labels نیز هستند



و پایان

ممنون از شما استاد گرامی