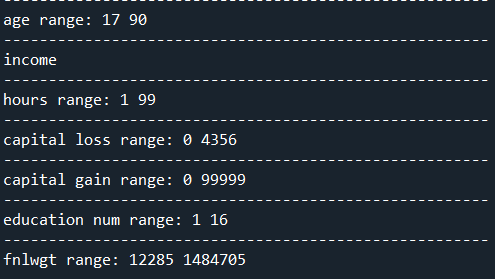
داده کاوی

تمرین سری دوم

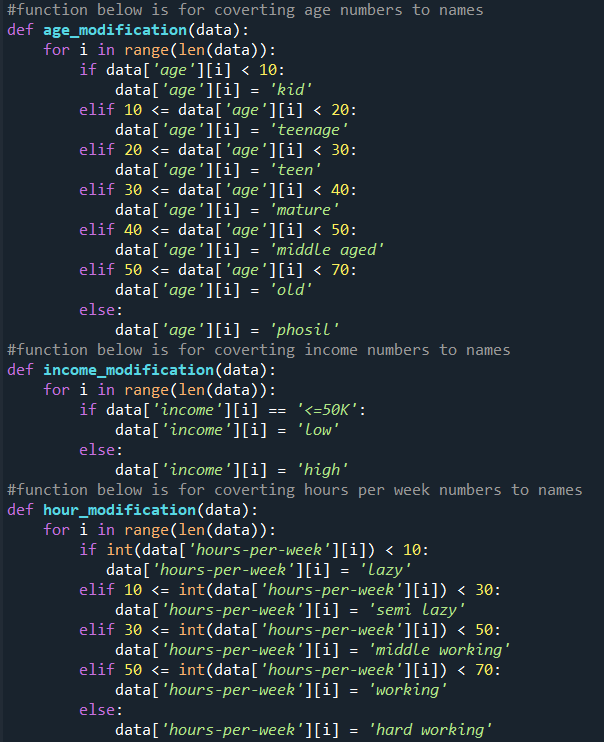
بخش عملی

سوال1- درخت تصمیم گیری)

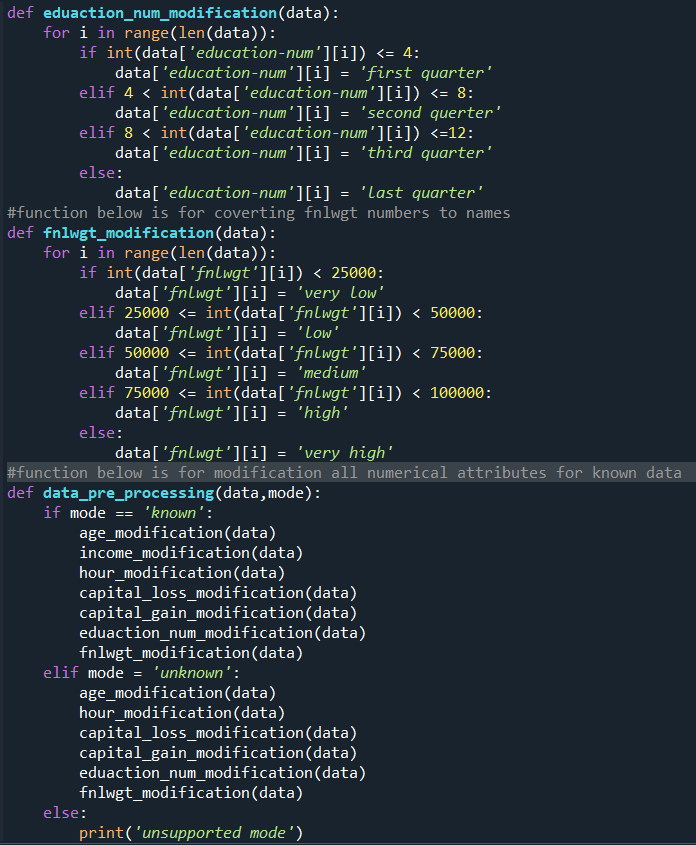
در ابتدا بازه های متغیر های عددی را پیدا کردم و بعد با استفاده از تابع data\_pre\_processing تبدیل داده عددی به فهرستی را انجام دادم.



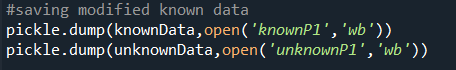
در شکل های زیر توابع مربوط به تبدیل فهرست عددی به اسمی را مشاهده میکنیم





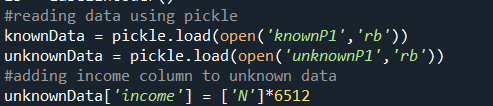


با استفاده از توابع بالا ما پیش پردازش داده را انجام میدهیم و در نهایت با استفاده از کتابخانه pickle ما عملیات ذخیره سازی را انجام میدهیم.

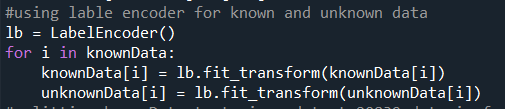


حال به پردازش داده اصلاح شده میپردازیم.

با استفاده از کتابخانه pickle داده های ذخیره شده را که اصلاح کرده بودیم میخوانیم و به داده های ناشناخته ستون درامد را اضافه میکنیم.

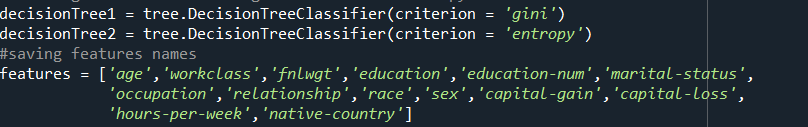


با استفاده از sklearn.preprocessing داده ها را یک دور دیگر برای استفاده‌ی درخت تصمیم پیش پردازش میکنیم.



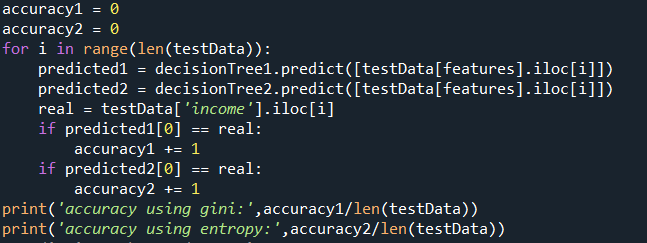
با استفاده از کتابخانه sklearn ما tree را import میکنیم.

در شکل زیر دو نوع درخت میسازیم یکی با معیار جینی دیگری با معیار انتروپی و ویژگی هارا در یک آرایه نگهداری میکنیم.



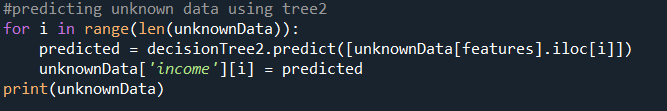
80 درصد داده ها برای اموزش است و 20 درصد برای تست که در واقع 20839 از داده ها برای یادگیری و الباقی برای تست میباشد.

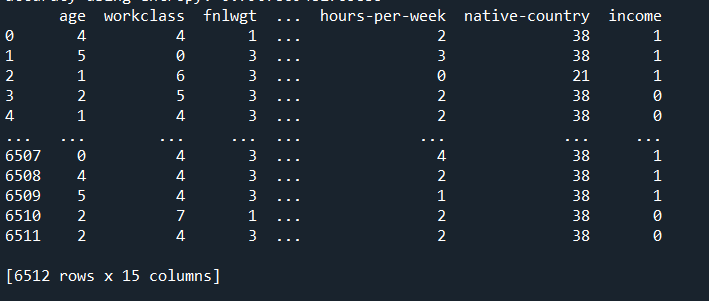
حال با استفاده از کد زیر پیش بینی را انجام میدهیم روی داده تست و دقت را محاسبه میکنیم





از انجایی که دقت برای انتروپی بیشتر است داده های ناشناخته را با درخت دوم پیش بینی میکنیم.





سوال 2-KNN)

همانند قست قبلی بازه های داده های عددی را پیدا میکنیم و آن هارا تبدیل به داده های فهرستی میکنیم.

با چک کردن داده ها متوجه میشویم که همه آن ها فهرستی هستند و داده عددی نداریم برای همین به سراغ پیاده سازی الگوریتم میرویم.

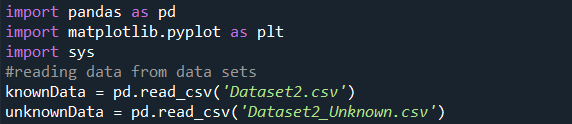
از آنجایی که knn یک الگوریتم lazy learning است نیازی به اموزش ندارد و کل هزینه را در زمان تست متحمل میشویم.

80 درصد از داده هارا برای مقایسه به کار میبریم و 20 درصد از داده هارا برای تست

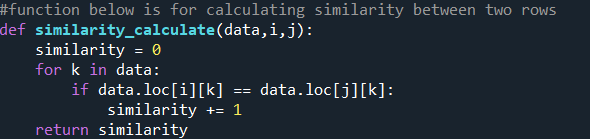
الگوریتم مورد استفاده توی این سوال به این گونه است که هر داده تست را با تک تک داده های های اکوزش مقایسه میکنم و معیار شباهت یه این صورت است که هر ویژگی که از داده تست با آموزش یکی باشد یکی به مقدار شباهت اضافه میشود.و بر اساس k های مختلف k تای برتر رای گیری میکنند و لیبل را مشخص میسازند.

تعداد کل داده ها 6499 میباشد که 5200 برای اموزش و 1299 برای تست استفاده میشود.

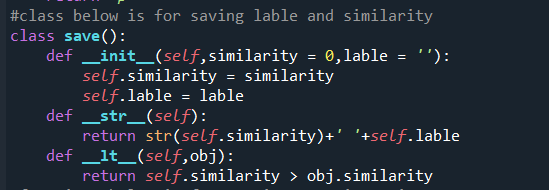
کتاب خانه های مورد نظر را ایمپورت میکنیم و داده هارا میخوانیم



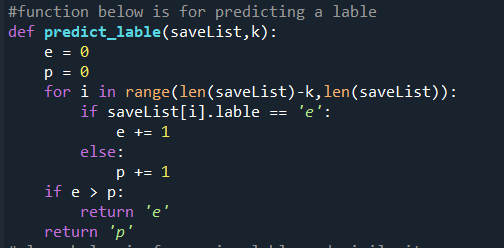
تابع محاسبه شباهت زیر برای مقایسه شباهت دو سطر از داده ما است به صورت که شباهت برابر است با تعداد ویژگی های یکسان



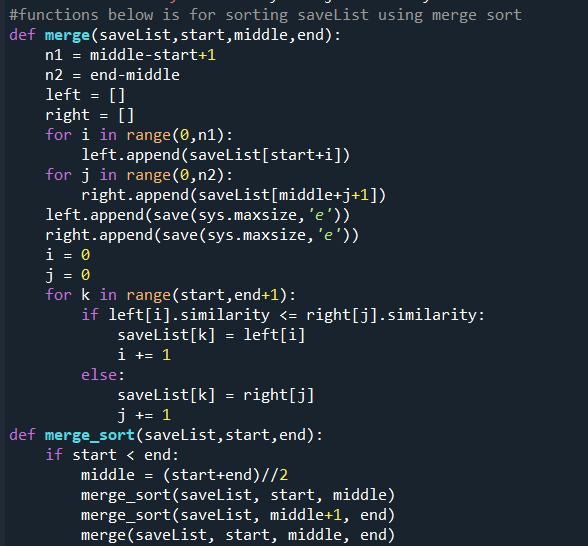
کلاس زیر برای نگهداری میزان شباهت و لیبل میباشد



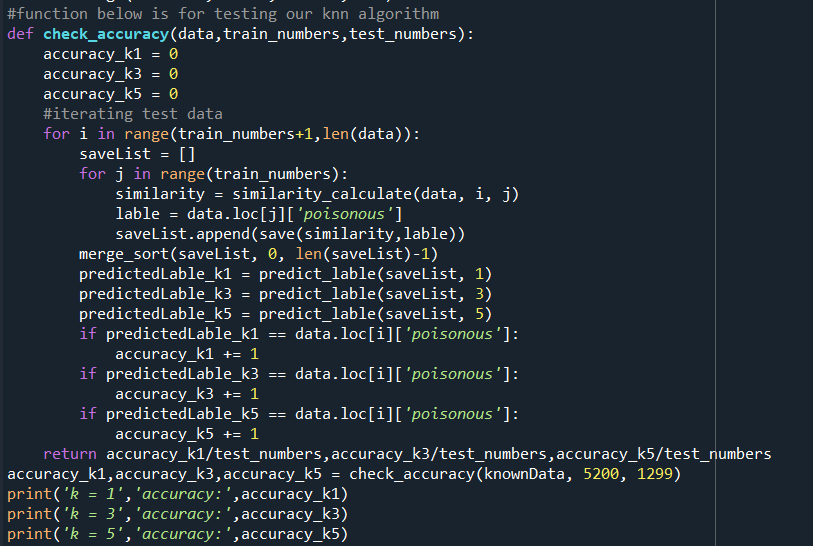
تابع زیر برای پیش بینی داده میباشد بر اساس k داده شده



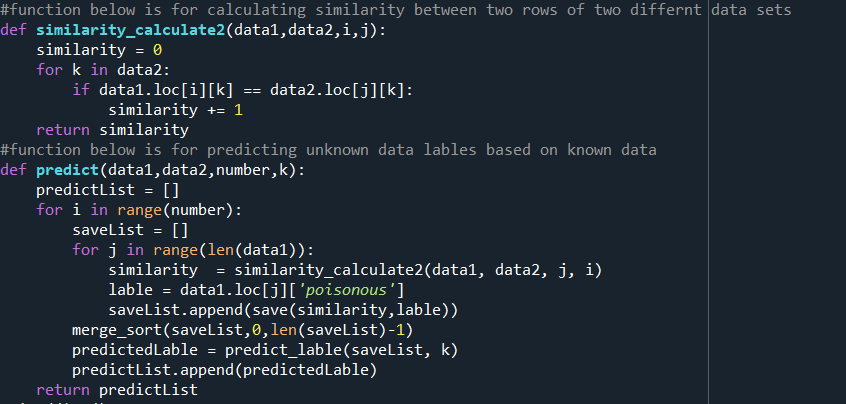
دلیل استفاده از merge sort این است که بار محاسبات کمتر شود این کار برای مرتب کردن لیستی از سیو هاست به این صورت که به صورت نزولی بر اساس شباهت مرتب میشوند



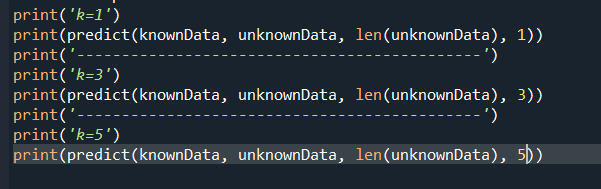
محاسبه و چاپ کردن دقت بر اساس 3 مقدار k



دو تابع زیر نیز برای محاسبه شباهت و پیش بینی میباشند اما با این تفاوت که بین دو داده مختلف میباشد



در نهایت چاپ کردن نتیجه به ازای 3 مقدار مختلف k



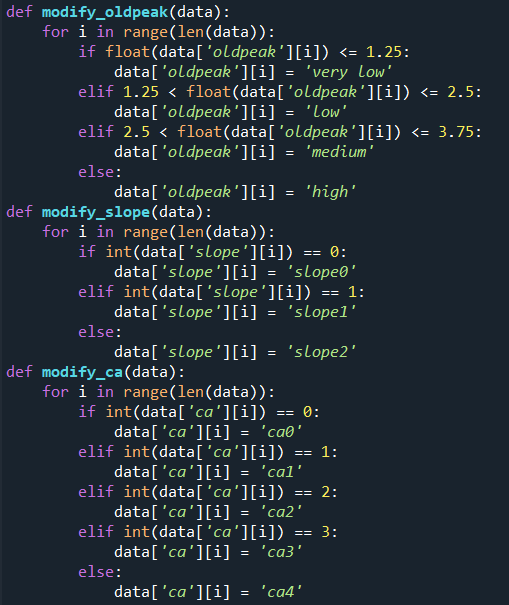
سوال 3-بیض ساده)

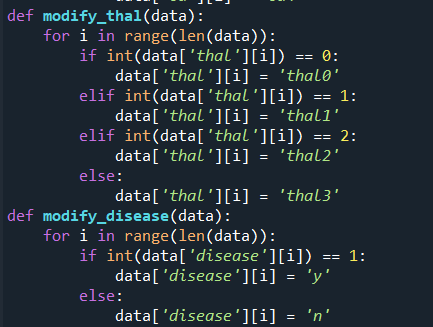
در قدم اول باید پیش پردازش داده را انجام دهیم به این صورت که داده های عددی را تبدیل به داده های فهرستی میکنیم

از توابع زیر برای انجام این کار استفاده میکنیم.

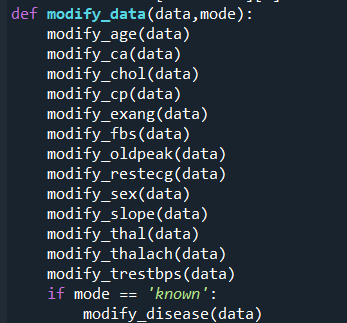








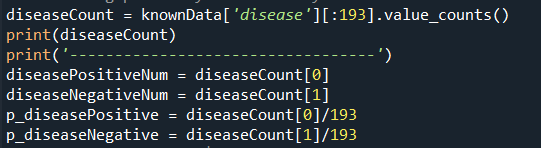
هرکدام از توابع بالا برای تبدیل داده عددی به فهرستی یک ستون از جدول هستند

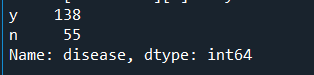


و تابع بالا کار نهایی اصلاح داده را انجام میدهد که اگر داده شناخته شده باشد مقدار ستون disease را نیز اصلاح میکند.(توجه شود که داده های ناشناخته ستون disease را ندارد لذا دستی این ستون را اضافه کردم.)



حال باید از 193 داده اول برای یادگیری مدل استفاده کنیم و بقیه داده ها برای تست با استفاده از قطعه کد زیر تعداد و احتمال حالت های بله/خیر بیماری را محاسبه میکنیم.

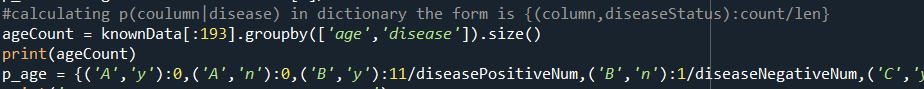


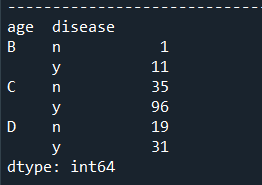


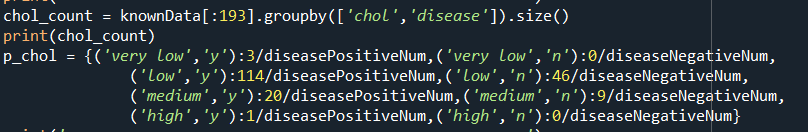
برای اعمال شرط با محدودیت روی دو ستون از دستور groupby استفاده میکنیم.

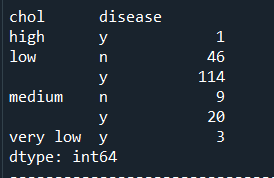
و مشاهده میکنیم که حالت های مختلف هر ستون به ازای y/n بودن ستون بیماری چه تعداد تکرار دارد و با استفاده از آن احتمال شرطی را مینیویسیم.

در شکل های زیر نمونه ای از محاسبات را مشاهده میکنیم.



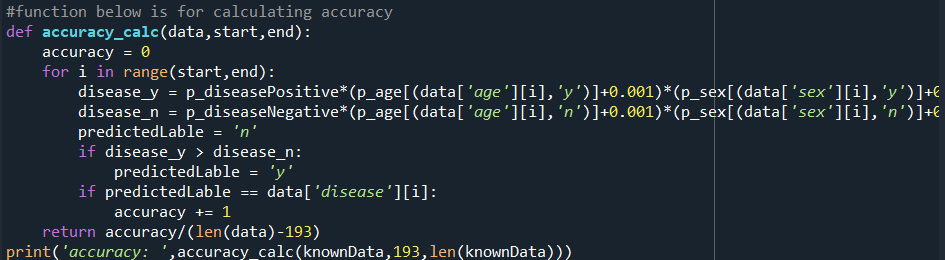




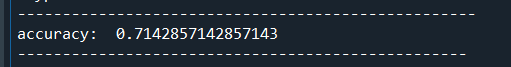


با توجه به شکل بالا تعداد حالت های مختلف هر ستون با ستون بیماری قابل مشاهده است که میتوان احتمال شرطی را با داشتن این اعداد نوشت.

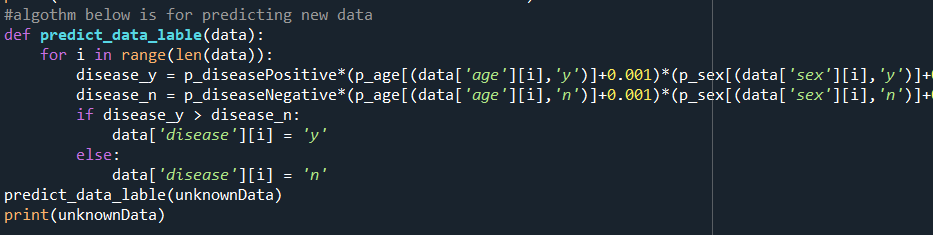
کد زیر نیز برای ارزیابی مدل با استفاده از داده های تست میباشد که از داده 193 به بعد میباشد که با حاصل ضرب احتمال ها به شرط بله/خیر هر کدام که احتمال بیشتری داشته باشد به عنوان لیبل پیش بینی شده در نظر گرفته میشود.(هر احتمال نیز با مقدار 0.001 جمع میشود تا اگر احتمالی برابر صفر باشد تاثیر منفی شدیدی روی احتمال نهایی نگذارد.)



حال مقدار دقت را مشاهده میکنیم که برابر با 0.71 میباشد.



کد زیر نیز برای پیش بینی لیبل داده ناشناخته میباشد.



در شکل زیر مشاهده میشود برای تمامی نمونه ها با استفاده از بیض ساده پیش بینی انجام شده است.

