به نام خدا

پروژه دوم هوش مصنوعی و سیستم های خبره

استاد : دکتر آرش عبدی

هانیه اسعدی

99521055

زمستان 1401

داده هایی در فایل diabetes.csv داریم که میخواهیم با استفاده از این داده ها، و اطلاعاتی که از فرد مشخصی داریم، ببرسی کنیم احتمال مبتلا بودن فرد با توجه به اطلاعاتی که از پیش داریم، به دیابت چقدر است.

به این منظور داده ها را به دو قسمت آموزشی و آزمایشی تقسیم میکنیم. در اینجا 80 درصد داده ها برای آموزش و 20 درصد برای آزمایش هستند. یعنی با استفاده از 80 درصد داده ها، درختی تولید میکنیم و با 20 درصد دیگر، درخت را پیمایش میکنیم و خروجی درخت را بررسی میکنیم. اگر خروجی درخت با خروجی داده یکسان بود، یعنی درخت درست حدس زده، در غیر این صورت، اشتباه بوده است. سپس درصد درستی را بدست می آوریم.

ابتدا، از بین داده ها، داده های آموزشی و آزمایشی را انتخاب میکنیم. برای اینکار هیچگونه اولویتی نداریم. سپس بر اساس داده های آموزشی، شروع به ساخت درخت میکنیم. چون کلا 8 ستون داریم، پس درختی با حداکثرعمق 7 خواهیم داشت.

ابتدا آنتروپی داده ها را بر اساس ستون outcome داده ها طبق فرمول آنتروپی حساب میکنیم. سپس به محاسبه مقدار information gain برای هر ستون میپردازیم. برای این کار، باید برای هر ستون، یک مقدار برای نقطه شکست یا نقطه دسته بندی در نظر بگیریم. که این مقدار را میتوان میانگین بین مقدار ماکسیمم و مینیمم اعداد موجود در هر ستون در نظر گرفت. حال مقدار آنتروپی برای هر ستون را با توجه به مقادیر ستون outcome بدست می آوریم.

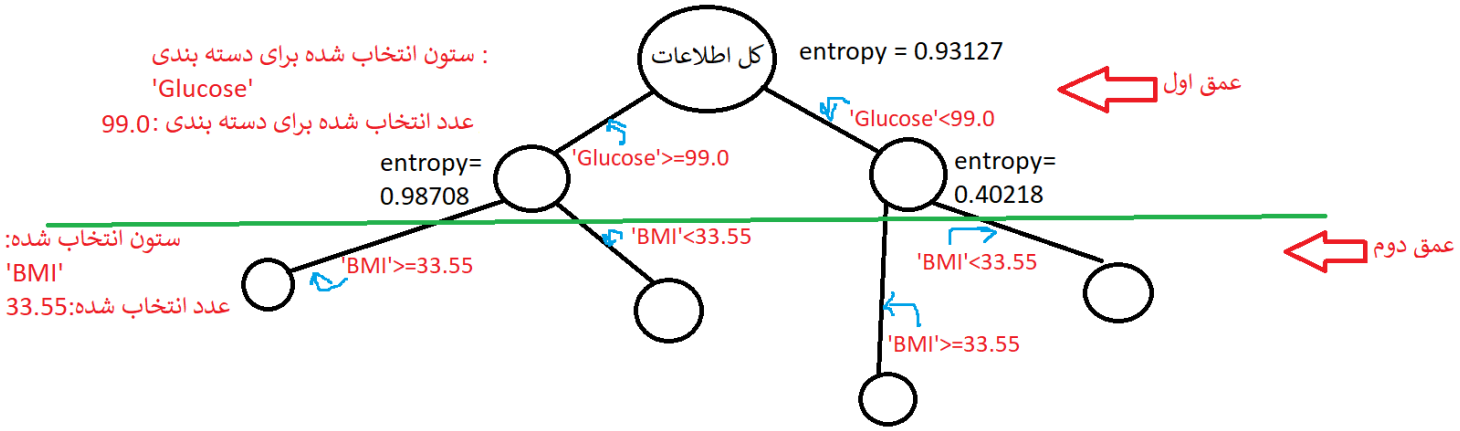
مقدار information gain هر ستون را از کم کردن مقدار آنتروپی آن ستون از آنتروپی داده ها بدست می‌آوریم. هر ستون که information gain بیشتری به ما بدهد برای انتخاب در عمق بعدی مناسب است.

مقادیر information gain ستون ها را از بیشترین به کمترین مرتب میکنیم. برای بدست آوردن عمق اول درخت، ستون بیشترین information gain را انتخاب میکنیم. برای دومین عمق، دومین ستون با بیشترین information gain و الی آخر.

پس ستون با information gain بیشتر را انتخاب میکنیم و داده ها را با توجه به نقطه دسته بندی آن تقسیم میکنیم. این کار را به صورت بازگشتی تا زمانی ادامه میدهیم که به برگ برسیم. شرط برگ بودن، این است که یا به حداکثر عمق رسیده باشیم و یا یکی از داده ها حداکثر 0.5 درصد از داده ها به خود اختصاص داده باشد. در برگ، مقدار خروجی را برابر قرار میدهیم با داده ای که در ستون outcome تکرار بیشتری داشته است. اگر در حساب کردن یکی از child های یک راس، داده های آن child به تعداد صفر رسید، آن را none قرار میدهیم.

پس از تمام شدن کار با داده آموزشی، به بررسی داده آزمایشی میپردازیم. برای چک کردن داده های آزمایشی، هر ردیف از داده ها را با توجه به درخت پیمایش میکنیم. اگر مقداری که درخت خروجی میدهد برابر با مقداری باشد که ستون outcome داده، داده است، پس جدس درخت تصمیم درست بوده و شمار حدس های درست را افزایش میدهیم. پس از بررسی تمام داده آزمایشی، از تقسیم تعداد حدس های درست بر تعداد داده ها، درصد درست حدس زدن دیابت توسط درخت را بدست می آوریم.

این درخت بر اساس عملکرد کد در دو عمق اول رسم شده است.



برای اجرای کد، فایل Test.py ران شود.

خروجی کد:

```
The order of attributes to branching tree is :
Glucose , BMI , Pregnancies , BloodPressure , Insulin , Age , DiabetesPedigreeFunction , SkinThickness
This program diagnoses diabetes with a probability of 68.182%
```

لینک کمکی