**به نام خدا**

**مریم ترکاشوند**

**کلاس مبانی و مفاهیم علم داده**

**استاد گرامی جناب آقای دکتر محتاط**

**بخش اول: فهم مساله کسب و کار:**

فایلی که در اختیار داریم داده های افراد مراجعه کننده به یک بیمارستان در هند می باشد که از طریق آزمایشات و کسب اطلاعات از مراجعین بدست آمده است.

**بخش دوم: درک داده**

تعداد بیماران 768 نفر است که اطلاعات دریافتی از آنها به شرح ذیل می باشد:

1- تعداد دفعات بارداری

2- غلظت گلوکز

3- فشار خون

4- ضخامت چین خوردگی سه سر بازو

5- انسولین

6- شاخص توده بدنی (BMI)

7- سابقه خانوادگی

8- سن

9-یک فیلد تارگت هم وجود دارد که نتیجه مثبت و منفی دیابت را با عدد صفر و یک نشان می دهد.

حال ما میخواهیم با استفاده از داده هایی که در اختیار دادیم داده ها را پاکسازی کنیم و یک دیتای تمیز استخراج کنیم تا بتوانیم بر اساس داده های موجود دیابت افراد را پیش بینی کنیم.

**بخش سوم: آماده سازی و پیش پردازش داده**

مراحل پاکسازی داده دیابت در نرم افزار ibm spss modeler:

بعد از ایمپورت فایل در نرم افزار اولین مرحله تعیین تایپ داده ها میباشد که در اینجا من ستون Pregnancies را اردینال و ستون outcome را فلگ و بقیه فیلدها را عددی تعیین کرد.

مرحله بعد تشخیص نوع داده ها می باشد جهت شناسائی داده پرت

برای این کار با ابزار simfit داده ها را شناسایی کرده، من در اینجا از روش کولموگروف استفاده کردم و 5 تا از فیچرها را این ابزار نرمال تشخیص داد.

با استفاده از ابزار data oudit داده های پرت را شناسایی کرده(standard divition جهت داده های نرمال و گزینه interquartile… جهت داده های غیرنرمال)

برای داده های normal داده های پرت را coerce کرده و داده های خیلی پرت را nullify کردم.

برای داده های lognormal نیز داده های پرت را coerce کرده و داده های خیلی پرت را nullify کردم.

در دیتای ما تعداد زیادی فیلد صفر وجود دارد که در واقعیت از نظر مفهومی این فیلدها نمیتوانند مقدار صفر داشته باشند و درواقع این داده ها null هستند. (فقط فیلد تعداد دفعات بارداری می تواند صفر داشته باشد)

در این مرحله ما باید داده های صفر را به null تبدیل کنیم که با استفاده از ابزار filler و ایجاد شرط برای فیلدهای مورد نظر، این عمل را انجام می دهیم.

مرحله بعد مدیریت داده های مفقوده می باشد که با استفاده از ابزار data audit شناسایی و مدیریت میشود.

در اینجا فیلدهای نرمالی که داده مفقوده دارند را متد random ، imput کردم. و یک فیچر DiabetesPedigreeFunction را که lognormal است را هم از روش random(uniform) ، جایگذاری کردم .

در اینجا داده کیفی نداریم و نیازی به تبدیل کردن داده کیفی به کمی وجود ندارد.

مرحله بعد، هم مقیاس کردن یا scale کردن داده می باشد که با استفاده از ابزار auto data prep و با استفاده از متد min/max هم مقیاس سازی را انجام می دهیم.

**دیتای پاکسازی شده ما جهت مراحل بعدی آماده می باشد.**

**مدلسازی:**

جهت مدلسازی داده ها یک نمونه کوچک به تعداد 5 ایتم با استفاده از selectاز بخش record ops جدا می کنیم که با نتیجه خروجی مدلهایمان بتوانیم مقایسه بکنیم.

جهت مدلسازی با روش test و train یکبار دیگر با استفاده از select مابقی داده ها را جدا می کنیم(با استفاده از دستور @INDEX>6

سپس جهت تعیین مقدار test , train از ابزار partishon از قسمت field ops استفاده می کنیم، در اینجا مقدار test=20% و train=80% را تعیین می کنیم.

بنده در این پروژه از مدلهای ذیل جهت مدلسازی استفاده نمودم:

KNN

CHIN

QUEST

C&T TREE

Neural Net

Random Tree

SVM

تقریبا برای همه مدلها دقت یکسانی بدست امد و نزدیک به 75% بود

در همه مدلها اگر مدل overfit میشد تعداد عمق درخت و همچنین لایه های درخت را تغییر میدادم و همچنین در partishon نیز seed را generate کردم و دوباره مدل را اجرا می کردم که بهینه ترین حالت را پیدا کنم ولی تقریبا تفاوتی نمی کرد.