**به نام پروردگار هستی بخش**

**مراحل پروژه دیابت**

**محمدسعید حیدری**

**نکته بسیار مهم**: اگر در فیچر های من مقداری ordinal به صورت high low normal وجود داشت آنها را توسط drive از کیفی به کمی تبدیل میکنم . حالا بعد از ساخت فیلد کمی جدید با درایو اگر مدل من svm یا knn باشد در قسمت data type مقدار رو روی countinuse میزارم اما اگر درخت تصمیم باشد چون به کیفی نیاز دارد روی گزینه Ordinal میزارم

# مراحل کار

**بخش اول: مرتب کردن داده و انتخاب نوع داده**(Data Type) **می باشد . داده را از طریق** var.file **فراخوانده و اگر فیلدی نیاز به کمی کردن یا کیفی کردن داشت. با** Drive **آنرا ایجاد و با** Filler **آن را آپدیت میکنیم.**

**بخش دوم : مراحل پاکسازی داده های پرت و نویزی و مدیریت کردن اینجور داده ها می باشد. و نرمال سازی و هم اسکیل کردن داده ها می باشد.**

**چون این پروژه دیابت دارای صفرهای نویزی هستند با نوشتن شرط زیر توسط** filler **صفرهای آنرا تبدیل به** null **میکنیم**

If @field = 0 then undef else @field endif

**لازم به ذکر است در قسمت فیلدهای اختیاری برای تبدیل صفر به** null **.** Glucose Bloodpre skin Insuline BMI **را قرار میدهیم.**

**حال با استفاده از** Data Audit **توزیع داده ها را بررسی کرده. اگر توزیع نرمال داشتند برای داده های** outlier **و مدیریت دادگان از روش** Z **ابزار استفاده میکنیم و اگر توزیع نرمال نداشتند از روش** IQR **استفاده میکنیم.**

**حال طبق مشاهدات فیلد های** Glucose BMI Skin BloodPresure **توزیع نرمال دارند و از روش** z **برای مدیریت دادگان پرت استفاده میکنیم .**

**داده های** outliers **را** Coerce **کرده و** Extreme **ها را** null **میکنیم.**

**همین رویکرد را برای فیلد هایی که توزیع نرمال ندارند نیز انجام میدهیم اما با روش** IQR **ابزار وارد بخش** Quality **می شویم. بعد از مدیریت کردن** ouliers**ها و** extreme**ها بار دوم قسمت** quality Data Audit **را باز کرده و دامنه ی** ouliers **ها و** extreme**ها را افزایش میدهیم تا مقدار دادگان پرت به کلی صفر شوند .**

**حال نوبت به مدیریت کردن** null**ها می رسد**

**ابتدا از** Data Audit **به صورت پیش فرض از مسیر** z **می رویم برای مدیریت کردن** null **ها و پر کردن آنها با روش هر خاص.**

**خب حالا از قسمت** Impute Missing **همه ی جاهای خالی رو روی گزینه** Blank&Null **قرار میدهیم حال برای پر کردن مقادیر به صورت زیر عمل میکنیم.**

**برای** Glucose **از روش میانگین**

**برای** BloodPresure **از روش میانه**

**برای** skin **از روش الگوریتم**

**برای** Insulin **از روش میانگین**

**برای** BMI **از روش میانه**

**برای** DiabetesPedidree Function **از روش الگوریتم** cart

**و برای** Age **از روش میانگین**

**برای پر کردن نال ها بهره جسته ام .**

**بعد** Generate **را میزنیم و از** 100 **درصد داده به ماشین برای محاسبه و پر کردن** null**ها استفاده میکنیم و این هم از گام مدیریت کردن داده های مفقوده .**

**حالا داده ها را هم اسکیل یا هم مقیاس میکنیم با استفاده از نود** Auto Data Prep

**نود را باز کرده و همه فیلد ها را غیر از** outcome **هم مقیاس میکنیم با روش** min max

**تیک همه ی قسمت های** prep input & Target **را برداشته و گزینه** min max **را انتخاب کرده و داده ها هم مقیاس می شوند.**

## بالانس کردن داده هامون

**با استفاده از تب گراف در پایین نود** Distribution **را فراخوانی میکنیم و آنرا به داده های هم اسکیل شده مان وصل میکنیم. سپس تارگت را** outcome **قرار داده تا داده های آنرا بالانس کنیم. یک بار با** BalanceNode(reduce) **بالانس میکنیم و یک بار با** BalanceNode(boost) **دقت هردو را بر روی مدل خواهیم سنجید . حال این عملیا ت به عنوان یک نود برای ما به پایان می رسد . که این نود رو به نودهای قبلی اتچ کرده و میریم برای مدلسازی.**

# **پایان مراحل** Data Preparation

# مدلسازی

**ابتدا قبل مدلسازی شکستن داده ها به دو قسمت** Train **و** Test **میباشد برای اینکار از قسمت** Field Ops **نود** partition **را انتخاب میکنیم . سپس وارد تنظیمات** partition **شده و سایز قسمت داده آموزشی را روی** 80 **درصد و سایز داده تست را روی** 20 **درصد میزاریم .**

**سپس** Random seed **را** Generate **کرده و با** seed 9896444 **داده هایم را به دو دسته** Train **و** Test **تقسیم کرده و تا آخر مدلسازی با همین** Random seed **پیش میریم. چون این مقدار نباید برای پروژه در مقاطع مختلف تغییر کند.**

**خب حالا چون این پروژه از نوع طبقه بندی هست پس با استفاده از الگوریتم ها ی طبقه بندی مدل خود را میسازیم**

**ابتدا از الگوریتم** KNN **شروع میکنیم . این الگوریتم با محاسبه فاصله همسایه ها پیش بینی را برای ما انجام می دهد.**

**از قسمت** Modeling **الگوریتم** KNN **را اضافه میکنیم و** partition **را به این الگوریتم وصل میکنیم .**

**حالا وارد تنظیمات** KNN **شده و از تب اول که** Objectives **نام دارد ما در قسمت اول از گزینه** Predict a target Field **استفاده میکنیم و در قسمت دوم گزینه** Balance speed and accuracy **را انتخاب میکنیم که مربوط به سرعت و دقت همسایه ها می باشد .**

**وارد تب سوم مدل** KNN **که** settings **هست می شیم . قسمت اول که** Model **نام دارد دو گزینه انتخاب** Use partition data **و** Build model for each split **دارد. که اولی یعنی از دیتای شکسته شده** Train **و** Test **استفاده کن و دومی هم داده های آموزشی و آزمایشی را می تواند برای دو جنسیت مجزا حساب کند. فقط تیک گزینه اول را فعال میکنیم .**

**وارد قسمت دوم که** Neighbors **یعنی همسایه ها می شویم . در اینجا نزدیک ترین همسایه را روی** 2 **و بیشترین را روی** 30 **میگذارم**

K flods **را پیش فرض روی** 10 **تنظیم کرده . تیک گزینه** Append all probabilities **فعال کرده و** run **را میزنیم.مدل ساخته میشود سپس با نود** Analysis **به بررسی مدلمان میپردازیم.**

**گزارش** KNN

## حالت اول

**دراین حالت با پارتیشن** 80 20 **و با** seed 9896444 **و همسایگی** 2 **تا** 40 **و بالانس** boost

K **بهینه** 21 **دقت آموزش** 77.61٪ **و دقتت تست** 79.34٪ **می باشد.**

## حالت دوم

**دراین حالت با پارتیشن** 80 20 **و با** seed 9896444 **و همسایگی** 2 **تا** 40 **و بالانس** reduce

K **بهینه** 16 **دقت آموزش** ٪77.19**و دقت تست** 77.57٪ **می باشد.**

**حالت سوم**

**دراین حالت با پارتیشن** 90 10 **و با** seed 9896444 **و همسایگی** 2 **تا** 20 **و بالانس** reduce

K **بهینه** 15 **دقت آموزش** 75.68% **و دقت تست** 73.57٪ **می باشد.** OVERFIT

**حالت چهارم**

**دراین حالت با پارتیشن** 90 10 **و با** seed 9896444 **و همسایگی** 2 **تا** 5 **و بالانس** boost

K **بهینه** 2 **دقت آموزش** ٪85.38**و دقت تست** 92.78٪ **می باشد.**

**گزارش** CRT

# حالت اول

**دراین حالت با پارتیشن** 80 20 **و با** seed 1509045785 **(**seed**خود درخت نه پارتیشن)**

**و بالانس** boost**. از تنظیمات** CRT **حالت اول را روی** new model Build **میزاریم . عمق درخت را** 5 **انتخاب کرده ام. گزینه هرس را تیک میزنیم.**

**دقت آموزش** 78.95**٪ و دقتت تست** 81.69**٪ می باشد.**

# حالت دوم

**دراین حالت با پارتیشن** 80 20 **و با** seed 1509045785 **و در حالت** bagging **و نیز از قسمت** Ensembles **برای** bagging **انتخاب با رای گیری از میانگین و روی عدد** 10 **تنظیم کرده ام و نیز عمق درخت را** 10 **انتخاب کرده ام. گزینه هرس را تیک میزنیم.**

**دقت آموزش** 81.55**٪ و دقت تست** 83.23**٪ می باشد.**

# حالت سوم

**دراین حالت با پارتیشن** 80 20 **و با** seed 1509045785 **و در حالت** boosting **و نیز عمق درخت را** 10 **انتخاب کرده ام. گزینه هرس را تیک میزنیم.**

**دقت آموزش** 92.38**٪ و دقت تست** 94.37**٪ می باشد.**

## گزارش C.5

**دراین حالت با پارتیشن** 80 20 **و در قسمت تنظیمات** C5 **تیک گزینه های** Croos-validate **و** boosting **را فعال کرده و هر دو را روی** 10 **میزارم** .

**دقت آموزش** 92.3٪ **و دقت تست** 95.28٪ **می باشد.**

## گزارش RandomForest

**به همان پارتیشن قبل وصل میکنیم. در قسمت** Basics **تنظیماتش تعداد درخت ها را روی** 150 **قرار میدهم سپس تیک گزینه استفاده از داده ها ی ایمبالانس رو هم فعال میکنم . حداکثر نود روی** 10000 **و عمق درخت روی** 10 **و حداقل انشعاب را روی** 7 **تنظیم میکنم**

**دقت آموزش** 90.09٪ **و دقت تست** 91.08٪ **می باشد.**

**کلا مدل رندم فارست مدل کندی می باشد و هنوز در ابزار به بلوغ کافی نرسیده است.**

## گزارش QUEST

### حالت اول

**به همان پارتیشن قبل وصل میکنیم. در قسمت** Basics **تنظیماتش عمق را روی** 4 **و گزینه هرس را فعال و مدل روی** Boosting **و عدد آن روی** 10 **می باشد**

**دقت آموزش** 76.91٪ **و دقت تست** 78.59٪ **می باشد.**

### حالت دوم

**در قسمت** Basics **تنظیماتش عمق را روی** 7 **و گزینه هرس را فعال و مدل روی** Bagging **و عدد آن روی** 10 **می باشد**.

**دقت آموزش** 74.49٪ **و دقت تست** 70.42٪ **می باشد .** OVERFIT

## گزارش CHAID

### حالت اول

**روی** NewModel **عمق** 5 **و گزینه هرس را فعال و تنظیمات پیش فرض**

**دقت آموزش** 78.04٪ **و دقت تست** 88.81٪ **می باشد.**

### حالت دوم

**روی عمق** 5 **و گزینه هرس را فعال و در حالت** Boosting **میباشد**

**دقت آموزش** 93.89٪ **و دقت تست** 91.61٪ **می باشد.** OVERFIT

**گزارش** SVM

## حالت اول

**تیک گزینه از پارتیشن بخون را فعال کرده سپس از قسمت** Exper **حالت** Simple **را فعال کرده گزینه** calculate predictor importance **را فعال کرده و سپس** run

**دقت آموزش** 73.95٪ **و دقت تست** 82.16٪ **می باشد.**

## حالت دوم

**تیک گزینه از پارتیشن بخون را فعال کرده سپس از قسمت** Expert **حالت** Expert **را فعال کرده** stoppingCriteria **را روی** 1.0E-4 **گذاشته با مقادیر پارامتری** 10 **و** 1

**نوع کرنل را روی خطی یا** Linear **می گذاریم. گزینه** calculate predictor importance **را فعال کرده و سپس** run

**دقت آموزش** 88.89٪ **و دقت تست** 90.65٪ **می باشد.**

**گزارش** Auto Classifier

**همه مدل هارو باهم داره و بررسی میکنه مثلا برای** C5 **از قسمت** Model parameters **مدل درخت را انتخاب میکنیم. در قسمت** Expert **در بحث هرس تعداد بچه هارو از** 2 **تا** 8 **میزارم. سپس در قسمت** pruning severity **از** 75 **تا** 100 **اضافه میکنم.**

**خب** 42 **مدل** C5 **برام تولید کرد. که دقت آموزش** 89 **و دقت تست** 77 **که همه** OVERFIT **هستند.**

**مرسی از زحمات شما استاد عزیز.شاد و سرافراز باشید**

**محمد سعید حیدری**