مقدمه فصل 3

كشف ارتبط بين دادهها یكی از مهمترین فعاليت ها درپژوهش های با حوزه یادگیری ماشین میباشد . كشف الگوهای مكرر ، مشاركت و همبستگی ها و مواردی از این دست، از جمله مواردی هستند که انتظار میرود در پژوهش های این حوزه به آنها پراخته شود.

داده ها اساسی ترین بخش سازنده یک پژوهش در حوزه یادگیری ماشینی و مسائل پیش بینی هستند در ادامه به معرفی دیتاست مورد استفاده در پژوهش خواهیم پرداخت

داده های مورد بررسی در این پژوهش از گردآوری اطلاعات بیش از 1000بیمار شرکت کننده در تحقیقات آینده نگر و مطالعات کوهورت بیماران کلیوی مربوط به بیمارستان شهر توکیو و 15 مرکز درمانی مرتبط با آن است است که بر اساس طبقه بندی بیماری کلیوی بین مراحل G2-G5 CKD قرار دارند که تحت دیالیز قرار نگرفته اند

که رضایتنامه بیماران برای مشارکت در طرح تحقیقاتی در ابتدا اخذ شده است

لازم به ذکر است مطالعه ای که اخیرا بر روی بیماران مورد بررسی قرار انجام شده است در دسته ی مطالعات آماری به حساب می­آید که موید مناسب بودن دیتاست برای اجرای الگوریتم های یادگیری ماشین است که در فصل دوم به آن پرداخنه شده است

شرکت کنندگان دز این مطالعه شرایط زیر را دارا بودند :

* افرادبا سن بالای 20 سال
* قرار داشتن که در یک مرحله از بیماری مزمن کلیوی
* نبودن تحت درمان و یا جراحی کلیه در 2سال گذشته

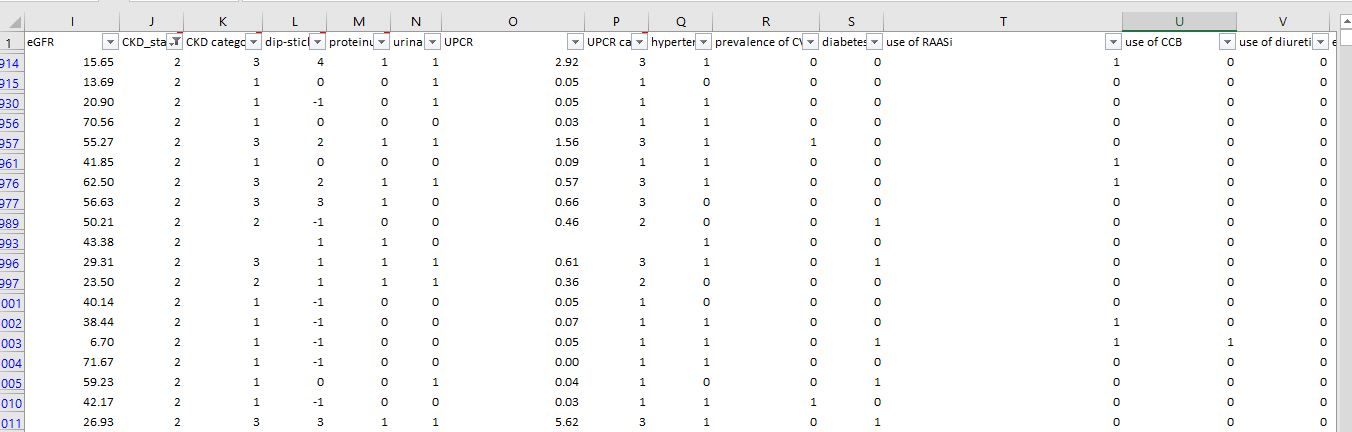
مشخصات دیتاست :

جدول (1-3)متغییرهای دیتاست بیماران مزمن کلیوی

|  |  |  |  |
| --- | --- | --- | --- |
| نام متغیر | مقادیر مشخصه | نوع مشخصه | توضیح فارسی |
| Age | >20 | numeric | سن |
| Gender | 1و0 | Binary | جنسیت |
| Systolic blood pressure | 139±22 | Numeric | فشارخون سیستولی |
| BMI | 23.8±3.9 | Numeric | شاخص توده بدنی |
| etiology of CKD | Nephrosclerosis  Glomerulonephritis  Diabetic  Others | Category | علت بیماری |
| Hb | 12.1±2.2 | Numeric | هموگلوبین |
| Alb | 3.9±0.6 | Numeric | آلبومین |
| Cr | 4.8±5.6 | Numeric | کراتینین |
| eGFR | 33.8±3.9 | Numeric | فیلتراسیون گلومری |
| CKD\_stage | 2/3/4/5 | Category | مرحله ckd |
| CKD category | /2/3/1  به زای هرمرحله ه | Category | دسته بندی  Ckd |
| dip-stick proteinuria | -1/0/1/2/3 | category | طیف پروتئین در ادرار |
| Proteinuria | 0/1 | binary | پروتئین ئوری |
| urinary occult blood | 0/1 | binary | هماچوری |
| UPCR | 2.7±1.72 | numeric | نسبت پروتئین به کراتنین ادراری |
| UPCR category | 1/2/3 | category | طبقه بندی کراتنین |
| Hypertension | 1/0 | binary | فشار خون |
| prevalence of CVD | 1/0 | binary | سابقه ی بیماری قلبی |
| Diabetes | 1/0 | binary | دیابت |
| use of RAASi | 1/0 | binary | داروی مهار کننده سیستم رنین آنزیوتانسین آلدسترون |
| use of CCB | 1/0 | Binary | بلاک کننده  کانال کلسیم |
| use of diuretics | 1/0 | binary | داروی ادرار آور |
| observational duration | 14±19 | Numeric | مدت زمان مشاهده |

|  |  |
| --- | --- |
| درصد همانندی دو ستون پیشرفت بیماری و بیماری قلبی | 0.049209139 |
| درصد همانندی مرگ و پیشرفت بیماری | 0.006151142 |
| درصد همانندی مرگ و بیماری قلبی | 0.049209139 |
| درصد بهبود در بیماران | 0.700351494 |
| درصد مرگ در بیماران | 0.068541301 |
| درصد پیشرفت بیماری | 0.246045694 |
| درصد بروز عوارض قلبی | 0.103690685 |

**شمای کلی دیتاست بیمارانckd**



: پيش پردازش

یكی از مهمترین مراحل کاربرد الگوریتم های یادگیری ماشین پیش پردازش است و تقریبا 80 ٪زمان صرف این مرحله میشود. اگر داده ی نامناسب و مغشوش به عنوان ورودی مدل استفاده شود نتایج قابل قبولی حاصل نخواهد شد

مرحله های پيش پردازش داده عبارتند از:

: • پاكسازی داده

• یكپارچه سازی داده

شكل داده به دادهی قابل تحليل توسط الگوریتمهای یادگيری ماشين •

كاهش دادن حجم داده به دادهی پاكسازی شده و قابل تحليل

مراحل پیش پردازش داده های بیماران ckd:

برای دسته بندی صحیح بیماران ابتدا بیمارانی که کم تر از 6 ماه تحت نظر بودند از دیتاست مورد بررسی جدا شدند که از جمع 1126رکورد ثبت شده برای بیماران 923رکورد برای بیمارانی که مدت زمان تحت نظر بودن آنها بالای 6 ماه بود برای مراحل بعدی از دیتای اصلی جدا شد

ساخت ستون مربوط به کلاس های اصلی :

ستون نهایی که به عنوان برچسب نهایی در نظر گرفته میشود از ترکیب 3 ستون مجزا در دیتاست که

* کاهش عملکرد کلیه

ثبات عملکرد کلیه

* مرگ
* بروز بیماری قلبی
* بروز بیماری قلبی با کاهش عملکرد کلیه

در ستون برچسب نهایی ثبت شد

شناسایی داده های پرت برای مشخصه های عددی با تابع نرمال و خارج کردن داده هایی که در مجدوده 3 سیگما قرار نمیگرفتند انجام گرفت

* داده های طبقه ای با مقادبر (dummy) جایگزین شدند ستون هایی که ماهیت طبقه ای چندگانه دارند تبدیل به ستون های دودویی می شوند
* مقادیر گم شده در ستون های efgr با درون یابی خطی پر شدند
* سایر مشخصه ها با میانگین برای مشخصه های عددی و برای مشخصه های طبقه ای با مد پر شدند
* داده های عددی نرمال سازی شدند

مفاهیم پایه :

تعریف مدل های دسته بند :

در مسائل دسته بندی هدف شناسایی مشخصه هایی است كه گروهی را كه هرمورد به آن تعلق دارد را نشان دهند. از این الگو میتوان هم برای فهم دادههای موجود و هم پيش بينی نحوه رفتار موارد جدید استفاده كرد. دادهكاوی مدلهای دسته بند را با بررسی دادههای دستهبندی شده قبلی ایجاد میكند و یك الگوی پيش بينی كننده را بصورت استقرایی مییابد(hand et al,2001)

در مسائل دستهبندی هدف شناسایی مشخصه هایی است كه گروهی را كه هرمورد به آن تعلق دارد را نشان دهند.

در الگوریتمهای دسته بندی مجموعه داده اوليه به دو مجموعه داده با عنوان مجموعه دادههای آموزشی1 و مجموعه

دادههای آزمایشی2 تقسيم میشود.

میدانيم هر رکورد شامل مجموعه ای از مشخصه هاست ، كه یكی از این

ها مشخصه دسته ناميده میشود. درمرحله آموزش؛ مجموعه دادههای آموزشی به یكی از الگوریتمهای

دسته بندی داده میشود تا بر اساس سایر مشخصه ها برای مقادیر مشخصه دسته، مدل ساخته شود .پس از

ساخت مدل، در مرحله ارزیابی؛ دقت مدل ساخته شده به كمك مجموعه دادههای آزمایشی ارزیابی خواهد شد.

در الگوریتمهای دسته بندی از آنجا كه مشخصه دسته مربوط به هر Case مشخص است لذا به صورت

الگوریتمهای با ناظر محسوب میشوند. بدیهی است كه تشخيص بر اساس دسته هایی است كه مدل در مرحله

آموزش با آنها روبرو شده است؛ بنابراین امكان تشخيص دسته جدید در كاربرد دسته بندی وجود نخواهد داشت.

گزارش اجرای کد های مرتبط در پایتون:

الگوریتم های دسته بند مرجع:

درخت تصمیم

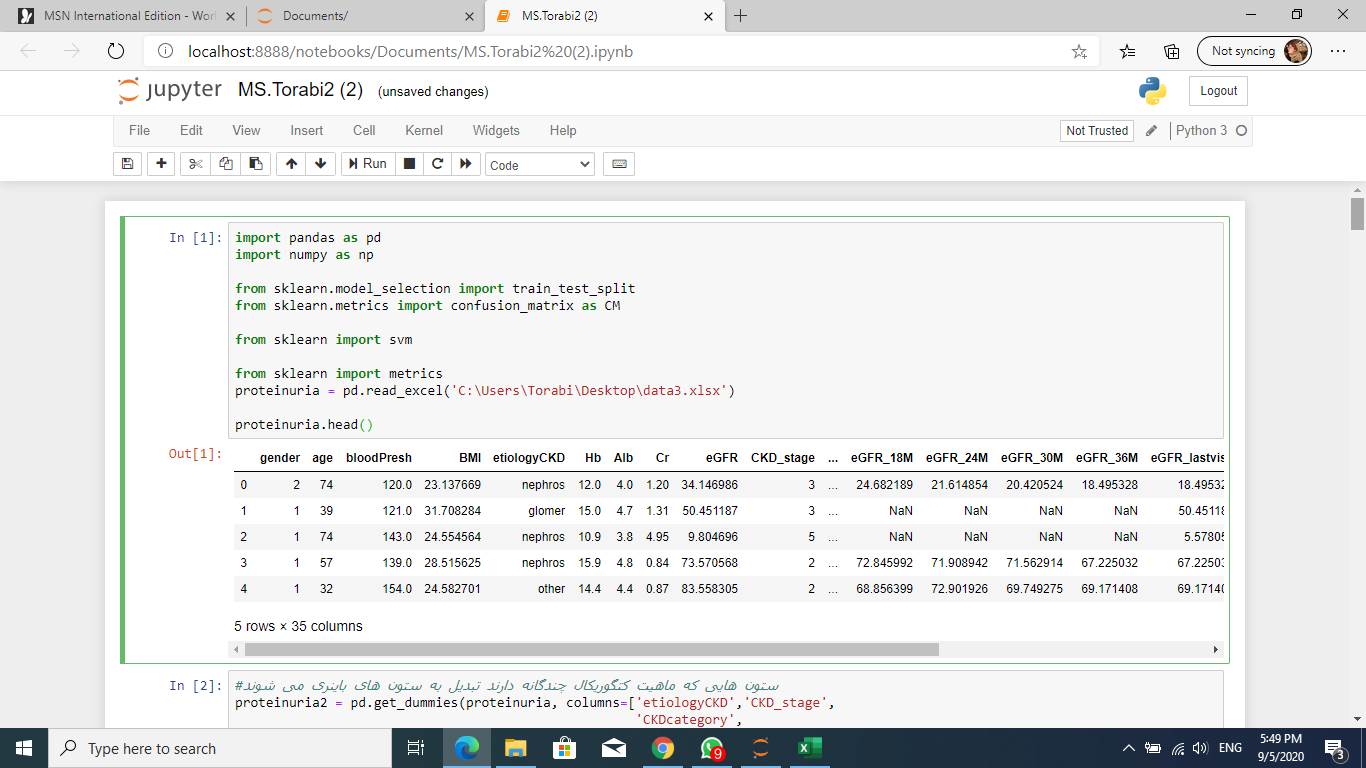
بردار پشتیبان

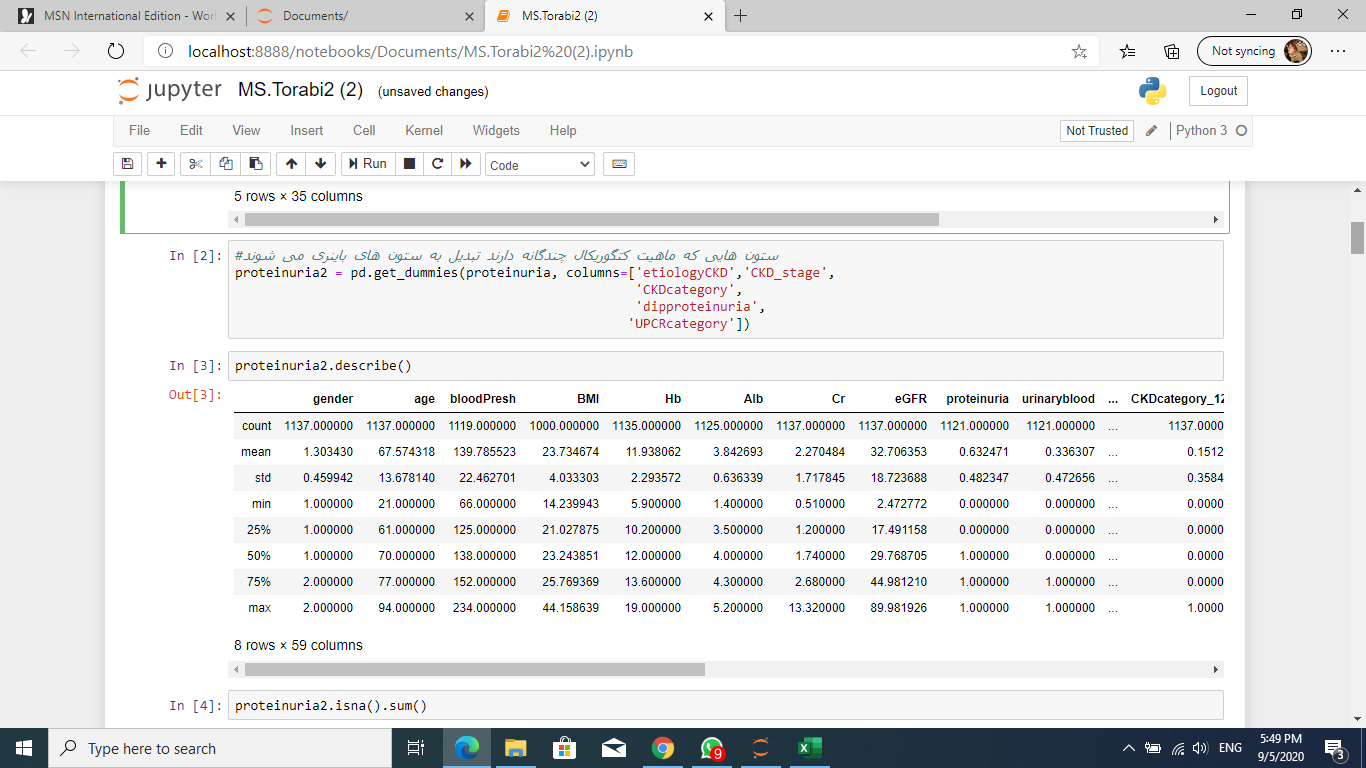
نایوبیز

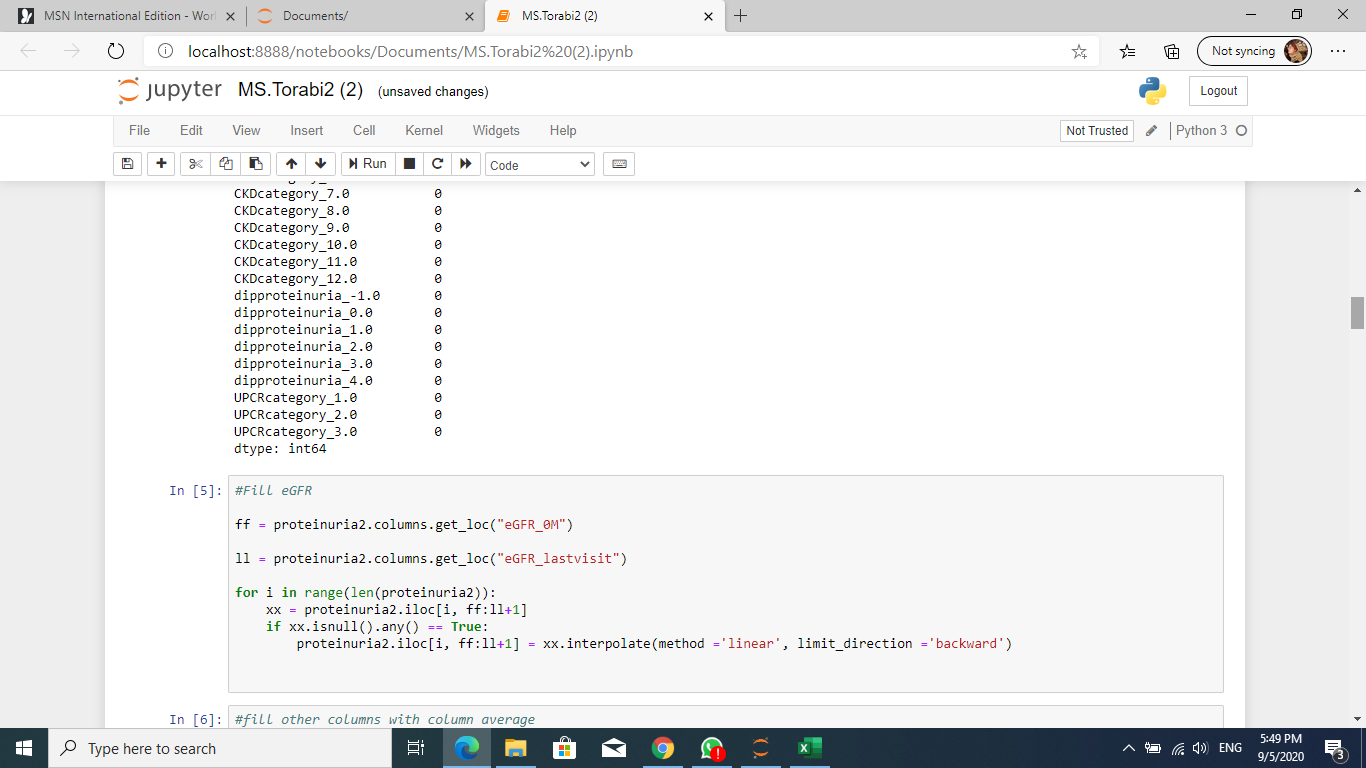
شبکه عصبی (ann)

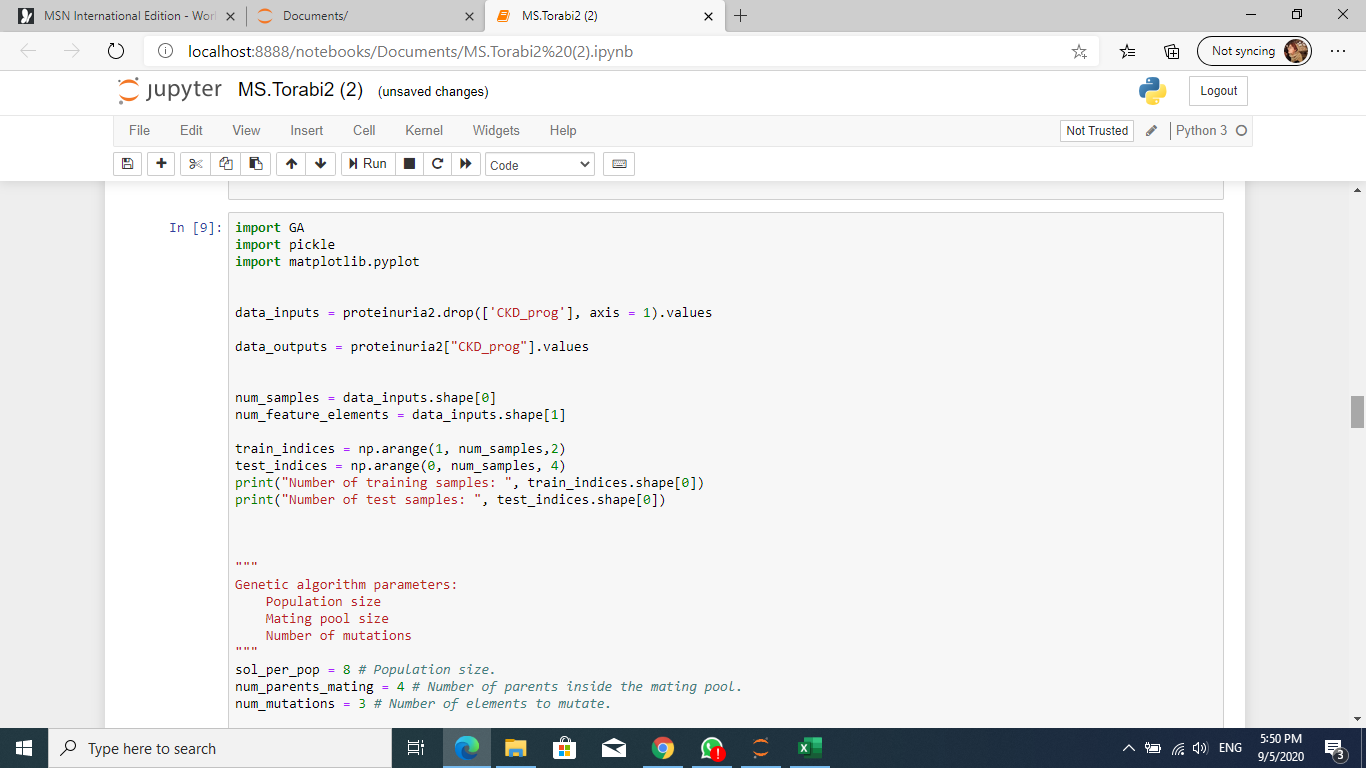
مراحل تحلیل داده ها در نرم افزار پایتون

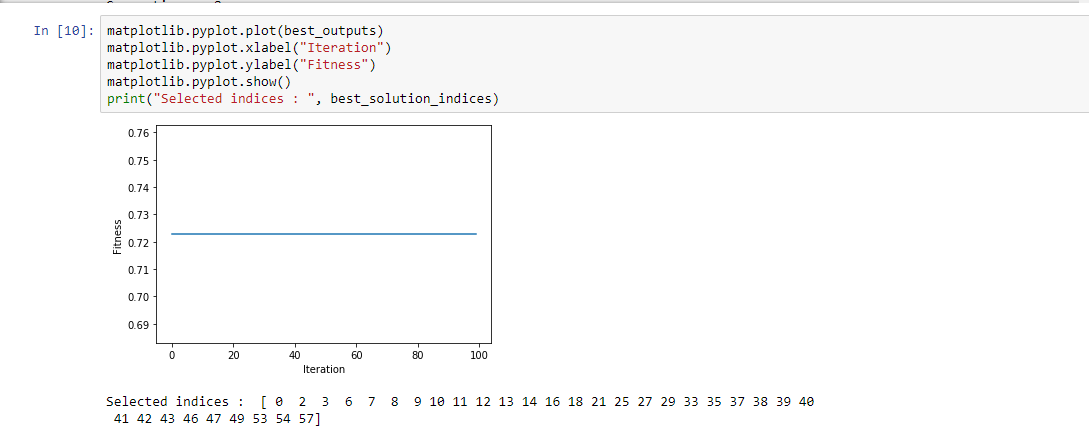
فاز 1)پیش پردازش





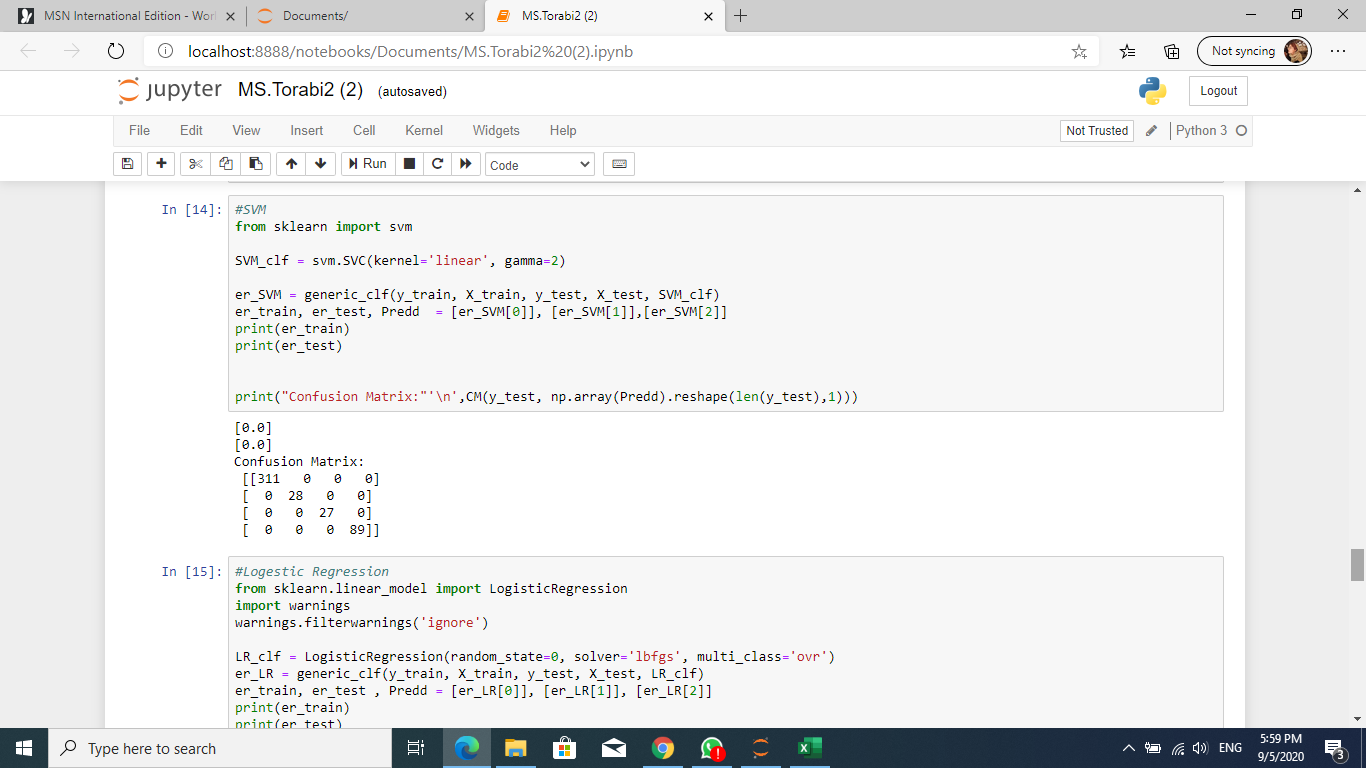


 انتخاب ویژگی بر اساس الگوریتم ژنتیک با cross-over

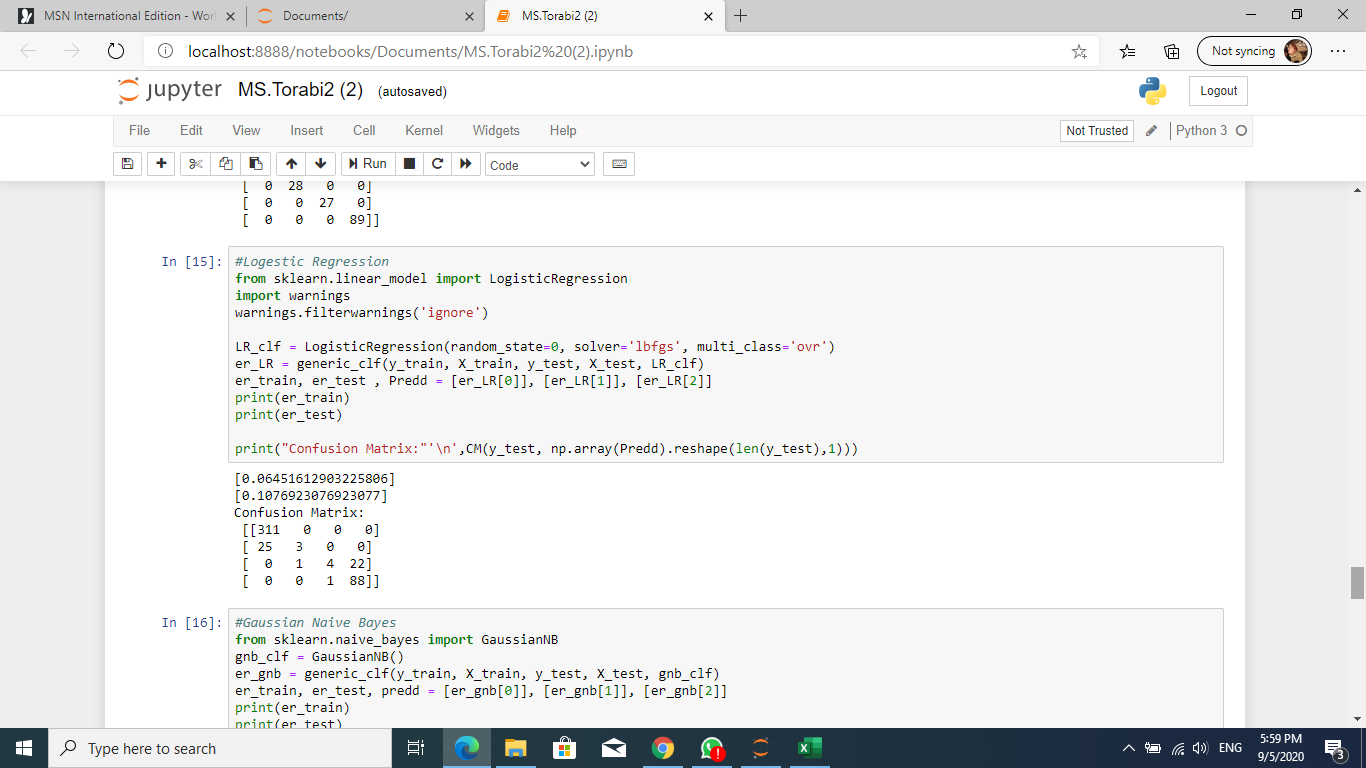


استخراج مشخصه ها با انطباق حدود 70 %

اجرای الگوریتم های پایه :



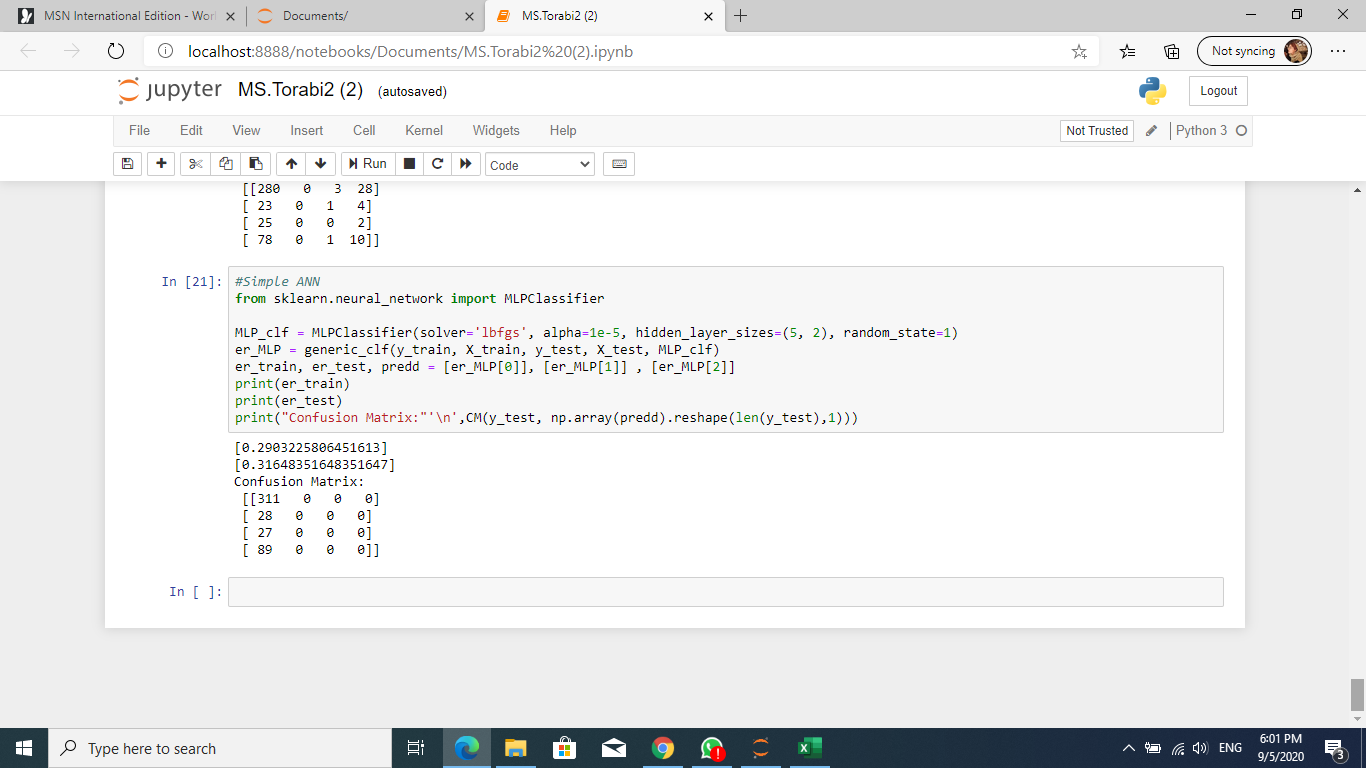
دقت 100درصدی مدل در اولین اجرا برای دسته بند svm



رگرسیون لجستیک:

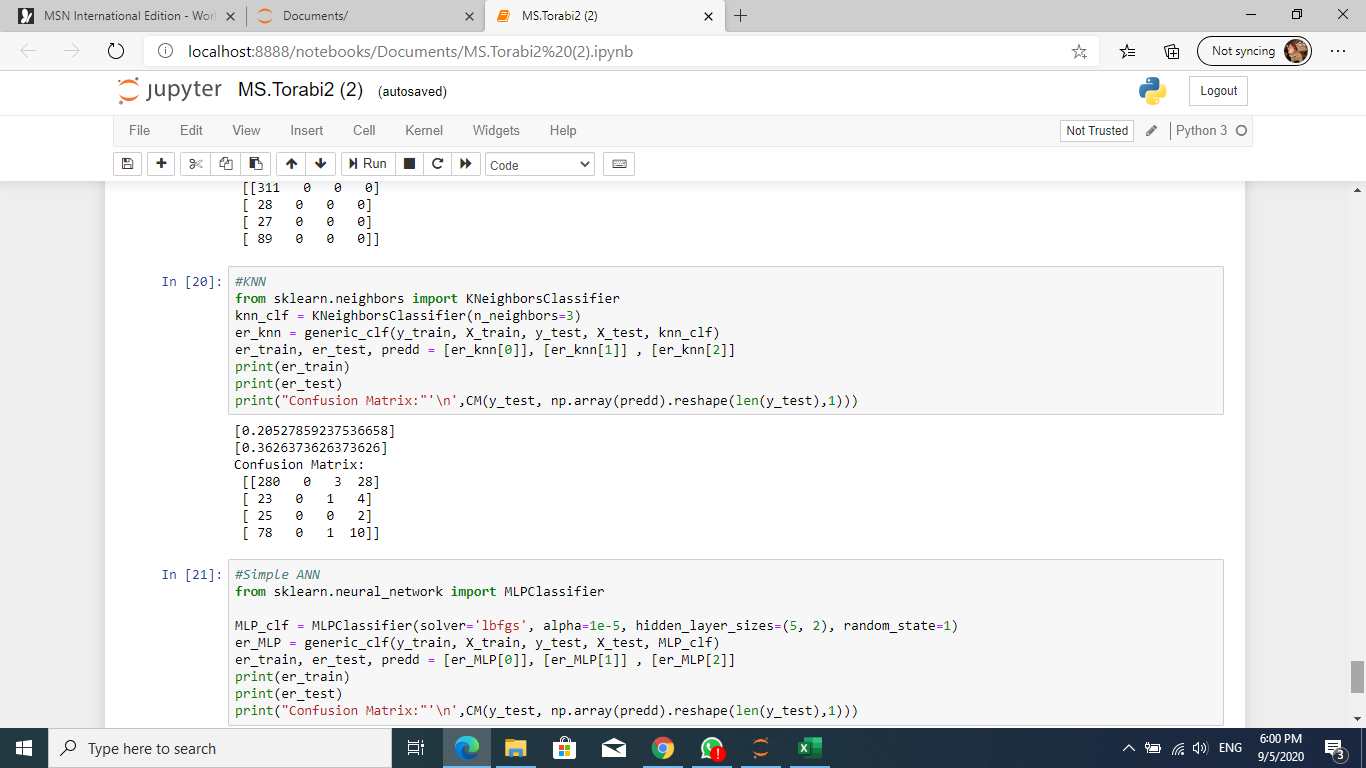
خطای دسته بند برای داده های آموزشی :0.06

خطای دسته بند برای داده های آزمایشی :0.1

شبکه عصبی پرسپترون چند لایه :

خطای دسته بند برای داده های آموزشی :0.29

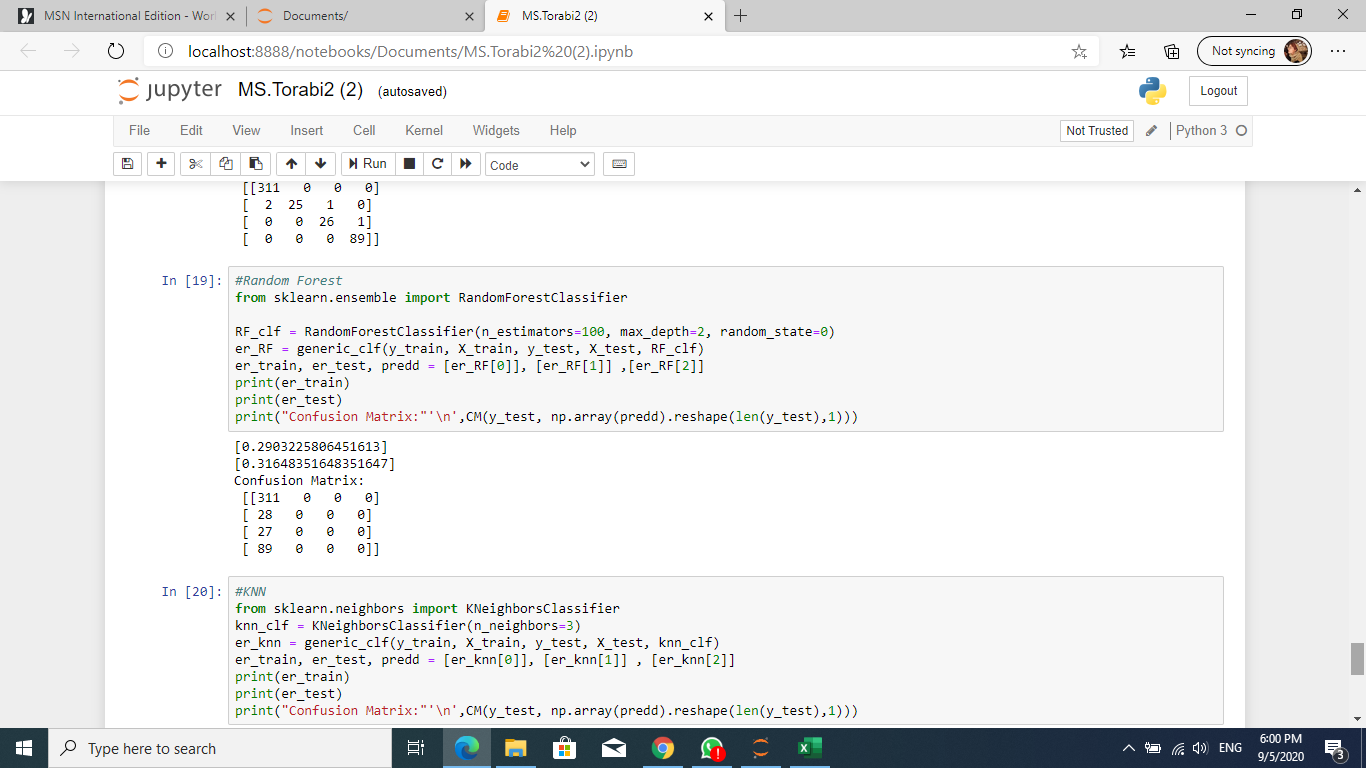
خطای دسته بند برای داده های آزمایشی :0.31



kنزدیک ترین همسایه :

خطای دسته بند برای داده های آموزشی :0.20

خطای دسته بند برای داده های آزمایشی :0.36



مدل دسته بند : جنگل تصادفی

خطای دسته بند برای داده های آموزشی :0.29

خطای دسته بند برای داده های آزمایشی :0.31

در این بین بردار ماشین پشتیبان بهترین عملکرد را بعد از اتتخاب ویژگی داراست