سوالات اصلی :

1. تعداد آنهایی که فشارخون دارند چقدر است؟
2. آیا فشارخون با مرگ رابطه مستقیم دارد؟
3. آیا دیابت و فشار خون بر مرگ افراد تاثیر مستقیم دارند؟

در این جا هیچ داده ای نال وجود نداشت. با نقسیم بندی دو قسمت intو float به نتایج جالبی رسیدم به طوریکه در شکل می بینی گروه بندی پیوسته به شکل زیر است.

|  | **age** | **platelets** | **serum\_creatinine** | **serum\_sodium** |
| --- | --- | --- | --- | --- |
| count | 299.000000 | 299.000000 | 299.00000 | 299.000000 |
| mean | 60.833893 | 263358.029264 | 1.39388 | 136.625418 |
| std | 11.894809 | 97804.236869 | 1.03451 | 4.412477 |
| min | 40.000000 | 25100.000000 | 0.50000 | 113.000000 |
| 25% | 51.000000 | 212500.000000 | 0.90000 | 134.000000 |
| 50% | 60.000000 | 262000.000000 | 1.10000 | 137.000000 |
| 75% | 70.000000 | 303500.000000 | 1.40000 | 140.000000 |
| max | 95.000000 | 850000.000000 | 9.40000 | 148.000000 |

نکته جالب در مورد پراکندگی انهاست به خوبی در داده دیده می شود.

در مرحله بعد نوبت به تقسیم بندی جنسیت براساس آقا و خانم و مرده و زنده در قسمت فوت شده است.

حالا براساس تعداد فشار خون های دسته بندی می کنیم که به این نتایج می رسیم.

|  |  | **age** | **anaemia** | **creatinine\_phosphokinase** | **diabetes** | **ejection\_fraction** | **platelets** | **serum\_creatinine** | **serum\_sodium** | **sex** | **smoking** | **time** | **DEATH\_EVENT** | **Marg** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **count** | **count** | **count** | **count** | **count** | **count** | **count** | **count** | **count** | **count** | **count** | **count** | **count** |
| **high\_blood\_pressure** | **sex1** |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 0 | Female | 61 | 61 | 61 | 61 | 61 | 61 | 61 | 61 | 61 | 61 | 61 | 61 | 61 |
| Male | 133 | 133 | 133 | 133 | 133 | 133 | 133 | 133 | 133 | 133 | 133 | 133 | 133 |
| 1 | Female | 44 | 44 | 44 | 44 | 44 | 44 | 44 | 44 | 44 | 44 | 44 | 44 | 44 |
| Male | 61 | 61 | 61 | 61 | 61 | 61 | 61 | 61 | 61 | 61 | 61 | 61 | 61 |

در پیوت تیبل نتایج عجیبی بدست آمده که نشان از دخالت سیگار کشیدن و دیابت است

|  |  | **age** | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | **diabetes** | **0** | | **1** | | **All** |
|  | **smoking** | **0** | **1** | **0** | **1** |  |
| **high\_blood\_pressure** | **sex1** |  |  |  |  |  |
| 0 | Female | 61.89 | nan | 56.41 | 60.00 | 58.98 |
| Male | 63.54 | 59.13 | 58.31 | 61.56 | 60.49 |
| 1 | Female | 60.15 | 71.00 | 61.13 | 50.00 | 60.88 |
| Male | 63.57 | 64.24 | 61.18 | 64.00 | 63.39 |
| All |  | 62.49 | 60.80 | 58.62 | 61.93 | 60.83 |

با گرفتن رابطه کوواریانس دادهها هیچکدام نمیتوان حذف کرد چون تاثیر زیر 0.2 درصد دارند.

خارج کردن outliear :

در این قسمت با استفاده از اماری z\_score دادهای پرت را خارج میکنم که به شکل زیر در میاید:

راه حل:

from scipy import stats

z\_score=np.abs(stats.zscore(df.iloc[:,:-2]))

df\_clean.shape= (15و280)

df.shape= (15و299)

بعد از اینکه دادههای پرت خارج شدن بازهم انچه که من با پیوت بدست آوردم درست بود.

حالا گروپ بای با مردن درست میکنم که میانه روز و میانگین روز به قرار زیر است.

ابتداد گروپ پیوسته دزست میکنیم با زمان:

["creatinine\_phosphokinase", "ejection\_fraction", "platelets", "serum\_creatinine", "serum\_sodium", "time"]

سپس این نتایج بدست می آید.

|  | **creatinine\_phosphokinase** | | **ejection\_fraction** | | **platelets** | | **serum\_creatinine** | | **serum\_sodium** | | **time** | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **mean** | **median** | **mean** | **median** | **mean** | **median** | **mean** | **median** | **mean** | **median** | **mean** | **median** |
| **Marg** |  |  |  |  |  |  |  |  |  |  |  |  |
| No | 467.18 | 231.00 | 40.15 | 38 | 261,565.74 | 262,500.00 | 1.14 | 1.00 | 137.36 | 137.00 | 159.14 | 174.00 |
| yes | 433.14 | 249.50 | 32.62 | 30 | 252,626.33 | 254,500.00 | 1.58 | 1.30 | 135.63 | 135.50 | 73.03 | 47.50 |

اگر بخواهیم تاثیر فشار خون بر روی مرگ و میر ببینم نشان میدهد 36 درصد خانم و 34 درصد آقایان که فشار خون داشتند به مرگ منتهی شده است.

|  | **Marg** | **No** | **yes** |
| --- | --- | --- | --- |
| **sex1** | **high\_blood\_pressure** |  |  |
| Female | 0 | 0.75 | 0.25 |
| 1 | 0.64 | 0.36 |
| Male | 0 | 0.70 | 0.30 |
| 1 | 0.66 | 0.34 |

استفاده از الگوریتم ماشین:

بااستفاده از درخت دیتا را به دوقسمت train وtest تبدیل میکنیم. سپس با درخت به بررسی آن می پردازیم که با توجه به auc در داده overfit داریم اما با محدود کردن عمق درخت به 3 میزان فاصله این دوتا را به 17 درصد میرسومنیم که به نسبت خوب است.

مهمترین فیچر در این دیتا برای تقسیم بندی درخت براساس ضریب جینی

array([0. , 0. , 0.02954982, 0. , 0. ,

0. , 0. , 0.18247646, 0. , 0. ,

0. , 0.78797372])

که فیچر زمان هست.

نتیجه گیری:

هر کدام به میزانی بر روی مرگ و میر اثر داشتند اما برای پیش بینی از تعداد فوت شدگان زمان را می توان مهمترین عامل برای زنده نگه داشتن براساس الگوریتم های ماشین دانست. اما اثر گذاری هر کدام فیچرهای انقدر بالا نبود که با استفاده از همبستگی بتوان یکی از آن ها را حذف کرد