باسمه تعالی

تکلیف سری دوم داده کاوی

سارا برادران (شماره دانشجویی : ۹۶۲۴۱۹۳)

سوال ۲)

در داخل کد نوشته شده مقادیر missing data هر ستون به صورت جداگانه مشخص شده است به علاوه در حالت کلی می توان گفت این مقادیر مفقود به صورت های NaN, 'MISS', '', '?' در داخل دیتاست قرار دارند.. به علاوه مقادیر ۰ در هیچ ستونی به عنوان داده مفقود در نظر گرفته نشده است.

در داخل کد نوشته شده ابتدا تمام مقادیر غیر عددی از ستون ها جداسازی شده و سپس در داخل یک مجموعه غیر تکراری از اعضا اضافه می شوند. و در نهایت این مجموعه برای هر ستون چاپ شده و تمام مقادیر درون آن در ستون مربوطه با مقدار NaN جاگذاری می شود.

تعداد اعضای مفقود هر ستون نیز به وسیله متد df.isna().sum() محاسبه می شود.

سوال ۳)

قسمت a) تمام مقادیر مفقود در داخل دیتاست با مقدار ثابت ۹۹۹ جایگذاری شده است.

قسمت b) تمام مقادیر مفقود در داخل دیتاست با مقدار نظیر همان ستون در رکورد اول دیتاست جایگذاری شده است. برای مثال اگر یک رکورد مقدار فیلد Glucose برای آن مفقود باشد به جای آن مقدار Glucose نظیر رکورد شماره ۰ قرار خواهد گرفت.

قسمت c) با توجه به اینکه مقدار میانگین داده های عددی یک عدد float می باشد لذا می بایست تایپ تمام ستون های دیتاست را به float تغییر داده و پس از محاسبه میانگین هر ستون مقادیر مفقود ستون ها را برابر با میانگین محاسبه شده لحاظ کنیم.

قسمت d) مد داده های یک ستون را می توان بدون نیاز به هر گونه تغییر تایپ داده ها محاسبه کرده و برای مقادیر مفقود نظیر همان ستون لحاظ کرد. در این قسمت ممکن است مد یک ستون بیش از یک عدد باشد در این صورت تمام مقادیر مفقود نظیر آن ستون با یکی از مد ها جایگذاری می شود.

قسمت e) برای رسم نمودار هیستوگرام نظیر ستون Glucose می بایست ابتدا تایپ این ستون را به int تغییر دهیم چرا که برای داده های غیر صحیح نمودار هیستوگرام قابل رسم نمی باشد.

توضیحات و تفسیر نمودار : از میان نمودار های کشیده شده در این مثال نمودار های نطیر جایگذاری با مد ، میانگین و یکی از اعضای دیتا ست مشابه هم رسم شده و نمودار نطیر جایگذاری با عدد ثابت ۹۹۹ متفاوت شده است. در حالت کلی جایگذاری با مقدار مد یا میانگین نسبت به دو روش دیگر گزینه های مناسب تر با عملکرد عموما بهتری هستند در این مثال به نظر میرسد رکورد شماره ۰ حاوی مقداری پرت برای فیلد Glucose نبوده است و مقدار آن چیزی نزدیک به مد و میانگین بوده است در غیر اینصورت نمودار رسم شده برای آن می توانست تفاوت بیشتری با دو نمودار دیگر داشته باشد.

همانطور که مشخص شده است مقدار میانگین برای فیلد Glucose حدودا برابر ۱۲۳ و مقدار مد برای آن مقدار ۱۰۰ است و مقدار فیلد Glucose برای رکورد شماره ۰ نیز ۱۴۸ است و همه این مقادیر در بازه ۱۰۰ تا ۲۰۰ قرار دارد و نزدیک به هم هستند حال اگر میزان Glucose برای رکورد اول ۰ یا داده بزرگی نظیر ۵۰۰ میبود آنگاه نمودار تفاوت چشمگیری با دو نمودار جایگذاری با مد و میانگین میداشت.

قسمت f)

Different Ways to Compensate for Missing Values In a Dataset

1- Do Nothing

می توان برای پر کردن missing data هیچ کاری نکرده و تمام کار را به الگوریتم واگذاری کنیم تا به بهترین نحو عمل imputation را انجام دهد تا کمترین میزان خطا یا از دست دادن دیتا را داشته باشیم. برخی از الگوریتم ها به کمک روش های یادگیری می توانند مقادیر مفقود را هندل کنند اما تمام الگوریتم های داده کاوی این کار را انجام نمی دهند و لذا ممکن است الگوریتم با خطا برخورد کند.

2- Imputation Using (Mean/Median) Values

پر کردن missing data ها با محاسبه و جایگذاری مقدار میانه یا میانگین داده های سالم. این روش فقط برای فیلد های عددی کاربرد دارد. این روش در عین سادگی معمولا برای دیتاست های عددی کوچک عملکرد مناسب و قابل قبولی دارد.

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

mean()

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

3- Imputation Using (Most Frequent)

پر کردن missing data ها با محاسبه عضو most frequent (عضو با بیشترین تعداد تکرار در داده ها) این روش هم برای فیلد های عددی و هم رشته ای کاربرد دارد.

4- Imputation Using (Zero/Constant) Values

پر کردن missing data ها با مقدار صفر یا مقدار ثابت

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

df.fillna(0)

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	0.0
1	9	0.0	9.0	0	7.0
2	19	17.0	0.0	9	0.0

5- Imputation Using k-NN

الگوریتم k nearest neighbours الگوریتم ساده ای است که به نوعی عملیات classification را به نحوی ساده پیاده سازی می کند. این الگوریتم بر مبنای feature similarity مقادیر missing data ها را پیش بینی می کند در حقیقت با توجه به اینکه یک رکورد حاوی missing data به کدام دسته از داده های یادگیری و ترینینگ نزدیک تر است مقدار مناسب برای فیلد missing data لحاظ می شود. این محاسبه بر اساس k رکورد همسایه و نزدیک به رکورد حاوی missing data صورت می پذیرد. این روش بسیار دقیق تر از روش های جایگذاری با mean, mode, most frequent است اما در عین حال به داده های پرت حساس است.

6- Imputation Using Multivariate Imputation by Chained Equation (MICE)

این روش missing data ها را به جای یکبار چندین بار جایگذاری می کند به دلیل عدم قطعیت این روش مناسب بوده و عملکرد خوبی خواهد داشت.

0	2	5.0	3.0	6.0	NaN
1	9	NaN	9.0	0.0	7.0
2	19	17.0	NaN	9.0	NaN
3	7	10.0	3.0	6.0	4.0
4	2	8.0	10.0	NaN	3.0

mice()

0	2.0	5.0	3.00	6.00	4.666667
1	9.0	10.0	9.00	0.00	7.000000
2	19.0	17.0	6.25	9.00	4.666667
3	7.0	10.0	3.00	6.00	4.000000
4	2.0	8.0	10.00	5.25	3.000000

مراجع و منابع :

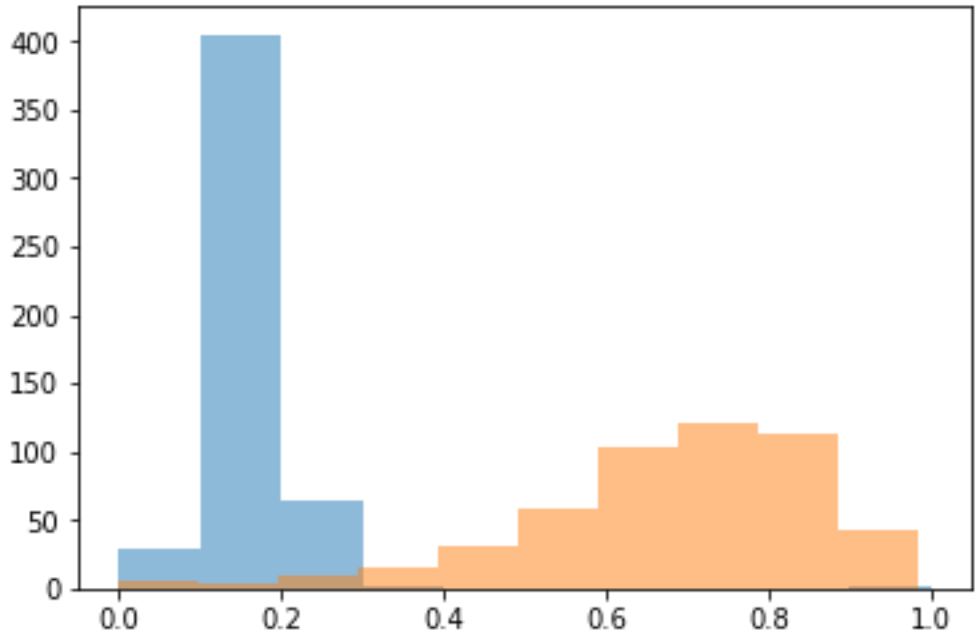
https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779

سوال ۴)

قسمت a) در صورتی که داده های داخل دیتاست نرمال سازی و استاندارد سازی نشوند اولا به دلیل range متفاوت داده ها می تواند الگوریتم های داده کاوی را دچار خطا کند در داخل همین دیتاست برای مثال فیلد **Glucose** حاوی مقادیری در بازه بزرگ ۰ تا حدود ۸۰۰ است در حالی که فیلد **DiabetesPedigreeFunction** حاوی مقادیری در بازه کوچک ۰ تا ۲ است. ممکن است الگوریتم داده کاوی ضمن مقایسه میان دو فیلد تاثیر فیلد با رنج کوچک تر را نادیده بگیرد و لذا صحت عملکرد آن به کلی زیر سوال می رود.

قسمت e) متد StandardScaler توزیع داده ها را هیچ تغییری نمی دهد و صرفا رنج داده ها را تغییر داده و به بازه کوچک حدود منفی ۲/۵ تا مثبت ۲/۵ تبدیل می کند. این درحالی است که متد Normalize توزیع داده ها را نیز تغییر داده و یک توزیع نرمال ایجاد می کند علاوه بر آنکه رنج مقادیر را نیز به بازه ۰ تا۱ تبدیل می کند.

قسمت d) توضیحات و تفسیر نمودار : همانطور که مشخص است نمودار استاندارد سازی نظیر متد MinMaxScaler توزیع داده ها را هیچ تغییری نداده و صرفا رنج داده ها را به بازه ۰ تا ۱ تبدیل کرده است در حالی که متد Normalize علاوه بر تبدیل رنج داده ها به بازه ۰ تا ۱ توزیع داده هارا نیز نرمال کرده است همانطور که مشخص است نمودار پس از نرمال سازی داری کجی چپ است.



سوال ۵)

در داخل کد داده های پرت تشخیص داده شده به وسیله روش IQR در هر ستون معین شده است و نمودار boxplot نیز رسم گشته است.

تعداد سطر های حذف شده در روش IQR تعداد ۱۰۸ سطر بوده است در حالی که در روش zscore تعداد ۵۵ سطر حاوی داده های پرت تشخیص داده شده است. برای بدست آوردن تعداد داده های پرت در روش IQR ابتدا رکورد های حاوی داده پرت برای تمام فیلد ها را ادغام کرده و سپس رکورد های تکراری از آن ها را حذف می کنیم چرا که ممکن است یک رکورد حاوی ۲ یا تعداد بیشتری فیلد باشد که مقدار پرت دارد. برای بدست آوردن تعداد رکورد های حذف شده در روش zscore نیز تعداد رکورد های اولیه را منهای تعداد رکورد های پس از حذف داده های پرت کرده و لذا تعداد رکورد های حذف شده محاسبه می گردد.

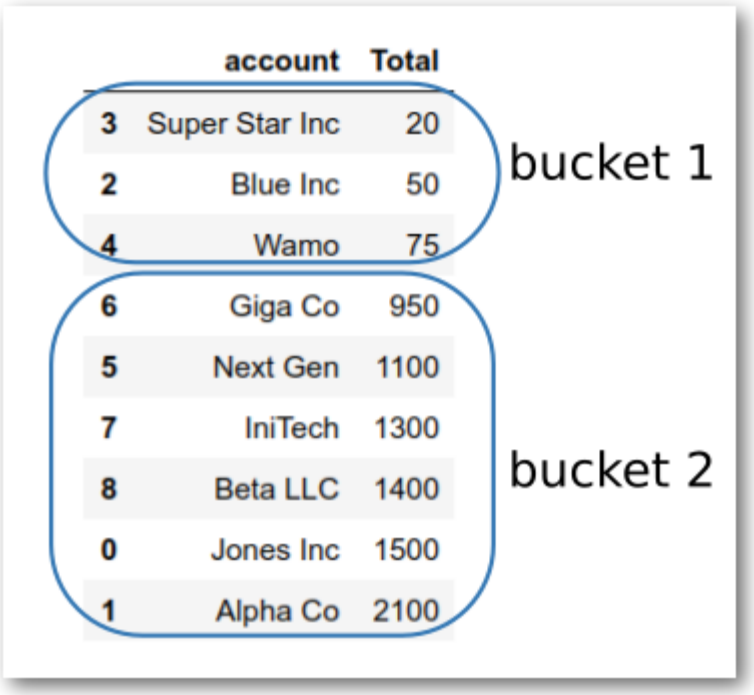
سوال ۶)

قسمت c)

jenks

به وسیله این روش در حقیقت رکورد های نظیر یک فیلد خاص به گونه ای دسته بندی می شوند که به طور طبیعی و خودکار داده هایی با بیشترین شباهت در یک دسته قرار گرفته و میان اعضای دسته های گوناگون بیشترین میزان تفاوت وجود داشته باشد. در حقیقت روش jenks برمبنای clustering عمل می کند.

یک نمونه ار عملکرد این روش در زیر آورده شده است.



همانطور که در تصویر مشخص است بدون نیاز به دانستن جزئیات الگوریتم می توان مشاهده کرد که اعداد ۲۰و ۵۰ و ۷۵ نزدیک به هم بوده و یک فاصله زیاد میان عدد ۷۵ تا عدد بعدی یعنی ۹۵۰ وجود دارد به طور واضح این گپ زیاد یک natural break ایجاد کرده و دسته اعداد بعد از ۹۵۰ را از دسته اعداد ۲۰و ۵۰ و ۷۵ جدا می کند. این عمل همان کاری است که الگوریتم jenks انجام می دهد و توسط یک رویکرد iterative محل natural break ها و جایی که دسته ها از هم مجزا می شوند تعیین می گردد این روش به گونه ای عمل می کند که علاوه بر شباهت اعضای یک دسته میان اعضای دسته های گوناگون حداکثر میزان تفاوت وجود داشته باشد به گونه می توان گفت واریانس میان گروه ها بیشینه می گردد.