# مقدمه

هدف این گزارش بررسی روش‌های پیش‌پردازش، مدل‌سازی، مقایسه و نتیجه‌گیری آن بر روی یک مجموعه داده ثابت است. در این پروژه از مجموعه‌داده Alzheimer جهت توصیف، پیش‌پردازش و مدل‌سازی استفاده شده است.

## توصیف آماری داده

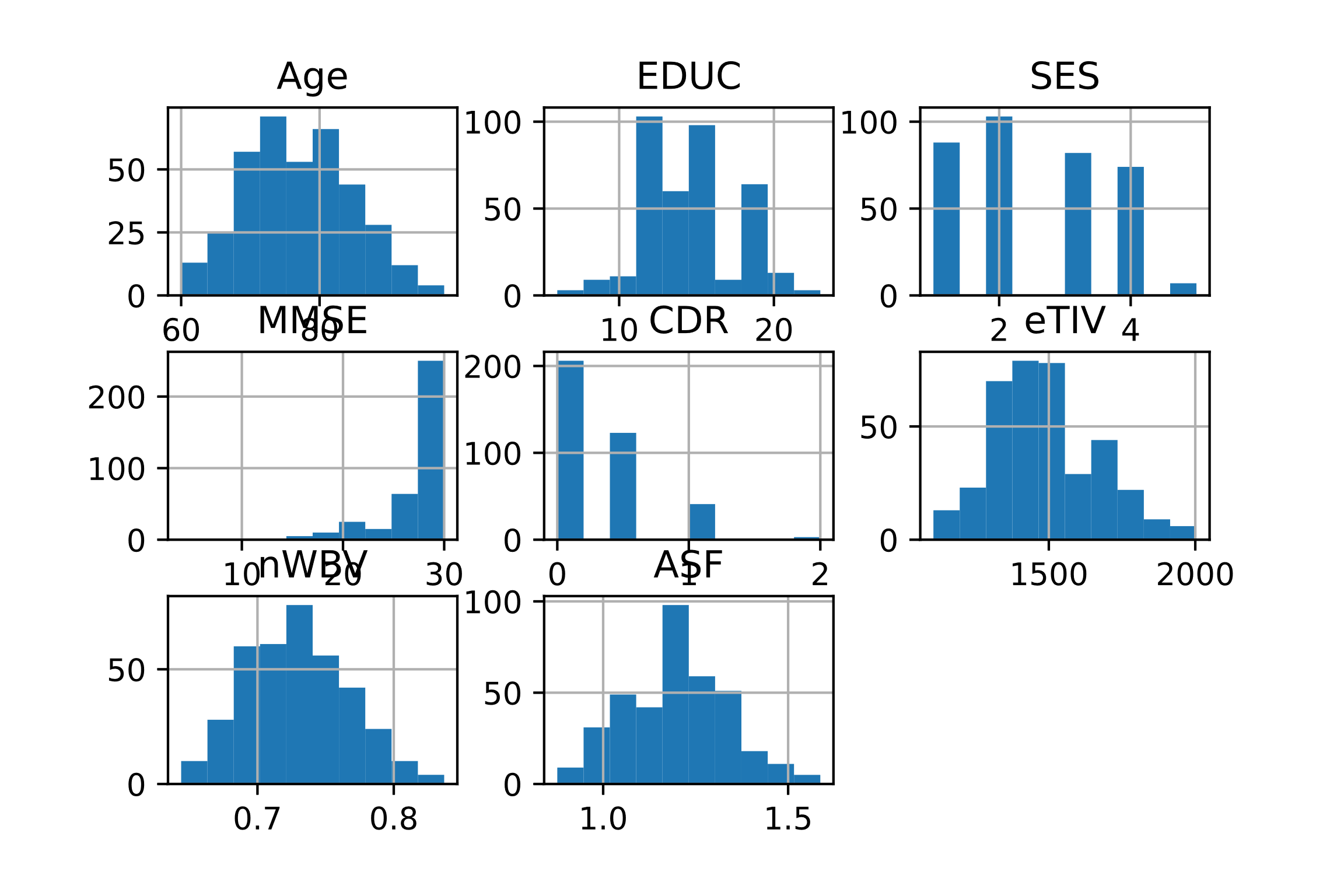
مجموعه داده Alzheimer دارای ده ستون می باشد که در زیر به توضیح هر ستون می پردازیم :

* ستون Group : برچسب و خروجی مجموعه داده است که دارای سه دسته با نام های Demented ، Nondemented و Converted می باشد.
* ستون Age  : سن افراد را نشان می دهد.
* ستون EDUC  : EDUC  سال های تحصیل افراد را نشان می دهد.
* ستون SES  : وضعیت اجتماعی و اقتصادی افراد را نشان می دهد که یک عددی بین 1 تا 5 را گرفته اند.
* ستون MMSE  : آزمون وضعیت ذهنی هست که افراد در آن شرکت کرده اند و نمره آنها در این ستون ثبت شده است.
* ستون CDR  : رتبه بندی دمانس بالینی افراد را نشان می دهد.
* ستون eTIV  : تخمین حجم کل داخل جمجمه افراد را نشان می دهد.
* ستون nWBV  : به نرمالسازی حجم کل مغز اشاره دارد.
* ستون ASF  : ضریب مقیاس پذیری اطلس را نشان می دهد.
* ستون M/F : جنسیت افراد ثبت شده است.

در این مجموعه داده، 373 نمونه داریم که هرکدام از این نمونه ها دارای 10 ویژگی هستند که در مورد ویژگی ها در بالا توضیح داده شده است.

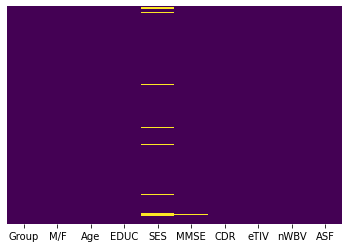
از میان ستون ها ، ستون های Group و M/F از نوع object ، ستون های Age ،EDUC و eTIV از نوع int64 و ستون های SES ، MMSE ، CDR ، nWBV و ASF از نوع float64 هستند. همچنین در بین داده‌های مجموعه، هیچ نمونه تکراری نداریم.

در زیر هیستوگرام مربوط به ستون های غیر قابل طبقه بندی را در یک پنجره نمایش می دهیم.



شکل 1- هیستوگرام ویژگی‌های غیر دسته‌ای

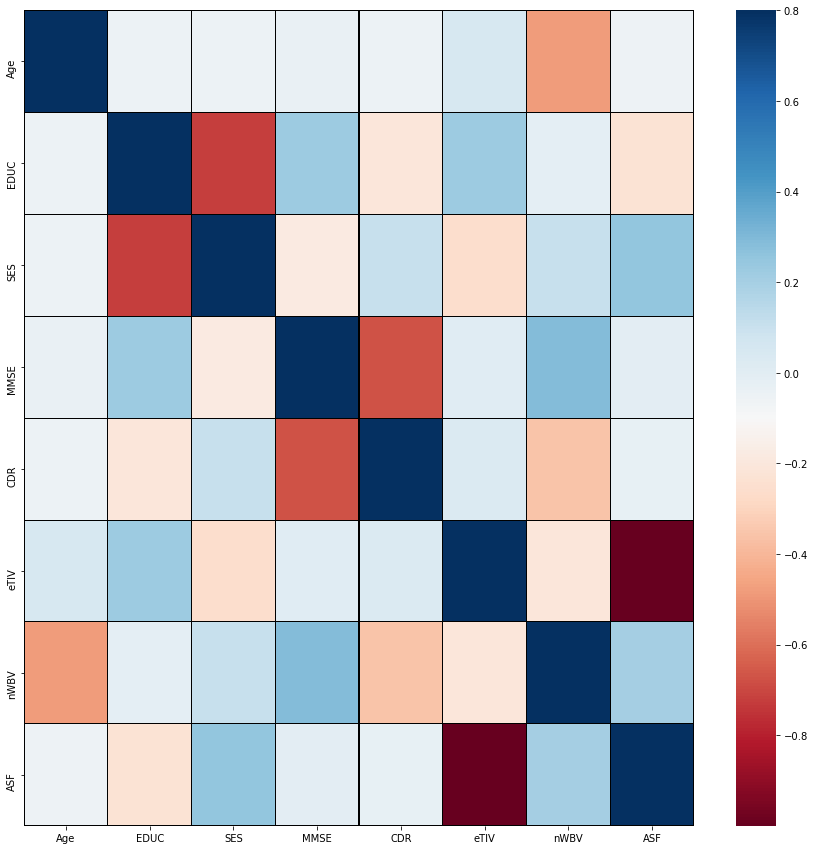
در زیر نقشه‌حرارتی مربوط به ویژگی های را نشان می دهیم تا بتوانیم تشخیص دهیم کدام ویژگی های مقادیر مفقوده دارند :



شکل 2- نقشه حرارتی ویژگی‌ها

طبق تصویر مشاهده می کنیم که ستون های مربوط به SES و MMSE مقادیر مفقوده دارند و باید برطرف شوند.

در این مرحله یک نقشه حرارتی دیگر برای بررسی همبستگی بین ویژگی‌ها استفاده می‌کنیم



شکل 3- نقشه حرارتی همبستگی بین متغیرها

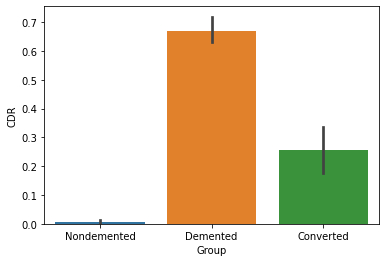
نمودار نقشه حرارتی همبستگی مثبت را با سایه های تیره و همبستگی منفی را با سایه های روشن تر نشان می دهد. مشاهده می کنیم که ستون های ASF و eTIV وابستگی نسبتا زیادی بهم دارند .همچنین ویژگی‌های SES و EDUC نسبت به هم همبستگی قابل توجهی دارند ولی به شدت دو ستون ASF و eTIV نیست. همچنین بین ویژگی‌هایCDR و MMSE ،Age و nWBV به نسبت ستون های باقیمانده همبستگی بیشتری برقرار است.

در این مرحله به پراکندگی ویژگی‌ها برحسب متغیر خروجی(Group) می‌پردازیم.



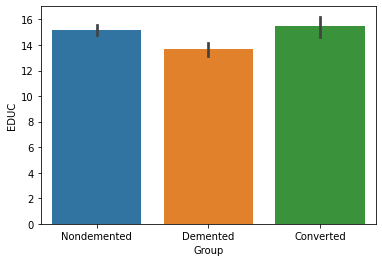
شکل 4- هیستوگرام SES نسب به متغیر خروجی

طبق نمودار مشاهده می کنیم که افرادی که Demented هستند به نسبت وضعیت اجتماعی و اقتصادی خوبی داشته اند.



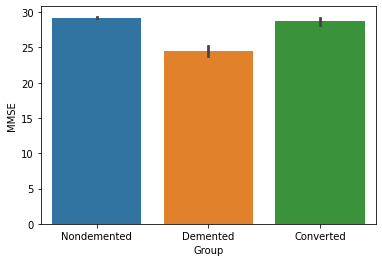
شکل 5- هیستوگرام CDR نسب به متغیر خروجی

طبق نمودار مشاهده می کنیم که افرادی که Demented هستند دارای علائم زوال عقل بالایی هستند.



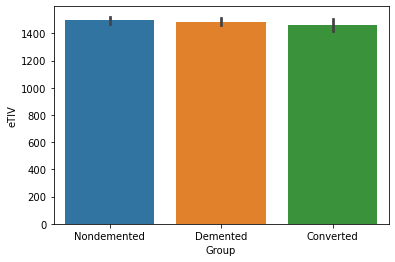
شکل 6- هیستوگرام EDUC نسب به متغیر خروجی

طبق نمودار مشاهده می کنیم که افرادی که converted هستند تحصیلات بالایی دارند.



شکل 7- هیستوگرام MMSE نسب به متغیر خروجی

طبق نمودار مشاهده می کنیم که افرادی که Nondemented و converted هستند به نسبت نمره ذهنی بالایی دارند.



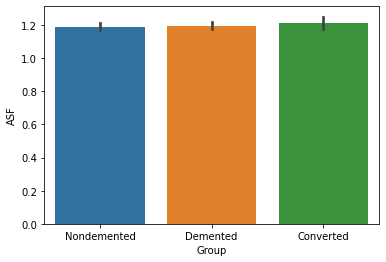
شکل 8- هیستوگرام eTIV نسب به متغیر خروجی

طبق نمودار مشاهده می کنیم که هر سه گروه ما تقریبا حجم کل جمجمه برابری دارند.



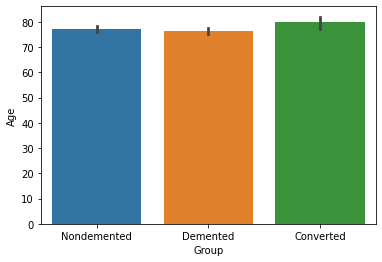
شکل 9- هیستوگرام nWBV نسب به متغیر خروجی

طبق نمودار مشاهده می کنیم که هر سه گروه ما تقریبا حجم کل مغز برابری دارند.



شکل 10- هیستوگرام ASF نسب به متغیر خروجی

طبق نمودار مشاهده می کنیم که هر سه گروه ما تقریبا ضریب مقیاس پذیری اطلس برابری دارند.

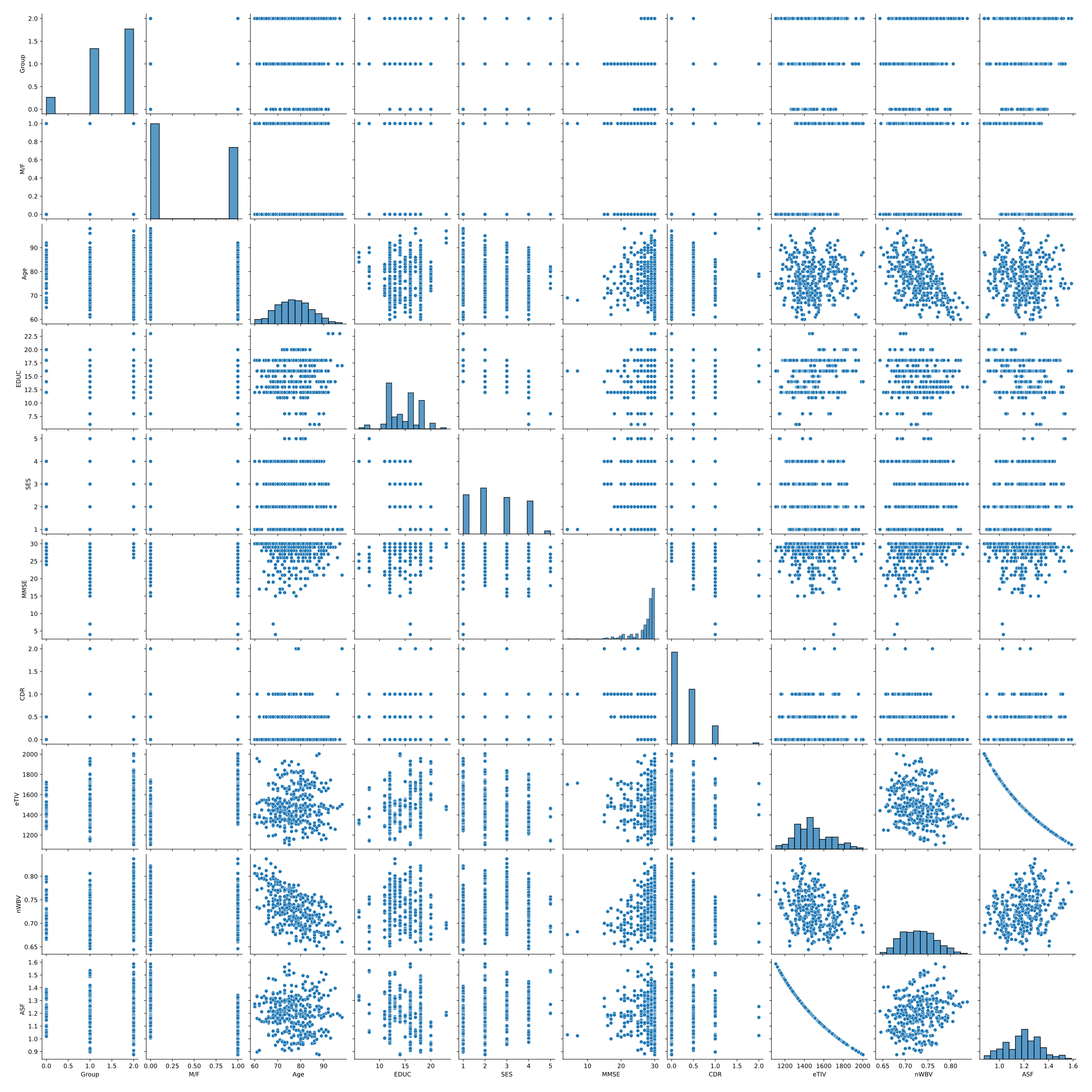


شکل 11- هیستوگرام Age نسب به متغیر خروجی

طبق نمودار مشاهده می کنیم که هر سه گروه ما تقریبا بازه ی سنی برابری دارند.

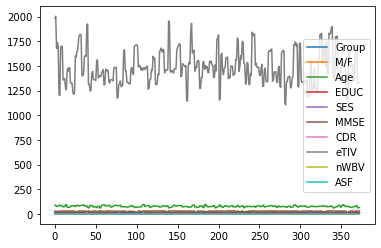
طبق بررسی های صورت گرفته در نمونه ها تعداد 213 مورد از نمونه های ما زن و160 مورد مرد هستند. همچنین 213 مورد ازنمونه های ما Nondemented ، 146 مورد Demented و 37 مورد Converted هستند.

در زیر با استفاده از کتابخانه ی seaborn یک نمودار مقایسه ای بین ویژگی ها رسم می کنیم :



شکل 12- همبستگی بین ویژگی‌ها به صورت دو به دو

برای فهم بهتر مجموعه داده‌ای که داریم تغییرات هریک از ویژگی ها را روی یک پنجره نمایش می دهیم.



شکل 13- تغییرات ویژگی‌ها در یک پنجره

طبق نموار بالا تغییرات هریک از ویژگی ها و واکنش هرکدام از ویژگی ها را در بازه های مختلف به طور کامل مشاهده می کنیم. مشاهده می کنیم که بازه ی تغییرات ویژگی eTIV به نسبت سایر ویژگی ها خیلی زیاد است.همچنین ویژگی هایی مثل Group و M/F که عددی نیستند با یک خط راست مشخص شده اند. سایر ویژگی ها نیز بازه‌ی تغییراتشان طبق نمودار کم هست.