به نام خدا

دانشگاه صنعتی خواجه نصیرالدین طوسی

دانشکده برق

مبانی سیستم های هوشمند

گزارشکار پروژه‌ی پایانی

عنوان پروژه: تشخیص بیماری قلبی با استفاده از شبکه عصبی چندلایه

مقایسه با شبکه هاپفیلد و مدل فازی

محدثه فیضی - 40007933

Google colab link:

https://colab.research.google.com/drive/1xL9PuvnkMd_97igqcaTTedhZBMgWf3SO?usp=drive_link

استاد : آقای دکتر مهدی علیاری

بهمن ۱۴۰۳

معرفی داده یا سیستم:

در این پروژه، از مجموعه داده Heart Disease Dataset استفاده می‌شود که شامل ویژگی‌های پزشکی بیماران (مانند فشار خون، کلسترول، ضربان قلب و غیره) و برچسب‌های مربوط به وضعیت بیماری قلبی آن‌ها است. هدف اصلی این تحقیق، طراحی یک سیستم هوشمند برای تشخیص بیماری قلبی بر اساس ویژگی‌های پزشکی بیماران است.

این دیتا شامل 14 ستون به شرح زیر است:

سن(age)

جنسیت(sex): 0 زن، 1 مرد

نوع درد قفسه سینه(cp): 1 درد غیرکلاسیک، 2 درد کلاسیک، 3 درد آتیپیک، 4 بدون درد

فشار خون در حالت استراحت(trestbps)

کلسترول خون(chol)

وضعیت قند خون ناشتا(fbs): 0 کمتر از 120، 1 بیشتر از 120

نتیجه آزمون الکتروکاردیوگرام در حالت استراحت(restecg): 0 نرمال، 1 معکوس st موج، 2 هیپرتروفی دیواره قلب

حداکثر ضربان قلب(thalach)

وجود درد قفسه سینه ناشی از ورزش(exang): 0 ندارد، 1 دارد

انحراف سطح st در اثر ورزش نسبت به حالت استراحت(oldpeak)

شیب بالای بخش بالایی خط ST (slope): 1 شیب نزولی، 2 شیب افقی، 3 شیب صعودی

تعداد رگ‌های خونی که در تصاویر رنگی قابل مشاهده‌اند(ca)

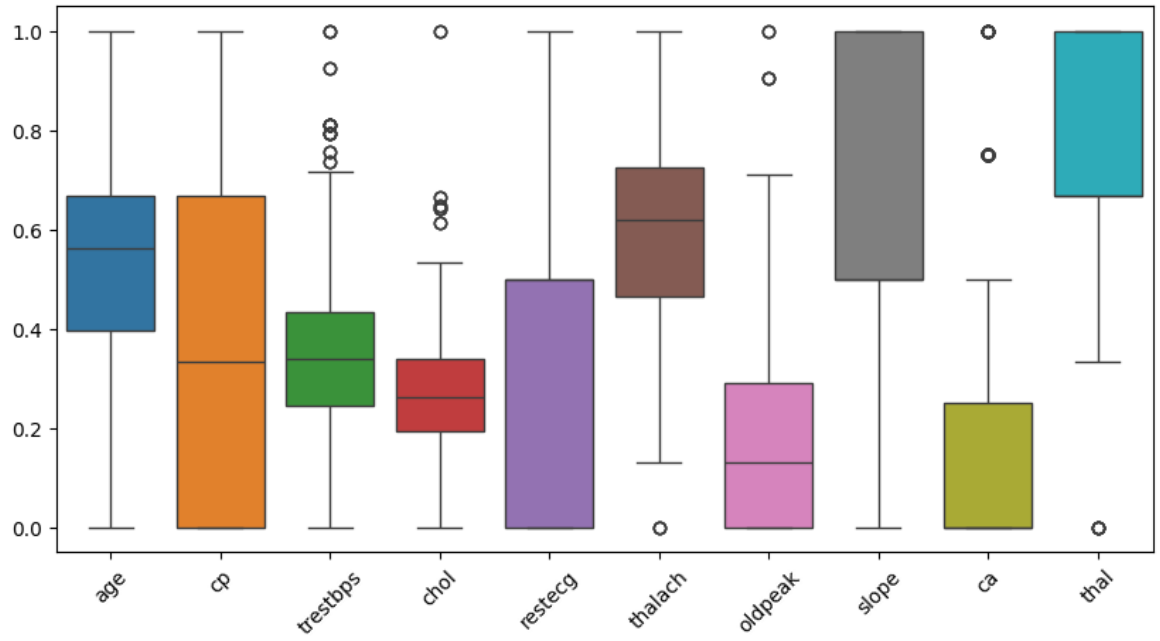
وضعیت جریان خون در قلب(thal): 0 نامشخص، 1 ثابت(مشکل دائمی در خون‌رسانی)، 2 برگشت‌پذیر(مشکل موقتی)، 3 نرمال(بدون مشکل)

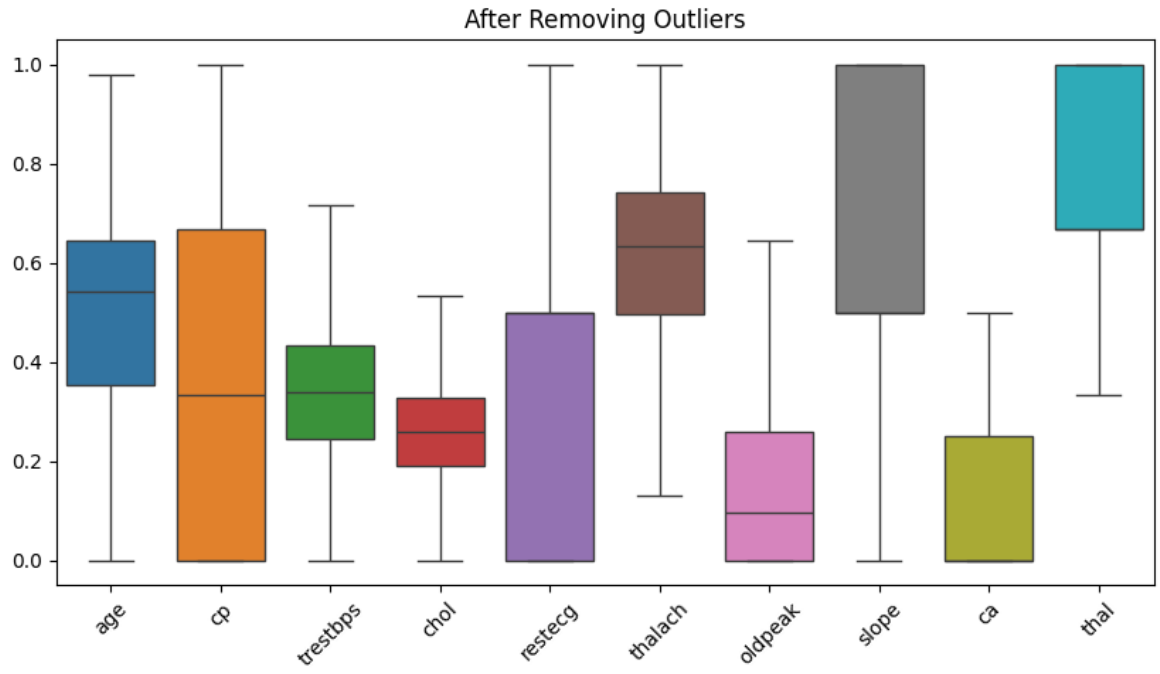
متغیر هدف برای پیش‌بینی بیماری قلبی(target): 0 بدون بیماری قلبی، 1 با بیماری قلبی

معرفی فرآیند و روش: برای تحلیل داده‌ها و تشخیص بیماری قلبی، مراحل زیر اجرا خواهند شد:

- **پیش‌پردازش داده‌ها**

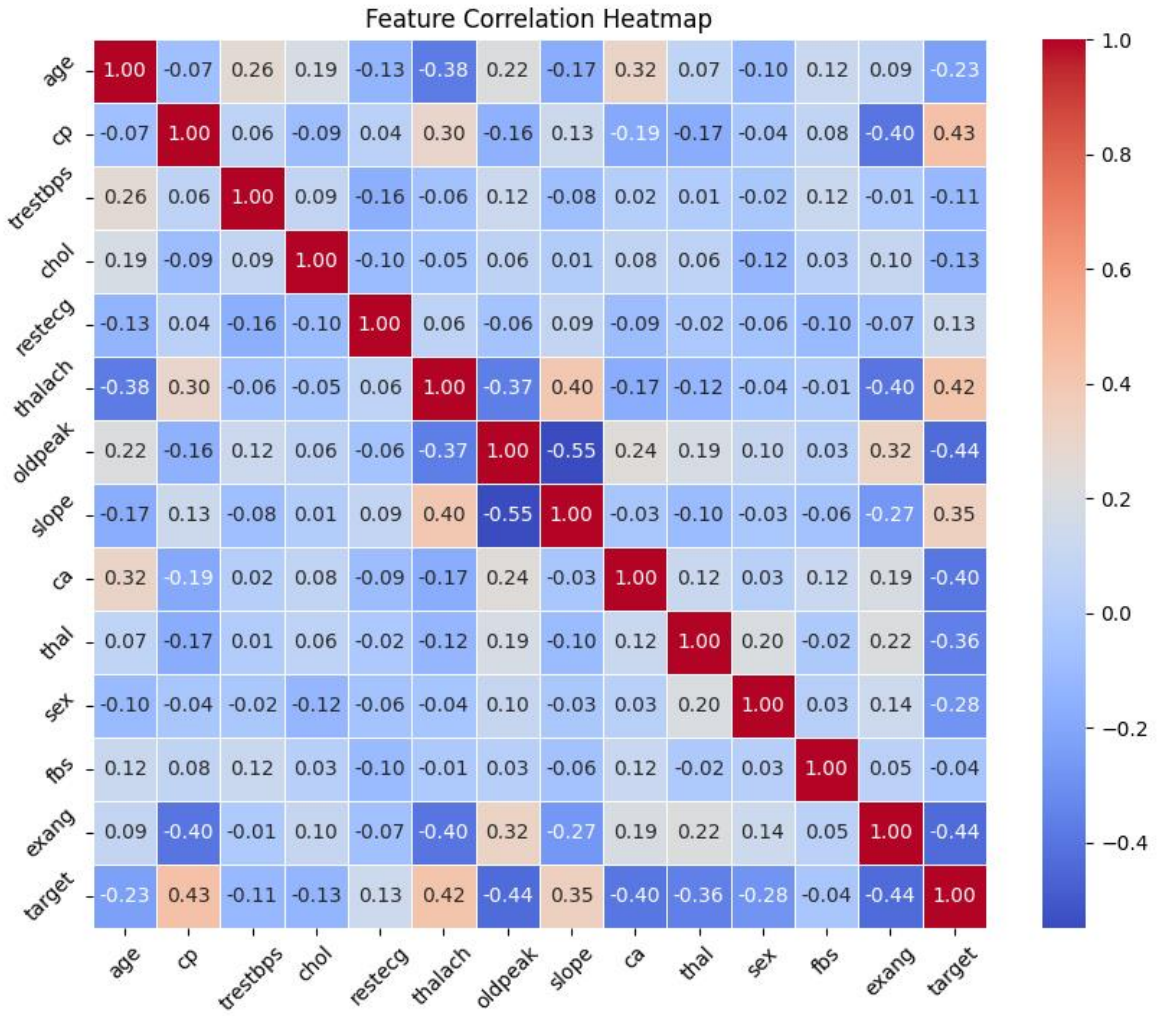
داده‌های پرت را با روش میانه، جایگزین می‌کنیم. (میانگین‌گیری خوب نیست چون ممکن است برای ویژگی‌ها در محدوده‌ی پرت باشد.)





Replaced 151 outlier values with median.

تحلیل همبستگی ویژگی‌ها با هدف با استفاده از heatmap:



طبق نقشه‌ی حرارتی ویژگی‌هایی که همبستگی مثبت دارند ینی هرچه مقدارشان افزایش یابد، احتمال ابتلا به بیماری قبلی بیشتر می‌شود و ویژگی‌هایی که همبستگی منفی دارند با افزایش مقدار آنها، احتمال ابتلا کمتر می‌شود.

cp، restecg، thalach و slope تاثیر مثبت و بقیه ویژگی‌ها تاثیر منفی دارند. ویژگی‌های کم اهمیت را می‌توان حذف کرد یا ویژگی‌های مهم را با وزن‌دهی تقویت کنیم. (مثلا fbs تاثیر قابل توجهی ندارد.) برای دقیق‌تر بودن نتیجه هیچکدام را حذف نمی‌کنیم.

بررسی داده‌های پرت و تقسیم داده‌ها به مجموعه‌های آموزش، اعتبارسنجی و تست:

Train size: 615
Test size: 205
Validation size: 205

- **مدلسازی با شبکه MLP(1)، Hopfield(2)، Fuzzy(3) و مقایسه‌ی آنها:**

1) Dense: لایه کاملاً متصل

لایه اول(128 نورون): یادگیری ویژگی‌های اولیه

لایه دوم(128 نورون): یادگیری یژگی‌های پیچیده‌تر

لایه سوم (64 نورون) :تمرکز روی ویژگی‌های اصلی و کاهش ابعاد

لایه چهارم(1 نورون، سیگموئید): خروجی نهایی با تابع سیگموئید با طبقه‌بندی باینری

L2 Regularization: کاهش وزن‌های بزرگ با تنظیم مقدار $\lambda=0.0005$

Batch Normalization: نرمال‌سازی مقدار خروجی لایه‌ی Dense برای پایداری بیشتر

ReLU: تابع فعال‌سازی برای افزایش توانایی مدل در یادگیری ویژگی‌های غیرخطی

Dropout: جلوگیری از بیش‌برازش با حذف تصادفی 25٪ یا 20٪ از نورون‌ها در هر مرحله

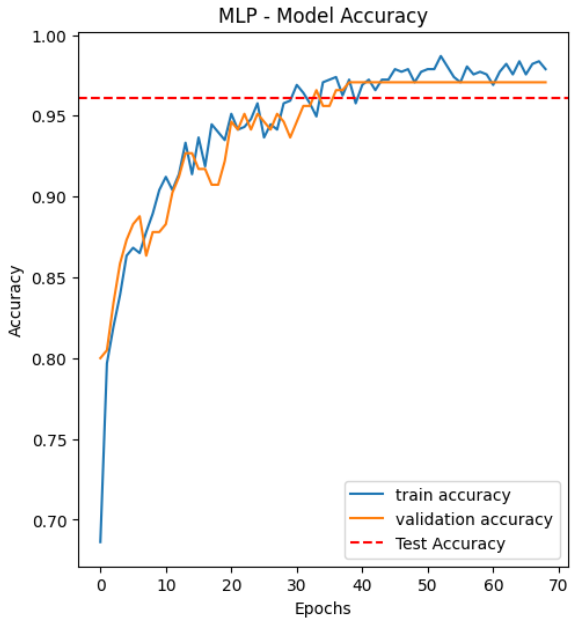
این ترکیب باعث می‌شود که مدل عمیق، پایدار، و مقاوم در برابر بیش‌برازش باشد.

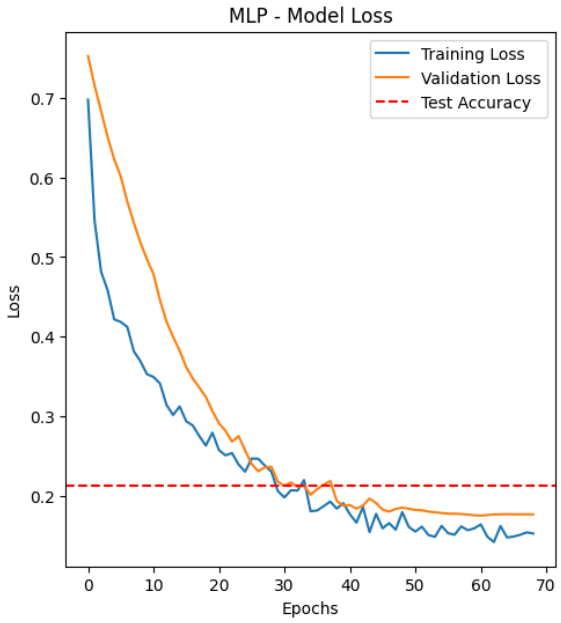
در ادامه از بهینه‌ساز Adam استفاده شده‌است.

EarlyStopping: اگر مقدار خطای داده‌ی اعتبارسنجی بعد از 8 دوره متوالی بهبود پیدا نکند، آموزش متوقف می‌شود.

ReduceLROnPlateau: اگر مقدار خطای داده‌ی اعتبارسنجی بعد از 3 دوره بهتر نشود،نرخ یادگیری به نصف کاهش پیدا می‌کند. این کار به مدل کمک می‌کندکه دقیق‌تر یاد بگیرد و نواسات شدید نداشته باشد.

مقادیر احتمال مدل بین صفر و یک است. برای پیش‌بینی، اگر مقدار احتمال بیشتر از 0.5 باشد آن شخص بیمار و در غیراینصورت سالم در نظر گرفته می‌شود.





MLP Accuracy: 96.1%

MLP Recall: 96.12%

MLP Precision: 96.12%

MLP MSE: 3.9%

2)

هر نورون به تمام نورون‌های دیگر متصل است (به جز خودش)

قانون یادگیری Hebbian: اگر دو نورون به‌طور همزمان فعال شوند، وزن بین آن‌ها افزایش می‌یابد.

شبکه هاپفیلد فقط با -1 و 1 کار می‌کند. پس داده‌ها به این مقادیر تبدیل می‌شوند. اگر مقدار خروجی پیش‌بینی شده بزرگتر از صفر باشد کلاس 1 یعنی بیمار و در غیر اینصورت کلاس 0 یعنی سالم است.

مدل بسیار سریع و ساده است ولی تعداد الگوهای قابل ذخیره محدود است و در شبکه‌های بزرگ دچار همپوشانی الگوها و خطا می‌شود.

Hopfield Accuracy: 50.24%

Hopfield Recall: 100%

Hopfield Precision: 50.24%

Hopfield MSE: 49.76%

Accuracy نزدیک 50٪ یعنی مدل تقریباً تصادفی عمل می‌کند و نیاز به بهبود دارد ولی هرچه سعی بر بهبود مدل شد، بالاترین درصدها همین‌ها بودند. شبکه هاپفیلد برای این داده مناسب نیست.

3)

تابع گوسی مشخص می‌کند که یک مقدار چقدر به یک مقدار میانگین نزدیک است. هرچه مقدار نزدیک‌تر باشد، مقدار عضویت بیشتر خواهد بود.

داده‌های عددی معمولی را به داده‌های **فازی** تبدیل شده و مقدار عضویت آنها مشخص می‌شود.

در این کد از Random Forest استفاده شده است. یک مدل یادگیری ماشین مبتنی بر چندین درخت تصمیم‌گیری است که خروجی آن میانگین یا اکثریت آراء درخت‌هاست. همچنین مدل با درخت تصمیم دقت مدل کمتر بود.

Fuzzy Accuracy: 94.63%

Fuzzy Recall: 99.03%

Fuzzy Precision: 91.07%

Fuzzy MSE: 5.37%

مقایسه مدل‌ها:

MLP Accuracy > Fuzzy Accuracy > Hopfield Accuracy

Hopfield Recall > Fuzzy Recall > MLP Accuracy

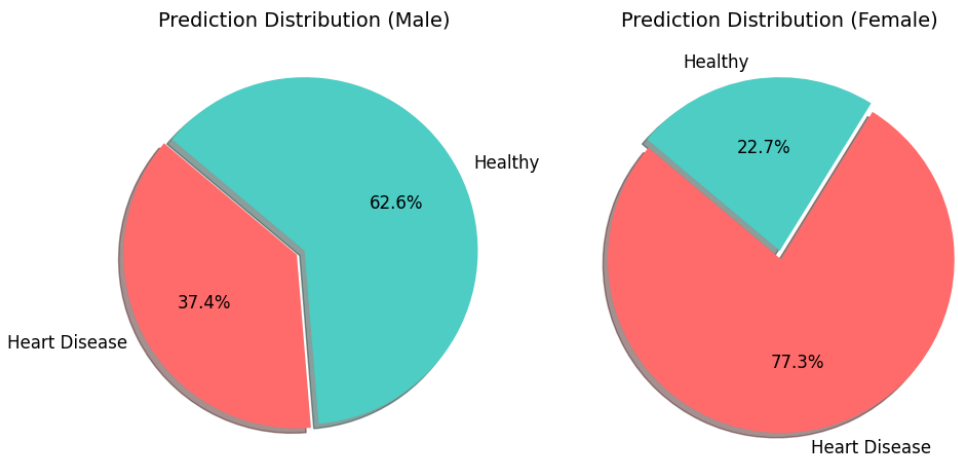
MLP Precision > Fuzzy Precision > Hopfield Precision

MLP MSE < Fuzzy MSE < Hopfield MSE

شبکه هاپفیلد مناسب برای داده‌های ساده و تعداد ویژگی کم است و هدف از آن یادآوری الگو و حافظه است، اما اگر داده‌های پیچیده‌تر، چندکلاسه و غیرخطی داشته باشیم، بهتر است از شبکه عصبی چندلایه استفاده کنیم که برای مسائل نظارتی مناسب‌تر است. همچنین، مدل‌های فازی زمانی که با ابهام، عدم قطعیت و پیچیدگی‌های داده‌ها روبه‌رو هستیم، انتخاب بهتری خواهند بود.

Accuracy نشان می‌دهد که MLP در شناسایی درست کلاس‌ها بسیار بهتر از شبکه هاپفیلد عمل می‌کند. Recall بیان می‌کند که مدل چقدر در شناسایی درست کلاس‌های مثبت موفق بوده است؛ هاپفیلد تمامی موارد مثبت را شناسایی کرده اما این لزوماً به معنای عملکرد خوب آن نیست، زیرا ممکن است این امر به قیمت کاهش دقت (Precision) باشد. MLP نیز عملکرد بالایی دارد. Precision نشان می‌دهد که چه مقدار از پیش‌بینی‌های مثبت مدل صحیح بوده‌اند. MLP احتمال بسیار بالایی دارد که پیش‌بینی‌های مثبتش درست باشند، در حالی که هاپفیلد عملکرد ضعیفی نشان داده است. MSE میزان خطای مدل را نمایش می‌دهد. هرچه مقدار آن کمتر باشد، مدل دقیق‌تر است. MSE در MLP نشان‌دهنده عملکرد بسیار خوب مدل است. در مقابل، در هاپفیلد بیانگر خطای بالا در پیش‌بینی‌های آن می‌باشد.
همچنین مدل فازی هم عملکرد بسیار خوبی داشته است. تقریباً تمام نمونه‌های مثبت را به درستی شناسایی کرده و همچنین، اکثر پیش‌بینی‌های مثبت آن صحیح بوده‌اند. MSE مدل فازی. از هاپفیلد بسیار کمتر و نزدیک به MLP می‌باشد.

ارزیابی اضافی: تاثیر جنسیت در بیماری قلبی



نتیجه گیری:

در مجموع، مدل MLP و مدل فازی عملکرد بهتری نسبت به هاپفیلد دارند و در حل مسائل پیچیده‌تر و دارای عدم قطعیت، انتخاب‌های مناسب‌تری هستند.

مدل MLP رو ذخیره می‌کنیم سپس یک برنامه تحت وب را با استفاده از Streamlit برای تشخیص بیماری قلبی پیاده‌سازی می‌کنیم. کاربر اطلاعات خود را وارد می‌کند، مدل از پیش آموزش‌دیده‌شده آن‌ها را پردازش کرده و نتیجه را نمایش می‌دهد.