

A Stacking-Based Model for Non-Invasive Detection of Coronary Heart Disease

Coronary arteriography (CAG) is an accurate invasive technique for the diagnosis of coronary heart disease (CHD).

Arash Mehrzadi, Hamidreza Hematyar
QIAU

فهرست

2	چکیده
3	مقدمه
6	دیتا
8	پیش پردازش داده ها
9	انتخاب ویژگی ها
10	مدل مقاله
14	مدل پیشنهادی
16	مقایسه نتایج پیش پردازش داده ها
18	مقایسه نتایج انتخاب ویژگی ها
22	نتایج برای روش پیشنهادی و روش های دیگر
24	پارامتر مدل ها
25	کد

چکیده

عروق کرونر (CAG) یک تکنیک تهاجمی دقیق برای تشخیص بیماری عروق کرونر قلب (CHD) است. با این حال، روش تهاجمی آن برای تشخیص CHD در معاینه فیزیکی سالانه مناسب نیست. با کاربرد موفقیت‌آمیز یادگیری ماشین (ML) در زمینه‌های مختلف، هدف ما انجام یکپارچه‌سازی انتخابی الگوریتم‌های چندگانه ML و تأیید اعتبار روش‌های انتخاب ویژگی با اطلاعات بالینی شخصی است که معمولاً در معاینه فیزیکی سالانه دیده می‌شود. در این مطالعه، یک مدل مبتنی بر انباشته دو سطحی طراحی شده است که در آن سطح ۱ سطح پایه و سطح ۲ متا سطح است. پیش‌بینی طبقه‌بندی‌کننده‌های سطح پایه به عنوان ورودی فراسطح انتخاب می‌شوند. ابتدا ضریب همبستگی پیرسون و حداکثر ضریب اطلاعات محاسبه می‌شود تا طبقه بندی‌کننده با کمترین همبستگی پیدا شود. سپس از الگوریتم شمارش برای یافتن بهترین طبقه‌بندی‌کننده‌های ترکیبی استفاده می‌شود که در پایان بهترین نتیجه را به دست می‌آورند. مجموعه داده Z-Alizadeh Sani CHD که ما استفاده می‌کنیم شامل ۳۰۳ مورد است که توسط CAG تأیید شده است.

نتایج تجربی نشان می‌دهد که مدل پیشنهادی، دقت، حساسیت و ویژگی به ترتیب ۹۵.۴۳٪، ۹۵.۸۴٪، ۹۴.۴۴٪ برای تشخیص CHD به دست می‌آورد. روش پیشنهادی می‌تواند به طور موثر به پزشکان کمک کند تا افراد دارای عروق کرونر طبیعی از مبتلایان به CHD را تشخیص دهند.

کلمات کلیدی: عروق کرونر، روش تهاجمی

مقدمه

بیماری عروق کرونر قلب (CHD) یکی از علل اصلی مرگ و میر قلبی عروقی در سطح جهان است. در حال حاضر روش های تشخیصی CHD را می توان به روش های تهاجمی و غیرتهاجمی تقسیم کرد. آنژیوگرافی عروق کرونر (CAG) یک تکنیک تشخیصی تهاجمی نسبتاً ایمن و قابل اعتماد است که به طور گسترده در عمل بالینی به عنوان استاندارد طلایی برای تشخیص CHD مورد استفاده قرار گرفته است. با این حال، ماهیت تهاجمی آن و هزینه عملیات نسبتاً گران، استفاده از آن در معاینه فیزیکی سالانه را دشوار می کند. الکتروکاردیوگرام (ECG) و اکوکاردیوگرافی روش های غیرتهاجمی هستند، اما دقت قابل اعتمادی ندارند.

بنابراین، یافتن روش های غیرتهاجمی جدید برای تشخیص CHD ضروری است.

در قلب و عروق بالینی، یادگیری ماشین (ML) یک روش موثر برای پیش بینی مرگ و میر ناشی از همه علل در بیماران مشکوک به CHD ثابت شده است. در اپیدمیولوژی قلبی عروقی تحت بالینی، ML می تواند پیش بینی بهتری نسبت به نمرات استاندارد خطر قلبی عروقی در ارتباط با نقاط داده فنوتیپی ارائه دهد. روش های ML به طور گسترده در برخورد با داده های موجود در پزشکی استفاده می شود. در سال های اخیر، تعدادی از الگوریتم های ML برای تشخیص CHD توسعه یافته اند. Feshki و Shijani تشخیص CHD را با یک الگوریتم تکاملی و یک شبکه عصبی پیشخور بهبود دادند. Davari و همکاران ویژگی های ECG را با روش های فرکانس و دامنه غیرخطی برای شناسایی علائم CHD با طبقه بندی کننده بردار پشتیبان (SVC) استخراج کرد.

Vernekar و همکاران ویژگی های مارکوف را به همراه سایر ویژگی های حوزه آماری و فرکانس از فونوکاردیوگرام (PCG) استخراج کرد و از مجموعه شبکه عصبی مصنوعی و درخت افزایش گرادیان برای آموزش مدل استفاده کرد.

Kumar و همکاران همچنین از سیگنال های ECG اما با تبدیل موجک تحلیلی انعطاف پذیر برای مشخص کردن CHD استفاده کرد.

یک روش ترکیبی پیشنهاد کرد که شامل شناسایی عوامل خطر با استفاده از انتخاب زیر مجموعه ویژگی مبتنی بر همبستگی با روش جستجوی بهینه‌سازی شنای ذرات و الگوریتم‌های خوشه‌بندی K-means بود [۹]. علیرزاده و همکاران از سه طبقه‌بندی کننده برای تشخیص تنگی سه شریان کرونری، یعنی نزولی قدامی چپ، سیرکومفلکس چپ و شریان کرونری راست استفاده کرد تا دقت بالاتری برای تشخیص CHD بدست آورد.

داوری و همکاران با پایگاه داده Long Term ST به دقت تشخیص ۹۹.۲٪ دست یافتند، اما پایگاه داده ای که آنها برای بیماران CHD استفاده کردند با تغییرات بخش ST مختلف همراه بود.

و در عمل بالینی، بسیاری از بیماران CHD قطعه ST طبیعی دارند. بنابراین، استفاده از پایگاه‌های اطلاعاتی بیماران مبتلا به بیماری عروق کرونر اما بخش‌های ST طبیعی ممکن است برای استفاده از مدل تشخیص CHD مبتنی بر هوش مصنوعی در موقعیت‌های پیچیده بالینی مفیدتر باشد. علاوه بر این، تحقیقات قبلی معمولاً تنها از یک نوع طبقه‌بندی کننده ML برای تشخیص خودکار CHD استفاده می‌کردند. با این حال، بسیاری از محققان ML به ویژه آنهایی که در مسابقات ML شرکت می‌کنند، با موفقیت از تکنیک‌های ترکیب طبقه بندی کننده برای بهبود دقت طبقه بندی کننده ها استفاده کرده اند .

تکنیک‌های ترکیب پیش‌بینی‌های به‌دست‌آمده از طبقه‌بندی‌کننده‌های چندگانه سطح پایه را می‌توان در سه چارچوب ترکیبی خلاصه کرد: رأی‌گیری (مورد استفاده در بسته‌بندی و تقویت)، پشته‌بندی و آبخاری. برای مجموعه داده های پیچیده تر، طبقه بندی کننده سنتی را می‌توان با انواع مختلفی از قوانین ترکیبی بهبود بخشید . در پشته‌بندی، پیش‌بینی‌های مجموعه‌ای از طبقه‌بندی‌کننده‌ها به عنوان ورودی‌های الگوریتم یادگیری سطح بعدی ارائه می‌شوند. سطح بعدی الگوریتم برای ارتباط بهینه پیش بینی های مدل و تشکیل سطح بعدی مجموعه نهایی پیش بینی ها آموزش داده شده است. روابط جفتی همیشه بین سطوح مختلف قبل از پیش بینی نهایی وجود دارد. ما روابط بین مدل‌ها را در سطح پایه تحلیل می‌کنیم و ترکیب بهینه مدل را با یک الگوریتم شمارش پیدا می‌کنیم.

به طور خلاصه، سهم اصلی این کار به شرح زیر است:

هشت روش انتخاب ویژگی برای ارزیابی عملکرد آنها برای تشخیص خودکار CHD بررسی شده است. ما متوجه شدیم که استراتژی یادگیری ماشینی RFECV بالاترین عملکرد پیش‌بینی را در اعتبارسنجی متقابل ده برابری مکرر به دست آورد. آن ویژگی‌هایی که با روش RFECV انتخاب می‌شوند برای متخصصان قلب در تشخیص بالینی CHD ارزش مرجع بالایی دارند.

در مجموع از ۱۰ روش طبقه‌بندی استفاده می‌شود. با تجزیه و تحلیل نتایج، مشخص شد که ترکیب مدلی که بهترین عملکرد را نشان می‌دهد را نمی‌توان با محاسبه مستقیم ضریب همبستگی پیرسون (PCC) و حداکثر ضریب اطلاعات (MIC) تعیین کرد. بنابراین، یک استراتژی جدید برای جستجوی ترکیب بهینه پیشنهاد می‌شود که در آن ابتدا مدلی انتخاب می‌شود که حداقل همبستگی را با سایر مدل‌ها داشته باشد و سپس با برشمردن هر ترکیب احتمالی مدل انتخاب‌شده با سایر مدل‌ها، ترکیب بهینه تعیین می‌شود. نتایج ما نشان می‌دهد که استراتژی پیشنهادی عملکرد رضایت بخشی دارد.

ترکیب مدل بهینه برای تشخیص خودکار CHD تعیین می‌شود. استفاده از استراتژی ترکیب مدل پیشنهادی در ۳ مجموعه داده دیگر نیز نتایج رضایت‌بخشی را نشان می‌دهد، که توانایی تعمیم استراتژی ترکیب مدل پیشنهادی ما را نشان می‌دهد. ادامه این مقاله به شرح زیر سازماندهی شده است. در بخش دوم، منبع داده و روش‌های پیش پردازش داده‌ها معرفی شده است. در بخش III، جزئیات فنی مدل مبتنی بر انباشته دو سطحی پیشنهادی ما شرح داده شده است. نتایج تجربی در بخش IV و سپس بحث در بخش V ارائه شده است.

دیتا

مجموعه داده Z-Alizadeh شامل ۲۱۶ بیمار CHD و ۸۷ فرد سالم است که توسط ۵۴ نوع مختلف نشان داده شده اند. ویژگی های بالینی و جمعیت شناختی همانطور که در جدول زیر نشان داده شده است.

مجموعه داده ها عدم تعادل بزرگی را در توزیع طبقات هدف نشان می دهد، زیرا تقریباً ۳ برابر بیشتر از افراد سالم بیماران CHD وجود دارد. در چنین حالتی، روش نمونه برداری بیش از حد اقلیت مصنوعی (SMOTE) برای حل مشکل عدم تعادل استفاده می شود. ایده اصلی روش SMOTE تجزیه و تحلیل کلاس های اقلیت و ترکیب کلاس های اقلیت جدید با نمونه برداری بیش از حد است.

داده های افراد عادی توسط SMOTE در طول اعتبارسنجی متقاطع و نه قبل از فرآیند اعتبارسنجی متقابل نمونه برداری می شود. داده های مصنوعی فقط برای مجموعه آموزشی ایجاد می شود بدون اینکه بر مجموعه آزمایشی تأثیر بگذارد. اگر یک ویژگی دارای واریانس باشد که مرتبه ای بزرگتر از سایر ویژگی ها باشد، ممکن است بر تابع هدف تأثیر بگذارد و باعث می شود برآوردگر نتواند همانطور که انتظار می رود از ویژگی های دیگر به درستی یاد بگیرد.

از آنجایی که ۵۴ ویژگی مجموعه داده شامل ۲۳ داده عددی و ۳۱ داده طبقه بندی می شود، از تکنیک حداکثر و حداقل نرمال سازی برای استانداردسازی این ویژگی ها استفاده می شود. حداکثر و حداقل نرمال سازی یک روش رایج پردازش داده ها است که می توان آن را به صورت (۱) تعریف کرد x . ویژگی ورودی، \max نشان دهنده حداکثر مقدار، \min نشان دهنده حداقل مقدار، و x^* نشان دهنده مقدار خروجی پس از عادی سازی است. در این مطالعه، ما از این رویکرد برای مقیاس بندی ۲۳ ویژگی استفاده می کنیم. یافتن روابط اهمیت بالقوه در میان ویژگی ها مفید است.

$$x^* = \frac{x - \min}{\max - \min}$$

در ادامه، فیچر های دیتاست ذکر شده است:

Feature type	Feature name	Range
Demographic	Age	30-86
	Weight	48-120
	Length	140-188
	Sex	Male,Female
	BMI(Body Mass Index)	18.1-40.9
	DM (Diabetes Mellitus)	Yes,No
	HTN (Hyper Tension)	Yes,No
	Current Smoker	Yes,No
	Ex-Smoker	Yes,No
	FH (Family History)	Yes,No
	Obesity	Yes,No
	CRF (Chronic Renal Failure)	Yes,No
	CVA (Cerebrovascular Accident)	Yes,No
	Airway Disease	Yes,No
	Thyroid Disease	Yes,No
Clinical	CHF (Congestive Heart Failure)	Yes,No
	DLP (Dyslipidemia)	Yes,No
	BP (Blood Pressure: mmHg)	90-190
	PR (Pulse Rate) (ppm)	50-110
	Edema	Yes,No
	Weak peripheral pulse	Yes,No
	Lung Rales	Yes,No
	Systolic murmur	Yes,No
	Diastolic murmur	Yes,No
	Typical Chest Pain	Yes,No
	Dyspnea	Yes,No
	Function Class	1,2,3,4
	Atypical	Yes,No
	Nonanginal	Yes,No
	Exertional CP (Exertional Chest Pain)	Yes,No
	LowTH Ang (low Threshold angina)	Yes,No
	Rhythm	Yes,No
	Q Wave	0,1
	ST Elevation	0,1
	ST Depression	0,1
	T inversion	0,1
	LVH (Left Ventricular Hypertrophy)	Yes,No
	Poor R Progression (Poor R Wave Progression)	Yes,No
	FBS (Fasting Blood Sugar) (mg/dl)	62-400
	Cr (creatinine) (mg/dl)	0.5-2.2
	TG (Triglyceride) (mg/dl)	37-1050
	LDL (Low density lipoprotein) (mg/dl)	18-232
	HDL (High density lipoprotein) (mg/dl)	15.9-111
	BUN (Blood Urea Nitrogen) (mg/dl)	6-52
	ESR (Erythrocyte Sedimentation rate) (mm/h)	1-90
	HB (Hemoglobin) (g/dl)	8.9-17.6
	K (Potassium) (mEq/lit)	3.0-6.6
	Na (Sodium) (mEq/lit)	128-156
	WBC (White Blood Cell) (cells/ml)	3700-18000
	Lymph (Lymphocyte) (%)	7-60
	Neut (Neutrophil) (%)	32-89
	PLT (Platelet) (1000/ml)	25-742
	EF (Ejection Fraction) (%)	15-60
	Region with RWMA a (Regional Wall Motion Abnormality)	0,1,2,3,4
	VHD (Valvular Heart Disease)	Normal, Mild, Moderate, Severe

پیش پردازش داده ها

با توجه به تنوع داده ها در انواع مختلف، در ابتدا نیاز داریم دیتای مورد نظر را به فرمی برسانیم که برای الگوریتم های ما قابل پذیرش و قابل پردازش میباشد.

دیتاست مورد بحث دارای انواع زیر میباشد:

• Integer

• Bool

• Enum

که برای هر یک از این داده ها نیاز به رویکردی متفاوت داریم، و برای این کار دیتاست را به سه بخش متفاوت تقسیم میکنیم تا بتوانیم فرایندهای مربوط به هر کدام از این تایپ را انجام دهیم و سپس دیتاست ها را به یکدیگر چسبانده و دیتافریمی یکپارچه تهیه میکنیم.

برای دیتاهایی با تایپ عددی از رویکرد MinMaxScaler برای اسکیل کردن مقادیر بین دو عدد ۰ و ۱ استفاده میکنیم.

همچنین برای دیتاهایی با تایپ bool از رویکرد binary استفاده میکنیم که مقادیر صحیح به مقدار ۱ و مقادیر نادرست به مقدار ۰ مپ میشوند

در نهایت برای مقادیر Enum نیز از تکنیک one hot encode استفاده میکنیم.

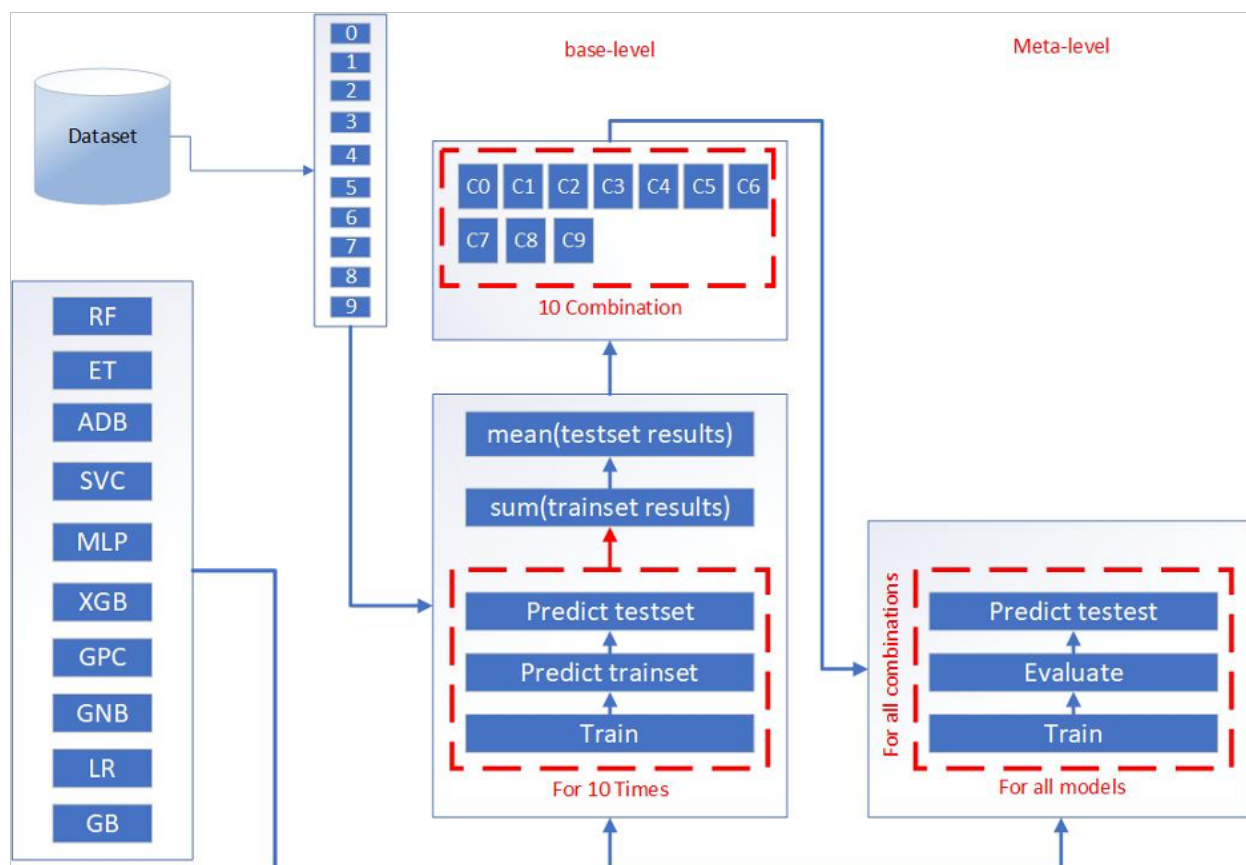
سپس با چسباندن بخش های جدا شده به یکدیگر، به دیتابیس واحدی میرسیم که قابل آرایه به مدل های تعریف شده میباشد.

انتخاب ویژگی ها

انتخاب ویژگی ها در برخورد با ویژگی های اضافی اهمیت زیادی دارد. سه معیار رایج انتخاب ویژگی شامل فیلتر، لفاف و تعبیه شده است.

روش های فیلتر، رابطه بین ویژگی ها و برچسب را با استفاده از ابزارهای آماری شامل واریانس، اطلاعات متقابل و آزمون مجذور کای (χ^2) محاسبه می کنند. روش های wrapper ارتباط نزدیکی با طبقه بندی کننده دارند. اصل روش wrapper انتخاب بهترین زیرمجموعه با توجه به عملکرد طبقه بندی کننده است.

علاوه بر این، حذف ویژگی بازگشتی با اعتبارسنجی متقاطع (RFECV) می تواند تأثیر تنظیم مصنوعی تعداد ویژگی های باقی مانده در مجموعه ویژگی را از بین ببرد. روش های تعبیه شده با فرآیند آموزش مدل ادغام می شوند تا ویژگی ها به طور خودکار انتخاب شوند. افزایش گرادیان شدید (XGB) به دلیل کارایی بالا به طور گسترده ای به عنوان یک روش انتخاب ویژگی تعبیه شده استفاده شده است.



مدل پیشنهادی عمدتاً از دو سطح تشکیل شده است که در آن سطح ۱ سطح پایه و سطح ۲ متا سطح است. پیش‌بینی طبقه‌بندی‌کننده‌های سطح پایه به عنوان ورودی فراسطح انتخاب می‌شوند. سطح پایه شامل ۱۰ مدل از scikit-learn، از جمله جنگل تصادفی (RF)، درختان اضافی (ET)، SVC، adaBoost (ADB)، پرسپترون چند لایه (MLP)، XGB، طبقه‌بندی فرآیند گاوسی (GPC)، گاوسی نیو بیز ساده (GNB)، رگرسیون لجستیک (LR)، تقویت گرادیان (GB) عملکرد طرح‌های انباشتگی تحت تأثیر تعداد طبقه‌بندی‌کننده‌های سطح پایه است [۳۷]. به طور کلی، طبقه‌بندی‌کننده‌های سطح پایه با پیش‌بینی‌های همبستگی ضعیف عملکرد خوبی دارند و MIC را می‌توان به عنوان معیاری برای تعیین کمیت ارتباط و افزونگی در بین ویژگی‌ها استفاده کرد، با نزدیک تر بودن به ۰ نشان دهنده همبستگی ضعیف تر است. سپس، از الگوریتم شمارش برای جستجوی بهترین طبقه‌بندی‌کننده‌های ترکیبی استفاده می‌کنیم.

ما دو الگوریتم را خلاصه می‌کنیم که می‌توانند فرآیند انباشته‌سازی و شمارش را نشان دهند. مجموعه داده ابتدا به صورت تصادفی مخلوط شده و به ۱۰ برابر تقسیم می‌شود. برای هر فولد، یک فولد به عنوان داده آزمایشی (S) و چین‌های باقیمانده به صورت R در نظر گرفته می‌شود. کل فرآیند ۱۰ بار تکرار می‌شود R و S ورودی $Alg.1$ هستند $Alg.1$. عمدتاً شامل دو حلقه است. حلقه اول فرآیند ساخت ده مدل سطح پایه است و حلقه دوم فرآیند اعتبارسنجی متقاطع ۱۰ برابری برای تولید داده‌های آموزشی و آزمایشی است R . نیز به ۱۰ تا تقسیم می‌شوند. یک فولد به عنوان مجموعه اعتبارسنجی (Rkv) و تاهای باقیمانده به عنوان مجموعه داده‌های آموزشی (Rkt) در نظر گرفته می‌شود Rkt . وارد مدل سطح پایه مورد استفاده برای آموزش مدل سطح پایه (ξ_l) می‌شود. Rkv برای تولید قطار استفاده می‌شود. بعداً S برای تولید $testl$ وارد مدل سطح پایه ξ_l می‌شود. از آنجایی که حلقه ۱۰ بار تکرار می‌شود، $Trainl$ دقیقاً برابر با مجموع ده برابر است و مجموعه داده‌های تست باید میانگین شود. در نهایت، مجموعه آموزش و آزمون اتحادیه تولید شده توسط ۱۰ مدل پایه مختلف به عنوان خروجی گرفته می‌شود.

Algorithm 1 The Process of Building Base-Level Model

Input: R , nine folds
Input: S , one fold(test data set)
 R_{kt} , training data set
 R_{kv} , validation data set
the model of base-level ξ_l , where
 $l = \{RF, ET, ADB, SVC, MLP, XGB, GPC, GNB, LR, GB\}$
//the model which can chose
forall the
 $l = \{RF, ET, ADB, SVC, MLP, XGB, GPC, GNB, LR, GB\}$ **do**
 forall the $k = 1, 2, \dots, 10$ **do**
 $\xi_l \leftarrow R_{kt}$
 //use R_{kt} to train ξ_l
 $train_l \leftarrow \xi_l \leftarrow R_{kv}$
 //use R_{kv} to get $train_l$
 $test_l \leftarrow \xi_l \leftarrow S$
 //input S to ξ_l to predict $test_l$
 end
 $train_l = (train_1 + train_2 + \dots + train_k)$
 $test_l = (test_1 + test_2 + \dots + test_k)/10$
 //calculate the mean of test data set
end
 $train = [train_{RF}, train_{ET}, \dots, train_{GB}]$
 $test = [test_{RF}, test_{ET}, \dots, test_{GB}]$
Output: $train, test$

خروجی ۱ Alg. به عنوان ویژگی های جدید سطح متا در نظر گرفته می شود. از آنجایی که استفاده مستقیم از همه ویژگی های جدید بدون فیلتر کردن عاقلانه نیست، Alg. ۲ برای جستجوی ترکیب بهینه استفاده می شود. Alg. ۲ عمدتاً شامل دو حلقه است. در حلقه اول، ۱۰ نوع ترکیب ممکن وجود دارد، از جمله C_{110} ، C_{210} ، C_{310} ، C_{410} ، C_{510} ، C_{610} ، C_{710} ، C_{810} ، C_{910} ، C_{1010} به عنوان ورودی حلقه دوم. همه ترکیبات ممکن (بدون تکرار آنها) به جای قرار دادن همه آنها در حلقه بعدی، شمارش می شوند. در حلقه دوم از ورودی قطار برای آموزش مدل Hm استفاده می شود. مدل LR برای کاهش پیچیدگی مدل استفاده می شود.

Algorithm 2 The Process of Searching the Best Combination

Input: $train$, training data set

Input: $test$, test data set

// H_m , the second model

forall the

$l = \{RF, ET, ADB, SVC, MLP, XGB, GPC, GNB, LR, GB\}$ **do**

forall the $m = \{LR\}$ **do**

$train \leftarrow C_{10}^l$

$test \leftarrow C_{10}^l$

$H_m \leftarrow train$

$result_m \leftarrow H_m \leftarrow test$

 //evaluate the performance of the model on the test
 data set

end

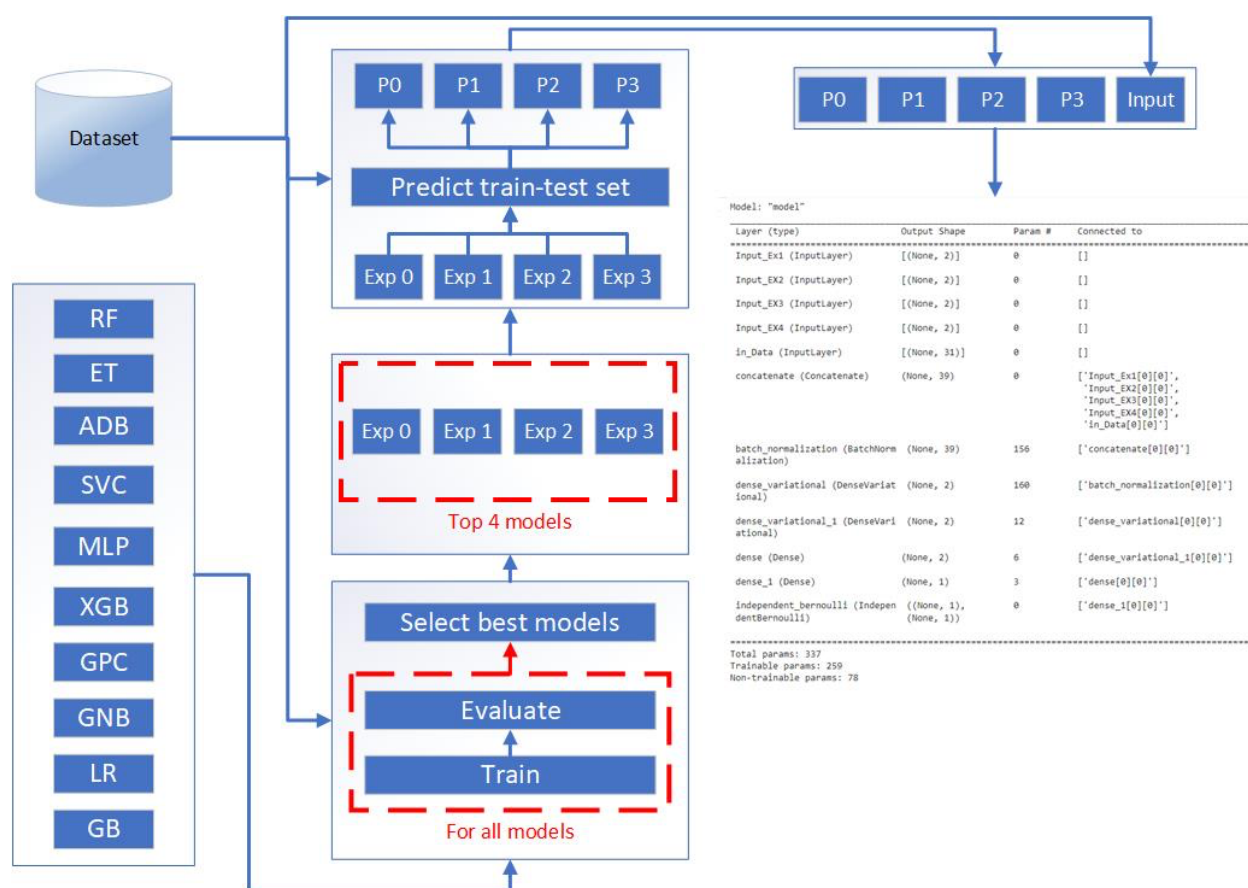
end

Output: $result_m$

سپس آزمون به مدل آموزش دیده (H_m) وارد می شود تا عملکرد مدل در مجموعه داده های آزمون ارزیابی شود. در نهایت ترکیب مدل با بالاترین دقت تعیین می شود.

مدل پیشنهادی

مدل پیشنهادی ما بر اساس انتخاب ۴ مورد از بهترین مدل های مورد بررسی در مقاله مورد بحث عمل میکند، به طوری که تمام مدل های موجود ابتدا آموزش دیده و سپس مورد ارزیابی قرار میگیرند و سپس بر اساس نتایج حاصل شده مدل های برتر انتخاب میشوند. دلیل استفاده ما از این تکنیک ، به دست آوردن نتایج بهینه تر بود و با توجه به بررسی متدولوژی مورد بحث در مقاله این نیازمندی را احساس کردیم که به جای انتخاب بهترین ترکیب پیشبینی ها و نهایتاً بهترین مدل سازگار با ترکیب ها ، از چند مدل استفاده کنیم و بتوانیم از پتانسیل فرضیه چندین مدل برای تصمیم گیری استفاده کنیم. استفاده از این روش به ما این امکان را میدهد که فضای درست حقیقی گسترده تری را پوشش دهیم و نهایتاً به پیشبینی های دقیق تری دست یابیم.



همانطور که در شکل ۲ مشاهده میکنید، مدل های موجود پس از آموزش، ارزیابی میشوند و سپس ۴ مورد از بهترین مدل ها برای استفاده در متدولوژی پیشنهادی ما انتخاب میشوند. مدل های آموزش دیده داده های ورودی آموزش و ارزیابی را پیشبینی میکنند و خروجی احتمال بین دو

کلاس موجود در مسئله به همراه داده های ورودی برای آموزش مدل Bayesian neural network (BNN) استفاده میشود. در مدل پیشنهادی از لایه های Dense variational استفاده شده است که از یک استنباط بیز برای محاسبه توزیع بر روی ماتریس هسته و بایاس استفاده میشود که توابع پیشین و پسین محاسبه این مقادیر را امکان پذیر میکند. در ادامه خروجی لایه های Dense variational پس از پردازش در دو لایه Dense نهایتاً به لایه Independent Bernoulli انتقال داده شده و نتیجه نهایی حاصل میشود.

مقایسه نتایج پیش‌پردازش داده‌ها

پس از پیش‌پردازش نشان داده شده در بخش دیتا، داده‌های خام استاندارد شده و "داده‌های پردازش شده" دارای محدوده ۰ تا ۱ هستند.

همانطور که در جدول ۲ نشان داده شده است، نتایج "داده‌های پردازش شده" عملکرد بهتری نسبت به "داده‌های خام" نشان می‌دهد. از طریق طبقه‌بندی کننده‌های مختلف برای «داده‌های پردازش شده»، مدل‌های LR و ADB دقت بالاتری نسبت به سایرین دارند. اما XGB و GNB به ترتیب از نظر حساسیت و ویژگی امتیازات بهتری نسبت به سایر مدل‌ها کسب می‌کنند. بنابراین، این تفاوت‌ها توسط ناهمگونی مدل ایجاد می‌شود که منعکس کننده پایه و اساس انباشته است. این نتایج با مدل پیشنهادی نیز همخوانی داشته و مدل‌های LR و RF دقت بالاتری نسبت به سایر مدل‌ها داشته‌اند و از نقطه نظر GNB امتیاز بالاتری را کسب کرده است.

نتایج جدول ۲ در صفحه‌ی بعدی با درج اختلافات نمایش داده شده است.

Table2 - paper results	Method	Raw data				Processed data		
		Ac	Se	Sp		Ac	Se	Sp
	LR	0.854	0.893	0.76		0.881	0.917	0.792
	RF	0.842	0.917	0.657		0.839	0.894	0.706
	GNB	0.734	0.679	0.872		0.428	0.217	0.953
	SVC	0.814	0.888	0.633		0.845	0.884	0.749
	DT	0.813	0.87	0.669		0.786	0.851	0.622
	KNN	0.681	0.875	0.197		0.805	0.879	0.622
	ADB	0.874	0.93	0.733		0.882	0.927	0.769
	GB	0.859	0.921	0.704		0.862	0.907	0.749
	ET	0.825	0.875	0.704		0.852	0.898	0.736
	MLP	0.547	0.617	0.357		0.852	0.898	0.735
	XGB	0.872	0.935	0.717		0.862	0.935	0.681
Table2 - our implementation results	Method	Raw data				Processed data		
		Ac	Se	Sp		Ac	Se	Sp
	LR	0.81	0.61	0.73		0.82	0.74	0.72
	RF	0.85	0.8	0.59		0.84	0.77	0.59
	GNB	0.77	0.57	0.92		0.57	0.38	0.92
	SVC	0.8	0.87	0.6		0.82	0.834	0.719
	DT	0.79	0.83	0.61		0.71	0.82	0.6
	KNN	0.69	0.86	0.18		0.895	0.81	0.612
	ADB	0.78	0.55	0.48		0.78	0.55	0.48
	GB	0.76	0.52	0.43		0.75	0.52	0.43
	ET	0.82	0.79	0.6		0.82	0.79	0.62
	MLP	0.62	0.06	0.3		0.76	0.79	0.73
	XGB	0.72	0.45	0.53		0.72	0.45	0.53
Table2 - results difference	Method	Raw data				Processed data		
		Ac	Se	Sp		Ac	Se	Sp
	LR	0.044	0.283	0.03		0.061	0.177	0.072
	RF	-0.008	0.117	0.067		-0.001	0.124	0.116
	GNB	-0.036	0.109	-0.048		-0.142	-0.163	0.033
	SVC	0.014	0.018	0.033		0.025	0.05	0.03
	DT	0.023	0.04	0.059		0.076	0.031	0.022
	KNN	-0.009	0.015	0.017		-0.09	0.069	0.01
	ADB	0.094	0.38	0.253		0.102	0.377	0.289
	GB	0.099	0.401	0.274		0.112	0.387	0.319
	ET	0.005	0.085	0.104		0.032	0.108	0.116
	MLP	-0.073	0.557	0.057		0.092	0.108	0.005
	XGB	0.152	0.485	0.187		0.142	0.485	0.151

مقایسه نتایج انتخاب ویژگی ها

جدول ۳ ویژگی های انتخاب شده توسط سه روش انتخاب ویژگی معمولی را نشان می دهد. ویژگی های انتخاب شده به پزشکان کمک می کند تا درک خود را از اهمیت متفاوت ویژگی های انتخاب شده بهبود بخشند. علاوه بر این، مداخلات درمانی مختلفی را می توان به طور خاص برای کاهش یا حتی حذف تأثیر مضر برخی از ویژگی های انتخاب شده انجام داد.

جدول شماره ۳:

Table3 - paper results		
CHI2	RFECV	XGB
Age	Age	Age
DM	DM	DM
HTN	HTN	HTN
CRE	Sex	EH
Diastolic Murmur	Typical Chest Pain	Atypical
Typical Chest Pain	Nonanginal	Typical Chest Pain
Dyspnea	O Wave	Lymph
Function Class	St Elevation	Nonanginal
Atypical	T inversion	St Depression
Nonanginal	HB	TG
O Wave	TG	VHD
St Elevation	Poor R Progression	PR
St Depression	EF-TTE	T inversion
T inversion	Current Smoker	FBS
Poor R Progression	DLP	ESR
FBS	Lung rales	EF-TTE
Airway disease	Dyspnea	Na
Weak Peripheral Pulse	PR	BMI
BP	Region RWMA	LDI.
Region RWMA		Region RWMA
		WBC
		K
		LDL
		HB
		Neut
		HDI.
		Length
		Weight
		CR
		PLT
		BP
		Obesity

Table3 - our implementation results		
CHI2	RFECV	XGB
Region RWMA	BMI	Typical Chest Pain
DM	PR	Age
HTN	Age	Region RWMA
Typical Chest Pain	TG	HTN
Atypical	HDL	Nonanginal
Nonanginal	BUN	Tinversion
Tinversion	ESR	TG
VHD_Severe	HB	PR
	EF-TTE	DM
	Region RWMA	WBC
	DM	K
	HTN	LDL
	Current Smoker	Na
	FH	Lymph
	DLP	BUN
	Lung rales	Length
	Typical Chest Pain	BMI
	Dyspnea	FBS
	Function Class	ESR
	Nonanginal	EF-TTE
	Q Wave	Neut
	St Elevation	CR
	Tinversion	HDL
		PLT

جدول ۴: دقت تشخیص CHD را برای روش های مختلف انتخاب ویژگی از جمله CHI2 ، اطلاعات متقابل، واریانس، RFE، SVC و LR با مقادیر k متفاوت نشان می دهد. الگوریتم طبقه بندی از SVC استفاده می کند. دقت ابتدا افزایش می یابد و سپس با افزایش مقدار k کاهش می یابد. مقادیر $k = 15, 17, 20$ و 22 زمانی برجسته می شوند که دقت مدل بالاتر از ۹۰٪ باشد. جدول ۴ نشان می دهد که وقتی $k = 15$ دقت ۹۱.۱٪ توسط LR به دست می آید.

دقت تشخیص CHD برای روش های مختلف انتخاب ویژگی با مقادیر k مختلف.

Table4 - paper results	K value	Feature selection methods					
		CHI2	Mutual	Variance	RFE	SVC	LR
	5	0.835	0.852	0.819	0.838	0.848	0.848
	10	0.852	0.825	0.838	0.888	0.881	0.858
	15	0.851	0.825	0.852	0.898	0.891	0.911
	20	0.868	0.865	0.852	0.904	0.888	0.878
	25	0.868	0.871	0.852	0.898	0.895	0.871
	30	0.858	0.875	0.881	0.898	0.897	0.898
	35	0.851	0.861	0.868	0.891	0.898	0.882
	40	0.848	0.838	0.868	0.891	0.881	0.875
	45	0.852	0.851	0.872	0.884	0.884	0.878
	50	0.865	0.865	0.855	0.855	0.855	0.861
	17	0.865	0.858	0.839	0.908	0.891	0.888
	22	0.861	0.845	0.865	0.901	0.898	0.868
Table4 - our implementation results	K value	Feature selection methods					
		CHI2	Mutual	Variance	RFE	SVC	LR
	5	0.82	0.78	0.78	0.86	0.78	0.82
	10	0.83	0.8	0.79	0.87	0.75	0.85
	15	0.82	0.76	0.76	0.86	0.73	0.84
	20	0.81	0.77	0.76	0.87	0.73	0.86
	25	0.81	0.76	0.74	0.87	0.72	0.86
	30	0.81	0.76	0.76	0.86	0.72	0.85
	35	0.81	0.77	0.75	0.87	0.71	0.86
	40	0.81	0.77	0.76	0.87	0.72	0.86
	45	0.8	0.76	0.74	0.84	0.71	0.85
	50	0.81	0.76	0.74	0.85	0.71	0.86
	17	0.81	0.77	0.76	0.87	0.73	0.86
	22	0.81	0.76	0.77	0.87	0.73	0.86
Table4 - results difference	K value	Feature selection methods					
		CHI2	Mutual	Variance	RFE	SVC	LR
	5	0.015	0.072	0.039	-0.022	0.068	0.028
	10	0.022	0.025	0.048	0.018	0.131	0.008
	15	0.031	0.065	0.092	0.038	0.161	0.071
	20	0.058	0.095	0.092	0.034	0.158	0.018
	25	0.058	0.111	0.112	0.028	0.175	0.011
	30	0.048	0.115	0.121	0.038	0.177	0.048
	35	0.041	0.091	0.118	0.021	0.188	0.022
	40	0.038	0.068	0.108	0.021	0.161	0.015
	45	0.052	0.091	0.132	0.044	0.174	0.028
	50	0.055	0.105	0.115	0.005	0.145	0.001
	17	0.055	0.088	0.079	0.038	0.161	0.028
	22	0.051	0.085	0.095	0.031	0.168	0.008

همانطور که در جدول ۵ نشان داده شده است، نتایج دو روش انتخاب ویژگی نماینده شامل رویکردهای پوششی و تعبیه شده با هم مقایسه شده است. جدول ۵ نشان می دهد که بهترین عملکرد توسط RFECV به دست می آید. و نتایج RFECV انحراف معیار کمتری دارند. بنابراین، RFECV به عنوان روش انتخاب ویژگی ما تصمیم گرفته می شود.

جدول پنجم

اجرای روش های مختلف تست با روش انتخاب ویژگی RFECV و XGB.

Table5 - paper results	Method	RFECV				XGB		
		Ac	Se	Sp		Ac	Se	Sp
	LR	0.901	0.93	0.826		0.868	0.921	0.736
	RF	0.832	0.862	0.76		0.849	0.898	0.725
	GNB	0.44	0.22	0.989		0.861	0.916	0.724
	SVC	0.911	0.926	0.875		0.872	0.907	0.783
	DT	0.799	0.865	0.633		0.789	0.856	0.622
	KNN	0.855	0.87	0.814		0.855	0.907	0.728
	ADB	0.881	0.926	0.768		0.859	0.903	0.747
	GB	0.876	0.922	0.761		0.859	0.917	0.715
	ET	0.829	0.871	0.726		0.833	0.875	0.729
	MLP	0.908	0.945	0.815		0.868	0.912	0.76
	XGB	0.868	0.944	0.681		0.865	0.93	0.703
Table5 - our implementation results	Method	RFECV				XGB		
		Ac	Se	Sp		Ac	Se	Sp
	LR	0.86	0.71	0.77		0.85	0.71	0.64
	RF	0.83	0.69	0.65		0.8	0.55	0.47
	GNB	0.49	0.34	0.97		0.82	0.6	0.36
	SVC	0.9	0.92	0.8		0.832	0.927	0.713
	DT	0.809	0.83	0.623		0.84	0.826	0.61
	KNN	0.825	0.8	0.811		0.82	0.9	0.71
	ADB	0.79	0.7	0.62		0.76	0.52	0.63
	GB	0.78	0.61	0.53		0.78	0.55	0.54
	ET	0.86	0.8	0.72		0.84	0.9	0.69
	MLP	0.88	0.8	0.71		0.88	0.84	0.74
	XGB	0.77	0.47	0.57		0.71	0.41	0.64
Table5 - results difference	Method	RFECV				XGB		
		Ac	Se	Sp		Ac	Se	Sp
	LR	0.041	0.22	0.056		0.018	0.211	0.096
	RF	0.002	0.172	0.11		0.049	0.348	0.255
	GNB	-0.05	-0.12	0.019		0.041	0.316	0.364
	SVC	0.011	0.006	0.075		0.04	-0.02	0.07
	DT	-0.01	0.035	0.01		-0.051	0.03	0.012
	KNN	0.03	0.07	0.003		0.035	0.007	0.018
	ADB	0.091	0.226	0.148		0.099	0.383	0.117
	GB	0.096	0.312	0.231		0.079	0.367	0.175
	ET	-0.031	0.071	0.006		-0.007	-0.025	0.039
	MLP	0.028	0.145	0.105		-0.012	0.072	0.02
	XGB	0.098	0.474	0.111		0.155	0.52	0.063

نتایج برای روش پیشنهادی و روش‌های دیگر

داده‌ها به ترتیب به مجموعه داده‌های آموزشی و مجموعه داده‌های آزمایشی با نسبت ۷:۳ و ۴:۶ تقسیم می‌شوند. سپس، مجموعه داده‌های آموزشی برای آموزش مدل و محاسبه PCC و MIC بین هر مدل به صورت جفتی استفاده می‌شود.

PCC دو نسبت متفاوت از یک داده را نشان می‌دهد. MIC دو نسبت متفاوت از یک داده را نشان می‌دهد. در نسبت‌های مختلف، PCC و MIC GNB همیشه حداقل مقادیر را دریافت می‌کنند. بنابراین مدل GNB را می‌توان به عنوان یکی از بهترین طبقه‌بندی‌کننده‌های ترکیبی انتخاب کرد. هفت مدل بهینه (GB, GNB, RF, ET, ADB, MLP, XGB) به عنوان مدل سطح پایه توسط الگوریتم ۱ و الگوریتم ۲ انتخاب شده است. جدول ۶ نتایج مقایسه‌ای روش ما را نشان می‌دهد و سایر روش‌های مختلف شامل روشی است که توسط ناشر مجموعه داده پیشنهاد شده است. نشان داده شده است که مدل مبتنی بر انباشتگی پیشنهادی ما تقریباً در تمام معیارهای نتایج پیشرفت‌های قابل توجهی را به دست می‌آورد. روش ما به دقت، حساسیت، ویژگی و امتیاز اف ۱ ۹۵.۴۳٪، ۹۵.۸۴٪، ۹۴.۴۴٪، ۹۶.۷۷٪ برای تشخیص CHD دست می‌یابد. پارامترهای مدلی که استفاده می‌کنیم در جدول ۷ آورده شده است.

جدول ششم

مقایسه عملکرد روش‌های مختلف تست:

Table6 - paper results	Method	Ac	Se	Sp	F1 - score
	LR	0.901	0.93	0.826	0.931
	RF	0.832	0.862	0.76	0.879
	GNB	0.44	0.22	0.989	0.336
	SVC	0.911	0.926	0.875	0.937
	DT	0.799	0.865	0.633	0.857
	KNN	0.855	0.87	0.814	0.895
	ADB	0.881	0.926	0.768	0.918
	GB	0.876	0.922	0.761	0.914
	ET	0.829	0.871	0.726	0.897
	MLP	0.908	0.945	0.815	0.936
	XGB	0.868	0.944	0.681	0.912
Table6 - our implementation results	Method	Ac	Se	Sp	F1 - score
	LR	0.86	0.71	0.77	0.69
	RF	0.83	0.69	0.65	0.63
	GNB	0.49	0.34	0.97	0.48
	SVC	0.9	0.92	0.8	0.931
	DT	0.809	0.83	0.623	0.82
	KNN	0.825	0.8	0.811	0.825
	ADB	0.79	0.7	0.62	0.57
	GB	0.78	0.61	0.53	0.52
	ET	0.86	0.8	0.72	0.68
	MLP	0.88	0.8	0.71	0.47
	XGB	0.77	0.47	0.57	0.69
Table6 - results difference	Method	Ac	Se	Sp	F1 - score
	LR	0.041	0.22	0.056	0.241
	RF	0.002	0.172	0.11	0.249
	GNB	-0.05	-0.12	0.019	-0.144
	SVC	0.011	0.006	0.075	0.006
	DT	-0.01	0.035	0.01	0.037
	KNN	0.03	0.07	0.003	0.07
	ADB	0.091	0.226	0.148	0.348
	GB	0.096	0.312	0.231	0.394
	ET	-0.031	0.071	0.006	0.217
	MLP	0.028	0.145	0.105	0.466
	XGB	0.098	0.474	0.111	0.222

پارامتر مدل ها

در جدول زیر پارامتر های ورودی برای هر یک از مدل های معرفی شده در مقاله با جزییات کافی برای پیاده سازی ذکر شده است.

برای پیاده سازی این مدل ها از کتابخانه **scikit-learn** استفاده شده است.

جدول ۷:

پارامتر مدل ها

Method	Model parameters
LR	C= 100, solver= newtoncg
RF	n_estimators x= 10, max_depth = None, min_samples_split = 2, random_state = 0
GNB	priors = None, var smoothing = 1e - 09
ADB	n_estimators = 50, learning_rate = 1.0, algorithm = SAMME.R
GB	loss = deviance, learning_rate = 0.1, n_estimators = 100
ET	n_estimators = 100, criterion = gini, min_samples_split = 2
MLP	hidden_layer_sizes = (100,), activation = relu, solver = adam, alpha = 0.0001
XGB	random_state = 1, learning rate = 0.05, n_estimators = 7, max depth = 5 eta = 0.05, objective = binary : logistic

کد

تمامی کد ها در مخزن https://github.com/hamidrezahy/AAI_Project در دسترس میباشد. برای اجرای کد ها نیاز به نصب کتابخانه `numpy`, `scikit-learn` و سایر نیازمندی های درج شده در مخزن میباشد.

در سطح پوشه بندی، پروژه دارای ۳ پوشه اساسی میباشد. پوشه اول کانفیگ که برای نگهداری پارامتر ها و تنظیم سریعتر و ساده تر مدل ها ایجاد شده است، پوشه دوم پوشه دیتا که دیتا ست اصلی در آنجا ذخیره شده و علاوه بر آن دیتا ست پیش پردازش شده نیز در آنجا ذخیره میشود. پوشه بعدی پوشه اکسپریمنت میباشد که نتایج در قالب فابل های اکسل پس از تولید در این پوشه نگهداری میشوند.

پوشه آخر پوشه سورس کد ها میباشد که تمامی سورس کد های مورد نیاز در این فولدر قابل دسترسی میباشد.

برای شروع سریعتر و سهولت انجام کار، نوت بومی تحت عنوان `all_fuction_run` تهیه شده است که با باز کردن و اجرای سلول های آن میتوان الگوریتم ها را اجرا کرده و خروجی ها را دریافت کرد.

عملکرد این نوت بوک به شرح زیر است: