پیش بینی بیماری دیابت با استفاده از روش SMOTEENN و الگوریتم جنگل تصادفی

علی کاویانی
دانشجوی کارشناسی ارشد هوش مصنوعی و رباتیک دانشگاه جامع امام حسین (ع)
محمدعلی جوادزاده
استادیار دانشگاه جامع امام حسین (ع) دانشکده و پژوهشکده هوش مصنوعی و علوم شناختی

چکیده
در دنیای پزشکی بیماری دیابت، یکی از بزرگترین دغدغه پزشکان و از چالش های سلامتی بشر بوده و هست. پیش بینی به هنگام و دقیق این بیماری می‌تواند از بسیاری از معضلات سلامتی که ممکن است در آینده روی سلامتی تاثیر بگذارد جلوگیری کند. در این پژوهش به کمک هوش مصنوعی تشخیص بیماری دیابت را به سهولت و دقت بیشتری رساندیم که علارقم تلاش های بسیار در این حوزه و این مجموعه داده در عدد ۹۲ درصد بود. ما با استفاده از روش ترکیبی Smoteenn جهت مقابله با عدم تعادل کلاس ها، همچنین، به‌منظور کاهش پیچیدگی مدل و انتخاب مهم‌ترین ویژگی‌ها، الگوریتم حذف بازگشتی ویژگی (RFE) با استفاده از جنگل تصادفی داده ها را استاندارد سازی کردیم. پس از آن، داده‌ها نرمال‌سازی و مدل جنگل تصادفی با بهره‌گیری از روش جستجوی شبکه‌ای (GridSearchCV) برای بهینه‌سازی هایپرپارامترها آموزش داده شد. در نهایت مدل پیشنهادی ما توانست با دقت ۹۷ درصدی بیماری دیابت را پیش بینی و تشخیص دهد. نکته قابل توجه در این پژوهش علاوه بر دستیابی به دقت بالا، به اهمیت پیش پردازش داده ها و انتخاب ویژگی مناسب در بهبود کارایی مدل های یادگیری ماشین اشاره دارد.

واژگان کلیدی: پیش بینی دیابت ، یادگیری ماشین ، جنگل تصادفی ، پیش پردازش داده ها

مقدمه

تشخیص زودهنگام بیماری‌ها، به‌ویژه بیماری‌های مزمن و خاموش مانند دیابت، نقشی کلیدی در بهبود کیفیت زندگی بیماران و کاهش هزینه‌های درمانی دارد. دیابت، که از جمله بیماری‌های متابولیکی جدی است، در صورت عدم تشخیص و مدیریت به‌موقع می‌تواند به ارگان‌های مختلف بدن آسیب‌های جبران‌ناپذیری وارد کند. با وجود پیشرفت‌های پزشکی، تشخیص سنتی دیابت همچنان با چالش‌هایی همچون عدم دقت کافی، زمان‌بر بودن فرایندها و وابستگی به تجربه‌ی پزشکان روبروست.

در این راستا،یادگیری ماشین به‌عنوان ابزاری قدرتمند برای تحلیل داده‌های پزشکی ظهور کرده‌اند. یادگیری ماشین با بهره‌گیری از الگوریتم‌های پیشرفته و تجزیه‌وتحلیل داده‌های پزشکی، می‌تواند دقت و سرعت در تشخیص بیماری‌ها را به طرز قابل‌توجهی بهبود بخشد (نصراله پور و خطیبی، ۱۴۰۲) .

در این تحقیق، هدف ارتقا سلامت در جامعه است که می‌توانیم با پیش بینی دقیق و به هنگام به آن دست یابیم. ما در این مقاله پیش‌بینی و تشخیص بیماری دیابت با استفاده از تکنیک‌های پیشرفته‌ی یادگیری ماشین و مجموعه داده‌ی معروف *Diabetes Dataset* مسئله تشخیص و پیش بینی را انجام خواهیم داد. این مجموعه داده شامل ویژگی‌هایی از جمله تعداد بارداری، غلظت گلوکز پلاسما، فشار خون، ضخامت چین پوستی، میزان انسولین، شاخص توده بدنی، سابقه خانوادگی دیابت، سن بیمار و متغیر کلاس (صفر یا یک) است که نمایانگر عدم وجود یا وجود دیابت در بیماران است.

برای بهبود دقت و کارایی مدل، از ترکیب تکنیک‌های SMOTEENN برای مدیریت عدم توازن داده‌ها و الگوریتم جنگل تصادفی برای طبقه‌بندی استفاده شده است. الگوریتمSMOTEENN ، با ترکیب روش‌های Oversampling و حذف داده‌های پرنویز، توازنی مؤثر بین نمونه‌های مثبت و منفی برقرار می‌کند. الگوریتم جنگل تصادفی نیز به دلیل توانایی بالا در پردازش داده‌های پیچیده و انتخاب ویژگی‌های مهم، یکی از بهترین گزینه‌ها برای این نوع مسائل است.

در ادامه مقاله، ابتدا چالش‌های موجود در داده‌های پزشکی و روش‌های مقابله با آن‌ها بررسی می‌شود. سپس، مراحل پیش‌پردازش داده‌ها، انتخاب ویژگی‌ها، و تنظیمات مدل شرح داده شده و در نهایت نتایج حاصل از ارزیابی مدل ارائه خواهد شد.

پیشینه تحقیق:

در حالی که تلاش‌های پیشین مانند پروژه‌های (Lara, 2024) و (Bin Helal, 2024) توانسته‌اند به نتایجی دست یابند، هنوز شکاف‌های قابل‌توجهی در این حوزه وجود دارد. به عنوان مثال، پروژه Lara چند الگوریتم را مورد بررسی قرار داده است. با استفاده از الگوریتم درخت تصمیم به دقت ۷۴.۶ درصد ، رگرسیون لجستیک ۷۵ درصد ، svm ۷۲ درصد ، Navie Bayes ۷۶.۶ و در جنگل تصادفی به ۷۲ درصد دقت رسیده است. از سوی دیگر، پروژه Bin Helal با به‌کارگیری الگوریتم جنگل تصادفی موفق به ثبت دقت ۹۲.۴ درصدی شده، اما ضعف‌های موجود در مرحله پیش‌پردازش داده‌ها، ارزیابی جامع‌تری از عملکرد مدل را دشوار کرده است.

این شکاف‌ها بر اهمیت استفاده از رویکردهای پیشرفته‌تر تأکید می‌کند. در این مطالعه، با بهره‌گیری از تکنیک SMOTEENN برای ترکیب نمونه‌سازی افزایشی و حذف نویز، سعی بر بهبود توزیع داده‌های آموزشی شده است. همچنین، الگوریتم جنگل تصادفی به دلیل توانایی بالای آن در مدیریت داده‌های پیچیده و انتخاب ویژگی‌های مهم، به عنوان مدل پیش‌بینی استفاده شده است. این ترکیب نوآورانه نه تنها بر چالش‌های پیشین غلبه می‌کند، بلکه انتظار می‌رود که معیارهای ارزیابی دقیق‌تری نیز ارائه دهد.

روش تحقیق

این مطالعه با استفاده از مجموعه داده دیابت و پیش‌پردازش دقیق شامل حذف مقادیر پرت و غیرواقعی، نرمال‌سازی داده‌ها و متعادل‌سازی کلاس‌ها با روشSMOTEENN ، تلاش کرده است عملکرد مدل پیش‌بینی دیابت را بهبود بخشد. الگوریتم جنگل تصادفی با تنظیم هایپرپارامترها به‌عنوان مدل اصلی انتخاب شد و توانست دقت بالایی (۹۶.۶٪) در پیش‌بینی دیابت ارائه دهد. تحلیل ویژگی‌ها نشان داد که گلوکز، BMI، سن، فشار خون و تعداد دفعات بارداری مهم‌ترین عوامل مؤثر در پیش‌بینی دیابت هستند. این مطالعه ترکیبی از تکنیک‌های یادگیری ماشین را برای ارائه یک مدل قوی و دقیق ارائه کرده است.

1- مجموعه داده و پیش‌پردازش داده‌ها

مطالعه حاضر از مجموعه داده دیابت (PIMA Diabetes Dataset) که از منبع UCI Machine Learning Repository استخراج شده است، استفاده می‌کند. این مجموعه شامل ۷۶۸ نمونه و ۹ ویژگی است. ویژگی‌های موجود شامل موارد زیر می‌باشند و میتوانید توزیع این ویژگی ها را در نمودار شماره ۲ مشاهده فرمائید:

تعداد دفعات بارداری‌ها (Pregnancies)

قند خون ناشتا (Glucose)

فشار خون (Blood Pressure)

ضخامت پوست (Skin Thickness)

سطح انسولین (Insulin)

شاخص توده بدنی (BMI)

سابقه خانوادگی دیابت

سن بیمار

وضعیت دیابت (Outcome): مقدار ۱ نشان‌دهنده وجود دیابت و مقدار ۰ نشان‌دهنده عدم دیابت است.

برای درک بهتر از شرایط داده های مجموعه داده تصویر ۱ heat map همبستگی ویژگی های مجموعه داده را میتوانید مشاهده نمایید.

بررسی اولیه داده‌ها نشان داد که کلاس‌ها نامتعادل هستند و همچنین برخی ویژگی‌ها دارای مقادیر غیرواقعی (مانند مقادیر صفر) بودند. برای بهبود کیفیت داده‌ها، اقدامات زیر انجام شد:

حذف داده‌های پرت (Outliers): با استفاده از معیار Z-Score، داده‌هایی که مقدار Z آن‌ها فراتر از ۲± بود، حذف شدند. این روش بر اساس استانداردسازی ویژگی‌ها انجام شد (Hodge & Austin, 2004).

حذف مقادیر غیرواقعی: ستون‌های گلوکز، فشار خون، ضخامت پوست، انسولین، و BMI بازبینی شدند و مقادیر صفر که غیرواقعی تلقی می‌شدند، حذف گردیدند.

نرمال‌سازی ویژگی‌ها: برای همگن‌سازی داده‌ها و بهبود عملکرد الگوریتم، ویژگی‌ها با استفاده از ابزار StandardScaler نرمال‌سازی شدند.

2- تکنیک SMOTEENN برای متعادل‌سازی داده‌ها

عدم تعادل کلاس‌ها در داده‌های این مطالعه می‌توانست منجر به کاهش دقت مدل در شناسایی کلاس اقلیت (بیماران دیابتی) شود. برای رفع این چالش، از روش SMOTEENN استفاده شد که ترکیبی از دو تکنیک زیر است:

SMOTE (Synthetic Minority Oversampling Technique): نمونه‌های مصنوعی برای کلاس اقلیت تولید کرد تا تعادل کلاس‌ها برقرار شود.

ENN (Edited Nearest Neighbor): نمونه‌های نویزی و نامعتبر را از کلاس اکثریت حذف نمود.

این روش به بهبود تعادل داده‌ها کمک کرد و باعث افزایش توانایی مدل در شناسایی نمونه‌های دیابتی گردید.

3- مدل جنگل تصادفی

الگوریتم جنگل تصادفی (Random Forest) به عنوان مدل اصلی برای طبقه‌بندی داده‌ها انتخاب شد. این الگوریتم به دلیل توانایی در ترکیب چندین درخت تصمیم‌گیری، عملکرد قوی و پایداری در برابر بیش‌برازش ارائه می‌دهد. (G. W. Brier, 2020) از مزایای این الگوریتم می‌توان به موارد زیر اشاره کرد:

- شناسایی خودکار ویژگی‌های مهم.
- مقاومت در برابر نویز و داده‌های نامتعادل.
- انعطاف‌پذیری در تنظیم هایپرپارامترها.
- برای بهینه‌سازی، هایپرپارامترهای زیر تنظیم شدند:

n_estimators: تعداد درختان جنگل، که مقدار بهینه آن با استفاده از جستجوی شبکه‌ای (GridSearchCV) تعیین شد.
max_depth: برای جلوگیری از پیچیدگی بیش از حد و کاهش خطا تنظیم شد.
min_samples_split و min_samples_leaf: حداقل تعداد نمونه‌های موردنیاز برای تقسیم گره‌ها و تشکیل برگ‌ها، بهینه‌سازی شدند.
برای افزایش اعتماد به نتایج، از روش اعتبارسنجی متقابل (K-Fold Cross-Validation) با 5بخش استفاده شد.
شایان ذکر است در مسیر طراحی این مدل ابتدا تعداد زیادی از الگوریتم های حوزه یادگیری ماشین مورد بررسی قرار گرفتند که الگوریتم جنگل تصادفی با توجه به نتیجه قابل قبول نسبت به سایر الگوریتم ها به عنوان الگوریتم پایه انتخاب و عملیاتی شد. نتیجه کاری الگوریتم های مورد بررسی را در جدول شماره ۱ میتوانید مشاهده کنید. در این جدول نتایج اجرای الگوریتم‌های مختلف روی این مجموعه داده نمایش داده شده است.

جمع‌بندی :

بر اساس تحلیل‌ها، مدل جنگل تصادفی با استفاده از تکنیک SMOTEENN عملکرد بهتری در مقایسه با سایر روش‌های یادگیری ماشین ارائه داد. نتایج نشان داد که این ترکیب به طور موثری توانسته است تعادل میان دقت و شناسایی کلاس‌های اقلیت را برقرار کند. همچنین، این مطالعه نشان می‌دهد که استفاده از تکنیک‌های پیش‌پردازش داده‌ها و متعادل‌سازی کلاس‌ها می‌تواند دقت و عملکرد مدل‌های یادگیری ماشین را به طور قابل توجهی بهبود بخشد.

یافته ها

- نتایج پیش‌پردازش داده‌ها

در مرحله پیش‌پردازش، عملیات زیر انجام شد:

- حذف داده‌های پرت: با استفاده از روشZ-Score ، تعداد ۸۵ داده پرت که ممکن بود بر عملکرد مدل تأثیر منفی بگذارند، شناسایی و حذف شدند.
- پاکسازی مقادیر صفر: مقادیر صفر غیرواقعی در ستون‌هایی نظیر گلوکز، فشار خون، ضخامت پوست، انسولین و BMI حذف شدند. این فرآیند منجر به کاهش تعداد نمونه‌های معتبر از ۷۶۸ به ۶۵۲ نمونه شد.
- متعادل‌سازی داده‌ها

به دلیل عدم تعادل در تعداد نمونه‌های دیابتی و غیردیابتی، از روش ترکیبی SMOTEENN استفاده شد. این تکنیک با ایجاد نمونه‌های جدید برای کلاس اقلیت و حذف داده‌های نویزی کلاس اکثریت، باعث تعادل کلاس‌ها شد. نتیجه این کار، بهبود توانایی مدل در پیش‌بینی نمونه‌های هر دو کلاس به طور همزمان بود.

- انتخاب ویژگی‌های کلیدی

با استفاده از روش حذف بازگشتی ویژگی (RFE) و مدل جنگل تصادفی، پنج ویژگی اصلی به عنوان عوامل تأثیرگذار در پیش‌بینی دیابت شناسایی شدند:

۱. گلوکز: بیشترین تأثیر را در پیش‌بینی دیابت داشت.
۲. BMI :نشان‌دهنده وضعیت سلامت جسمانی بیماران.
۳. سن: عامل کلیدی در پیش‌بینی بیماری‌های مرتبط با دیابت.
۴. فشار خون: متغیری مرتبط با وضعیت قلبی و عروقی بیماران.
۵. تعداد دفعات بارداری: تأثیری غیرمستقیم در پیش‌بینی دیابت به‌ویژه در زنان.

- مقایسه عملکرد مدل پیشنهادی با روش‌های دیگر

مدل پیشنهادی بر اساس جنگل تصادفی با بهینه‌سازی هایپرپارامترها توانست به دقت بالاتری نسبت به سایر الگوریتم‌های پیشین دست یابد. برای مقایسه:

- Naive Bayes: با دقت ۷۶.۶۲٪، عملکرد نسبتاً ضعیفی داشت.
- جنگل تصادفی (Bin Helal, 2024) : با دقت ۹۲.۴٪، اما بدون پیش‌پردازش کافی، نتایج مطلوبی در معیارهای دیگر مانند F1-Score ارائه نکرد.
- مدل پیشنهادی :با دقت ۹۶.۶٪ و میانگین انحراف معیار ۰.۸٪ در اعتبارسنجی متقابل، عملکرد بهتری در تمام معیارها داش که میتوانید نتایج را در جدول شماره ۲ مشاهده کنید.

تجزیه و تحلیل نتایج :

- دقت بالا :نتایج مدل پیشنهادی نشان داد که ترکیب پیش‌پردازش دقیق، متعادل‌سازی داده‌ها و انتخاب ویژگی‌های کلیدی، منجر به بهبود عملکرد مدل شده است.
- تعادل در پیش‌بینی کلاس‌ها: استفاده از روش SMOTEENN باعث شد مدل به طور مؤثری دیابت را در بیماران شناسایی کند، بدون آنکه دقت در پیش‌بینی نمونه‌های غیردیابتی کاهش یابد.
- اهمیت ویژگی‌ها در پیش‌بینی

تحلیل اهمیت ویژگی‌ها نشان داد که متغیر گلوکز تأثیر حیاتی در پیش‌بینی دیابت دارد. ویژگی‌هایی مانند BMI و سن نیز به دلیل ارتباط مستقیم با شرایط سلامتی بیماران، در مدل بسیار تأثیرگذار بودند. این نتایج نشان‌دهنده تأیید مجدد یافته‌های پیشین و کارایی الگوریتم RFE در شناسایی مهم‌ترین ویژگی‌ها است.

- جمع‌بندی :

مدل پیشنهادی با استفاده از تکنیک‌های پیشرفته یادگیری ماشین و پیش‌پردازش دقیق، نه تنها به دقت بالاتری نسبت به مدل‌های مشابه دست یافت، بلکه با ایجاد تعادل در کلاس‌ها، توانست عملکرد پایداری در پیش‌بینی دیابت ارائه دهد. نتایج این مطالعه می‌تواند به عنوان پایه‌ای برای بهبود ابزارهای تشخیصی در پزشکی مورد استفاده قرار گیرد.

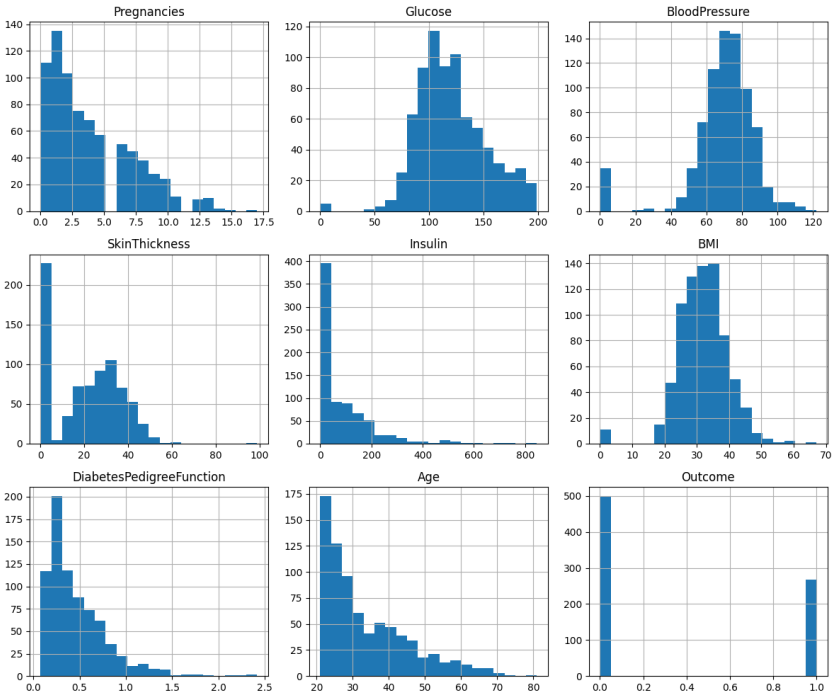
جداول، شکل ها و نمودارها

جدول شماره ۱- مقایسه خروجی الگوریتم های مختلف روی مجموعه داده

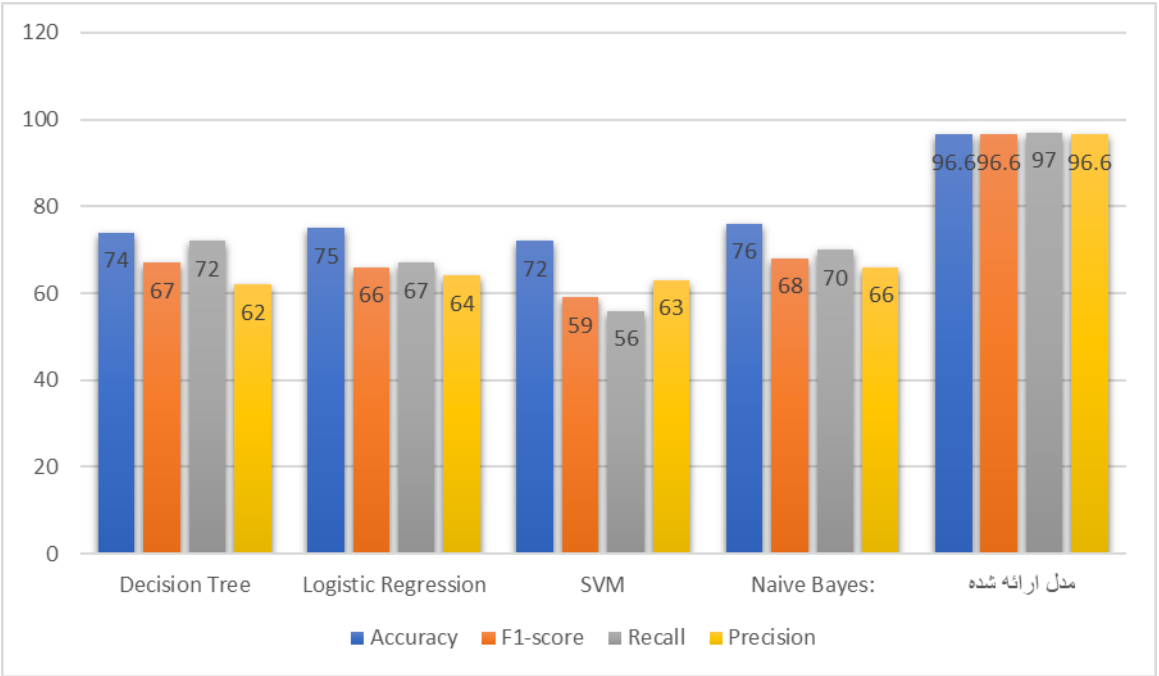
Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score
BernoulliNB	0.74	0.73	0.73	0.74
LinearSVC	0.77	0.72	0.72	0.76
CalibratedClassifierCV	0.77	0.72	0.72	0.76
LogisticRegression	0.77	0.72	0.72	0.76
RandomForestClassifier	0.76	0.72	0.72	0.75
LinearDiscriminantAnalysis	0.76	0.72	0.72	0.75
RidgeClassifierCV	0.76	0.72	0.72	0.75
DecisionTreeClassifier	0.73	0.71	0.71	0.73
BaggingClassifier	0.75	0.71	0.71	0.74
RidgeClassifier	0.75	0.71	0.71	0.74
NearestCentroid	0.71	0.70	0.70	0.71
AdaBoostClassifier	0.75	0.70	0.70	0.73
SGDClassifier	0.74	0.70	0.70	0.73
LGBMClassifier	0.73	0.70	0.70	0.72
NuSVC	0.74	0.70	0.70	0.73
ExtraTreesClassifier	0.74	0.70	0.70	0.73
SVC	0.74	0.70	0.70	0.73
GaussianNB	0.71	0.68	0.68	0.71
ExtraTreeClassifier	0.70	0.68	0.68	0.70
KNeighborsClassifier	0.71	0.67	0.67	0.70
LabelSpreading	0.69	0.66	0.66	0.68
QuadraticDiscriminantAnalysis	0.69	0.66	0.66	0.69
LabelPropagation	0.68	0.65	0.65	0.68
Perceptron	0.60	0.57	0.57	0.60
PassiveAggressiveClassifier	0.56	0.52	0.52	0.55
DummyClassifier	0.62	0.50	0.50	0.47

جدول شماره ۲ – خروجی مدل معرفی شده

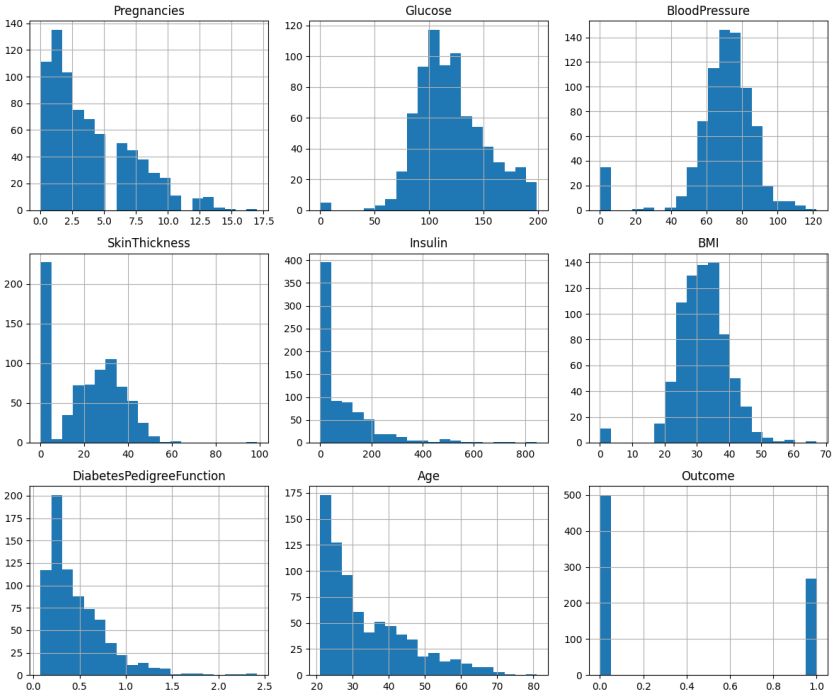
Accuracy	F1-Score	Recall	Precision
۰.۹۶۶۶	۰.۹۶۶۶	۹۷	۹۷



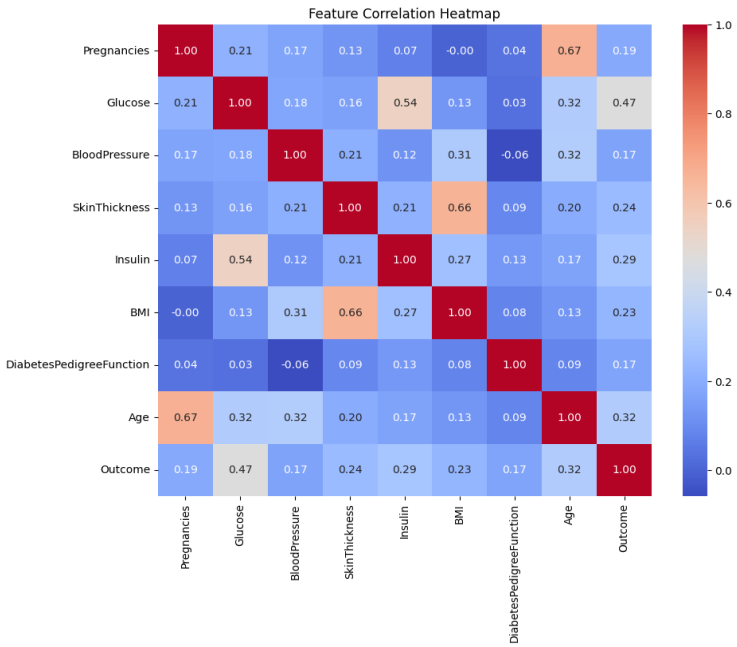
نمودار شماره ۱- مقایسه مدل ارائه شده و کارهای انجام شده قبلی



نمودار شماره ۲ – توزیع ویژگی های مجموعه داده



تصویر شماره ۱– heat map همبستگی ویژگی‌ها



بحث و نتیجه‌گیری

در این پژوهش، مدل جنگل تصادفی با بهره‌گیری از تکنیک‌های پیشرفته‌ای همچون پیش‌پردازش داده‌ها، انتخاب ویژگی(RFE) ، و متعادل‌سازی کلاس‌ها با SMOTEENN توانست به دقت بالایی معادل 96.7٪ در پیش‌بینی دیابت دست یابد که در نمودار شماره ۱ میتوانید آن را با مدل های قبلی مقایسه نمایید. این نتایج نشان‌دهنده توانایی بالای مدل در تفکیک بیماران دیابتی و غیردیابتی است. این نتایج هم‌راستا با تحقیقات پیشین تأکید می‌کند که تعادل‌سازی داده‌ها و انتخاب ویژگی‌های کلیدی نقش مهمی در بهبود عملکرد مدل‌های یادگیری ماشین دارند. به‌ویژه، یافته‌های ما حاکی از آن است که استفاده از روش SMOTEENN به عنوان راهکاری مؤثر برای مقابله با چالش عدم تعادل کلاس‌ها، به طور قابل‌توجهی کیفیت پیش‌بینی را افزایش داده است.

مطالعات قبلی نظیر (Bin Helal, 2024) و (Lara, 2024) نیز تأکید داشتند که الگوریتم جنگل تصادفی در طبقه‌بندی داده‌های پزشکی عملکرد خوبی دارد، اما با پیش‌پردازش ناکافی یا عدم تعادل کلاس‌ها مواجه بودند. مقایسه نتایج این پژوهش با آن‌ها نشان می‌دهد که ترکیب دقیق روش‌های پیش‌پردازش، انتخاب ویژگی، و متعادل‌سازی می‌تواند منجر به بهبود معنادار دقت و عملکرد کلی مدل شود.

نتیجه‌گیری نهایی:

۱. خلاصه یافته‌ها: مدل پیشنهادی با دقت ۹۶.۷٪ عملکرد قابل‌توجهی در پیش‌بینی دیابت ارائه داد.
۲. تأکید بر اهمیت SMOTEENN و جنگل تصادفی: استفاده از این تکنیک‌ها نشان داد که می‌توان چالش‌های مرتبط با داده‌های نامتعادل را به طور مؤثری برطرف کرد.
۳. نقش پیش‌پردازش داده‌ها: حذف داده‌های پرت و پاکسازی مقادیر غیرواقعی نقش حیاتی در افزایش دقت مدل داشت.
۴. چشم‌اندازهای آینده: این مطالعه می‌تواند پایه‌ای برای تحقیقات بیشتر در استفاده از تکنیک‌های یادگیری ماشین در حوزه پزشکی باشد.

با وجود محدودیت‌های مطالعه، یافته‌ها نشان می‌دهند که ترکیب روش‌های پیشرفته با مدل‌های قوی مانند جنگل تصادفی می‌تواند گامی مؤثر در بهبود سیستم‌های تشخیصی در حوزه سلامت باشد.

پیشنهادات :

برای تحقیقات آتی، پیشنهاد می‌شود موارد زیر بررسی شوند:

۱. استفاده از مدل‌های ترکیبی (ensemble) دیگر، مانند XGBoost یا LightGBM ، برای مقایسه عملکرد آن‌ها با جنگل تصادفی.
۲. اعمال تکنیک‌های پیشرفته‌تر پیش‌پردازش داده‌ها، به‌ویژه در مواقعی که داده‌ها شامل نویز یا مقادیر گمشده باشند.
۳. طراحی سیستمی مبتنی بر این مدل برای غربالگری سریع بیماران دیابتی در مراکز درمانی.

در نهایت، این مطالعه نشان داد که ترکیب روش‌های یادگیری ماشین و پیش‌پردازش داده‌ها می‌تواند به‌عنوان ابزاری قوی برای پیش‌بینی بیماری‌های پیچیده‌ای مانند دیابت مورد استفاده قرار گیرد و در راستای بهبود سلامت جامعه تأثیرگذار باشد.

منابع (فونت B Nazanin – اندازه ۱۲ – پررنگ)

مقاله منبع	فارسی	انگلیسی
یک نویسنده		(G. W. Brier, 2020)
دو نویسنده	(نصراله پور و خطیبی، ۱۴۰۲)	.(Hodge , Austin .2004)

– منابع انتهای مقاله:

۱- نصراله پور احمد ، خطیبی توکتم .(۱۴۰۲) . پیش بینی بیماری دیابت با رویکرد یادگیری ماشین . کنفرانس بین المللی

مهندسی صنایع و سیستم ها (۹) . https://civilica.com/doc/1772820/

2- G. W. Brier . .(2020). The random forest algorithm for statistical learning.sage journal . https://doi.org/10.1177/1536867X20909688

3- Victoria Hodge & Jim Austin .(2004). A Survey of Outlier Detection Methodologies . Volume 22, pages 85–126 . DOI: 10.1023/B:AIRE.0000045502.10941.a9

Prediction of Diabetes Using the SMOTEENN Method and the Random Forest Algorithm

Ali Kavyani
Master's Student in Artificial Intelligence and Robotics
Imam Hossein University
Mohammadali javadzade
Assistant Professor at Imam Hossein Comprehensive University, Faculty and Research Center of Artificial Intelligence and Cognitive Sciences

Abstract

In the medical field, diabetes has long been one of the greatest concerns for physicians and a major health challenge for humanity. Timely and accurate prediction of this disease can prevent numerous health complications that may affect individuals in the future. In this study, we leveraged artificial intelligence to enhance the accuracy and ease of diabetes diagnosis, surpassing the prior benchmark of 92% achieved in this dataset. Using the combined SMOTEENN method to address class imbalance, as well as the Recursive Feature Elimination (RFE) algorithm with Random Forest to reduce model complexity and select the most important features, we standardized the data. Subsequently, the data was normalized, and the Random Forest model was trained with optimized hyperparameters using GridSearchCV. Ultimately, our proposed model achieved a remarkable 97% accuracy in predicting and diagnosing diabetes. Notably, this study highlights the significance of data preprocessing and feature selection in improving the performance of machine learning models, in addition to achieving high accuracy.

Keywords: Diabetes prediction, machine learning, random forest, data preprocessing