Karafan
Quarterly Scientific Journal of National University of Skills
Spring 2025, Vol. 22, No. 1, p.150-182
Journal Homepage: https://karafan.tvu.ac.ir/?lang=en
10.48301/kssa.2025.448406.2932

Effective Feature Identification for Type-2 Diabetes Prediction Using Novel Wrapper-Based Random Feature Selection Methods

Hamed SabbaghGol[1], Hamid Saadatfar [2*], Mahdi Khazaiepoor[3]

1. Faculty Member, Department of Computer Engineering, Payame Noor University, Tehran, Iran.

2. Associate Professor, Department of Computer Engineering, Faculty of Electrical and Computer Engineering, University of Birjand, Birjand, Iran.

3. Assistant Professor, Department of Computer, Bi.C., Islamic Azad University, Birjand, Iran.

ARTICLE INFO

Received: 29.06.2024
Revised: 11.12.2024
Accepted: 13.04.2025

Keyword:
Type-2 Diabetes
Dimension Reduction, Feature Selection
Machine Learning
Classification

***Corresponding Author:**
Hamid Saadatfar

Email: saadatfar@birjand.ac.ir

ABSTRACT

Diabetes mellitus Type-2 is a chronic metabolic disorder characterized by hyperglycemia resulting from insulin resistance or deficiency. According to estimates, in 2021, approximately 537 million adults had diabetes, a significant portion of which is attributed to type 2 diabetes. This highlights the critical need to focus on preventive strategies, early diagnosis, and management of type 2 diabetes. This study investigates the performance of different novel feature selection methods in machine learning models for predicting type 2 diabetes. In this research, various wrapper-based feature selection methods were employed to identify the most significant features. Classification algorithms including KNN, decision tree, SVM, random forest, and MLP were evaluated on two standard datasets: Pima Indian Diabetes and Mendeley Diabetes. The results were compared and evaluated using evaluation criteria such as accuracy, specificity, precision, sensitivity, F1-measure and ROC curve. The selected features in the Pima dataset include glucose, body mass index, age and blood pressure, and in the Mendeley dataset include HbA1c, BMI and cholesterol. These features showed the highest accuracy with values of 77.3% and 98% using the ERSFS feature selection method in the Pima and Mendeley datasets, respectively. The present study reveals the potential of feature selection methods in improving the classification performance of type 2 diabetes and can help clinicians and researchers in developing and using more accurate diagnostic tools for this disease. In addition, this study provides valuable insight into the most important factors affecting the prediction of type 2 diabetes.

E-ISSN: 2538-4430
ISSN: 2382-9796

EXTENDED ABSTRACT

Introduction

Type 2 diabetes is a chronic disease characterized by high blood sugar levels. It is the most common type of diabetes and is associated with insulin resistance. Individuals with type 2 diabetes may require medication or lifestyle changes to manage their blood sugar levels.

Early diagnosis of type 2 diabetes is crucial for preventing complications. Machine learning algorithms can be utilized to improve the early detection of type 2 diabetes. However, the performance of these algorithms can be affected by irrelevant features in the data. Feature selection can be employed to identify a subset of relevant features that enhance the prediction accuracy and interpretability of the model.

In this study, various Wrapper-based feature selection methods were applied to standard type 2 diabetes datasets. This study aimed to compare the impact of these methods on the performance of machine learning models in predicting type 2 diabetes. Experiments were conducted on two standard datasets, Pima Indian Diabetes and Mendeley Diabetes.

Initially, different Wrapper-based feature selection methods were employed to identify effective features for type 2 diabetes classification. Subsequently, the performance of these features was evaluated using various classification algorithms, including KNN, decision tree, SVM, random forest, and artificial neural networks. Finally, the results were compared based on a set of evaluation metrics including accuracy, F1-measure, specificity, precision, sensitivity, and the relative operating characteristic (ROC) curve.

Methodology

This study encompassed four main phases:

1.Data Preprocessing: In this phase, the data is prepared for removing noise, missing values, and anomalies.

2.Feature Selection: In this phase, feature selection techniques, including Forward selection, Backward elimination, ERSFS, SFE, and SPFSR are employed to identify a subset of relevant features from the high-dimensional datasets. This is done to improve model accuracy and reduce computational complexity.

3.Machine Learning: In this phase, various machine learning algorithms including K-Nearest Neighbors (KNN) classifier, Support Vector Machine (SVM), J48 decision tree, random forest, and Multilayer Perceptron (MLP) neural networks are trained on the selected feature subsets.

4.Evaluation: In this phase, the performance of the machine learning algorithms is assessed using various metrics.

Results and Discussion

In the present study, five-fold cross-validation was employed to evaluate the models. In this method, the data is divided into five random subsets. Four subsets were used to train the model, and one subset was used to evaluate it. This process was repeated five times, and the model performance was reported based on the average of the evaluation results. Various metrics including accuracy, F1-measure, specificity, precision, sensitivity, and the relative operating characteristic (ROC) curve, were used to evaluate the performance of the models. These metrics were calculated based on the confusion matrix, which shows the number of correctly and incorrectly classified samples by the model. F1-Score and AUC-ROC are more comprehensive measures of model performance and represent their overall value.

Two datasets, Pima Indian Diabetes and Mendeley Diabetes, were utilized to identify effective features for the early diagnosis of type 2 diabetes. Initially, machine learning algorithms were executed on the datasets without applying feature selection methods. Subsequently, five feature selection methods including SFS, RFE, ERSFS, SFE, and SPFSR were employed. Wrapper-based feature selection methods can improve the performance of machine learning algorithms for type 2 diabetes diagnosis.

In the Pima dataset, ERSFS and SPFSR methods each select 4 features, while the other methods select 5 features. None of the algorithms selects the same features. The highest accuracy was achieved by selecting 4 features in the ERSFS method in the SVM model. The features selected by the ERSFS method (i.e., glucose, BMI, age, and blood pressure) provided the best performance in various machine learning models and can play an effective role in the early diagnosis of type 2 diabetes. Following ERSFS, the RFE and SFS feature selection methods demonstrated promising results for decision tree, random forest, and MLP models. For instance, the SFS method in the MLP model achieved the highest performance metrics among all feature selection techniques.

In the Mendeley dataset, ERSFS and SFE methods select 3 features, SFE and SPFSR methods select 4 features, and SFS and RFE methods select 5 features. None of the algorithms selects the same features. The features selected by SFS, RFE, and SPFSR methods include the features selected by ERSFS (i.e., HbA1c, BMI, and Chol). The highest performance, especially accuracy, was achieved in the ERSFS method using decision tree and random forest algorithms. Following ERSFS, the SFE, RFE, and SFS feature selection methods also provided acceptable results for the decision tree, random forest, and MLP models. For instance, the SFS and RFE methods achieved the highest performance metrics among all feature selection techniques after ERSFS in the MLP model. In the RF and decision tree models, the results of SPFSR are promising, particularly in terms of accuracy, sensitivity, and F1-measure.

Machine learning models can utilize these features to identify individuals at risk of type-2 diabetes and for early diagnosis of the disease.

Conclusions

This study demonstrates that Wrapper-based feature selection methods can be a valuable tool for improving the performance of machine learning algorithms in type 2 diabetes diagnosis.

The identified features, glucose, BMI, and blood pressure in the Pima dataset and HbA1c, BMI, and cholesterol in the Mendeley dataset, can be employed as tools for faster and more accurate diagnosis of individuals at risk of type 2 diabetes, as well as for early disease detection.

Future studies should focus on collecting larger and more comprehensive datasets in addition to developing more accurate machine learning algorithms for type 2 diabetes diagnosis.

کارافن
فصلنامه علمی دانشگاه ملی مهارت
بهار۱٤۰٤، دوره ۲۲، شماره ۱، ۱۸۲-۱۵۰
آدرس نشریه: https://karafan.nus.ac.ir
10.48301/kssa.2025.448406.2932

شناسایی ویژگی‌های موثر برای پیش‌بینی دیابت نوع ۲ با استفاده از روش‌های نوین انتخاب ویژگی تصادفی مبتنی بر Wrapper

حامد صباغ گل[۱]، حمید سعادت‌فر[۲*]، مهدی خزاعی‌پور[۳]

۱- عضو هیئت علمی، گروه مهندسی کامپیوتر، دانشگاه پیام‌نور، تهران، ایران.

۲. دانشیار، گروه مهندسی کامپیوتر، دانشکده مهندسی برق و کامپیوتر، دانشگاه بیرجند، بیرجند، ایران

۳. استادیار، گروه کامپیوتر، واحد بیرجند، دانشگاه آزاد اسلامی، بیرجند، ایران.

اطلاعات مقاله

دریافت مقاله: ۱۴۰۳/۰۴/۰۹
بازنگری مقاله: ۱۴۰۳/۰۹/۲۱
پذیرش مقاله: ۱۴۰۴/۰۱/۲۴

کلید واژگان:
دیابت نوع۲
کاهش ابعاد
انتخاب ویژگی
یادگیری ماشین
طبقه‌بندی

***نویسنده مسئول:** حمید سعادت‌فر
پست الکترونیکی:
saadatfar@birjand.ac.ir

چکیده

دیابت ملیتوس نوع ۲ یک اختلال متابولیک مزمن است که با هایپرگلیسمی ناشی از مقاومت به انسولین یا کمبود آن مشخص می‌شود. بر اساس برآوردها، در سال ۲۰۲۱ حدود ۵۳۷ میلیون بزرگسال دچار دیابت بودند که بخش قابل توجهی از آن به دیابت نوع۲ نسبت داده می‌شود. این موضوع نشان‌می‌دهد که تمرکز بر راهکارهای پیشگیری، تشخیص زودهنگام و مدیریت دیابت نوع ۲ بسیار حیاتی است. این پژوهش به بررسی عملکرد روش‌های مختلف انتخاب ویژگی در مدل‌های یادگیری ماشین برای پیش‌بینی بیماری دیابت نوع ۲ می‌پردازد. در این تحقیق، از روش‌های مختلف و نوین انتخاب ویژگی مبتنی‌بر wrapper برای شناسایی مهم‌ترین ویژگی‌ها استفاده شده است. الگوریتم‌های طبقه‌بندی شامل KNN، درخت تصمیم، SVM، جنگل تصادفی و MLP روی دو مجموعه‌داده استاندارد Pima Indian Diabetes و Mendeley Diabetes مورد ارزیابی قرار گرفته‌اند. نتایج با استفاده از معیارهای ارزیابی مانند دقت، ویژگی، صحت، حساسیت، F1-measure و منحنی ROC مقایسه و بررسی می‌شوند. ویژگی‌های انتخاب‌شده در مجموعه‌داده Pima شامل گلوکز، شاخص توده بدنی، سن و فشار خون، و در مجموعه‌داده Mendeley شامل HbA1c، BMI و کلسترول هستند. این ویژگی‌ها بالاترین میزان دقت را به‌ترتیب با مقادیر ٪۷۷,۳ و ٪۹۸ توسط روش انتخاب ویژگی ERSFS در مجموعه‌داده‌های Pima و Mendeley نشان می‌دهند. پژوهش حاضر پتانسیل روش‌های انتخاب ویژگی را در بهبود عملکرد طبقه‌بندی دیابت نوع ۲ آشکار می‌سازد و می‌تواند به پزشکان و محققان در توسعه و استفاده از ابزارهای تشخیصی دقیق‌تر برای این بیماری کمک کند. همچنین، این تحقیق بینش ارزشمندی درباره مهم‌ترین عوامل مؤثر بر پیش‌بینی ابتلا به دیابت نوع۲ ارائه می‌دهد.

مقدمه

دیابت ملیتوس یک اختلال متابولیک مزمن است که با هایپرگلیسمی یا افزایش قند خون مشخص می‌شود. این بیماری به سه دسته اصلی تقسیم می‌شود: دیابت نوع ۱، دیابت نوع ۲ و سایر انواع خاص دیابت [۱]. دیابت نوع ۱ که قبلاً به عنوان دیابت وابسته به انسولین شناخته می‌شد، یک بیماری خودایمنی است که در آن سیستم ایمنی بدن به اشتباه سلول‌های بتای تولیدکننده انسولین در پانکراس را از بین می‌برد. افراد مبتلا به دیابت نوع ۱ برای ادامه حیات به تزریق انسولین نیاز دارند [۱, ۲].

دیابت نوع ۲، شایع‌ترین نوع دیابت است و با مقاومت به انسولین مرتبط است. مقاومت به انسولین به این معنی است که سلول‌ها به‌طور مؤثر به انسولین پاسخ نمی‌دهند، که منجر به افزایش قند خون می‌شود. افراد مبتلا به دیابت نوع ۲ ممکن است برای کنترل قند خون خود نیاز به دارو یا تغییرات در سبک زندگی داشته باشند [۱]. این شکل از دیابت حدود ۹۰ درصد از موارد را تشکیل می‌دهد [۳]. عوامل خطر آن شامل چاقی، سبک زندگی کم‌تحرک، سن بالا، سابقه خانوادگی دیابت و نژاد یا قومیت است [۴]. بیماران مبتلا به دیابت نوع ۲ در معرض خطر بالایی از عوارض مزمن از جمله بیماری‌های قلبی عروقی، نفروپاتی، نوروپاتی و رتینوپاتی قرار دارند [۵].

سایر انواع خاص دیابت، از جمله دیابت بارداری است که به افزایش قند خون در زنان باردار اشاره دارد و معمولاً پس از زایمان برطرف می‌شود [۶].

بر اساس برآورد سازمان جهانی بهداشت (WHO)[1] در سال ۲۰۱۴، حدود ۴۲۲ میلیون نفر در سطح جهان به دیابت مبتلا بوده‌اند و سالانه ۱,۵ میلیون مرگ و میر به‌طور مستقیم به این بیماری نسبت داده‌می‌شود [۷]. گزارش فدراسیون بین‌المللی دیابت در سال ۲۰۲۱، ابعاد گسترده‌تر این چالش را نشان می‌دهد. این گزارش حاکی از آن است که ۵۳۷ میلیون بزرگسال (۲۰ تا ۷۹ ساله) با دیابت زندگی می‌کردند که ۱۰ درصد از جمعیت بزرگسال را شامل می‌شد. پیش‌بینی‌ها نشان‌می‌دهد که این آمار رو به افزایش است و تا سال ۲۰۳۰ به ۶۴۳ میلیون و تا سال ۲۰۴۵ به ۷۸۳ میلیون نفر خواهد رسید. این بیماری در سال ۲۰۲۱، ۶,۷ میلیون مرگ (تقریباً هر ۵ ثانیه یک مرگ) را به خود اختصاص داده و با هزینه ۹۶۶ میلیارد دلار، بار اقتصادی سنگینی را به جوامع تحمیل می‌کند که بخش عمده‌ای از آن مربوط به دیابت نوع ۲ می‌باشد [۸].

اهمیت این موضوع ضرورت تمرکز بر راهکارهای پیشگیری، تشخیص و مدیریت دیابت نوع ۲ را به‌منظور مقابله با این چالش بزرگ جهانی آشکار می‌کند. به‌طور سنتی، تشخیص بر اساس علائم بالینی و میزان قند خون ناشتا انجام می‌شود. با این‌حال، این روش‌ها ممکن است نتوانند بیماری را در مراحل اولیه آن تشخیص دهند. بنابراین، نیاز فزاینده‌ای برای روش‌های کارآمد و قابل‌اعتماد برای تشخیص زودهنگام این بیماری وجود دارد.

امروزه، الگوریتم‌های یادگیری ماشین به‌طور گسترده‌ای در زمینه‌های مختلف ازجمله: تشخیص نفوذ [۹]، سیستم‌های قدرت [۱۰]، تحلیل احساسات [۱۱]، پیش‌بینی بازار سهام [۱۲] و تشخیص پزشکی به‌کار گرفته می‌شوند. این روش‌ها پتانسیل خوبی برای تجزیه و تحلیل مجموعه‌داده‌های بزرگ و پیچیده جهت بهبود تشخیص الگو و تصمیم‌گیری دارند [۱۳]. طبقه‌بندی یکی از روش‌های رایج تحت نظارت در داده‌کاوی است که داده‌ها را به کلاس‌ها تقسیم می‌کند و به فرد امکان می‌دهد انواع مختلف داده، از داده‌های پیچیده گرفته تا داده‌های ساده را سازماندهی کند [۱۴]. با استفاده از این مدل‌های یادگیری ماشین، می‌توان بیماری دیابت نوع ۲ را در مراحل ابتدایی تشخیص داد.

[1] World Health Organization

الگوریتم‌های طبقه‌بندی که در این مطالعه مورد استفاده قرار می‌گیرد، طبقه بند KNN، درخت تصمیم، ماشین‌های بردار پشتیبان (SVM)، جنگل تصادفی و شبکه عصبی پرسپترون چندلایه (MLP) را شامل می‌شود [۱۵].

اگرچه یادگیری ماشین برای تشخیص طیف گسترده‌ای از بیماری‌ها ضروری است، اما عملکرد این الگوریتم‌ها ممکن است تحت تأثیر حضور ویژگی‌های بی‌ربط یا زائد در داده‌ها قرار گیرد، که منجر به افزایش پیچیدگی محاسباتی، برازش بیش از حد، و کاهش قابلیت تفسیر مدل می‌شود [۱۶]. روش‌های انتخاب ویژگی به‌دنبال شناسایی زیرمجموعه ای از ویژگی‌های پرمعنی و مرتبط می‌باشند؛ در نتیجه دقت پیش‌بینی، کارایی محاسباتی و قابلیت تفسیر مدل‌های طبقه‌بندی را بهبود می‌بخشند [۱۷]. بنابراین، انتخاب یک زیرمجموعه مناسب از ویژگی‌های اصلی برای افزایش عملکرد طبقه‌بندی و همچنین غلبه بر "نفرین ابعاد"[1] ضروری است [۱۸].

در زمینه طبقه‌بندی دیابت، مجموعه‌داده‌های مختلفی که حاوی ویژگی‌های فیزیولوژیکی، بالینی و جمعیت شناختی هستند به‌طور گسترده‌ای مورد مطالعه قرار گرفته‌اند [۲۱-۱۹]. با این حال، انتخاب ویژگی‌های متمایزکننده به‌دلیل ابعاد نسبتا بالا، اهمیت تشخیص سریع بیماری و احتمال ازدحام اطلاعات، چالش‌برانگیز است [۱۷]. بنابراین به‌کارگیری روش‌های انتخاب ویژگی ضروری به نظر می‌رسد.

روش‌های انتخاب ویژگی به سه نوع اصلی دسته‌بندی می‌شود: فیلتر، Wrapper و تعبیه‌شده[2] [۲۲]. روش‌های فیلتر، ارتباط ویژگی‌ها را با در نظر گرفتن ویژگی‌های ذاتی یا رابطه آنها با متغیر هدف، مستقل از الگوریتم طبقه‌بندی ارزیابی می‌کنند [۲۳]. این روش‌ها ویژگی‌ها را بر اساس معیارهای آماری مانند هم‌بستگی یا اطلاعات متقابل رتبه‌بندی می‌کنند. نمونه‌هایی از روش‌های فیلتر عبارتند از Relief و mRMR (حداقل ارتباط افزونگی حداکثر) [۲۴]. از سوی دیگر، روش‌هایWrapper، زیرمجموعه‌های ویژگی را با آموزش یک الگوریتم طبقه‌بندی خاص و ارزیابی عملکرد مدل آموزش‌دیده با استفاده از مجموعه اعتبارسنجی، ارزیابی می‌کنند [۲۵]. این روش‌ها معمولاً عملکرد بهتری در انتخاب ویژگی‌های مرتبط و افزایش دقت مدل طبقه‌بندی دارند. اما وقتی‌که این روش‌ها با داده‌های با ابعاد بالا سروکار دارند، از نظر محاسباتی گران می‌باشند. نمونه‌هایی از روش‌های Wrapper شامل انتخاب ویژگی رو به جلو[3]،حذف به عقب[4] و رویکردهای مبتنی بر فراابتکاری مانند الگوریتم‌های ژنتیک و بهینه‌سازی ازدحام ذرات است [۲۶, ۲۷]. در روش‌های تعبیه‌شده[5]عملیات انتخاب ویژگی و برازش مدل به‌صورت همزمان انجام می‌شوند و ویژگی‌های مهم را در حین ساخت مدل یادگیری ماشین، شناسایی می‌کنند [۲۸]. نمونه‌هایی از روش‌های تعبیه‌شده عبارتند از LASSO [۲۹]، درخت‌های تصمیم‌گیری و جنگل‌های تصادفی [۳۰].

الگوریتم‌های انتخاب ویژگی wrapper، دسته‌ای محبوب و کارآمد می‌باشند که با گرفتن بازخورد از مدل یادگیری، سعی در انتخاب بهترین زیرمجموعه از ویژگی‌ها دارند. برخلاف رویکردهای فیلتر، این روش‌ها امکان تشخیص بهتر تعاملات احتمالی بین ویژگی‌ها را فراهم می‌کنند. معمولا این روش‌ها در مقایسه دو دسته دیگر ویژگی‌های بهتری را انتخاب می‌نمایند. در ادامه به معرفی چند روش wrapper محبوب می‌پردازیم.

الگوریتم انتخاب ویژگی رو به جلو (مستقیم)، یک روش ترتیبی انتخاب ویژگی است که با یک مجموعه اولیه خالی از ویژگی‌ها آغاز می‌شود. در هر مرحله، ویژگی‌هایی که بهترین عملکرد را برای مدل یادگیری به ارمغان می‌آورند، به

---

1 The curse of dimensionality
2 Embedded
3 Forward feature selection
4 Backward elimination
5 Embedded

این مجموعه اضافه می‌شوند [۲۲]. برای هر ویژگی جدید، ارزش آن با استفاده از یک طبقه‌بند و معیار مناسب مانند دقت بررسی می‌شود. سپس ویژگی‌هایی که عملکرد مدل را بیشینه می‌کنند، به مجموعه اضافه می‌شوند. این الگوریتم تکرار می‌شود تا زمانی که یک معیار توقف مشخص مانند تعداد مورد نیاز ویژگی‌ها یا آستانه عملکرد، برآورده شود [۳۱].

الگوریتم انتخاب ویژگی حذف به عقب (معکوس) نیز یک روش ترتیبی است که رویکردی مخالف را دنبال می‌کند و با یک مجموعه اولیه شامل تمام ویژگی‌ها آغاز می‌شود. سپس در هر مرحله، ویژگی‌هایی که حذف آن کمترین کاهش در عملکرد، کارایی و یا دقت طبقه‌بند را به‌همراه دارد (بدترین ویژگی‌ها)، حذف می‌شوند [۳۲]. این الگوریتم نیز تا زمانی که یک معیار توقف مشخص برآورده شود، به‌صورت تکراری اجرا می‌شود. الگوریتم انتخاب معکوس می‌تواند به‌طور موثرتری نسبت به الگوریتم انتخاب رو به جلو تعاملات و ازدحام ویژگی‌ها را شناسایی کند، زیرا سهم هر ویژگی را در سایر ویژگی‌های باقیمانده ارزیابی می‌کند. با این حال، باید توجه داشت که برای مجموعه‌داده‌های با ابعاد بالا، این الگوریتم ممکن است از نظر محاسباتی پرهزینه باشد [۳۱].

الگوریتم بهبودیافته انتخاب تصادفی زیرمجموعه‌ای از ویژگی‌ها در سال ۲۰۲۴ توسط صباغ‌گل و همکاران [۳۳]، برای مسائل طبقه‌بندی ارائه شده‌است. این الگوریتم یک روش انتخاب ویژگی wrapper است که هدف آن شناسایی زیرمجموعه‌ای از ویژگی‌های مفید در یک مسئله طبقه‌بندی است. الگوریتم چهار مرحله دارد: پیش‌پردازش داده‌ها، انتخاب تصادفی زیرمجموعه ویژگی، ارزیابی زیرمجموعه با استفاده از طبقه‌بندی‌کننده، و انتخاب بهترین ویژگی‌ها بر اساس نتایج ارزیابی. الگوریتم ارائه‌شده بر تقویت فرآیند جستجو، اصلاح و ارتقای کیفیت زیرمجموعه‌های ویژگی، ارزیابی مستمر ویژگی‌های انتخاب‌شده برای جلوگیری از گیر افتادن در بهینه محلی، و توجه به مفهوم هم‌افزایی و تعامل بین ویژگی‌ها در تکرارهای مختلف تمرکز دارد. هدف از این الگوریتم دستیابی به‌سرعت همگرایی بالاتر، نرخ انتخاب ویژگی کمتر و دقت طبقه‌بندی بالاتر است. عملکرد الگوریتم با استفاده از ۲۰ مجموعه‌داده استاندارد مورد ارزیابی قرار گرفت و نشان داد که با انتخاب ویژگی‌های کمتر، دقت طبقه‌بندی بالاتری نسبت به روش‌های دیگر دارد. در این پژوهش، این روش به نامERSFS معرفی شده است.

احدزاده و همکاران [۳۴] در سال ۲۰۲۳ الگوریتم جدیدی برای انتخاب ویژگی به نام SFE معرفی کردند. این الگوریتم از یک عامل جستجو و دو اپراتور غیرانتخابی و انتخابی برای انجام جستجوی جامع به منظور شناسایی و حذف ویژگی‌های نامربوط، زائد و نویزی استفاده می‌کند. در نهایت، SFE ویژگی‌هایی را برمی‌گزیند که تأثیر قابل‌توجهی بر نتایج طبقه‌بندی دارند. این مطالعه همچنین الگوریتم ترکیبی SFE-PSO (بهینه‌سازی ازدحام ذرات) را برای یافتن یک زیرمجموعه ویژگی بهینه پیشنهاد می‌کند. این الگوریتم زمانی به کار گرفته می‌شود که پس از کاهش ابعاد مجموعه‌داده، عملکرد SFE به‌طور قابل‌توجهی بهبود نیابد. الگوریتم SFE-PSO روش SFE را با الگوریتم PSO ترکیب می‌کند تا یک زیرمجموعه ویژگی بهینه را در فضای جستجوی کاهش‌یافته جستجو کند. نتایج نشان می‌دهد که الگوریتم‌های SFE و SFE-PSO از نظر کارایی و اثربخشی در مقایسه با سایر الگوریتم‌های انتخاب ویژگی، به‌ویژه برای انتخاب ویژگی‌ها در مجموعه‌داده‌های با ابعاد بالا، عملکرد مطلوبی دارند.

اکمان و همکاران در سال ۲۰۲۳ [۳۵] الگوریتم انتخاب ویژگی SPFSR را ارائه کردند. این الگوریتم، یک رویکرد تقریب تصادفی جدید برای رتبه‌بندی همزمان k-بهترین ویژگی و انتخاب ویژگی بر اساس تقریب تصادفی آشفتگی همزمان (SPSA) با دستاوردهای غیریکنواخت BB[1] است. الگوریتم SPFSR یک روش مبتنی بر wrapper

[1] Simultaneous Perturbation Stochastic Approximation

است که می‌توان آن را در کنار هر طبقه‌بند یا رگرسیون و با هر معیار عملکردی استفاده نمود. این روش شامل میانگین‌گیری گرادیان و هموارسازی برای کاهش نویز است. در آزمایش‌هایی که بر روی ۴۷ مجموعه‌داده عمومی انجام شد، SPFSR در بیش از ۸۰ درصد آزمایش‌های طبقه‌بندی و بیش از ۸۵ درصد آزمایش‌های رگرسیون، نسبت به تکنیک‌های کاهش ویژگی موجود، بهبود معنی‌داری یا عملکرد معادل را نشان داده است. علاوه بر این، SPFSR به طور متوسط در مقایسه با استفاده از کل مجموعه ویژگی، به دقت طبقه‌بندی بهتری دست یافت. با این حال، SPFSR دو محدودیت اصلی دارد که عبارتند از هزینه محاسباتی بالا در مجموعه‌داده‌های با ابعاد بالا و نادیده گرفتن روابط بین ویژگی‌ها.

چندین مطالعه به بررسی عوامل خطر مرتبط با دیابت نوع ۲ و ارائه مدل‌های پیش‌بینی بر اساس این عوامل پرداخته‌اند. پژوهش پن و همکاران [۳۶] بر توسعه یک مدل پیش‌بینی خطر برای رتینوپاتی دیابتی در بیماران چینی مبتلا به دیابت نوع ۲ تمرکز دارد و عواملی چون HbA1c، مدت ابتلا به دیابت، قند خون پس از غذا، سن و فشار خون سیستولیک (SBP) را به‌عنوان عوامل مهم شناسایی کرده است. این یافته‌ها با عوامل خطر شناخته‌شده دیابت نوع ۲ همخوانی دارند و بر اهمیت آن‌ها در مدیریت بیماری و پیشگیری از عوارض تأکید می‌کنند. مطالعه [۳۷]، نقش مداخلات تغذیه‌ای در کنترل گلیسمی افراد مبتلا به پیش‌دیابت را بررسی کرده و بر تأثیر رژیم غذایی و انتخاب‌های سبک زندگی بر مقاومت به انسولین و خطر دیابت نوع ۲ تأکید دارد. این مطالعه نقش محافظتی رژیم‌های غذایی پر فیبر و غلات کامل را برجسته کرده و از شیوع رو به افزایش پیش‌دیابت به‌عنوان پیش‌زمینه مهمی برای دیابت نوع ۲ یاد می‌کند. پژوهش اسلام وهمکارانش [۳۸] از یادگیری ماشین برای شناسایی عوامل خطر با استفاده از داده‌های NHANES استفاده کرده و عواملی چون سن، سطح تحصیلات، وضعیت تأهل، SBP، وضعیت استعمال دخانیات و شاخص توده بدنی (BMI) را به‌عنوان پیش‌بینی‌کننده‌های کلیدی معرفی کرده است. این عوامل بر تعامل بین تعیین‌کننده‌های اجتماعی-اقتصادی، رفتارهای سبک زندگی و سلامت متابولیک در توسعه دیابت نوع ۲ تأکید دارند.

مطالعات دیگر به تعامل بین دیابت نوع ۲ و شرایط مرتبط پرداخته‌اند. فیتریانی و همکاران در پژوهش [۳۹]، دیابت نوع ۲ را در بیماران مبتلا به بیماری کبد چرب غیرالکلی (NAFLD) بررسی کرده و نشان داده است که شاخص‌های عملکرد کبد (ALT، AST، ALP و GGT) و SBP از پیش‌بینی‌کننده‌های کلیدی هستند، که به ارتباط نزدیک بین اختلال عملکرد کبد و دیابت نوع ۲ اشاره دارد. به‌طور مشابه، مطالعه[۴۰] به بررسی NAFLD و سندرم متابولیک پرداخته و نقش چاقی مرکزی (احشایی)، دیس‌لیپیدمی، هایپرگلیسمی و فشار خون بالا را به‌عنوان عوامل خطر مشترک برای هر دو شرایط مورد تأکید قرار داده است. همچنین به عدم فعالیت بدنی به‌عنوان یکی از عوامل مؤثر در دیابت نوع ۲ و سندرم متابولیک اشاره شده است. پژوهش [۴۱] با استفاده از تکنیک‌های داده‌کاوی بر روی مجموعه‌داده BRFSS2015، عواملی مانند سن، سلامت عمومی، جنسیت، مصرف زیاد الکل، فشار خون بالا، کلسترول بالا، دسترسی محدود به مراقبت‌های بهداشتی و مصرف سبزیجات را به‌عنوان پیش‌بینی‌کننده‌های کلیدی دیابت نوع ۲ شناسایی کرده است. این یافته‌ها بر چندبعدی بودن دیابت نوع ۲ تأکید دارند و عوامل رفتاری، نشانگرهای بالینی و اجتماعی-اقتصادی را در پروفایل خطر آن در نظر می‌گیرند.

همچنین پژوهش‌های [۴۴-۴۲] روش‌های یادگیری ماشین را برای تشخیص دیابت نوع ۲ با استفاده از داده‌های Pima Indians Diabetes بررسی می‌کنند. میتیلی و همکاران [۴۳] الگوریتمی ترکیبی از ویژگی‌های برتر الگوریتم‌های موجود مانند Chi-square و خوشه‌بندی پیشرفته را پیشنهاد می‌کند که هدف آن دستیابی به دقت بالا در تشخیص است. چهار ویژگی سن، BMI، فشار خون و قند خون را به عنوان مهم‌ترین فاکتورهای تشخیصی معرفی

[1] Non-monotone Barzilai–Borwein (BB) search method

می‌کند. همچنین صفایی و همکارش [۴۲] رویکردی مبتنی بر انتخاب بهینه زیرمجموعه‌ای از ویژگی‌ها با الگوریتم بهینه‌سازی ملخ را معرفی می‌کند. سپس ماشین بردار پشتیبان با این زیرمجموعه بهینه آموزش داده می‌شود. در پژوهش آن‌ها، ویژگی‌های دفعات بارداری، گلوکز، BMI و سن بیشترین تأثیر را در تشخیص دارند. همچنین در پژوهش [۴۴] از الگوریتم درخت تصمیم C4.5 بر روی مجموعه‌داده Pima Indians Diabetes برای توسعه مدل‌های پیش‌بینی استفاده شد. عوامل خطر کلیدی شناسایی‌شده در این مطالعه، شامل سطح بالای قند خون، تعداد بالای بارداری‌ها، سن، فشار خون دیاستولیک بالا، سابقه خانوادگی و شاخص توده بدنی بالا هستند.

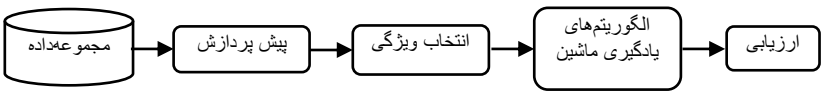
این مطالعات به‌طور جمعی بر پیچیدگی عوامل خطر دیابت نوع ۲ تأکید دارند که ابعاد متابولیک، رفتاری و اجتماعی-اقتصادی را شامل می‌شود. تنوع و تعدد این عوامل نشان‌دهنده اهمیت شناسایی دقیق‌تر ویژگی‌های مؤثر برای بهبود پیش‌بینی و مدیریت بیماری است. در این زمینه، استفاده از روش‌های نوین انتخاب ویژگی، مانند رویکردهای مبتنی بر Wrapper، می‌تواند با تمرکز بر شناسایی زیرمجموعه‌های بهینه ویژگی‌ها، به کاهش پیچیدگی مدل‌ها و افزایش دقت آن‌ها کمک کند. پژوهش حاضر، با توسعه و ارزیابی روش‌های نوین انتخاب ویژگی تصادفی، به مقایسه تأثیر این روش‌ها در مدل‌های یادگیری ماشین پرداخته و با استفاده از مجموعه‌داده‌های استاندارد، نتایج قابل‌توجهی در پیش‌بینی دیابت نوع ۲ ارائه کرده است. مقاله ارائه شده شامل موارد قابل توجه زیر است:

- بکارگیری روش‌های مختلف انتخاب ویژگی مبتنی بر Wrapper بر روی مجموعه‌داده‌های استاندارد بیماری دیابت نوع ۲.
- شناسایی ویژگی‌های موثر در طبقه‌بندی دیابت نوع ۲ و ارزیابی عملکرد آنها با استفاده از الگوریتم‌های مختلف طبقه‌بندی مانند KNN، درخت تصمیم، SVM، جنگل تصادفی و شبکه‌های عصبی مصنوعی.
- مقایسه و بررسی نتایج بر اساس مجموعه جامعی از معیارهای ارزیابی از جمله دقت، F1-measure، Specificity، Precision، حساسیت و منحنی مشخصه عملکرد (ROC).

ساختار این مقاله به شرح زیر است: بخش «مقدمه» شامل معرفی مسئله تحقیق و بررسی ادبیات مرتبط می‌باشد. در بخش «روش‌شناسی» به بیان جزییات در خصوص روش‌های تحقیق پرداخته شده است. در بخش «روش انجام آزمایش» نحوه انجام آزمایش‌ها و در بخش «نتایج»، نتایج تحقیق مورد بررسی و ارزیابی قرار گرفته‌اند. در بخش «بحث و نتیجه‌گیری»، این دستآوردها مورد تجزیه و تحلیل قرار گرفته و نتیجه‌گیری‌های مهم بیان شده است. در پایان مقاله، محدودیت‌های روش و پیشنهادات آینده، بیان و بررسی شده‌اند.

روش‌شناسی

این پژوهش، همان‌گونه که در فلوچارت شکل ۱ نشان داده شده‌است، به چهار مرحله کلی تقسیم می‌شود. در مرحله اول، داده‌ها پیش‌پردازش می‌شوند، سپس از تکنیک‌های انتخاب ویژگی بر روی مجموعه‌داده موردنظر استفاده می‌شود. پس از آن، بهترین زیرمجموعه انتخاب و الگوریتم‌های یادگیری ماشین مورد نظر بر روی آن اعمال می‌شوند. در نهایت، عملکرد الگوریتم‌های انتخاب ویژگی با استفاده از معیارهای مختلف ارزیابی می‌شود.



شکل ۱. فلوچارت روش

مجموعه‌داده‌ها

۱- مجموعه‌داده Pima Indian Diabetes [۴۵]، یکی از مجموعه‌داده‌های کلاسیک و پرکاربرد در حوزه یادگیری ماشین و داده‌کاوی برای مسائل طبقه‌بندی دوتایی است. این مجموعه‌داده از پایگاه داده معتبر یادگیری ماشین ایروین دانشگاه کالیفرنیا (UCI) [۴۶] استخراج شده است. مجموعه‌داده شامل ۷۶۸ نمونه با ۸ ویژگی برای هر نمونه و دو کلاس است که از جمعیت زنان بالغ ۲۱ ساله و بالاتر از قبیله هندی پیما جمع‌آوری شده است. هدف، پیش‌بینی ابتلا یا عدم ابتلا به دیابت نوع ۲ بر اساس ویژگی‌هایی مانند سن، فشار خون، چربی بدن و سایر معیارهای آزمایشگاهی است. این مجموعه‌داده دارای ویژگی‌هایی با مقادیر گمشده است. جزئیات ویژگی‌ها در جدول ۱ آمده است.

جدول ۱. جزئیات ویژگی‌های مجموعه‌داده Pima Indian diabetes

ردیف	هاویژگی	مشخصات	نوع	مقدار
۱	Pregnancies	تعداد دفعات بارداری	صحیح	۰ -۱۷
۲	Glucose	قند خون دو ساعت بعد صبحانه (Mg/dl)	صحیح	۴۴ - ۱۹۹
۳	Blood Pressure	فشار خون دیاستولیک $(mmHg)$	صحیح	۲۴ – ۱۲۲
۴	Skin Thickness	ضخامت پوست عضله سه سربازو	صحیح	۷ - ۹۹
۵	Insulin	میزان انسولین ناشتا در خون$(\mu U/ml)$	صحیح	۸۴۶ ۱۴-
۶	BMI	شاخص توده بدنی	اعشاری	۱۸,۲ – ۶۷,۱
۷	Diabetes Pedigree Function	سابقه خانوادگی	اعشاری	۰,۰۷۸ – ۲,۴۲
۸	Age	سن (سال)	صحیح	81 ۲۱ -

2- مجموعه‌داده Mendeley Diabetes [۴۷]، که شامل اطلاعات بیماران دیابتی در عراق است، از آزمایشگاه بیمارستان مدیکال سیتی، مرکز تخصصی غدد درون‌ریز و بیمارستان آموزشی دیابت الکندی جمع‌آوری شده است. این مجموعه‌داده شامل ۱۰۰۰ نمونه با ۱۱ ویژگی برای هر نمونه و ۳ کلاس (دیابتی، غیردیابتی و پیش‌دیابتی) است. جزئیات ویژگی‌ها در

۳- جدول **۲** ارائه شده است.

جدول ۲. جزئیات ویژگی‌های مجموعه‌داده Mendeley Diabetes.

ردیف	هاویژگی	مشخصات	نوع	مقدار
۱	Gender	جنسیت	باینری	مرد / زن
۲	AGE	سن	عددی	۲۰ - ۷۹
۳	Urea	اوره (Mg/dl)	عددی	۰,۵-۳۸,۹
۴	Cr	نسبت کراتینین $(\mu mol/L)$	عددی	۶ –۸۰۰
۵	HbA1c	HbA1c$(mmol/L)$	عددی	۰,۹ – ۱۶
۶	Chol	کلسترول $(mmol/L)$	عددی	۰,۵ – ۱۰,۳
۷	TG	گلیسیرید $(mmol/L)$تری	عددی	۰,۳ – ۱۳,۸
۸	HDL	HDL$(mmol/L)$کلسترول	عددی	۰,۲ – ۹,۹
۹	LDL	LDL$(mmol/L)$کلسترول	عددی	۰,۳ – ۹,۹
۱۰	VLDL	VLDL$(mmol/L)$	عددی	۰,۱ – ۳۵
۱۱	BMI	شاخص توده بدنی	عددی	۱۹ -۴۷,۵

پیش‌پردازش

برای پیش‌پردازش داده‌ها، گام‌های زیر را می‌توان انجام داد: ۱- کدگذاری داده‌های طبقه‌بندی شده[1]۲- پر کردن داده‌های گمشده[2]۳- نرمالیزه کردن داده‌ها و ۴- حذف داده‌های پرت.

1 Encoding Categorical Data

در ابتدا، برای داده‌های طبقه‌بندی‌شده غیرعددی، آنها را به فرمت عددی تبدیل کردیم تا الگوریتم‌های یادگیری ماشین بتوانند از آنها استفاده کنند [۴۸].

در مجموعه‌داده‌های واقعی، وجود داده‌های گمشده امری رایج است. برای داده‌های عددی، می‌توان از روش‌هایی مانند میانگین یا میانه برای پر کردن داده‌های گمشده استفاده‌کرد. برای داده‌های طبقه‌بندی شده، می‌توان از روش‌های جایگزینی مانند مد (مقدار رایج‌ترین) یا الگوریتم‌های پیچیده‌تر مانند K نزدیک‌ترین همسایه (KNN) یا رگرسیون لجستیک بهره برد [۴۹]. در این پژوهش، از روش KNN با مقدار K برابر با ۵ برای جایگزینی داده‌های گمشده در مجموعه‌داده Pima Indian Diabetes استفاده شد.

علاوه براین، داده‌های عددی ممکن است در محدوده‌های مختلفی قرار داشته باشند که می‌تواند بر عملکرد الگوریتم‌های یادگیری ماشین تأثیر بگذارد. برای رفع این مشکل، می‌توان از تکنیک‌های نرمالیزه‌کردن مانند استانداردسازی (Z-score normalization) یا Min-Max scaling استفاده کرد [۵۰]. در این مطالعه، از روش Z-score برای نرمالیزه‌کردن تمامی مجموعه‌داده‌ها استفاده شد.

همچنین۱۷۴ نمونه تکراری از مجموعه داده دیابت Mendeley که شامل اطلاعات بیماران دیابتی بود، حذف شد و ۸۲۶ نمونه باقی ماند.

انتخاب ویژگی

مرحله بعد از پیش‌پردازش داده‌ها، انتخاب ویژگی است. هدف از این مرحله، شناسایی و انتخاب زیرمجموعه‌ای از ویژگی‌های مرتبط و مهم برای ارتقای عملکرد مدل طبقه‌بندی است. روش‌های انتخاب ویژگی wrapper، یک الگوریتم یادگیری ماشین را به‌عنوان یک جعبه سیاه در نظر می‌گیرند و با استفاده از آن، زیرمجموعه‌ای از ویژگی‌ها را در یک فرآیند جستجوی هدایت‌شده ارزیابی می‌کنند. این روش‌ها به‌طور کلی عملکرد بهتری نسبت به روش‌های فیلتر ارائه می‌دهند، اما از نظر محاسباتی پرهزینه‌تر هستند. روش‌های wrapper اثرات متقابل بین ویژگی‌ها و وابستگی آنها به الگوریتم یادگیری خاص را در نظر می‌گیرند. این روش‌ها همچنین دارای دقت، انعطاف‌پذیری و قابلیت تفسیر بالایی هستند [۵۱]. به‌دلیل مزایای فوق، در این پژوهش از روش‌های wrapper برای انتخاب ویژگی استفاده می‌کنیم. روش‌های مورد استفاده در این پژوهش عبارتند از: انتخاب رو به جلو [۲۲]، حذف معکوس [۳۲]، ERSFS [۳۳]، SFE [۳۴] و SPFSR [۳۵].

الگوریتم‌های یادگیری ماشین

1 Missing Value Imputation

2 Data Normalization

3 Outlier Removal

در این پژوهش، جهت بررسی جامع‌تر از انواع مدل‌های یادگیری ماشین شامل طبقه‌بند K-نزدیک‌ترین همسایه (KNN)، ماشین بردار پشتیبان (SVM)، درخت تصمیم J48، جنگل تصادفی[1] و شبکه‌های عصبی چندلایه پرسپترون (MLP) استفاده شده است.

الگوریتم K نزدیک‌ترین همسایه (KNN) یکی از ساده‌ترین و در عین حال کارآمدترین روش‌های یادگیری ماشین برای مسائل طبقه‌بندی است. این الگوریتم، داده‌های جدید را بر اساس شباهت به نمونه‌های آموزشی موجود دسته‌بندی می‌کند. به این منظور، KNN داده جدیدی را که قصد طبقه‌بندی آن را دارد، با داده‌های موجود در مجموعه‌داده آموزشی مقایسه می‌کند و بر اساس شباهت یا فاصله آن نمونه با K تعداد نزدیک‌ترین همسایه در مجموعه آموزشی، آن را در یکی از طبقات موجود جای می‌دهد [۵۲].

ماشین بردار پشتیبان (SVM) [۱۱]، یکی دیگر از الگوریتم‌های رایج و کارآمد برای طبقه‌بندی داده‌ها در یادگیری ماشین است. این روش بر پایه ایده جداسازی خطی دسته‌های داده با استفاده از یک سطح (هایپرصفحه) تصمیم‌گیری بهینه استوار است. روش SVM تلاش می‌کند این هایپرصفحه را به‌گونه‌ای انتخاب کند که بیشترین فاصله را با نزدیک‌ترین نمونه‌های آموزشی، که به‌عنوان بردارهای پشتیبان شناخته می‌شوند، داشته باشد. این امر باعث می‌شود SVM از مقاومت خوبی در برابر نویز برخوردار گردد [۵۳].

جنگل‌های تصادفی مجموعه‌ای از مدل‌های مبتنی بر درخت هستند که در آن، پیش‌بینی‌های مختلفی از طریق درخت‌های مستقل از یکدیگر، بر پایه مقادیر یک بردار تصادفی انجام می‌شود. این بردار تصادفی با توزیعی که برای همه درخت‌های درون جنگل یکسان است، محاسبه می‌شود [۳۰].

طبقه‌بندی‌کننده J48، نسخه‌ای از الگوریتم درخت تصمیم طبقه‌بندی C4.5 است که درخت‌های دودویی تولید می‌کنند. در این روش، یک درخت تصمیم‌گیری بر اساس مجموعه‌داده‌های آموزشی ساخته می‌شود. هر گره در درخت، یک ویژگی را نمایش می‌دهد و شاخه‌های آن مقادیر مختلف آن ویژگی را نشان می‌دهند. الگوریتم به‌طور متوالی بهترین ویژگی را بر اساس معیاری مانند آنتروپی یا گین اطلاعاتی برای جداسازی دسته‌ها انتخاب می‌کند. این فرآیند تا زمانی که تمام نمونه‌ها طبقه‌بندی شوند یا معیاری برای توقف ساخت درخت برآورده شود،

ادامه می‌یابد. برای طبقه‌بندی نمونه‌های جدید، الگوریتم از ریشه درخت شروع می‌کند و با بررسی ویژگی‌های آن نمونه در گره‌های مختلف، از شاخه‌های مرتبط پایین می‌آید تا به یک برگ (گره نهایی) برسد که دسته مورد نظر را مشخص می‌نماید [۵۴].

شبکه‌های عصبی چندلایه یا MLP[2] نوعی از شبکه‌های عصبی پیش‌خور[1] هستند که برای آموزش الگوریتم‌های یادگیری عمیق به کار می‌روند. این شبکه‌ها از چندین لایه گره یا نورون محاسباتی تشکیل شده‌اند که به‌صورت

1 Random Forest

2Multilayer Perceptron

پی‌درپی به هم متصل هستند. به‌دلیل وجود لایه‌های متعدد در یک MLP، می‌توان آن را به‌عنوان یک رویکرد یادگیری عمیق در نظر گرفت که معمولاً برای حل مسائل یادگیری تحت نظارت استفاده می‌شود [۵۵].

روش انجام آزمایش

در این پژوهش، مدل‌ها با استفاده از روش اعتبارسنجی متقاطع پنج‌برابر (5-Fold Cross-Validation) ارزیابی می‌شوند و عملکرد آنها با استفاده از معیارهای مختلفی مانند دقت[۲]،حساسیت[۳]،ویژگی[۴]،صحت[۵]،F1-score و مساحت زیر منحنی ROC (AUC-ROC) مورد بررسی قرار می‌گیرد [۱۴]. این معیارها بر اساس ماتریس آشفتگی محاسبه می‌شوند که یک ماتریس دو بعدی است و مقادیر واقعی کلاس‌ها را با پیش‌بینی‌های مدل مقایسه می‌کند. در این ماتریس، مثبت واقعی (TP[۶]) به تعداد نمونه‌های مبتلا به بیماری که به‌درستی تشخیص داده شده‌اند، مثبت کاذب (FP[۷]) به تعداد نمونه‌های سالم که به‌اشتباه مبتلا تشخیص داده شده‌اند، منفی کاذب (FN[۸]) به تعداد نمونه‌های مبتلا به بیماری که به‌اشتباه سالم تشخیص داده شده‌اند و منفی واقعی (TN[۹]) به تعداد نمونه‌های سالم که به‌درستی تشخیص داده شده‌اند، اشاره دارد [۹]. شاخص‌های F1-Score و AUC-ROC ارزیابی‌های جامع‌تری از عملکرد مدل‌ها ارائه می‌دهند و ارزش کلی آنها را مشخص می‌کنند [۵۶]. در این مقاله، الگوریتم‌های انتخاب ویژگی، یادگیری ماشین و ارزیابی با استفاده از زبان برنامه‌نویسی پایتون پیاده‌سازی شدند.

جدول ۳. شاخص‌های ارزیابی [۱۴].

معیار عملکرد	محاسبه
دقت (Accuracy)	$(TP + TN) / (TP + FP + TN + FN)$
صحت(Precision)	$TP / (TP + FP)$
حساسیت(Sensitivity)	$TP / (TP + FN)$
ویژگی(Specificity)	$TN / (FP + TN)$
امتیاز F1 (F1-score)	$2 * (\text{Precision} * \text{Sensitivity}) / (\text{Precision} + \text{Sensitivity})$

[۱] Feedforward Neural Network
[۲] Accuracy
[۳] Sensitivity
[۴] Specificity
[۵] Precision
[۶] True Positive
[۷] False Positive
[۸] False Negative
[۹] True Negative

نتایج

در این پژوهش، از دو مجموعه‌داده استاندارد Pima Indian Diabetes و Mendeley Diabetes برای ارزیابی عملکرد پنج روش انتخاب ویژگی wrapper در تشخیص بیماری دیابت نوع ۲ استفاده گردید. ابتدا، پنج روش انتخاب ویژگی شامل SFS، RFE، ERSFS، SFE و SPFSR بر روی هر دو مجموعه‌داده اعمال شد. سپس، عملکرد ویژگی‌های انتخاب‌شده توسط هر روش با استفاده از پنج روش یادگیری ماشینی مختلف شامل KNN، SVM، J48، RF، MLP و با استفاده از تکنیک اعتبارسنجی متقاطع پنج برابری ارزیابی گردید.

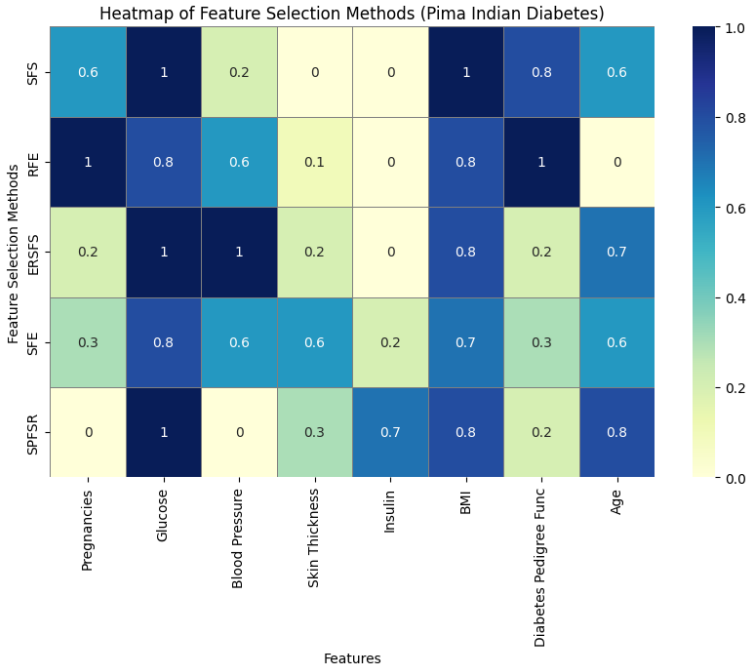
پس از اجرای تمامی روش‌های انتخاب ویژگی و ارزیابی بر اساس معیار F1، بهترین نتایج با توجه به تعداد ویژگی‌های انتخاب‌شده بر اساس میانگین ۲۰ اجرا تعیین و در جدول ۴ و جدول ۵ ارائه شد. نتایج نشان داد که الگوریتم ERSFS به‌طور میانگین کمترین تعداد ویژگی را انتخاب می‌کند، در حالی که روش‌های SFS و RFE به‌طور میانگین بیشترین تعداد ویژگی را انتخاب می‌کنند.

جدول ۴. نتایج انتخاب ویژگی برای مجموعه‌داده ۱ (Pima Indian diabetes)

روش انتخاب ویژگی	تعداد ویژگی های انتخاب شده	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age
SFS	۵	❊	❊				❊	❊	❊
RFE	۵	❊	❊	❊			❊	❊	
ERSFS	۴		❊	❊			❊		❊
SFE	۵		❊	❊	❊		❊		❊
SPFSR	۴		❊			❊	❊		❊

با توجه به یافته‌های جدول ۴، گلوکز، شاخص توده بدنی، سن و فشار خون به عنوان مهم‌ترین ویژگی‌ها برای تشخیص بیماری دیابت نوع ۲ در مجموعه‌داده Pima شناخته شده‌اند. این ویژگی‌ها توسط اکثر روش‌های انتخاب ویژگی انتخاب شده‌اند، که نشان‌دهنده اهمیت آنها در تمایز افراد مبتلا به دیابت از افراد سالم است. علاوه بر این، ویژگی‌های دیگری نیز وجود دارند که می‌توانند در تشخیص دیابت نوع ۲ نقش داشته باشند. این ویژگی‌ها شامل سابقه خانوادگی، تعداد دفعات بارداری، میزان انسولین و ضخامت پوست می‌شوند.

نمودار حرارتی ویژگی‌های انتخاب‌شده بر اساس میانگین ۲۰ بار اجرا در شکل ۲ نشان داده شده است.



شکل ۲. نمودار حرارتی نتایج انتخاب ویژگی برای مجموعه‌داده ۱ (Pima Indian diabetes).

نمودار حرارتی شکل ۲ نشان می‌دهد که روش‌های مختلف انتخاب ویژگی تمایل به انتخاب ویژگی‌های خاصی دارند که می‌تواند نشان‌دهنده اهمیت آن‌ها در پیش‌بینی دیابت نوع ۲ باشد. ویژگی گلوکز، مستقیماً با سطح قند خون و عملکرد انسولین مرتبط است و در تمامی روش‌ها انتخاب شده است، که بر اهمیت بالای آن در پیش‌بینی بیماری تأکید دارد. به‌طور مشابه، ویژگی BMI نیز به دلیل ارتباط آن با چاقی و مقاومت به انسولین، نقش مهمی در تشخیص و پیش‌بینی دیابت نوع ۲ ایفا می‌کند که نشان‌دهنده تأثیر قوی این عوامل در پیش‌بینی دیابت است. سن به‌عنوان یک عامل خطر مهم، نشان‌دهنده افزایش احتمال دیابت با بالاتر رفتن سن و تغییرات متابولیکی است. در مقابل، ویژگی‌هایی مانند ضخامت پوست و میزان انسولین کمتر انتخاب شده‌اند، که ممکن است به نقش کم‌اهمیت‌تر آن‌ها یا همبستگی بالای آن‌ها با سایر ویژگی‌ها اشاره داشته باشد.

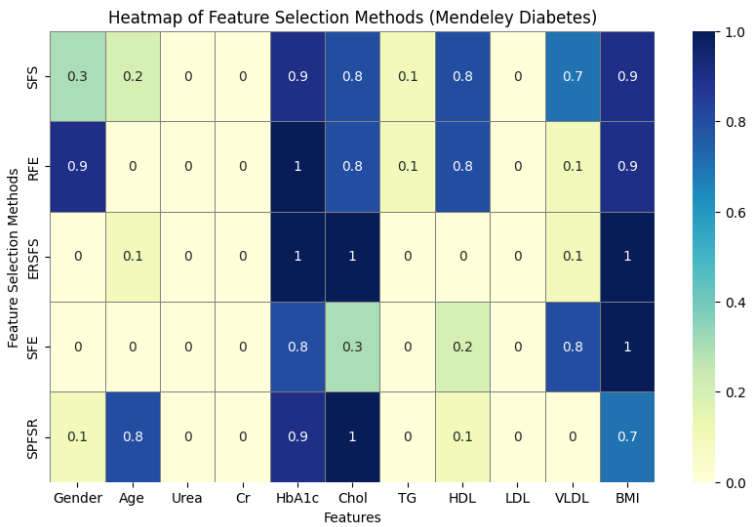
این تحلیل نشان می‌دهد که برخی ویژگی‌ها به دلیل اهمیت پیش‌بینی بیشتر، اولویت بالاتری در انتخاب توسط الگوریتم‌ها دارند، در حالی که برخی دیگر ممکن است تأثیر محدودی داشته باشند.

جدول ۵. نتایج انتخاب ویژگی برای مجموعه‌داده ۲ (Mendeley Diabetes).

روش انتخاب ویژگی	تعداد ویژگی های شده‌انتخاب	Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI
SFS	۵					❋	❋		❋		❋	❋
RFE	۵	❋				❋	❋		❋			❋
ERSFS	۳					❋	❋					❋
SFE	۳					❋					❋	❋
SPFSR	۴		❋			❋	❋					❋

طبق نتایج جدول ۵، در مجموعه‌داده Mendeley، ویژگی‌های HbA1c و شاخص توده بدنی (BMI) توسط تمامی روش‌های انتخاب ویژگی و کلسترول (Chol) توسط ۴ روش انتخاب شده‌اند که نشان‌دهنده اهمیت بالای آنها برای پیش‌بینی و تشخیص دیابت نوع ۲ است. علاوه بر این، ویژگی‌های مهم دیگری نیز وجود دارند که شامل VLDL، HDL، جنسیت و سن می‌شوند که توسط تعداد کمتری از الگوریتم‌ها انتخاب شده‌اند.

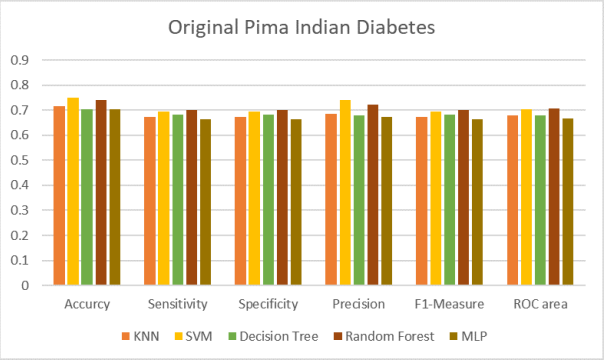
نمودار حرارتی ویژگی‌های انتخاب‌شده بر اساس میانگین ۲۰ بار اجرا در شکل ۳ نشان داده شده است.



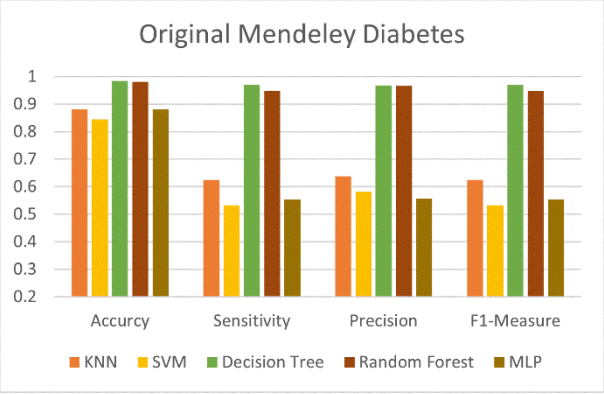
شکل ۳. نمودار حرارتی نتایج انتخاب ویژگی برای مجموعه‌داده ۲ (Mendeley Diabetes).

در شکل ۳ که نشان‌دهنده انتخاب ویژگی‌ها توسط روش‌های مختلف است، مشاهده می‌شود که ویژگی‌هایی مانند HbA1c، BMI و کلسترول به طور مکرر توسط اکثر روش‌های انتخاب ویژگی (مانند SFS، RFE، SFE، و SPFSR) برگزیده شده‌اند. این تکرار نشان‌دهنده اهمیت بالای این ویژگی‌ها در پیش‌بینی دیابت نوع ۲ است، زیرا این ویژگی‌ها به طور مستقیم با شاخص‌های متابولیکی و ریسک بیماری ، مانند مقاومت به انسولین و اختلالات قند خون، مرتبط هستند. در مقابل، ویژگی‌هایی مانند سن و VLDL کمتر توسط این روش‌ها انتخاب شده‌اند، که احتمالاً به دلیل نقش کم‌اثرتر آن‌ها در پیش‌بینی بیماری یا وجود همبستگی بالا با سایر ویژگی‌ها است. این تحلیل نشان می‌دهد که روش‌های مختلف انتخاب ویژگی بر شناسایی عواملی تمرکز دارند که اطلاعات افزوده بیشتری نسبت به سایرین ارائه می‌دهند و ارتباط قوی‌تری با متغیر هدف دارند، در حالی که ویژگی‌های کم‌اثر یا تکراری را حذف می‌کنند.

در شکل ۴، نتایج ارزیابی عملکرد الگوریتم‌های یادگیری ماشین قبل از انتخاب ویژگی برای هر دو مجموعه‌داده ارائه شده‌است.



الف) مجموعه‌داده Pima Indian

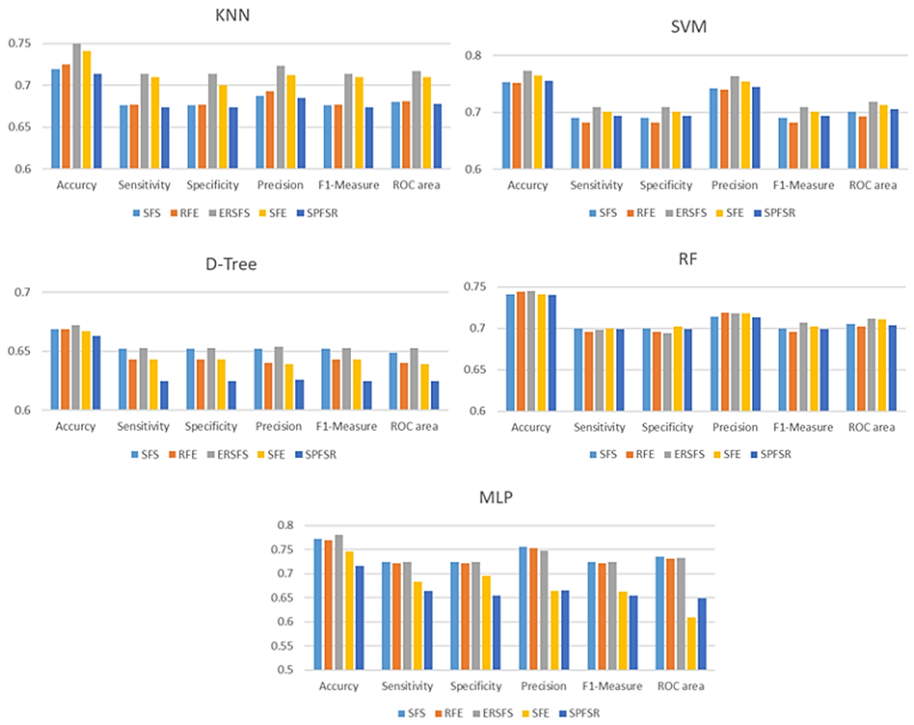


ب) مجموعه‌داده Mendeley

شکل ۴. نتایج اجرای الگوریتم‌های یادگیری ماشین قبل از انتخاب ویژگی.

در مجموعه‌داده Pima، الگوریتم SVM با دقت (Accuracy) ٪۷۵ و صحت (Precision) ٪۷۴ عملکرد قابل قبولی دارد. همچنین، الگوریتم Random Forest با F1-measure معادل ٪۶۹,۹ عملکرد بهتری نسبت به سایر الگوریتم‌ها نشان می‌دهد. در مجموعه‌داده Mendeley نیز، الگوریتم درخت تصمیم و جنگل تصادفی بهترین عملکرد را از خود نشان می‌دهند.

شکل ۵ نمایانگر عملکرد مختلف روش‌های یادگیری ماشین پس از اعمال پنج روش مختلف انتخاب ویژگی بر روی مجموعه‌داده Pima است.

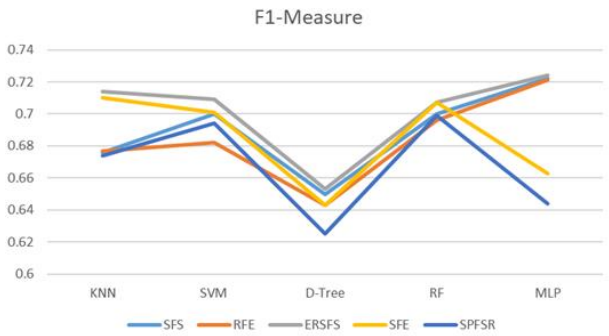


شکل ۵. نتایج حاصل از اعمال الگوریتم‌های مختلف یادگیری ماشین پس از انتخاب ویژگی بر روی مجموعه‌داده Pima Indian diabetes.

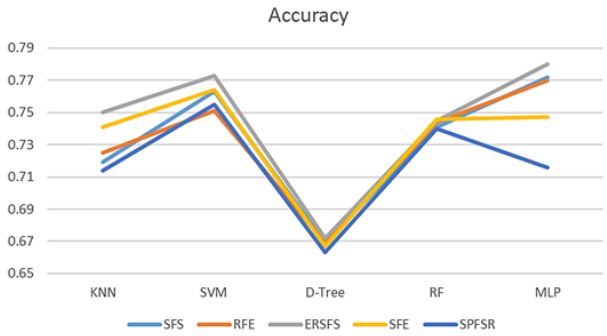
بررسی روش‌های مختلف انتخاب ویژگی طبق شکل ۵ نشان می‌دهد که روش انتخاب زیرمجموعه ویژگی تصادفی (ERSFS) مؤثرترین است. نتایج نشان می‌دهند که ویژگی‌های انتخاب‌شده توسط این روش (گلوکز،

شاخص توده بدنی، سن و فشار خون) برای تمامی الگوریتم‌های یادگیری ماشین عملکرد بسیار خوبی را به ارمغان می‌آورند. به طوری که ERSFS بالاترین رتبه شاخص‌های عملکرد را در مدل‌های KNN، SVM و درخت تصمیم کسب می‌کند. همچنین، این روش بالاترین سطح دقت ۷۸٪ و اندازه F1 معادل ۷۱٪ را در میان تمامی مدل‌ها ارائه می‌دهد.

پس از ERSFS، روش‌های انتخاب ویژگی RFE و SFS برای مدل‌های درخت تصمیم، جنگل تصادفی و MLP نتایج امیدوارکننده‌ای را ارائه می‌دهند. به‌عنوان مثال، روش SFS در مدل MLP بالاترین رتبه شاخص‌های عملکرد را در میان تمام تکنیک‌های انتخاب ویژگی به‌دست می‌آورد. در مدل RF نیز نتایج حاصل، به‌ویژه در شاخص‌های حساسیت و specificity، با مقدار ۷۱٪ قابل قبول است. همچنین روش‌های SFE و SPFSR نیز عملکرد مطلوبی، به‌ویژه در مدل‌های SVM و جنگل تصادفی از خود نشان می‌دهند. بنابراین، سایر ویژگی‌های انتخابی این روش‌ها (سابقه خانوادگی، تعداد دفعات بارداری، میزان انسولین و ضخامت پوست) نیز می‌توانند احتمال تشخیص دیابت نوع ۲ را در مراحل اولیه افزایش دهند. پژوهش‌های پیشین [۴۲, ۴۳] در مجموعه داده Pima، یافته‌های این پژوهش که شامل انتخاب ویژگی‌های گلوکز، فشار خون، BMI و سن است را تأیید می‌کنند.



الف) نتایج محاسبه F1-Measeare

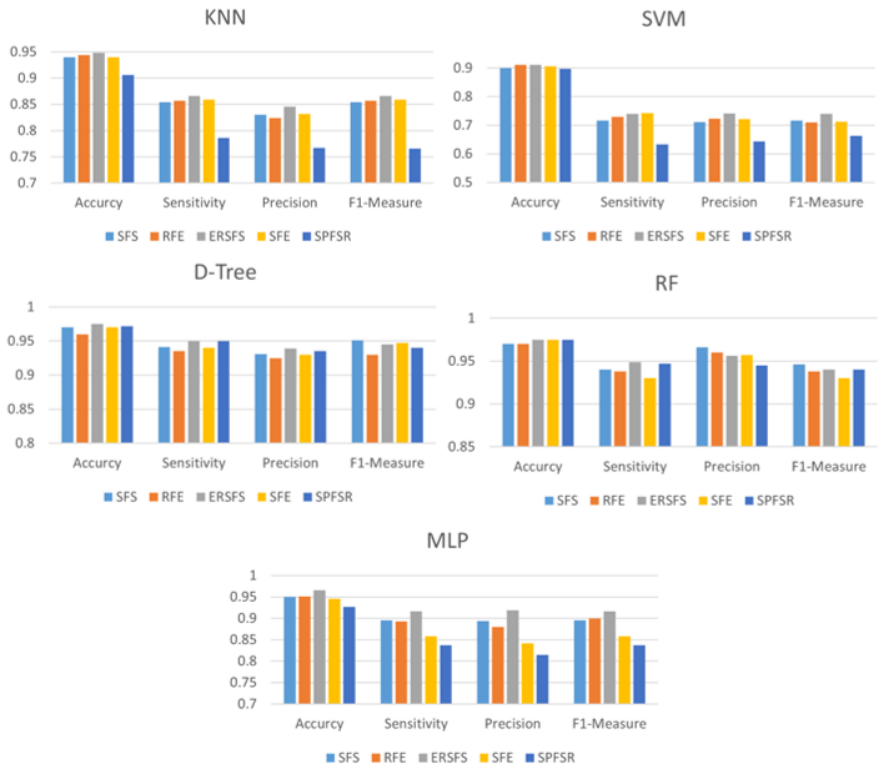


ب) نتایج محاسبه Accuracy

شکل ۶. نتایج F1-Measure و Accuacy پس از انتخاب ویژگی در مجموعه‌داده Pima Indian Diabetes.

شکل ۶ دقت و F1-measure مدل‌های مختلف یادگیری ماشین را پس از انتخاب ویژگی نشان می‌دهد. بر اساس نتایج به دست آمده از شکل ۵ و شکل ۶، می‌توان نتیجه گرفت که ویژگی‌های انتخاب‌شده توسط روش ERSFS (گلوکز، فشار خون، سن و BMI) می‌توانند به عنوان مؤثرترین ویژگی‌ها برای تشخیص دیابت نوع ۲ در مجموعه‌داده Pima Indian Diabetes در نظر گرفته شوند.

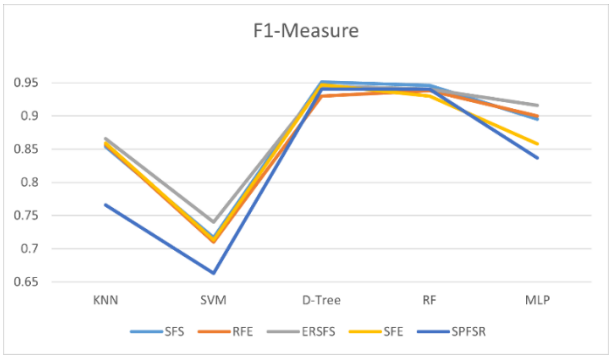
شکل ۷ کارایی روش‌های مختلف انتخاب ویژگی را بر روی مجموعه‌داده Mendeley به نمایش می‌گذارد. این شکل نشان می‌دهد که فرآیند انتخاب ویژگی به منظور تشخیص دقیق‌تر دیابت نوع ۲، نتایج قابل‌توجهی را به همراه دارد.



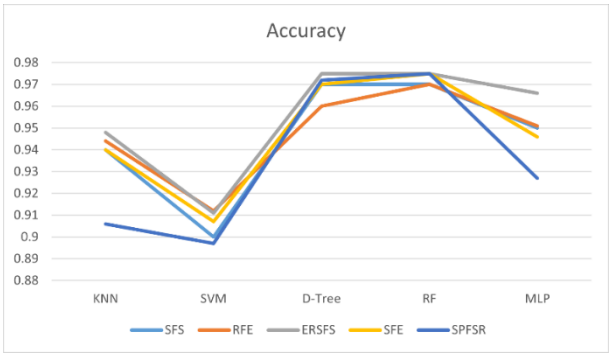
شکل ۷. نتایج حاصل از اعمال الگوریتم‌های مختلف یادگیری ماشین پس از انتخاب ویژگی بر روی مجموعه‌داده Mendeley

با بررسی روش‌های مختلف انتخاب ویژگی طبق شکل ۷، نتایج نشان می‌دهد که ویژگی‌های انتخاب‌شده توسط روش ERSFS (HbA1c، BMI و کلسترول) عملکرد بسیار خوبی را برای طبقه‌بندی‌کننده‌های KNN، SVM و MLP ارائه می‌دهند و در تمامی معیارها بهترین نتایج را به دست می‌آورند. این روش حتی بالاترین دقت ۹۷٫۵٪ و حساسیت ۹۵٪ را در مدل درخت تصمیم به دست آورده است. پس ازERSFS، روش‌های انتخاب ویژگی SFE، RFE و SFS نیز برای مدل‌های درخت تصمیم، جنگل تصادفی و MLP نتایج قابل قبولی ارائه می‌دهند. به عنوان مثال، روش‌های SFS و RFE در مدل MLP بالاترین رتبه را در شاخص‌های عملکرد در میان تمام تکنیک‌های انتخاب ویژگی پس از ERSFS به دست آورده‌اند. در مدل‌های RF و درخت تصمیم، نتایج حاصل از SPFSR بسیار خوب بوده و به ویژه در شاخص‌های دقت، حساسیت و F1-measure نتایج قابل قبولی دارند. همچنین، روش SFE نیز عملکرد مطلوبی، به ویژه در مدل‌های SVM وKNN، از خود نشان

می‌دهد. بنابراین می‌توان سایر ویژگی‌های انتخابی این روش‌ها را (VLDL، HDL، جنسیت و سن) به عنوان پیش‌بینی‌کننده‌های بالقوه‌ی دیابت نوع ۲ در نظر گرفت.



الف) نتایج محاسبه F1-Measeare



ب) نتایج محاسبه Accuracy

شکل ۸. نتایج محاسبه F1-Measure و Accuracy پس از انتخاب ویژگی در مجموعه‌داده Mendeley.

شکل ۸، دقت و F1-measure مدل‌های مختلف یادگیری ماشین را پس از انتخاب ویژگی نشان می‌دهد. این شکل نشان می‌دهد که روش ERSFS بالاترین سطح دقت و F1-measure را در بین تمام مدل‌ها ارائه می‌دهد. همچنین، مدل‌های درخت تصمیم و جنگل تصادفی بهترین نتایج را در مجموعه‌داده Mendeley به دست می‌دهند. با توجه به یافته‌ها، HbA1c، Chol و BMI بهترین ویژگی‌ها برای طبقه‌بندی افراد مبتلا به دیابت نوع ۲ براساس مجموعه‌داده Mendeley هستند. دقت بالای روش‌های طبقه‌بندی موردبحث، گواهی بر انتخاب مناسب این ویژگی‌ها می‌باشد.

بحث و نتیجه‌گیری

این مطالعه به بررسی اثر روش‌های انتخاب ویژگی Wrapper بر عملکرد الگوریتم‌های یادگیری ماشین می‌پردازد. دو مجموعه‌داده Pima Indian Diabetes و Mendeley Diabetes برای شناسایی ویژگی‌های موثر در تشخیص سریع‌تر دیابت نوع ۲ مورد استفاده قرار گرفتند. ابتدا، الگوریتم‌های یادگیری ماشین بدون اعمال روش‌های انتخاب ویژگی بر روی مجموعه‌داده‌ها اجرا شدند. نتایج نشان داد که الگوریتم‌های SVM و جنگل تصادفی در مجموعه‌داده Pima به ترتیب با دقت‌های ۷۵ و ۷۵,۱ درصد و الگوریتم‌های درخت تصمیم و جنگل تصادفی در مجموعه‌داده Mendeley با دقت‌های ۹۸ و ۹۷,۸ درصد عملکرد بهتری داشتند. سپس، از پنج روش انتخاب ویژگی شامل SFS، RFE، ERSFS، SFE و SPFSR استفاده شد.

در مجموعه‌داده Pima، روش‌های ERSFS و SPFSR هر کدام ۴ ویژگی انتخاب کرده‌اند، در حالی که سایر روش‌ها ۵ ویژگی را برگزیده‌اند. هیچ‌یک از الگوریتم‌ها ویژگی‌های کاملاً یکسانی را انتخاب نکرده‌اند. روش ERSFS بالاترین دقت را با انتخاب ۴ ویژگی در مدل SVM و میزان ۷۷,۳٪ به دست آورده است. همچنین، این روش بالاترین حساسیت و ویژگی (Specificity) را با مقدار ۰,۷۲۴، و بالاترین مقادیر F1 و ROC را به‌ترتیب با ۰,۷۲۴ و ۰,۷۳۲ در مدل MLP ثبت کرده است. علاوه بر این، روش ERSFS بالاترین مقدار Precision را در مدل SVM با میزان ۰,۷۶۳ ارائه داده است. با این حال، در سایر روش‌ها با انتخاب تعداد بیشتر (۵ ویژگی)، هیچ بهبودی در عملکرد مشاهده نمی‌شود و حتی برخی از آن‌ها عملکرد را کاهش می‌دهند. روش SFE چهار ویژگی انتخابی مشابه ERSFS دارد و انتخاب ویژگی پنجم (ضخامت پوست عضله سه سربازو) هیچ بهبودی در عملکردها ایجاد نمی‌کند. همچنین، روش SPFSR نیز دارای ۳ ویژگی انتخابی مشابه ERSFS (BMI، Chol و Age) و یک ویژگی متفاوت (Insulin) است که باعث کاهش عملکرد آن در مقایسه با روش ERSFS می‌شود. در تمامی روش‌های انتخاب ویژگی، حداقل ۳ ویژگی انتخابی آنها با روش ERSFS مشترک است که اهمیت ویژگی‌های انتخابی این روش را تأکید می‌کند. بنابراین، می‌توان نتیجه گرفت که ویژگی‌های انتخابی توسط روش ERSFS (یعنی گلوکز، شاخص توده بدنی، سن و فشار خون) بهترین عملکرد را در مدل‌های مختلف یادگیری ماشین ایجاد می‌کنند و می‌توانند نقش مؤثری در تشخیص سریع‌تر بیماری دیابت نوع ۲ داشته باشند.

در مجموعه‌داده Mendeley، روش‌های ERSFS و SFE به‌ترتیب ۳ ویژگی، روش‌های SFE و SPFSR، ۴ ویژگی، و روش‌های SFS و RFE نیز ۵ ویژگی انتخاب کرده‌اند. هیچ‌یک از الگوریتم‌ها ویژگی‌های کاملاً مشابهی را برنگزیده‌اند. بالاترین سطح عملکرد، به‌ویژه دقت، توسط روش ERSFS و با استفاده از الگوریتم‌های درخت تصمیم و جنگل تصادفی به‌ترتیب با مقادیر ۹۷,۸٪ و ۹۸٪ به دست آمده است. همچنین، این روش بالاترین حساسیت را با مقدار ۰,۹۵ و بالاترین مقدار F1 را با مقدار ۰,۹۴۵ در مدل درخت تصمیم، و بالاترین ویژگی (Specificity) را در مدل جنگل تصادفی با مقدار ۰,۹۵۶ ثبت کرده است. با این حال، در سایر روش‌ها با انتخاب تعداد بیشتر (۵ و ۴ ویژگی)، هیچ بهبودی در عملکرد مشاهده نشد و حتی شاهد کاهش میزان عملکرد

بودیم. ویژگی‌های انتخاب‌شده توسط روش‌های SFS، RFE و SPFSR شامل ویژگی‌های انتخاب‌شده توسط ERSFS (HbA1c، BMI و Chol) می‌شوند. روش SFE، دو ویژگی مشابه ERSFS دارد و با انتخاب ویژگی VLDL، نتایج بسیار نزدیکی به ERSFS ارائه می‌دهد. روش SPFSR نیز دارای ۳ ویژگی مشابه ERSFS و یک ویژگی دیگر (سن) است که عملکرد خوبی را به ویژه در مدل‌های یادگیری ماشین درخت تصمیم و جنگل تصادفی به ارمغان آورده است. در تمام روش‌های انتخاب ویژگی به کار گرفته شده، حداقل ۳ ویژگی انتخابی با ERSFS مشترک است که نشان‌دهنده اهمیت ویژگی‌های انتخاب‌شده توسط این روش است.

به طور کلی، می‌توان نتیجه گرفت که ویژگی‌های انتخاب‌شده شامل گلوکز، BMI، سن و فشار خون در مجموعه‌داده Pima و HbA1c، BMI و کلسترول در مجموعه‌داده Mendeley، در مدل‌های مختلف یادگیری ماشین عملکرد بهتری ایجاد می‌کنند. این ویژگی‌ها می‌توانند نقش مؤثری در تشخیص سریع‌تر بیماری دیابت نوع ۲ داشته باشند. این امر به این دلیل است که این ویژگی‌ها اطلاعات مهمی در مورد متابولیسم گلوکز، چاقی و سلامت قلب و عروق ارائه می‌کنند که همگی با خطر ابتلا به دیابت نوع ۲ مرتبط هستند.

به عنوان مثال، گلوکز و HbA1c نشانگرهای اصلی قند خون هستند. سطوح بالای گلوکز خون با مقاومت به انسولین و دیابت نوع ۲ مرتبط است. پارامتر HbA1c میانگین قند خون را در طول چند ماه گذشته نشان می‌دهد و برای نظارت بر کنترل قند خون در افراد مبتلا به دیابت استفاده می‌شود. همچنین BMI شاخصی از وزن بدن نسبت به قد است. چاقی یک عامل خطر شناخته شده برای دیابت نوع ۲ است. سن نیز یک عامل خطر است، زیرا خطر ابتلا به دیابت نوع ۲ با افزایش سن افزایش می‌یابد. فشار خون بالا نیز با خطر ابتلا به دیابت نوع ۲ مرتبط است.

از طرفی یافته‌های پژوهش حاضر، با نتایج مطالعات دیگر همخوانی قابل توجهی دارد. به عنوان مثال، پژوهش پن و همکاران [۳۶] نیز عوامل HbA1c، مدت ابتلا به دیابت، گلوکز پس از غذا، سن و فشار خون سیستولیک را به‌عنوان عوامل کلیدی شناسایی کرده‌اند که با انتخاب ویژگی‌های ما در مجموعه‌داده Pima همخوانی دارد. همچنین، مطالعه اسلام و همکاران [۳۸] که از یادگیری ماشین برای شناسایی عوامل خطر استفاده کرده‌اند، عوامل سن، سطح تحصیلات، وضعیت تأهل، SBP، وضعیت استعمال دخانیات و BMI را به‌عنوان پیش‌بینی‌کننده‌های کلیدی معرفی کرده‌اند که BMI و سن در هر دو پژوهش مشترک هستند. علاوه بر این، مطالعات میتیلی و همکاران [۴۳] و صفایی و همکاران [۴۲] نیز به اهمیت ویژگی‌های گلوکز، BMI و سن در تشخیص دیابت نوع ۲ اشاره کرده‌اند که نشان‌دهنده تأکید مداوم بر این عوامل در تحقیقات مختلف است. این تطابق میان یافته‌های پژوهش ما و مطالعات پیشین، اعتبار نتایج به‌دست‌آمده را تقویت می‌کند و نشان می‌دهد که ویژگی‌های انتخاب‌شده در مدل‌های پیش‌بینی دیابت نوع ۲ دارای اهمیت و قابلیت تعمیم بالایی هستند. بنابراین، این همخوانی با تحقیقات دیگر نشان‌دهنده یکپارچگی علمی و پشتیبانی از اهمیت این ویژگی‌ها در پیش‌بینی دقیق‌تر دیابت نوع ۲ می‌باشد.

مدل‌های یادگیری ماشین می‌توانند از این ویژگی‌ها برای شناسایی افراد در معرض خطر ابتلا به دیابت نوع ۲ و همچنین برای تشخیص زودهنگام این بیماری استفاده‌کنند. به‌عنوان مثال، مدل‌ها می‌توانند برای شناسایی افرادی که سطح قند خون بالا یا BMI بالا دارند استفاده شوند. سپس این افراد می‌توانند برای غربالگری بیشتر دیابت ارجاع داده شوند. تشخیص زودهنگام دیابت نوع ۲ مهم است، زیرا این بیماری را می‌توان با تغییرات سبک زندگی و دارو به‌طور مؤثر مدیریت کرد. همچنین تشخیص زودهنگام بیماری می‌تواند به جلوگیری از عوارض دیابت، مانند بیماری قلبی، سکته مغزی و نارسایی کلیه کمک کند. تمامی نتایج بدست آمده توسط این پژوهش توسط فرد متخصص مورد بررسی و تایید قرارگرفته است.

این مطالعه دارای چند محدودیت است. اول، این مطالعه ویژگی‌های اجتماعی-اقتصادی و سایر ویژگی‌های بالینی مرتبط با سبک زندگی افراد، مانند سیگارکشیدن و فعالیت بدنی را در نظر نگرفته است. این عوامل می‌توانند خطر ابتلا به دیابت نوع ۲ را تحت‌تأثیر قرار دهند و عدم در نظر گرفتن آنها ممکن است بر دقت یافته‌های مطالعه تأثیر بگذارد. به‌عنوان مثال، مطالعات نشان داده‌اند که افراد با وضعیت اجتماعی-اقتصادی پایین‌تر بیشتر در معرض خطر ابتلا به دیابت نوع ۲ هستند. علاوه بر این، افراد سیگاری و افراد کم‌تحرک نیز در معرض خطر ابتلا به این بیماری هستند.

دوم، این مطالعه از یک مجموعه‌داده نسبتاً کوچک استفاده کرده است. این امر ممکن است توانایی تعمیم یافته‌های مطالعه را به جمعیت کلی محدود کند. مطالعات آینده باید از مجموعه‌داده‌های بزرگ‌تر با طیف گسترده‌تری از ویژگی‌ها برای افزایش قابلیت تعمیم‌یافته‌ها استفاده‌کنند.

با وجود این محدودیت‌ها، این مطالعه سهم ارزشمندی در درک ما از عوامل خطر ابتلا به دیابت نوع ۲ دارد. مطالعات آینده باید بر رفع این محدودیت‌ها و بهبود بیشتر درک ما از این بیماری تمرکز کنند.

منابع

[1] Association, A.D. (2014). *Diagnosis and classification of diabetes mellitus. Diabetes care, 37*, 81-90. https://doi.org/10.2337/dc14-S081

[2] Atkinson, M.A., G.S. *Eisenbarth, and A.W. Michels.* (2014). Type 1 diabetes. *The lancet, 383*(9911), 82-69 . https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(13)60591-7/fulltext

[3] Control, C.f.D. and Prevention. (2020). *National diabetes statistics report, 2020. Atlanta, GA: centers for disease control and prevention, US dept of health and human services; 2020.* https://www.cdc.gov/diabetes/php/data-research/index.html

[4] Chatterjee, S., K. Khunti, and M.J. Davies. (2017). Type 2 diabetes. *The lancet*,*389*(10085),2239-2251. https://www.thelancet.com/journals/lancet/article/piiS0140-6736(17)30058-2/fulltext

[5] Forbes, J.M. and M.E. Cooper. (2013). Mechanisms of diabetic complications. *Physiological reviews*, *93*(1), 137-188. https://www.ncbi.nlm.nih.gov/pubmed/23303908

[6] Care, D. (2019). Care in diabetesd2019. *Diabetes care*, *42*(1), S13-S28 . https://pubmed.ncbi.nlm.nih.gov/30559228/

[7] WHO, (2023). *World Health Organization*. Website Name. Retrieved from https://www.who.int/news-room/fact-sheets/detail/diabetes

[8] Federation, I.D. (2022). *Diabetes around the world in 2021*. Retrieved from https://diabetesatlas.org/atlas-reports/

[9] Namjouye Rad, A.a.D., Mahdi (2021). *Detection of network penetration by data mining and using machine learning via SVM algorithm. Karafan Quarterly Scientific Journal*, *17*(4), 13-34. https://karafan.tvu.ac.ir/article_128393_ceb8bbb84a290af623e3744516a42921.pdf

[10] Alipour, M. and M. Jafari. (2022). *Estimating the Dynamic Margin of Voltage Stability in Power Systems Using Machine Learning*. *Karafan Quarterly Scientific Journal*, *19*(3), 221-245. https://karafan.tvu.ac.ir/article_143524_12280c7e82a5860ec12f63c83d2d3df4.pdf

[11] Basiri, M. and F .Fathnejad. (2023). *Presenting a Framework for Intelligent Sentiment Analysis Using a Novel Method of feature Combination and Meta-Initiative in Particle Swarm Optimization. Karafan Quarterly Scientific Journal*, *20*(3), 531-551. https://karafan.tvu.ac.ir/article_178537_64cc67fe36c313c328b09464a29b65e4.pdf

[12] Bahmani, M., M.E. Pourzarandi, and M. Minoei. (2022). *Factors Affecting the Forecast of Stock Returns using Delphi-Fuzzy Knowledge Analysis and Technique. Karafan Quarterly Scientific Journal*, *19*(2) 431-453. https://karafan.tvu.ac.ir/article_148742_241883f182a19b66c85376f090483227.pdf

[13] Gao, L.A., et al. (2022). *Prokaryotic innate immunity through pattern recognition of conserved viral proteins*. *Science*, *377*(6607). https://www.science.org/doi/full/10.1126/science.abm4096

[14] Ghaffarian, H. and A. Bamohabbat. (2023). *Classification and Prediction of Customer Categories Using Combination of LRFM Method, Quartiles and Multi-Class Data Mining Methods*. *Karafan Quarterly Scientific Journal*, *20*(1), 511-532. https://karafan.tvu.ac.ir/article_150022_ed4e7ca8ea509e6d752cb0250d32fc7e.pdf

[15] Latha, C.B.C. and S.C. Jeeva. (2019). *Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked*, *16*(100203). https://www.sciencedirect.com/science/article/pii/S235291481830217X

[16] Ali, Y.A., E.M. Awwad, M. Al-Razgan, and A. Maarouf. (2023). *Hyperparameter search for machine learning algorithms for optimizing the computational complexity*. *Processes*, *11*(2), 349. https://doi.org/10.3390/pr11020349

[17] Pudjihartono, N., T. Fadason, A.W. Kempa-Liehr, and J.M. O'Sullivan. (2022). *A review of feature selection methods for machine learning-based disease risk prediction*. *Frontiers in Bioinformatics*, *2*(927312) . https://www.frontiersin.org/journals/bioinformatics/articles/10.3389/fbinf.2022.927312/full

[18] Crespo Márquez, A. (2022). *The Curse of Dimensionality. In (Ed.), Digital Maintenance Management: Guiding Digital Transformation in Maintenance. Springer International Publishing: Cham*. 67-86. https://link.springer.com/book/10.1007/978-3-030-97660-6

[19] Organization, W.H. (2019). *Classification of diabetes mellitus*. https://apps.who.int/iris/bitstream/handle/10665/325182/9789241515702-eng.pdf

[20] Alaguselvi, R. and K. Murugan. (2024). *A Systematic Review for the Classification and Segmentation of Diabetic Retinopathy Lesion from Fundus. Artificial Intelligence and Machine Learning Techniques in Image Processing and Computer Vision*, 54-74. https://www.taylorfrancis.com/chapters/edit/10.1201/9781003425700-5/systematic-review-classification-segmentation-diabetic-retinopathy-lesion-fundus-alaguselvi-kalpana-murugan

[21] Tigga, N.P. and S. Garg. (2020). Prediction of type 2 diabetes using machine learning classification methods. *Procedia Computer Science*, *167*(706-716). https://doi.org/10.1016/j.procs.2020.03.336

[22] Guyon, I. and A. Elisseeff. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), 1157-1182. https://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf?ref=driverlayer.com/web

[23] Chandrashekar, G. and F. Sahin. (2014). A survey on feature selection methods. *Computers & electrical engineering*, *40*(1), 16-28. https://doi.org/10.1016/j.compeleceng.2013.11.024

[24] Peng, H., F. Long, and C. Ding. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, *27*(8), 1226-1238. https://www.ncbi.nlm.nih.gov/pubmed/16119262

[25] Canayaz, M. (2022). Classification of diabetic retinopathy with feature selection over deep features using nature-inspired wrapper methods. *Applied Soft Computing*, *128*(109462). https://doi.org/10.1016/j.asoc.2022.109462

[26] Gnana, D.A.A., S.A.A. Balamurugan, and E.J. Leavline. (2016). Literature review on feature selection methods for high-dimensional data. *International Journal of Computer Applications*, *136*(1), 9-17 . https://doi.org/10.1109/TEVC.2015.2504420

[27] Xue, B., M. Zhang, W.N. Browne, and X. Yao. (2015). A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on evolutionary computation*, *20*(4), 606-626 .https:// https://doi.org/10.1109/TEVC.2015.2504420

[28] Siham, A., S. Sara, and A. Abdellah. (2021). *Feature selection based on machine learning for credit scoring: An evaluation of filter and embedded methods*. 2021 International conference on innovations in intelligent systems and applications (INISTA), IEEE. https://doi.org/10.1109/INISTA52262.2021.9548410

[29] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267-288 . https://www.jstor.org/stable/2346178

[30] Breiman, L. (2001). Random forests. *Machine learning*, *45*(5-32). https://link.springer.com/article/10.1023/A:1010933404324

[31] Déjean, S., R.T. Ionescu, J. Mothe, and M.Z. Ullah. (2020) .*Forward and backward feature selection for query performance prediction*. Proceedings of the 35th annual ACM symposium on applied computing, https://dl.acm.org/doi/abs/10.1145/3341105.3373904

[32] Kohavi, R. and G.H. John. (1997). Wrappers for feature subset selection. *Artificial intelligence*, *97*(1-2), 273-324. https://www.sciencedirect.com/science/article/pii/S000437029700043X

[33] SabbaghGol, H., H. Saadatfar, and M. Khazaiepoor. (2024). Evolution of the random subset feature selection algorithm for classification problem. *Knowl Based Syst*, *285*(111352). https://doi.org/10.1016/j.knosys.2023.111352

[34] Ahadzadeh, B., et al. (2023). SFE: A Simple, Fast, and Efficient Feature Selection Algorithm for High-Dimensional Data. *IEEE Trans Evol Comput*, *27*(6), 1896-1911 .https://doi.org/10.1109/TEVC.2023.3238420

[35] Akman, D.V., et al. (2023). k-best feature selection and ranking via stochastic approximation. *Expert Syst Appl*, *213*(118864). https://doi.org/10.1016/j.eswa.2022.118864

[36] Pan, H., et al. (2023). A risk prediction model for type 2 diabetes mellitus complicated with retinopathy based on machine learning and its application in health management. *Frontiers in Medicine*, *10*(1136653). https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2023.1136653/full

[37] Thipsawat, S. (2023). Dietary Consumption on Glycemic Control Among Prediabetes: A Review of the Literature. *SAGE Open Nursing*, *9*. https://journals.sagepub.com/doi/abs/10.1177/23779608231218189

[38] Islam, M.M., et al .(2023) .Identification of the risk factors of type 2 diabetes and its prediction using machine learning techniques. *Health Systems*, *12*(2), 243-254. https://www.tandfonline.com/doi/abs/10.1080/20476965.2022.2141141

[39] Fitriyani, N.L., et al. (2023). Performance Analysis and Assessment of Type 2 Diabetes Screening Scores in Patients with Non-Alcoholic Fatty Liver Disease. *Mathematics*, *11*(10), 2266. https://www.mdpi.com/2227-7390/11/10/2266

[40] Zohara, Z., et al. (2023). The prospect of non-alcoholic fatty liver disease in adult patients with metabolic syndrome: a systematic review. *Cureus*, *15*(7), https://pmc.ncbi.nlm.nih.gov/articles/PMC10427027/

[41] Halias, A.F., et al. (2023). *Type 2 Diabetes Mellitus Prediction Using Data Mining Approach*. 2023 IEEE International Conference on Computing (ICOCO), IEEE. https://ieeexplore.ieee.org/abstract/document/10398078/

[42] Safai, M., Safai, Alireza. (2021). *Improving the diagnosis of type 2 diabetes and identifying its effective indicators with the feature selection approach*. In (Ed.), *The 5th International Conference on Electrical Engineering, Electronics and Smart Networks*. https://civilica.com/doc/1257205

[43] R., M., A. Banu.W, and D. Mavaluru. (2020). An efficient feature selection algorithm for health care data analysis. *Bulletin of Electrical Engineering and Informatics*. https://beei.org/index.php/EEI/article/view/1744

[44] Sabbagh Gol, H. (2018). A Detection of Type2 Diabetes using C4.5 Decision Tree. *Journal of Health and Biomedical Informatics*, *5*(2) ,293-303 .http://jhbmi.ir/article-1-281-en.html

[45] Repository, U.M.L. (2017). *Pima Indians Diabetes Database*. In (Ed.), https://www.kaggle.com/uciml/pima-indians-diabetes-database

[46] Dua, D. and C. Graff. (2019). UCI machine learning repository . University of California. *School of Information and Computer Science, Irvine, CA*, https://archive.ics.uci.edu/ml/datasets.php

[47] Rashid, A. (2020). *Diabetes Dataset*. In (Ed.), Mendeley Data: https://doi.org/10.17632/wj9rwkp9c2.1

[48] Cerda, P. and G. Varoquaux, (2020). Encoding high-cardinality string categorical variables. *IEEE Transactions on Knowledge and Data Engineering*, *34*(3), 1164-1176. https://doi.org/10.1109/TKDE.2020.2992529

[49] Donders, A.R.T., G.J. Van Der Heijden, T. Stijnen, and K.G. Moons. (2006). A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, *59*(10), 1087-1091. https://www.ncbi.nlm.nih.gov/pubmed/16980149

[50] Patro, S. and K.K. Sahu. (2015). Normalization: A preprocessing stage. *arXiv preprint arXiv:1503.06462*, https://doi.org/10.48550/arXiv.1503.06462

[51] Bouchlaghem, Y., Y. Akhiat, and S. Amjad. (2022). *Feature selection: a review and comparative study*. E3S Web of Conferences, EDP Sciences. https://doi.org/10.1051/e3sconf/202235101046

[52] Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. *Ann Transl Med*, *4*(11), 218. https://www.ncbi.nlm.nih.gov/pubmed/27386492

[53] Ben-Hur, A. and J. Weston. (2010). A user's guide to support vector machines. *Methods Mol Biol*, *609*(2), 23-39 . https://www.ncbi.nlm.nih.gov/pubmed/20221922

[54] Maimon, O.Z. and L. Rokach. (2014). *Data mining with decision trees: theory and applications* World scientific. https://doi.org/10.1142/9097

[55] Goodfellow, I., Y. Bengio, and A. Courville. (2016). *Deep learning* MIT press . https://www.deeplearningbook.org/

[56] Sanyal, D., N. Bosch, and L. Paquette. (2020). Feature Selection Metrics: Similarities, Differences, and Characteristics of the Selected Models. *Int Edu Data Mining Soci*, https://eric.ed.gov/?id=ED607910