دانشگاه صنعتی امیرکبیر
(پلی‌تکنیک تهران)
دانشکده ریاضی و علوم کامپیوتر

پروژه پنجم

عنوان :

تحلیل داده‌های مجموعه WUSTL EHMS 2020 برای امنیت سایبری اینترنت اشیاء پزشکی(IoMT)

نگارش:
الهه مدرس

استاد درس :
دکتر مهدی قطعی

استاد کارگاه :
دکتر بهنام یوسفی مهر

چکیده

در این مقاله به بررسی داده‌هایی مربوط به یک مجموعه داده ای به نام WUSTL EHMS 2020 پرداخته شده که برای تحقیق در حوزه امنیت سایبری در اینترنت اشیاء پزشکی (IoMT) استفاده می‌شود .این مجموعه داده از یک سیستم نظارت بر سلامت پیشرفته (EHMS) جمع‌آوری شده است که در زمان واقعی داده‌های بیماران را جمع‌آوری می‌کند. در این مقاله سعی کرده ایم بهترین روش برای پردازش و طبقه بندی داده ها پیدا کنیم .

واژه‌های کلیدی: WUSTL EHMS 2020/ حوزه امنیت سایبری در اینترنت اشیاء پزشکی/ سیستم نظارت بر سلامت پیشرفته / طبقه بندی

فهرست مطالب **صفحه**

فصل اول

مقدمه

مقدمه

1. معرفی مسئله

- در دنیای امروز، تحلیل داده‌ها به ابزاری کلیدی برای استخراج اطلاعات مفید از داده‌ها تبدیل شده است. در این پروژه، داده‌های مربوط به حملات و وضعیت‌های پزشکی بیماران مورد بررسی قرار گرفته‌اند. هدف این پروژه پیش‌بینی و شناسایی حملات در داده‌های سلامت به کمک مدل‌های یادگیری ماشین است. این مسئله در زمینه‌های پزشکی، مراقبت‌های بهداشتی و امنیت بیمارستان‌ها بسیار حائز اهمیت است، چرا که شناسایی به موقع حملات می‌تواند در نجات جان بیماران موثر باشد.

2. اهداف پروژه

- هدف اصلی این پروژه، طراحی و پیاده‌سازی مدل‌های یادگیری ماشین برای پیش‌بینی حملات و وضعیت‌های اضطراری در داده‌های سلامت است. در این راستا:
- پیش‌پردازش داده‌ها، شناسایی ناهنجاری‌ها و ارزیابی مدل‌ها از اهمیت ویژه‌ای برخوردار است.
- انتخاب مدل‌های بهینه با استفاده از ابزارهایی مانند LazyPredict و ارزیابی عملکرد آن‌ها با استفاده از معیارهای مختلف دقت، F1-Score، و Accuracy انجام خواهد شد.

3. اهمیت و کاربرد پروژه

- پیش‌بینی حملات پزشکی می‌تواند به ویژه در شرایط بحرانی و در بیمارستان‌ها اهمیت زیادی داشته باشد. این پروژه به محققان و پزشکان کمک می‌کند تا سریع‌تر و دقیق‌تر مشکلات و حملات را شناسایی کرده و در نتیجه اقدامات درمانی به موقع انجام دهند.

فصل دوم

بررسی بیشتر داده و توضیح اعضای داده

مقایسه روش‌های پیش‌بینی در داده‌های سلامت

در حوزه سلامت، تحلیل‌های پیش‌بینی‌کننده نقش مهمی در بهبود کیفیت خدمات درمانی، پیشگیری از بیماری‌ها و مدیریت منابع بهداشتی ایفا می‌کنند. انتخاب روش مناسب برای پیش‌بینی در داده‌های سلامت اهمیت بسیاری دارد، زیرا دقت و کارایی مدل‌های پیش‌بینی می‌تواند تأثیر زیادی بر تصمیم‌گیری‌های کلینیکی و مدیریتی داشته باشد. در این بخش، به مقایسه روش‌های مختلف پیش‌بینی در داده‌های سلامت پرداخته می‌شود.

رگرسیون لجستیک (Logistic Regression)

درخت تصمیم (Decision Tree)

جنگل تصادفی (Random Forest)

ماشین‌های بردار پشتیبان (Support Vector Machines - SVM)

شبکه‌های عصبی مصنوعی (Artificial Neural Networks - ANN)

گرادیان بوستینگ (Gradient Boosting)

K-نزدیک‌ترین همسایه (K-Nearest Neighbors - KNN)

مقایسه روش‌ها:

KNN	گرادیان بوستینگ	شبکه‌های عصبی	SVM	جنگل تصادفی	درخت تصمیم	رگرسیون لجستیک	معیار مقایسه
ساده	پیچیده‌تر	پیچیده	پیچیده‌تر	پیچیده‌تر	نسبتاً ساده	بسیار ساده	سادگی پیاده‌سازی
پایین	متوسط	پایین	پایین	متوسط	بالا	بالا	قابلیت تفسیر
محدود	بسیار خوب	بسیار خوب	خوب	خوب	متوسط	محدود	توانایی مدیریت داده‌های بزرگ
کم‌مقاومت	بسیار بالا	کم‌مقاومت	متوسط	بالا	کم‌مقاومت	کم‌مقاومت	مقاومت در برابر بیش برازش
متوسط	بسیار بالا	بسیار بالا	بالا	بالا	پایین‌تر	متوسط	دقت پیش‌بینی
بسیار سریع	کندتر	کند	کندتر	سریع‌تر	سریع	بسیار سریع	سرعت آموزش
کم‌مقاومت	بالا	بالا	بالا	متوسط	کم‌مقاومت	کم	نیاز به تنظیم پارامترها
ضعیف	بسیار خوب	بسیار خوب	بسیار خوب	خوب	متوسط	ضعیف	مناسب برای داده‌های غیرخطی

تحلیل جامع:

رگرسیون لجستیک:

مزایا: مدل ساده و قابل تفسیر، سریع در آموزش و پیش‌بینی.

معایب: فرض خطی بودن روابط، عملکرد کمتر در داده‌های پیچیده و غیرخطی.

درخت تصمیم:

مزایا: قابل تفسیر، نیاز به پیش‌پردازش کم، قابلیت مدیریت داده‌های غیرخطی.

معایب: مستعد بیش برازش، دقت پایین‌تر نسبت به روش‌های پیشرفته‌تر.

جنگل تصادفی:

مزایا: مقاومت بالا در برابر بیش برازش، دقت بالا، قابلیت مدیریت داده‌های بزرگ و متنوع.

معایب: پیچیدگی بیشتر، کاهش قابلیت تفسیر نسبت به درخت تصمیم واحد.

ماشین‌های بردار پشتیبان (SVM):

مزایا: عملکرد خوب در داده‌های با ابعاد بالا، مناسب برای داده‌های غیرخطی با استفاده از کرنل‌ها.

معایب: زمان آموزش بالا، نیاز به تنظیم دقیق پارامترها، قابلیت تفسیر پایین.

شبکه‌های عصبی مصنوعی (ANN):

مزایا: قدرت بالا در یادگیری الگوهای پیچیده، مناسب برای داده‌های بزرگ و غیرخطی.

معایب: نیاز به منابع محاسباتی بالا، پیچیدگی در تنظیم و آموزش، قابلیت تفسیر پایین.

گرادیان بوستینگ:

مزایا: دقت بسیار بالا، مقاومت خوب در برابر بیش برازش، قابلیت مدیریت داده‌های مختلف.

معایب: زمان آموزش طولانی، نیاز به تنظیم دقیق پارامترها.

K-نزدیک‌ترین همسایه (KNN):

مزایا: ساده در پیاده‌سازی، بدون نیاز به مدل‌سازی پیچیده.

معایب: عملکرد کمتر در داده‌های بزرگ، حساسیت به مقیاس داده‌ها، قابلیت تفسیر پایین.

نتیجه‌گیری:

انتخاب روش پیش‌بینی مناسب در داده‌های سلامت بستگی به معیارهای خاص پروژه دارد. برای کاربردهایی که نیاز به مدل‌های قابل تفسیر و سریع دارند، رگرسیون لجستیک یا درخت تصمیم می‌تواند گزینه مناسبی باشد. در پروژه‌هایی که دقت و توانایی مدیریت داده‌های پیچیده اهمیت بیشتری دارد، روش‌هایی مانند جنگل تصادفی، گرادیان بوستینگ یا شبکه‌های عصبی مصنوعی پیشنهاد می‌شوند. همچنین، ترکیب چندین روش (Ensemble Methods) می‌تواند به بهبود عملکرد مدل‌های پیش‌بینی در داده‌های سلامت کمک شایانی کند.

فصل سوم

روش پیشنهادی

روش پیشنهادی

پیش‌پردازش داده‌ها

پیش‌پردازش داده‌ها یکی از مراحل حیاتی در تحلیل داده‌ها است. در این پروژه، داده‌های موجود شامل ستون‌هایی با مقادیر گم‌شده و ناهنجاری‌هایی هستند که باید قبل از مدل‌سازی اصلاح شوند. مراحل پیش‌پردازش عبارتند از:

شناسایی و پرکردن مقادیر گم‌شده با استفاده از میانگین برای داده‌های عددی و بیشترین مقدار (Mode)برای داده‌های غیرعدد.

جایگزینی مقادیر صفر با میانگین مقادیر همان ستون برای جلوگیری از خطاهای محاسباتی.

نرمال‌سازی داده‌ها با استفاده از .MinMaxScaler

انتخاب ویژگی‌ها

در این مرحله، ویژگی‌های داده‌ها که به شناسایی حملات و وضعیت‌های پزشکی مربوط هستند انتخاب می‌شوند. برای بهبود دقت مدل‌ها، از یک-داغ‌سازی (One-Hot Encoding)برای ویژگی‌های دسته‌ای استفاده می‌شود. همچنین، ویژگی‌هایی که ارتباط کمی با هدف دارند حذف می‌شوند.

مدل‌سازی و ارزیابی

پس از پیش‌پردازش داده‌ها، از مدل‌های مختلف یادگیری ماشین برای پیش‌بینی استفاده می‌شود. این مدل‌ها عبارتند از:

LazyPredictبرای مقایسه اولیه مدل‌ها

جنگل تصادفی (Random Forest)

رگرسیون لجستیک (Logistic Regression)

SVM

پس از آموزش مدل‌ها، از معیارهای مختلف مانند دقت (Accuracy)، F1-Score، دقت و یادآوری برای ارزیابی عملکرد مدل‌ها استفاده می‌شود.

شناسایی ناهنجاری‌ها

برای شـــناســـایی داده‌های غیرعادی و حذف آن‌ها از داده‌های آموزشـــی، از الگوریتم‌های Isolation Forestو One-Class SVMاستفاده می‌شود. این ناهنجاری‌ها می‌توانند اثرات منفی بر روی مدل‌سازی داشته باشند و حذف آن‌ها می‌تواند عملکرد مدل را بهبود بخشد

فصل چهارم

بررسی بهترین الگوریتم

1. **Random Forest Classifier:**
   - **دقت (Accuracy): 0.9963 نشان می‌دهد که مدل بسیار دقیق عمل کرده و تقریباً تمامی نمونه‌ها را به درستی دسته‌بندی کرده است.**
   - **گزارش طبقه‌بندی (Classification Report): مقدارهای بالا برای precision ، recall، و f1-score برای همه کلاس‌ها نشان‌دهنده عملکرد عالی مدل است.**
   - **ماتریس سردرگمی (Confusion Matrix): بررسی این ماتریس نشان می‌دهد که مدل خطای بسیار کمی در پیش‌بینی دارد.**
   - **منحنی ROC: مساحت زیر منحنی (AUC) نزدیک به 1 است، که نشان‌دهنده قدرت بالای تفکیک مدل بین کلاس‌هاست.**
2. **AdaBoost Classifier:**
   - **دقت 1.0است که نشان می‌دهد مدل هیچ خطایی نداشته است. این ممکن است به دلیل ساختار داده یا قدرت این روش باشد.**
   - **ماتریس سردرگمی و منحنی ROC مشابه Random Forest نشان‌دهنده نتایج بدون نقص است.**
3. **Gradient Boosting Classifier:**
   - **دقت نیز 1.0گزارش شده است، که این مدل هم مشابه AdaBoost عمل کرده است.**
4. **تحلیل کلی:**

- **تمام مدل‌ها عملکرد بسیار خوبی دارند Random Forest ،. AdaBoostو Gradient Boostingهر سه می‌توانند گزینه‌های مناسبی برای استفاده باشند.**
- **برای سناریوهای واقعی، ممکن است بررسی بیشتری روی مقاومت مدل‌ها نسبت به نویز یا داده‌های ناشناخته لازم باشد.**

فصل پنجم

جمع بندی و نتیجه گیری

نتیجه‌گیری:

تحلیل نتایج نشـــان می‌دهد که مدل‌های مورد اســـتفاده(Random Forest ، AdaBoost، و Gradient Boosting)عملکرد بسیار خوبی در دسته‌بندی داده‌ها دارند. دقت بالا، مقادیر مطلوب برای معیارهایprecision ، recall، وf1-score ، و نتایج قوی در منحنی ROC بیانگر این است که مدل‌ها در پیش‌بینی داده‌های آزمون بسیار موثر هستند.

نکات کلیدی:

Random Forest:

قدرت تفکیک بالا.

حساس به پارامترهایی مانند تعداد درخت‌ها و عمق درخت.

AdaBoost:

دقت کامل و انعطاف‌پذیری در مدیریت داده‌های مختلف.

ممکن است در برابر نویز حساس‌تر باشد.

Gradient Boosting:

عملکرد مشابه AdaBoost با توانایی یادگیری خطاهای مدل‌های قبلی.

در صورت وجود داده‌های بزرگ‌تر، ممکن است زمان اجرا بیشتر شود.

1-2- پیشنهادات آتی:

ارزیابی بیشتر برای قابلیت تعمیم‌دهی:

تست مدل‌ها روی مجموعه داده‌های کاملاً متفاوت یا مستقل برای ارزیابی توانایی تعمیم‌دهی.

استفاده از روش‌های k-fold cross-validation برای کاهش اثرات داده‌های خاص در ارزیابی.

بهینه‌سازی بیشتر مدل‌ها:

اســـتفاده از روش‌های پیشـــرفته‌تر مانند Bayesian Optimizationیا Random Searchبرای بهینه‌سازی هایپرپارامترها.

استفاده از مدل‌های قدرتمندتر مانند XGBoostیا LightGBM.

تحلیل ویژگی‌ها:

بررسی اهمیت ویژگی‌ها با استفاده از ابزارهایی مثل SHAP یاPermutation Importance.

حذف ویژگی‌های کم‌اثر برای بهبود عملکرد و کاهش پیچیدگی مدل.

مدیریت داده‌های جدید و نویز:

شبیه‌سازی سناریوهای واقعی شامل داده‌های نویزی و بررسی عملکرد مدل‌ها در این شرایط.

اضافه کردن داده‌های واقعی یا مصنوعی برای بررسی پایداری مدل.

گسترش استفاده از مدل‌ها:

ایجاد سیستم پیش‌بینی بلادرنگ (real-time prediction) با استفاده از مدل بهینه‌شده.

پیاده‌سازی مدل‌ها در یک سرویس REST یا اپلیکیشن برای استفاده عملیاتی.

.

منابع و مراجع

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.

Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5-32.

Cortes, C., & Vapnik, V. (1995). *Support-Vector Networks*. Machine Learning, 20(3), 273-297.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Friedman, J. H. (2001). *Greedy Function Approximation: A Gradient Boosting Machine*. Annals of Statistics, 29(5), 1189-1232.