भारतीय वैज्ञानिक एवं औद्योगिक अनुसंधान पत्रिका
वर्ष 32 अंक (2) दिसम्बर 2024 पृ. 69-79
DOI: 10.56042/bvaap.v32i2.1757

विभिन्न मशीन लर्निंग तकनीकों द्वारा मधुमेह की पूर्व-सूचना

वैशाली गुप्ता[1] एवं रूचि पटेल[2]
[1]आई.पी.एस. एकैडमी, आई.ई.एस., इंदौर, मध्य प्रदेश, भारत
[2]ज्ञान गंगा इंस्टीट्यूट ऑफ टेक्नोलॉजी एंड साइंसेज, जबलपुर, मध्य प्रदेश
ई-मेलः [1]vaishali.gupta77@gmail.com, [2]ruchipatel294@gmail.com

सारांश

मधुमेह एक जानलेवा बीमारी है जो असामान्य रूप से उच्च रक्त शर्करा के स्तर से पहचानी जाती है। यह दुनिया में मौत का प्रमुख कारण है। हाल के वर्षों में बढ़ती बीमारी के अनुसार, विश्व स्तर पर मधुमेह के रोगियों की संख्या साल 2040 तक 642 मिलियन या औसतन हर दस में से एक व्यक्ति तक पहुंच जाएगी। यह सच है कि इस पर बहुत ध्यान देने की जरूरत है। इस रोग के जोखिम की भविष्यवाणी करने के लिए मधुमेह डेटासेट पर कई डेटा माइनिंग और मशीन लर्निंग तकनीकों का उपयोग किया गया है। इस शोध पत्र का लक्ष्य फीचर-आधारित डेटासेट का उपयोग करके मधुमेह वर्गीकरण, प्रारंभिक चरण की पहचान और पूर्वानुमान के लिए कई मशीन लर्निंग एल्गोरिद्म की जांच करना है। हमने प्रायोगिक मूल्यांकन के लिए एक बेंचमार्क पीआईएमए (PIMA) इंडियन डायबिटीज (Indian Diabetes) डेटासेट का उपयोग किया है, जिसमें 768 रोगी शामिल हैं, जिनमें से 268 मधुमेह रोगी हैं और जिनमें से 500 व्यक्ति मधुमेह रोगी नहीं हैं। अंत में, विभिन्न मशीन लर्निंग तकनीकों की सटीकता का आकलन किया गया है।

मुख्य शब्दः मधुमेह की पूर्व-सूचना, ग्लूकोज स्तर की पूर्वानुमान, मशीन लर्निंग, वर्गीकरण, लॉजिस्टिक रिग्रेशन, रैंडम फ़ॉरेस्ट

Prediction of Diabetes Using Various Machine Learning Techniques

Vaishali Gupta[1] and Ruchi Patel[2]
[1]IPS Academy, IES, Indore
[2]Gyan Ganga Institute of Technology and Sciences, Jabalpur, MP, India
E-mail: [1]vaishali.gupta77@gmail.com, [2]ruchipatel294@gmail.com

Abstract

Diabetes is a life-threatening disease marked by unusually high blood sugar levels. It is the leading cause of death in the globe. According to rising morbidity in recent years, the number of diabetic patients globally will reach 642 million by 2040, or approximately one out of every ten persons. It is true that this requires a lot of focus. On the diabetes dataset, a number of data mining and machine learning techniques were utilized to predict disease risk. The goal of this work is to investigate several machine learning algorithms for diabetes categorization, early-stage identification, and prediction using a feature-based dataset. A benchmark PIMA Indian Diabetes dataset is used for experimental evaluation, which includes 768 patients, 268 of whom are diabetic and 500 of whom are not. At the end, the accuracy of various machine learning approaches is measured in order to assess their performance.

Keywords: Diabetic prediction, Glucose level prediction, Machine Learning, Classification, Logistic Regression, Random Forest

1. प्रस्तावना

मधुमेह मेलिटस (डीएम) को आमतौर पर मधुमेह कहा जाता है। यह एक स्वास्थ्य समस्या है जो गंभीर और जटिल है। जब अग्न्याशय पर्याप्त इंसुलिन का उत्पादन नहीं करता है, तब रक्त शर्करा बढ़ जाती है और यह विभिन्न अंगों, विशेष रूप से आंखों, गुर्दे, तंत्रिकाओं को प्रभावित करता है[1]। यही कारण है कि मधुमेह को साइलेंट किलर कहा

जाता है। आमतौर पर तीन प्रकार के मधुमेह मौजूद हैं जो कि चित्र 1 में भी प्रदर्शित हैः टाइप-1 मधुमेह, टाइप-2 मधुमेह और गर्भकालीन मधुमेह[2]।

टाइप-1 मधुमेह या यहां तक कि इंसुलिन नहीं होने की स्थिति में अग्न्याशय बहुत कम इंसुलिन का उत्पादन करता है। मोटे तौर पर सभी मधुमेह का 5 से 10% टाइप-1 होता है और यह न केवल यौवन या शैशवावस्था में होता है, बल्कि वयस्कता में भी होता है। टाइप-2 मधुमेह तब होता है जब शरीर द्वारा इंसुलिन पर्याप्त रूप से जारी नहीं होती है। दुनिया में मधुमेह के लगभग 90% रोगी टाइप-2 मधुमेह के हैं। तीसरे प्रकार के मधुमेह, गर्भकालीन मधुमेह मेलिटस (जीडीएम) के समान है। कई मायनों में इसमें इंसुलिन के तुलनात्मक रूप से अपर्याप्त स्राव के मिश्रण की आवश्यकता होती है। सभी गर्भवती महिलाओं में से लगभग 2-10% गर्भावधि मधुमेह से प्रभावित होती हैं, प्रसव के बाद यह बढ़ भी सकती है या गायब भी हो सकती है।

मधुमेह के लिए जोखिम कारक

टाइप-1 मधुमेह के लिए सबसे महत्वपूर्ण जोखिम कारकों में से एक पारिवारिक इतिहास है। इसके अलावा, टाइप-1 मधुमेह पर्यावरणीय कारणों और कुछ वायरल संक्रमणों से जुड़ा हुआ है[3]। टाइप-2 मधुमेह निम्नलिखित जोखिम कारकों के कारण होता है।

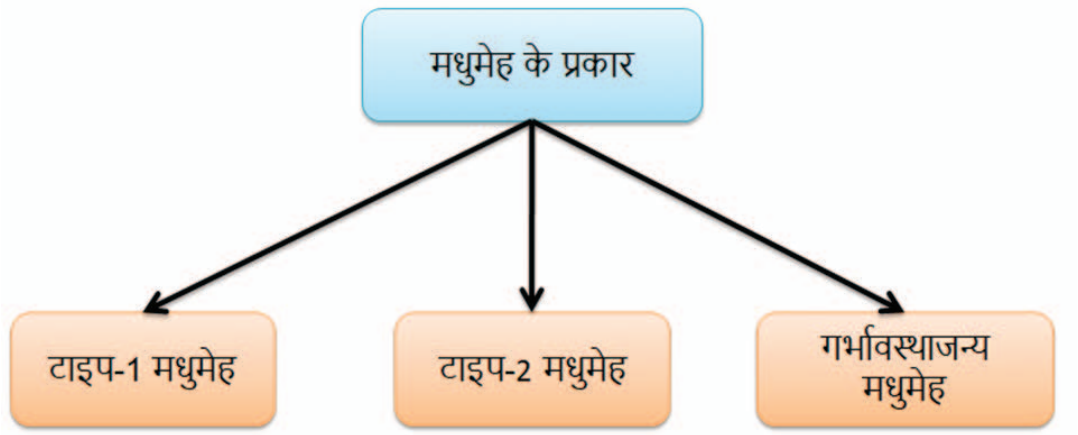
1. तनाव 2. मधुमेह का पारिवारिक इतिहास 3. आयु 4. मोटापा 5. शारीरिक गतिविधि की कमी 6. उच्च रक्तचाप 7. असंतुलित आहार 8. गर्भकालीन मधुमेह 9. वंश

मधुमेह मेलिटस के दीर्घकालीन परिणाम होते हैं। एक मधुमेह रोगी को विभिन्न प्रकार की स्वास्थ्य समस्याओं के विकसित होने का भी उच्च जोखिम होता है। हालांकि मधुमेह का सटीक कारण स्पष्ट नहीं है, वैज्ञानिक इस बात से सहमत हैं कि आनुवंशिक कारकों के साथ-साथ पर्यावरणीय जीवन शैली एक महत्वपूर्ण भूमिका निभाती है। इस तथ्य के बावजूद कि यह लाइलाज है, इसे चिकित्सा और दवा के साथ प्रबंधित किया जा सकता है। मधुमेह के रोगियों को दिल की विफलता और तंत्रिका क्षति सहित माध्यमिक स्वास्थ्य समस्याओं के विकास का खतरा होता है। नतीजतन, मधुमेह का शीघ्र निदान और उपचार जटिलताओं से बचने और गंभीर स्वास्थ्य समस्याओं के जोखिम को कम करने में मदद करेगा।

मधुमेह रोग निदान और मधुमेह डेटा की व्याख्या करना कठिन समस्या है। स्वास्थ्य संबंधी समस्याओं से निपटने के लिए विभिन्न मशीन लर्निंग विधियों का उपयोग किया जाता है जो प्राकृतिक तौर में विशिष्ट हैं। अधिकांश चिकित्सा डेटा में गैर-रैखिकता, गैर-सामान्यता और एक अंतर्निहित सहसंबंध संरचना होती है। इसलिए पारंपरिक और व्यापक रूप से वर्गीकरण तकनीकें जैसे कि एसवीएम, रैण्डेम फ़ॉक्एसट और डीसीजण ट्री आदि इस्तेमाल की जाती हैं, लेकिन ये डेटा को ठीक से वर्गीकृत नहीं करते हैं[4-5]। इस शोध का मुख्य लक्ष्य मधुमेह रोगियों के डेटासेट को देखना और उसमें विभिन्न मशीन लर्निंग एल्गोरिदम लागू करना है। सटीकता और दक्षता मेट्रिक्स को विभिन्न एल्गोरिदम की तुलना के लिए उपयोग किया जाता है। इस अध्ययन का उद्देश्य मधुमेह के पूर्वानुमान के लिए तैयार करने में विभिन्न मशीन लर्निंग एल्गोरिदम का उपयोग करना है।

2. साहित्य की समीक्षा

इस खंड में, हमने स्वास्थ्य देखभाल में मधुमेह के पूर्वानुमान के लिए मशीन लर्निंग आधारित वर्गीकरण एल्गोरिदम पर चर्चा की है। चिकित्सा डेटा विश्लेषण और निदान करने के लिए स्वास्थ्य पेशेवरों को आसानी प्रदान करके मशीन लर्निंग, स्वास्थ्य सेवा उद्योग में एक आवश्यक भूमिका निभाता है। यहाँ कुछ समीक्षित साहित्य इस प्रकार प्रस्तुत हैं-



चित्र 1. मधुमेह के प्रकार

कुमारी और अन्य ने एक सॉफ्ट कंप्यूटिंग-आधारित मधुमेह पूर्वानुमान प्रणाली प्रस्तुत की जो कि तीन व्यापक रूप से उपयोग की जाने वाली पर्यवेक्षित मशीन लर्निंग एल्गोरिदम का एक संयोजन तरीके से उपयोग करती है[6]। उन्होंने मूल्यांकन उद्देश्यों के लिए पीआईएमए और स्तन कैंसर डेटासेट का इस्तेमाल किया। उन्होंने रैंडम फ़ॉरेस्ट, लॉजिस्टिक रिग्रेशन और नैव बेयस का इस्तेमाल किया और अपने प्रदर्शन की तुलना अत्याधुनिक व्यक्तिगत और कलाकारों के दृष्टिकोण से की, और उनका सिस्टम 79% सटीकता के साथ बेहतर प्रदर्शन करता है। इस्लाम और अन्य प्रारंभिक अवस्था में मधुमेह का पूर्वानुमान करने के लिए डेटा माइनिंग तकनीकों, यानी रैंडम फ़ॉरेस्ट, लॉजिस्टिक रिग्रेशन और नैवे बेयस एल्गोरिथम का उपयोग किया था[7]। उन्होंने प्रशिक्षण उद्देश्यों के लिए 10-गुना क्रॉस-सत्यापन और प्रतिशत विभाजन तकनीकों का उपयोग किया। उन्होंने प्रश्नावली के माध्यम से बांग्लादेश के एक अस्पताल से 529 व्यक्तियों से मधुमेह और गैर-मधुमेह डेटा एकत्र किया। प्रयोगात्मक परिणाम बताते हैं कि रैंडम फ़ॉरेस्ट अन्य एल्गोरिदम की तुलना में बेहतर प्रदर्शन करते हैं। हालांकि, अत्याधुनिक तुलना गायब है और हासिल की गई सटीकता को स्पष्ट रूप से रिपोर्ट नहीं किया गया है। मलिक और अन्य महिलाओं में प्रारंभिक और आरंभिक मधुमेह मेलिटस पूर्वानुमान में डेटा माइनिंग और मशीन लर्निंग तकनीकों का तुलनात्मक विश्लेषण किया[8]। उन्होंने मधुमेह पूर्वानुमान ढांचे का प्रस्ताव करने के लिए पारंपरिक मशीन लर्निंग एल्गोरिदम का फायदा उठाया। प्रस्तावित प्रणाली का मूल्यांकन जर्मनी के एक अस्पताल के मधुमेह डेटासेट पर किया गया था। अनुभवजन्य परिणाम अन्य पारंपरिक एल्गोरिदम की तुलना में के-निकटतम पड़ोसी, रैंडम फ़ॉरेस्ट और निर्णय वृक्ष (decision tree) की श्रेष्ठता दिखाते हैं। क़ौक़ज़ेह और अन्य ने मधुमेह वर्गीकरण के लिए फोटोप्लेथिस्मोग्राम विश्लेषण पर आधारित एक लॉजिस्टिक रिग्रेशन मॉडल का प्रस्ताव रखा था[9]। इन्होंने प्रशिक्षण के लिए 459 रोगियों के डेटा और मॉडल के परीक्षण और सत्यापन के लिए 128 डेटा बिंदुओं का उपयोग किया। इस प्रस्तावित प्रणाली ने 552 व्यक्तियों को गैर-मधुमेह के रूप में सही ढंग से वर्गीकृत किया और 92% की सटीकता हासिल की। हालांकि, प्रस्तावित तकनीक की तुलना अत्याधुनिक तकनीकों से नहीं की गई थी।

पेथुनाचियार ने मशीन लर्निंग एल्गोरिथम का उपयोग करके मधुमेह मेलिटस वर्गीकरण प्रणाली प्रस्तुत की थी[10]। मुख्य रूप से, उन्होंने यूसीआई मशीन रिपोजिटरी से विभिन्न कार्यों और मधुमेह डेटा के साथ एक सपोर्ट वेक्टर मशीन का उपयोग किया था। उन्होंने एसवीएम को लीनियर फंक्शन के साथ नैव बेयस, डिसीजन ट्री और न्यूरल नेटवर्क की तुलना में अधिक कुशल पाया। फिर भी, अत्याधुनिक तुलना नहीं थी और पैरामीटर चयन विस्तृत नहीं थी। गुप्ता और अन्य ने नेव बेयस का शोषण किया और मधुमेह वर्गीकरण के लिए सपोर्ट वेक्टर मशीन एल्गोरिदम का समर्थन किया था[11]। उन्होंने पीआईएमए इंडियन डायबिटीज डेटासेट का इस्तेमाल किया। इसके अलावा, उन्होंने मॉडल की सटीकता में सुधार के लिए फीचर चयन आधारित दृष्टिकोण और के-फोल्ड क्रॉस-सत्यापन का उपयोग किया। इसमें प्रयोगात्मक परिणामों ने नैव बेयस मॉडल पर सपोर्ट वेक्टर मशीन की सर्वोच्चता को दिखाया। चौबे और अन्य ने मधुमेह वर्गीकरण के लिए वर्गीकरण तकनीकों का तुलनात्मक विश्लेषण प्रस्तुत किया[12]। उन्होंने यूसीआई मशीन लर्निंग रिपोजिटरी और एक स्थानीय मधुमेह डेटासेट से एकत्र किए गए पीआईएमए भारतीय डेटा का इस्तेमाल किया। साथ ही रोगियों को मधुमेह के रूप में डेटासेट से वर्गीकृत करने के लिए AdaBoost, K-निकटतम पड़ोसी प्रतिगमन, और रेडियल आधार फ़ंक्शन का उपयोग किया। इसके अलावा, उन्होंने फीचर इंजीनियरिंग के लिए पीसीए और एलडीए का इस्तेमाल किया, और यह निष्कर्ष निकाला कि दोनों सटीकता में सुधार और अवांछित सुविधाओं को हटाने के लिए वर्गीकरण एल्गोरिदम के साथ उपयोगी हैं। के. विजया कुमार और अन्य ने मधुमेह के पूर्वानुमान के लिए रैंडम फ़ॉरेस्ट एल्गोरिथम आधारित एक ऐसी प्रणाली विकसित की जो मशीन लर्निंग तकनीक में उपयोग करके उच्च सटीकता के साथ एक रोगी के लिए मधुमेह की प्रारंभिक पूर्वानुमान कर सकता है[13]। प्रस्तावित मॉडल मधुमेह के पूर्वानुमान के लिए सर्वोत्तम परिणाम देता है और परिणाम से पता चलता है कि भविष्यवाणी प्रणाली प्रभावी ढंग से, कुशलतापूर्वक और सबसे महत्वपूर्ण रूप से तत्काल मधुमेह रोग की भविष्यवाणी करने में सक्षम है।

तेजस एन जोशी और अन्य ने मशीन लर्निंग तकनीकों का उपयोग करके मधुमेह की भविष्यवाणी प्रस्तुत की, जिसका उद्देश्य तीन अलग-अलग पर्यवेक्षित मशीन सीखने के तरीकों के माध्यम से मधुमेह की भविष्यवाणी करना हैः एसवीएम, लॉजिस्टिक रिग्रेशन, एएनएन[14]। यह परियोजना मधुमेह रोग का पहले पता लगाने के लिए एक प्रभावी तकनीक का प्रस्ताव करती है। डीराज शेट्टी और अन्य ने डेटा माइनिंग का उपयोग करते हुए प्रस्तावित मधुमेह रोग की भविष्यवाणी बुद्धिमान मधुमेह रोग भविष्यवाणी प्रणाली को इकट्ठा करती है जो मधुमेह रोगी के डेटाबेस का उपयोग करके मधुमेह की बीमारी का विश्लेषण देती है[15]। इस प्रणाली में, वे मधुमेह रोगी के डेटाबेस पर लागू करने के लिए बायेसियन और केएनएन (के-निकटतम पड़ोसी) जैसे एल्गोरिदम के उपयोग का प्रस्ताव करते हैं और मधुमेह रोग की भविष्यवाणी के लिए मधुमेह की विभिन्न विशेषताओं को लेकर उनका विश्लेषण करते हैं। यहां, कुछ संबंधित कार्य निम्नलिखित तालिका 1 में मौजूद हैं-

तालिका 1: पिछले संबंधित कार्य का सारांश

संदर्भ	कार्य	पद्धति/दृष्टिकोण	डेटासेट	परिणाम
परवीन एवं अन्य (2016)[16]	मधुमेह की भविष्यवाणी के लिये डेटा माइनिंग वर्गीकरण तकनीकों का विश्लेषण	एडबूस्ट, बैगिंग और जे48 डिसीजन ट्री	कनाडा की आबादी के तीन आयु वर्ग का डेटा	बैगिंग विधि द्वारा उत्पादित उच्च सटीकताः एयूसी (AUC)=0.98
सेल्वकुमार एवं अन्य (2017)[17]	वर्गीकरण आधारित डेटा माइनिंग तकनीकों का उपयोग करके मधुमेह निदान की भविष्यवाणी	बाइनरी लॉजिस्टिक रिग्रेशन मल्टीलेयर परसेप्ट्रोन K-नियरेस्ट नेबोर	बहुआयामी स्वास्थ्य देखभाल डेटासेट	शुद्धता = 69% शुद्धता = 71% शुद्धता = 80%
सिसोदिया एवं अन्य (2018)[18]	वर्गीकरण एल्गोरिद्म का उपयोग करके मधुमेह की भविष्यवाणी	डिसीजन ट्री, सपोर्ट वेक्टर मशीन (SVM) एवं नैवे बयेस	पिमा इंडियंस मधुमेह डेटाबेस	उच्चतम सटीकता = 76.30%
ऐश्वर्या एवं अन्य (2019)[19]	मशीन लर्निंग एल्गोरिदम का उपयोग करके मधुमेह की भविष्यवाणी	लॉजिस्टिक रिग्रेशन ग्रेडइएंट डिसेंट बैगिंग रैंडम फोरेस्त डिसीजन ट्री K-नियरेस्ट नेबोर	मैन्युअल रूप से एकत्रित मधुमेह डेटासेट जिसमें 800 रिकॉर्ड हैं	शुद्धता = 96% शुद्धता = 93% शुद्धता = 90% शुद्धता = 91% शुद्धता = 86% शुद्धता = 90%
ज़्हौ एवं अन्य (2020)[20]	एक उन्नत गहरे तंत्रिका नेटवर्क पर आधारित मधुमेह भविष्यवाणी मॉडल	कृत्रिम तंत्रिका मॉडल	पिमा इंडियंस मधुमेह डेटाबेस	बेस्ट ट्रेनिंग शुद्धता = 94.02
अहमद एवं अन्य (2021)[21]	मशीन लर्निंग का उपयोग करके मधुमेह की भविष्यवाणी पर स्वास्थ्य संबंधी विशेषताओं और उनके प्रभाव की जांच करना	लॉजिस्टिक रिग्रेशन, डिसीजन ट्री, एसवीएम, रैंडम फॉरेस्ट और एनसेंबल	HbA1c-लेबल वाला डेटासेट FPG-लेबल डेटासेट	F1-स्कोर = 82.05% SVM का उपयोग कर F1-स्कोर = 87.72% रैंडम फ़ॉरेस्ट का उपयोग कर

3. प्रस्तावित कार्यप्रणाली

चित्र 2. मधुमेह भविष्यवाणी मॉडल के लिए प्रस्तावित कार्य का प्रवाह दिखाता है। इस मॉडल में पांच अलग-अलग मॉड्यूल हैं। इन मॉड्यूल में शामिल हैं i. डेटासेट संग्रह ii. डेटा प्री-प्रोसेसिंग iii. निर्माण मॉडल iv. सर्वोत्तम हाइपर-पैरामीटर सेट करें v. मूल्यांकन

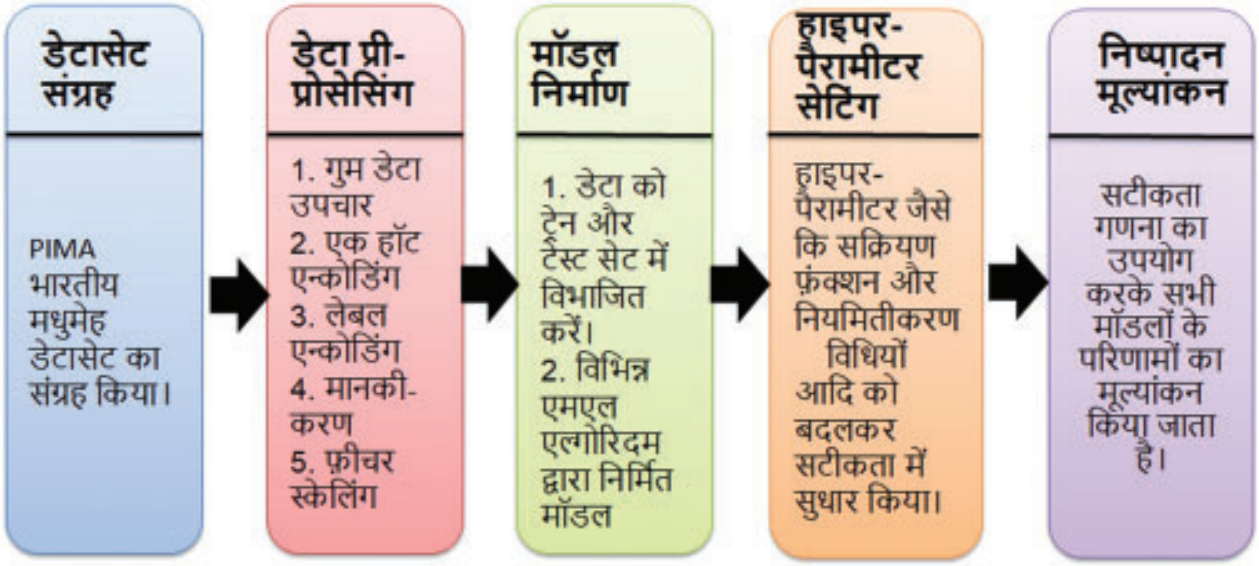
3.1 मशीन लर्निंग के तरीके

मधुमेह के वर्गीकरण के लिए, हमने व्यापक रूप से उपयोग की जाने वाली छह अत्याधुनिक तकनीकों का इस्तेमाल किया है। मुख्य रूप से, किसी व्यक्ति को मधुमेह की किसी भी श्रेणी में वर्गीकृत करने के लिए प्रस्तावित तकनीकों के बीच तुलनात्मक विश्लेषण किया जाता है। प्रस्तावित मधुमेह भविष्यवाणी तकनीकों का विवरण इस प्रकार है।

(i) लॉजिस्टिक रिग्रेशन (Logistic Regression)

एक पर्यवेक्षित शिक्षण वर्गीकरण एल्गोरिथ्म, लॉजिस्टिक प्रतिगमन भी एक पर्यवेक्षित शिक्षण एल्गोरिथ्म है। इसका उपयोग यह पता लगाने के लिए किया जाता है कि एक या अधिक भविष्यवक्ताओं पर आधारित द्विआधारी प्रतिक्रिया की कितनी संभावना है। वे प्रकृति में निरंतर या असतत हो सकते हैं। जब हम कुछ डेटा ऑब्जेक्ट्स को श्रेणियों में वर्गीकृत या अलग करना चाहते हैं, तो हम लॉजिस्टिक रिग्रेशन लागू करते हैं। यह डेटा को बाइनरी रूप में वर्गीकृत करता है, यानी सिर्फ 0 और 1 में, जिसका उपयोग रोगियों को मधुमेह के सकारात्मक या नकारात्मक के रूप में वर्गीकृत करने के लिए किया जाता है। लॉजिस्टिक रिग्रेशन का मुख्य लक्ष्य सबसे अच्छा फिट खोजना है, जो लक्ष्य और भविष्यवक्ता चर के बीच संबंध का वर्णन करता है। लॉजिस्टिक रिग्रेशन एक मॉडल है जो लीनियर रिग्रेशन पर आधारित है। लॉजिस्टिक रिग्रेशन मॉडल सकारात्मक और नकारात्मक वर्ग की संभावना की भविष्यवाणी करने के लिए सिग्मॉइड फ़ंक्शन का उपयोग करता है। लॉजिस्टिक रिग्रेशन एल्गोरिथम के चरण-

1. डेटा प्री-प्रोसेसिंग चरण
2. प्रशिक्षण सेट के लिए लॉजिस्टिक रिग्रेशन फिटिंग



चित्र 2. प्रस्तावित कार्य का प्रवाह

3. परीक्षा परिणाम की भविष्यवाणी करना
4. परिणाम की परीक्षण सटीकता मापना
5. परीक्षण सेट परिणाम की कल्पना करना

(ii) ग्रेडिएंट बूस्ट क्लासिफायर (Gradient Boost Classifier)

ग्रेडिएंट बूस्टिंग एक वर्गीकरण दृष्टिकोण है जो भविष्यवाणी के लिए सबसे शक्तिशाली पहनावा तकनीकों में से एक है। यह भविष्यवाणी के लिए शक्तिशाली शिक्षार्थी मॉडल बनाने के लिए कई सप्ताह के शिक्षार्थियों को जोड़ती है। यह डिसीजन ट्री कॉन्सेप्ट पर आधारित है। यह कठिन डेटा सेटों को वर्गीकृत करने के लिए एक अत्यंत सफल और सामान्य तरीका है। ग्रेडिएंट बूस्टिंग मॉडल के प्रदर्शन में प्रत्येक पुनरावृत्ति के साथ सुधार होता है। ग्रेडिएंट बूस्ट एल्गोरिथम के चरण-

1. लक्ष्य मानों के एक नमूने पर P . के रूप में विचार करें।
2. लक्ष्य मूल्यों में त्रुटि का अनुमान लगाएं।
3. त्रुटि एम को कम करने के लिए वजन को अद्यतन और समायोजित करें।
4. पी [एक्स] = पी [एक्स] + अल्फा एम [एक्स]
5. मॉडल शिक्षार्थियों का विश्लेषण और गणना हानि फ़ंक्शन 'एफ' (F) द्वारा की जाती है।
6. वांछित और लक्ष्य परिणाम 'पी' तक चरणों को दोहराएं।

(iii) एडाबूस्ट क्लासिफायर (AdaBoost Classifier)

एडाबूस्ट (AdaBoost) पहली पूरी तरह से सफल बाइनरी क्लासिफिकेशन बूस्टिंग तकनीक थी। AdaBoost, या अनुकूली बूस्टिंग, एक लोकप्रिय बूस्टिंग तकनीक है जो कई ''कमजोर क्लासिफायर'' को एक ''मजबूत क्लासिफायरियर'' में जोड़ती है। बूस्टिंग बड़ी संख्या में कमजोर लोगों से एक मजबूत क्लासिफायरियर बनाने के लिए एक पहनावा मॉडलिंग दृष्टिकोण है। यह कमजोर मॉडलों की एक श्रृंखला से एक मॉडल को एक साथ रखकर किया जाता है। शुरू करने के लिए, एक मॉडल विकसित करने के लिए प्रशिक्षण डेटा का उपयोग किया जाता है। मूल मॉडल की खामियों को ठीक करने के लक्ष्य के साथ दूसरा मॉडल विकसित किया गया है। यह दृष्टिकोण तब तक दोहराया जाता है जब तक या तो पूरे प्रशिक्षण डेटा सेट का ठीक से अनुमान नहीं लगाया जाता है या अधिकतम संख्या में मॉडल जोड़े जाते हैं। एडाबूस्ट एल्गोरिथम के चरण-

1. सबसे पहले, एडाबूस्ट (Adaboost) यादृच्छिक रूप से एक प्रशिक्षण सबसेट चुनता है।
2. यह अंतिम प्रशिक्षण की सटीक भविष्यवाणी के आधार पर प्रशिक्षण सेट को चुनकर एडाबूस्ट (AdaBoost) मशीन लर्निंग मॉडल को पुनरावृत्त रूप से प्रशिक्षित करता है।
3. यह गलत तरीके से वर्गीकृत प्रेक्षणों को अधिक महत्व देता है ताकि उनके अगले पुनरावृत्ति में वर्गीकृत होने की अधिक संभावना हो।

4. यह क्लासिफायरियर की सटीकता के आधार पर प्रत्येक पुनरावृत्ति में प्रशिक्षित क्लासिफायरियर को वजन भी आवंटित करता है। अधिक सटीक क्लासिफायरियर को अधिक भार दिया जाएगा।
5. इस दृष्टिकोण को तब तक दोहराएं जब तक कि सभी प्रशिक्षण डेटा पूरी तरह से फिट न हो जाएं या जब तक अनुमानकर्ताओं की अधिकतम संख्या न हो जाए।
6. वर्गीकृत करने के लिए, आपके द्वारा बनाए गए सभी शिक्षण एल्गोरिद्म पर वोट करें।

(iv) डिसीजन ट्री क्लासिफायर (Decision Trees Classifier)

डिसीजन ट्री में, प्रत्येक आंतरिक नोड एक फीचर टेस्ट का प्रतिनिधित्व करता है (उदाहरण के लिए, क्या एक सिक्का लिप हेड या टेल लैंड करेगा), प्रत्येक लीफ नोड एक क्लास लेबल (सभी सुविधाओं की गणना का परिणाम) का प्रतिनिधित्व करता है, और ब्रांचिंग उन फीचर संयोजनों को इंगित करता है जो उन तक ले जाते हैं वर्ग लेबल। जड़ से पत्ती तक के पथ वर्गीकरण नियमों का वर्णन करते हैं। डिसीजन ट्री एल्गोरिथम के चरण-

1. इनपुट फीचर के रूप में नोड्स के साथ ट्री का निर्माण करें।
2. इनपुट फीचर से आउटपुट की भविष्यवाणी करने के लिए फीचर का चयन करें जिसका सूचना लाभ सबसे अधिक है।
3. उच्चतम सूचना लाभ की गणना पेड़ के प्रत्येक नोड में प्रत्येक विशेषता के लिए की जाती है।
4. उपरोक्त नोड में उपयोग नहीं की जाने वाली सुविधा का उपयोग करके एक सबट्री बनाने के लिए चरण 2 दोहराएं।

(v) गौस्सिअन नैव बयेस (Gaussian Naïve Bayes)

गौस्सिअन नैव बयेस एक नैव बयेस (Naive Bayes) की भिन्नता है जो निरंतर डेटा की अनुमति देती है और गाऊसी सामान्य वितरण का अनुसरण करती है। हमने गौस्सिअन नैव बयेस (Naive Bayes) की अवधारणा को देखा और एक उदाहरण दिया। आरंभ करने से पहले, आइए एक नज़र डालते हैं कि नैव बयेस (Naive Bayes) क्या है। बयेस प्रमेय पर्यवेक्षित मशीन लर्निंग वर्गीकरण एल्गोरिद्म के संग्रह के लिए आधार प्रदान करता है जिसे नैव बयेस (Naive Bayes) कहा जाता है। यह बहुत अधिक शक्ति के साथ एक सरल वर्गीकरण तकनीक है। वे तब उपयोगी होते हैं जब इनपुट की आयामिता अधिक होती है। नैव बयेस (Naive Bayes) क्लासिफायर का उपयोग जटिल वर्गीकरण मुद्दों को हल करने के लिए भी किया जा सकता है।

(vi) रैंडम फॉरेस्ट (Random Forest)

यह एक विधि है जिसका उपयोग वर्गीकरण और प्रतिगमन के लिए किया जा सकता है। अन्य मॉडलों की तुलना में, यह अधिक सटीकता प्रदान करता है। इस रणनीति के साथ बड़े डेटासेट कोई समस्या नहीं हैं। लियो ब्रेमेन रैंडम फ़ॉरेस्ट के निर्माता हैं। यह एक प्रसिद्ध पहनावा सीखने की विधि है। भिन्नता को कम करके, यादृच्छिक वन निर्णय वृक्ष के प्रदर्शन में सुधार करता है। यह प्रशिक्षण के दौरान बड़ी संख्या में निर्णय वृक्षों का निर्माण करके काम करता है और फिर उस वर्ग को आउटपुट करता है जो कि अलग-अलग पेड़ों की कक्षाओं, वर्गीकरण, या औसत भविष्यवाणी (प्रतिगमन) का तरीका है। रैंडम फॉरेस्ट एल्गोरिथम के चरण-

1. पहला कदम कुल सुविधाओं ''एम" से ''आर" सुविधाओं का चयन करना है जहां आर<<एम
2. ''आर" सुविधाओं में, सबसे अच्छा विभाजन बिंदु का उपयोग करने वाला नोड।
3. सर्वोत्तम विभाजन का उपयोग करके नोड को उप-नोड्स में विभाजित करें।
4. 'ए' से 'सी' चरणों को तब तक दोहराएं जब तक कि ''एल" नोड्स की संख्या तक नहीं पहुंच गया हो।
5. पेड़ों की ''एन" संख्या बनाने के लिए ''ए" संख्या के लिए कदम ए से डी दोहराकर जंगल का निर्माण किया।

3.2 प्रतिरूप निर्माण

यह सबसे महत्वपूर्ण चरण है जिसमें मधुमेह की भविष्यवाणी के लिए मॉडल निर्माण शामिल है। इसमें हमने मधुमेह की भविष्यवाणी के लिए विभिन्न मशीन लर्निंग एल्गोरिद्म को लागू किया है। इन एल्गोरिद्म में, लॉजिस्टिक रिग्रेशन, ग्रेडिएंट बूस्टिंग, एडबूस्ट, डिसीजन ट्री, गॉसियन नैव बयेस और रैंडम फॉरेस्ट एल्गोरिथम एल्गोरिथम शामिल हैं।

प्रस्तावित कार्यप्रणाली की प्रक्रिया

चरण 1: आवश्यक पुस्तकालय आयात करें, मधुमेह डेटासेट आयात करें।

चरण 2: लापता डेटा को हटाने के लिए डेटा को प्री-प्रोसेस करें।

चरण 3: डेटासेट को प्रशिक्षण सेट के रूप में और 20% को परीक्षण सेट के रूप में विभाजित करने के लिए 80% का प्रतिशत विभाजन करें।

चरण 4: मशीन लर्निंग एल्गोरिदम यानी लॉजिस्टिक रिग्रेशन (Logistic regression), ग्रेडिएंट बूस्टिंग (Gradient boosting),

एडाबूस्ट (Adaboost), डिसीजन ट्री (Decision Tree), गॉसियन नैव बयेस (Gaussian NB) और रैंडम फॉरेस्ट एल्गोरिथम (Random Forest algorithm) का चयन करें।

चरण 5: प्रशिक्षण सेट के आधार पर उल्लिखित मशीन लर्निंग एल्गोरिदम के लिए क्लासिफायर मॉडल बनाएं।

चरण 6: हाइपर-पैरामीटर में कुछ परिवर्तनों का उपयोग करके सटीकता में सुधार करें।

चरण 7: परीक्षण सेट के आधार पर उल्लिखित मशीन लर्निंग एल्गोरिदम के लिए क्लासिफायर मॉडल का परीक्षण करें।

चरण 8: प्रत्येक क्लासिफायरियर के लिए प्राप्त प्रयोगात्मक प्रदर्शन परिणामों की तुलना मूल्यांकन करें।

चरण 9: विभिन्न उपायों के आधार पर विश्लेषण करने के बाद सर्वोत्तम प्रदर्शन करने वाले एल्गोरिथम का निष्कर्ष निकालें।

4. प्रयोगिक व्यवस्था

निम्नलिखित अनुभागों ने मशीन लर्निंग एल्गोरिदम के साथ प्रयोग करने के लिए सेटअप के बारे में विस्तार से बताया। यहां, PIMA भारतीय डेटासेट का उपयोग सीखने के उद्देश्य के लिए किया जाता है और प्रदर्शन मूल्यांकन का उपयोग विशेष मशीन लर्निंग तकनीकों की समानता की जांच के लिए किया जाता है।

4.1 डेटासेट

पिमा इंडियंस डायबिटीज (पीआईडी) डेटासेट[22] में पाया जा सकता है और 1990 में जॉन्स हॉपकिन्स यूनिवर्सिटी के एप्लाइड फिजिक्स लेबोरेटरी के सदस्य विन्सेंट सिगिलिटो द्वारा दान किया गया था। यह 768 रोगियों की चिकित्सा निदान रिपोर्ट का संकलन है। रोगियों में से कई पीमा भारतीय महिलाएं हैं जो कम से कम 21 वर्ष की हैं और संयुक्त राज्य अमेरिका में फीनिक्स, एरिजोना के पास रहती हैं। कक्षा 1 (सकारात्मक मधुमेह परीक्षण) में 268 मामले और कक्षा 0 (नकारात्मक मधुमेह परीक्षण) में 500 मामले हैं, जो कुल डेटासेट का 34.9 प्रतिशत और 65.1 प्रतिशत है। निम्नलिखित आठ विशेषताएँ (प्लस वर्ग) हैं जिनका वर्णन किया जा सकता हैः

1. गर्भवती होने की संख्या
2. मौखिक ग्लूकोज सहिष्णुता परीक्षण में प्लाज्मा ग्लूकोज एकाग्रता 2 घंटे
3. डायस्टोलिक रक्तचाप (मिमी एचजी)
4. ट्राइसेप्स त्वचा की तह मोटाई (मिमी)
5. 2-घंटे सीरम इंसुलिन (एमयू यू/एमएल)
6. बॉडी मास इंडेक्स (किलो में वजन/(ऊंचाई मीटर में)^2)
7. मधुमेह वंशावली समारोह 8. आयु (वर्ष) 9. वर्ग चर (0 या 1)

4.2 मूल्यांकन पैरामीटर

यह भविष्यवाणी मॉडल का अंतिम चरण है। यहां, हम विभिन्न मूल्यांकन मेट्रिक्स जैसे वर्गीकरण सटीकता, भ्रम मैट्रिक्स और f1-स्कोर का उपयोग करके भविष्यवाणी परिणामों का मूल्यांकन करते हैं।

वर्गीकरण शुद्धता- यह इनपुट नमूनों की कुल संख्या के लिए सही भविष्यवाणियों की संख्या का अनुपात है। इसे समीकरण 1 के रूप में दिया गया हैः

शुद्धता = (सही भविष्यवाणियों की संख्या/की गई भविष्यवाणियों की कुल संख्या)*100 समीकरण. 1

5. परिणाम

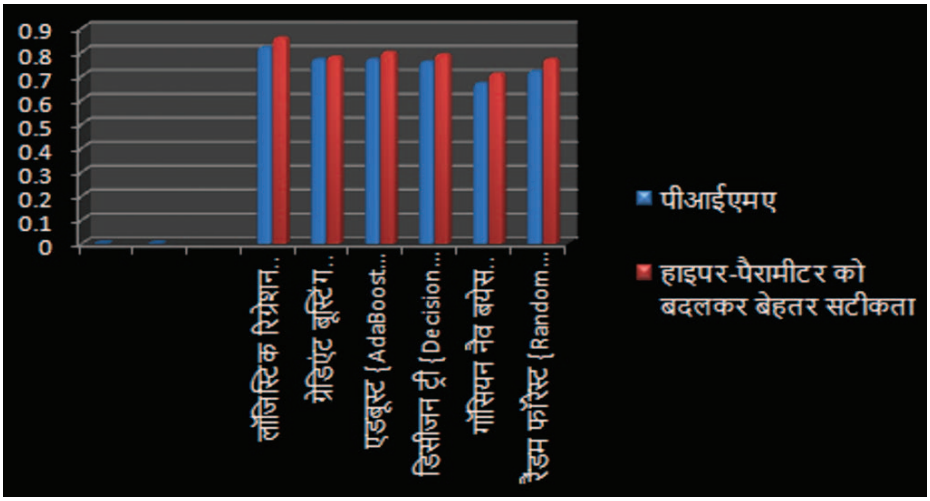
लॉजिस्टिक रिग्रेशन का उपयोग वर्गीकरण की समस्याओं को हल करने के लिए भी किया जा सकता है। सामान्य तौर पर, लॉजिस्टिक रिग्रेशन क्लासिफायर सिग्मॉइड फ़ंक्शन के तर्क के रूप में एक से अधिक फीचर वैल्यू या व्याख्यात्मक चर के रैखिक संयोजन का उपयोग कर सकता है। सिग्मॉइड फ़ंक्शन (Sigmoid Function) का संगत आउटपुट 0 और 1 के बीच की संख्या है। मध्य मान को यह स्थापित करने के लिए थ्रेसहोल्ड माना जाता है कि कक्षा 1 और कक्षा 0 में से क्या संबंधित है। विशेष रूप से, 0.5 से अधिक परिणाम उत्पन्न करने वाले इनपुट को कक्षा 1 के लिए संबंधित माना जाता है। इसके विपरीत, यदि आउटपुट 0.5 से कम है, तो संबंधित इनपुट को 0 वर्ग से संबंधित के रूप में वर्गीकृत किया जाता है। यहां, हाइपर-पैरामीटर सेटिंग तालिका 2 में प्रदर्शित है। डेटासेट पर विभिन्न मशीन लर्निंग एल्गोरिद्म को लागू करने के बाद, पाई गई सटीकता तालिका 3 में उल्लिखित है। लॉजिस्टिक रिग्रेशन 86% की उच्चतम सटीकता देता है। हम चित्र 3 में बार चार्ट के माध्यम से सटीकता भी देख सकते हैं।

तालिका 2: हाइपर-पैरामीटर सेटअप

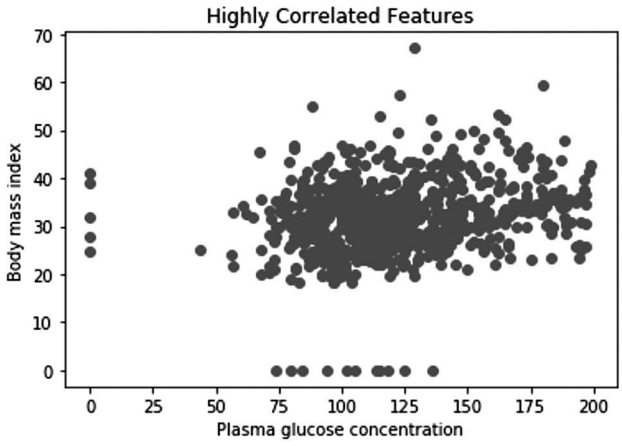
हाइपर-पैरामीटर (Hyper-parameter)	**मूल्यांकन (Value)**
अनुकूलक (Optimizer)	एडम (ADAM)
लॉस फंकशन (Loss function)	श्रेणीबद्ध क्रॉस एंट्रोपी (Categorical Cross Entropy)
नियमितीकरण (Regularization)	रिज/लसो (Ridge/ Lasso)
साधारण कम से कम वर्ग (Ordinary Least Square-OLS)	प- मूल्यांकन (P-value < 0.05)
सक्रियण फंक्शन (Activation Function)	सिग्मोआईड (Sigmoid)

तालिका 3: विभिन्न एल्गोरिदम की शुद्धता

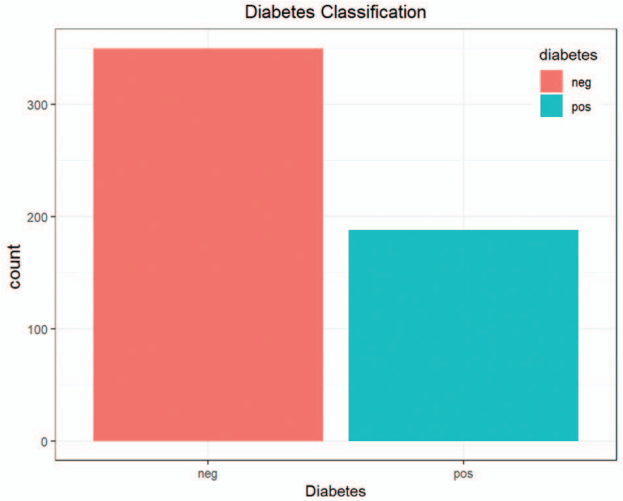
मशीन लर्निंग एल्गोरिदम	पीआईएमए (PIMA) के साथ सटीकता डेटासेट	हाइपर-पैरामीटर को बदलकर बेहतर सटीकता
लॉजिस्टिक रिग्रेशन (Logistic Regression)	82%	86%
ग्रेडिएंट बूस्टिंग (Gradient Boost Classifier)	77%	78%
एडाबूस्ट (AdaBoost Classifier)	77%	80%
डिसीजन ट्री (Decision Trees Classifier)	76%	79%
गॉसियन नैव बयेस (Gaussian NB)	67%	71%
रैंडम फॉरेस्ट (Random Forest)	72%	77%



चित्र 3. विभिन्न एल्गोरिदम की सटीकता के लिए चार्ट



चित्र 4. प्लाज्मा ग्लूकोज और बीएमआई के बीच संबंध

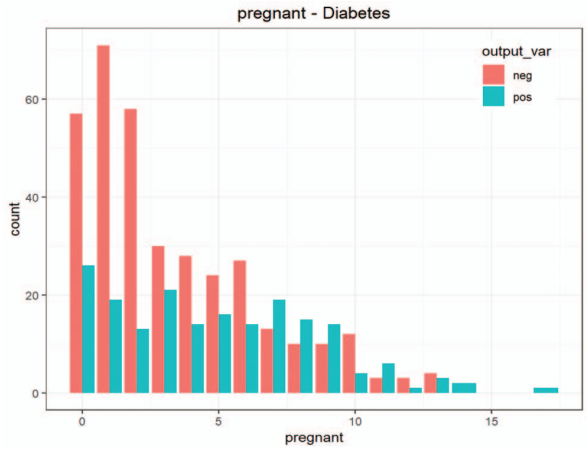


चित्र 5. मधुमेह वर्गीकरण

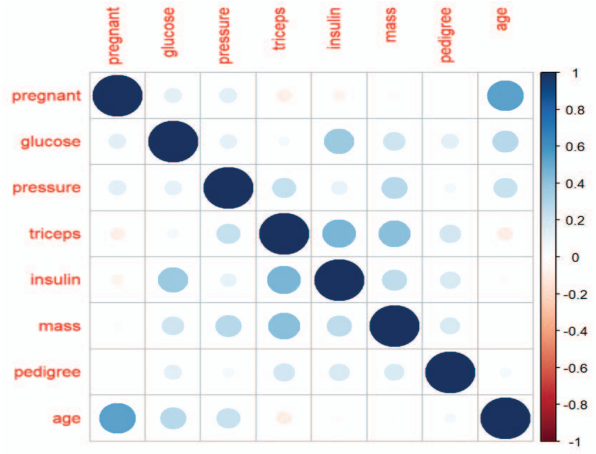
यहाँ हमने पिमा भारतीय डेटासेट को भी विस्तृत रूप में बताया है। निम्म्लिखित तालिका 4 पिमा भारतीय डेटा सेट के सांख्यिकीय विश्लेषण का प्रतिनिधित्व करती है। इसके अलावा, चित्र 4,5,6 और 7, तालिका 4 में उल्लिखित डेटा के बीच संबंध को दर्शाता है।

6. निष्कर्ष और भविष्य का दायरा

इस शोध पत्र में विभिन्न मशीन लर्निंग विधियों की समीक्षा एक ही स्थान पर उनके मूल्यांकन के साथ प्रस्तुत की गई है। लेखक मशीन



चित्र 6. गर्भवती महिलाओं का मधुमेह वर्गीकरण



चित्र 7. सहसंबंध मानचित्र

तालिका 4: पिमा इंडियन डेटासेट सांख्यिकीय विश्लेषण

	संख्या (Count)	मध्य (Mean)	मानक विचलन (Standard Deviation)	श्रेणी (Range)
गर्भवती की संख्या (Number of times pregnant)	768.0	3.84	3.36	0-17
प्लाज्मा ग्लूकोज एकाग्रता (Plasma glucose concentration)	768.0	120.89	31.97	0-199
डायस्टोलिक रक्त दबाव (Diastolic blood pressure)	768.0	69.10	19.35	0-122
ट्राइसेप्स स्किन फोल्ड मोटाई (Triceps skin fold thickness)	768.0	20.53	15.95	0-99
2 घंटे का सीरम इंसुलिन (2-Hour serum insulin)	768.0	79.79	115.24	0-846
बॉडी मास इंडेक्स (Body mass index)	768.0	31.99	7.88	0-67
मधुमेह वंशावली समारोह (Diabetes pedigree function)	768.0	0.471	0.33	0.078-2.4
उम्र (Age)	768.0	33.24	11.76	21-81
परिणाम (Outcomes)	768.0	0.348	0.47	Yes/No (1/0)

अर्निंग के तरीकों और मधुमेह डेटासेट के लिए डोमेन निर्भर या गैर-डोमेन निर्भर सुविधाओं के साथ इन विधियों के उपयोग को समझ सकते हैं। इन विधियों का मूल्यांकन भी दिखाया गया है, जिससे भविष्य में अन्य अनुप्रयोगों के लिए सर्वोत्तम परिणाम के तरीकों का उपयोग किया जा सकता है। इस मूल्यांकन के अनुसार, पिमा इंडियन डेटासेट के लिए लॉजिस्टिक रिग्रेशन ने डिसीजन ट्री, रैंडम फ़ॉरेस्ट आदि की तुलना में सबसे अच्छा प्रदर्शन किया है। पिमा इंडियन डेटासेट के लिए लॉजिस्टिक रिग्रेशन की सटीकता 86% है जिसे दूसरे तकनीकों की तुलना में सर्वश्रेष्ठ स्कोर मिला है। हमने मशीन लर्निंग एल्गोरिथम की सटीकता की तुलना देखी है। यह स्पष्ट है कि मॉडल मौजूदा डेटासेट की तुलना में इस डेटासेट के साथ मधुमेह की भविष्यवाणी की सटीकता और सटीकता में सुधार करता है। इसके अलावा इस काम को आगे बढ़ाया जा सकता है ताकि यह पता लगाया जा सके कि अगले कुछ वर्षों में गैर-मधुमेह लोगों को मधुमेह होने की कितनी संभावना है।

शोध पत्र में प्रयुक्त अंग्रेजी शब्दावलियों की तालिका:

English	Hindi
Algorithm	कलन विधि
Analysis	विश्लेषण

AdaBoost Classifier	एडाबूस्ट
Computer	संगणक
Classifier	वर्गीकरणकर्ता
Diabetes	मधुमेह
Decision Tree	डिसीजन ट्री
Gradient Boosting	ग्रेडिएंट बूस्टिंग
Gaussian Naïve Bayes	गॉसियन नैव बयेस
Hyperparameter	हाइपर-पैरामीटर
Logistic Regression	लॉजिस्टिक रिग्रेशन
Machine Learning	यंत्र अधिगम (मशीन लर्निंग)

संदर्भ

1. चो, एन.; शॉ, जे.; करुरांगा, एस.; हुआंग, वाई.; फर्नांडीस, जे. डी.आर.; ओह्लोग, ए.; मालांडा, बी. आईडीएफ डायबिटीज एटलसः 2017 के लिए डायबिटीज की व्यापकता का वैश्विक अनुमान और 2045 के लिए अनुमान। डायबिटीज रिस. क्लिन. पीआर. 2018, 138, 271-281।
2. सान्ज़, जे.ए.; गैलर, एम.; जुरियो, ए.; ब्रुगोस, ए.; पैगोला, एम.; बुस्टिंस, एच. अंतराल-मूल्यवान फ़ज़ी नियम-आधारित वर्गीकरण प्रणाली का उपयोग करके हृदय रोगों का चिकित्सा निदान। एप्लाइड सॉफ्ट कंप्यूट। 2014, 20, 103-111।
3. कंधासामी, जे.पी.; बालमुरली, एस विज्ञान 2015, 47, 45-51।
4. अय्यर, ए.; जयलता, एस.; सुंबाली, आर. वर्गीकरण खनन तकनीकों का उपयोग कर मधुमेह का निदान। अंतर्राष्ट्रीय जर्नल डेटा मिन. ज्ञान प्रबंधन प्रक्रिया। 2015, 5, 1-14।
5. रज़ावियन, एन.; ब्लेकर, एस.; श्मिट, ए.एम.; स्मिथ-मैकलेलन, ए.; निगम, एस.; सोनटैग, डी. दावा डेटा और जोखिम कारकों के विश्लेषण से टाइप 2 मधुमेह की जनसंख्या-स्तर की भविष्यवाणी। बिग डेटा 2015, 3, 277-287।
6. एस. कुमारी, डी. कुमार, और एम. मित्तल, ''सॉफ्ट वोटिंग क्लासिफायर का उपयोग करके मधुमेह मेलेटस के वर्गीकरण और भविष्यवाणी के लिए एक समूह दृष्टिकोण,'' इंटरनेशनल जर्नल 2, 2021।
7. एम.एम.एफ. इस्लाम, आर. फिरदौसी, एस. रहमान, और एच. वाई. बुशरा, ''डेटा माइनिंग तकनीकों का उपयोग करके प्रारंभिक अवस्था में मधुमेह की संभावना का पूर्वानुमान,'' मेडिकल इमेज एनालिसिस में कंप्यूटर विज़न और मशीन इंटेलिजेंस में, पीपी. 113-125, स्प्रिंगर, सिंगापुर, 2020।
8. एस. मलिक, एस. हारौस, और एच.ई. सईद, ''महिलाओं में मधुमेह के प्रारंभिक पूर्वानुमान के लिए मशीन लर्निंग एल्गोरिदम का तुलनात्मक विश्लेषण,'' मॉडलिंग और जटिल प्रणालियों के कार्यान्वयन पर अंतर्राष्ट्रीय संगोष्ठी की कार्यवाही में, पीपी. 95-106, स्प्रिंगर, बटना, अल्जीरिया, अक्टूबर 2020।
9. वाई.के. कौक़ज़ेह, ए.एस. बजाहज़ार, एम. जेम्माली, एम.एम. ओटूम, और ए.-अलजौई, ''फोटोप्लेथिस्मोग्राम (पीपीजी) का उपयोग करके मधुमेह का वर्गीकरण तरंग विश्लेषणः लॉजिस्टिक रिग्रेशन मॉडलिंग,'' बायोमेड रिसर्च इंटरनेशनल, वॉल्यूम। 2020, आर्टिकल आईडी 3764653, 6 पेज, 2020।
10. जी. ए. पेथुनाचियार, ''कर्नेल आधारित सपोर्ट वेक्टर मशीनों का उपयोग करके मधुमेह रोगियों का वर्गीकरण,'' 2020 इंटरनेशनल कॉन्फ्रेंस ऑन कंप्यूटर कम्युनिकेशन एंड इंफॉर्मेटिक्स (ICCCI) की कार्यवाही में, पीपी. 1-4, IEEE, कोयंबटूर, भारत, जनवरी 2020।
11. एस. गुप्ता, एच. के. वर्मा, और डी. भारद्वाज, ''एक तकनीक के रूप में नैवे बेयस और सपोर्ट वेक्टर मशीन का उपयोग करके मधुमेह का वर्गीकरण,'' ऑपरेशंस मैनेजमेंट एंड सिस्टम इंजीनियरिंग, स्प्रिंगर, सिंगापुर, पीपी. 365-376, 2021।
12. डी. के. चौबे, एम. कुमार, वी. शुक्ला, एस. त्रिपाठी, और वी. के. ढांढनिया, ''मधुमेह के लिए पीसीए और एलडीए के साथ वर्गीकरण विधियों का तुलनात्मक विश्लेषण 16, सं. 8, पृ. 833-850, 2020।
13. के. विजिय कुमार, बी. लावण्या, आई. निर्मला, एस. सोफिया कैरोलीन, ''मधुमेह की भविष्यवाणी के लिए रैंडम फ़ॉरेस्ट एल्गोरिदम''। सिस्टम कम्प्यूटेशन ऑटोमेशन और नेटवर्किंग पर अंतर्राष्ट्रीय सम्मेलन की कार्यवाही, 2019।
14. तेजस एन. जोशी, प्रो. प्रमिला एम. चव्हाण, ''मशीन लर्निंग तकनीकों का उपयोग करके मधुमेह की भविष्यवाणी''। अंतर्राष्ट्रीय जर्नल ऑफ़ इंजीनियरिंग रिसर्च एंड एप्लीकेशन, खंड 8, अंक 1, (भाग-II) जनवरी 2018, पृ.-09-13।
15. दीराज शेट्टी, किशोर रीत, सोहेल शेख, निकिता पाटिल, ''डेटा माइनिंग का उपयोग करके मधुमेह रोग की भविष्यवाणी''। सूचना, एम्बेडेड और संचार प्रणालियों में नवाचारों पर अंतर्राष्ट्रीय सम्मेलन (ICIIECS), 2017।

16. परवीन, एस., शाहबाज, एम., गुएर्गाची, ए. और केशवजी, के., 2016। मधुमेह की भविष्यवाणी करने के लिए डेटा माइनिंग वर्गीकरण तकनीकों का प्रदर्शन विश्लेषण। प्रोसीडिया कंप्यूटर साइंस, 82, पीपी. 115-121।

17. सेल्वाकुमार, एस., कन्नन, के.एस. और गोथाई नचियार, एस., 2017। वर्गीकरण आधारित डेटा माइनिंग तकनीकों का उपयोग करके मधुमेह निदान की भविष्यवाणी। सांख्यिकी और प्रणालियों के अंतर्राष्ट्रीय जर्नल, 12(2), पीपी. 183-188।

18. सिसोदिया, डी. और सिसोदिया, डी.एस., 2018। वर्गीकरण एल्गोरिद्म का उपयोग करके मधुमेह की भविष्यवाणी। प्रोसीडिया कंप्यूटर साइंस, 132, पीपी. 1578-1585।

19. मुजुमदार, ए. और वैदेही, वी., 2019। मशीन लर्निंग एल्गोरिदम का उपयोग करके मधुमेह की भविष्यवाणी। प्रोसीडिया कंप्यूटर साइंस, 165, पीपी. 292-299।

20. झोउ, एच., मिर्ज़ाशोवा, आर. और झेंग, आर., 2020। एक उन्नत डीप न्यूरल नेटवर्क पर आधारित मधुमेह भविष्यवाणी मॉडल। वायरलेस संचार और नेटवर्किंग पर EURASIP जर्नल, 2020 (1), पीपी. 1-13।

21. अहमद, एच.एफ., मुख्तार, एच., अलक़ैल, एच., सेलियामन, एम. और अलहुमाम, ए., 2021। मशीन लर्निंग का उपयोग करके मधुमेह की भविष्यवाणी पर स्वास्थ्य संबंधी विशेषताओं और उनके प्रभाव की जांच करना। एप्लाइड साइंसेज, 11(3), पृ. 1173।

22. http://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes/pima-indians-diabetes.data