मशीन लर्निंग आधारित डायबिटीज संसूचन : वर्गीकरणकारी तकनीकों का एक तुलनात्मक अध्ययन

Machine Learning Based Diabetes Detection - A Comparative Study of Classifiers

श्रीमती सोफिया गोयल[1], डॉ. सुधांश शर्मा[2]

[1]रिसर्च स्कॉलर,सोसिस, इग्नू, नई दिल्ली, इंडिया
[2]असिस्टेंट प्रोफेसर,सोसिस, इग्नू, नई दिल्ली, इंडिया

[1]sofiagoel@gmail.com; [2]sudhansh@ignou.ac.in

सारांश

मधुमेह पूरी दुनिया के स्वास्थ्य के लिए एक चिंता जनक विषय है, इसका निदान और इलाज चिकित्सा क्षेत्र के लिए प्रमुख चुनौती है, क्योंकि इसे नियंत्रित किया जा सकता है लेकिन इसे ठीक नहीं किया जा सकता है। इस रोग का निदान जितना जल्दी होगा, रोगी के लिए उतना ही अच्छा होगा, और इस बीमारी के निदान के लिए हम मशीन लर्निंग (एमएल–ML) की तकनीक का उपयोग कर सकते हैं। हम जानते हैं कि मधुमेह का समय पर निदान हो जाए तो रोगी को विभिन्न प्राण घातक रोगों से बचाया जा सकता है. इस कारण,प्रस्तुत पेपर (आर्टिकल) में हम मशीन लर्निंग की विभिन्न वर्गीकरण तकनीकों का तुलनात्मक अध्ययन प्रस्तुत करने का प्रयास कर रहे है. मशीन लर्निंग (एमएल) में विभिन्न वर्गीकरण तकनीक उपलब्ध हैं, जैसे कि सपोर्ट वेक्टर मशीन (Support Vector Machine-SVM) (एस.वी.एम), के–नियरेस्ट नेबर (K-Nearest Neighbor-KNN) (के.एन.एन),एल्गोरिथ्म, डिसिशनट्री (Decision Tree), आर्टिफिशियल न्यूरल नेटवर्क्(Artificial Neural Network-ANN)(ए.एन.एन), नैव बैस क्लासिफायर (Naïve Bayes Classifier) आदि। लेकिन सवाल यह है कि मधुमेह की पहचान करने में वर्गीकरण तकनीकों में से कौन सी तकनीक सटीक है, क्योंकि निदान की सटीकता अधिक महत्वपूर्ण है। प्रस्तुत अध्ययन में वर्गीकरण तकनीकों की सटीकता जांचने के लिए हमने UCI रिपॉजिटरी द्वारा जारी किए गए पीमा इंडियन (PIMA INDIAN) नामक डाटासेट पर डाटा इंप्यूटेशन तकनीक का उपयोग किया है। उसके बाद तुलनात्मक अध्ययन के लिए हमने इस डाटा सेट पर मशीन लर्निंग की विभिन्न वर्गीकरण तकनीकों का प्रयोग किया और मधुमेह निदान की सटीकता जाँची है। तुलनात्मक अध्ययन के फलस्वरुप यह पाया गया कि आर्टिफिशियल न्यूरल नेटवर्क तकनीक की सटीकता सर्वोत्तम, 88.6% है, जो अन्य तकनीकों के मुकाबले काफी ज्यादा है।सटीकता का यह स्तर इंगित करता है कि आर्टिफिशियल न्यूरल नेटवर्क में डेटा को वर्गीकृत करने की उच्च क्षमता है, जो इसकी हाई कंप्यूटिंग आर्किटेक्चर से संबंध रखता है, जिसके कारण बिना फीचर इंजीनियरिंग के इसको प्रोग्रामित किया जा सकता है।

Abstract

Diabetes is a matter of concern for the health of the entire world, its diagnosis and cure are among the prime challenges for the medical fraternity, because it can be controlled but can't be cured,sooner the diagnosis the better it will be for the patient.Thus,use of machine learning for

earlier classification of diabetes plays a vital role to protect patient from the life threatening complications in future, various classification techniques are available in Machine Learning (ML) viz. Support Vector Machines (SVM), K-Nearest Neighbor (KNN) algorithm, Decision Trees, Artificial Neural Networks (ANN), Naïve Bayes Classifier etc. But the question is which of the classification techniques is quite accurate in identifying the Diabetes, accuracy of diagnosis is more important. Thus, in the present work the assurance of accuracy level is strengthened by applying the data imputation techniques on the dataset i.e.dataset of Pima Indian women of Arizona named as PIMA INDIAN DATASET from UCI repository.Thereafter the said ML techniques were compared, and it is found that among the said classifiers ANN shows high performance as compared to other conventional classifiers. Highest accuracy achieved with ANN is 88.6%. This level of accuracy indicates that neural network has high potential to classify the data, which relates to its high computing layer architecture that can be programmed without any featured engineering.

मुख्य शब्द : सपोर्ट वेक्टर मशीन (एस.वी.एम), के–नियरेस्ट नेबर (के.एन.एन) एल्गोरिथ्म, डिसीशन ट्रीज, आर्टिफिशियल न्यूरल नेटवर्क्स (ए.एन.एन), नैव बैस क्लासिफायर, डेटा इंप्यूटेशन, पीमा इंडियन नामक डाटा सेट।

Keywords : Support Vector Machines (SVM), K-Nearest Neighbor (KNN) algorithm, Decision Trees, Artificial Neural Networks (ANN), Naïve Bayes Classifier, Data imputation, PIMA Indian dataset.

परिचय :

मधुमेह एक घातक बीमारी है, जो दुनिया भर में बहुत तेजी से फैल रही है। इस बीमारी मे शरीर पर्याप्त इंसुलिन का उत्पादन करने में असमर्थ हो जाता है, जिससे रक्त शर्करा (ब्लड शुगर) का स्तर बढ़ जाता है, जो आगे चलकर खतरनाक बीमारियों [2] का कारण बनता है। यह अनियंत्रित रक्त शर्करा कई जटिलताओं और विकारों का मूल कारण है, जैसे – दृष्टि विकार, हृदय संबंधी समस्याएं, धमनी उच्च रक्तचाप, त्वचा की समस्याएं, स्ट्रोक, तंत्रिका क्षति, गुर्दे की विफलता यहां तक कि दिल का दौरा भी, अनियंत्रित रक्त शर्करा से हो सकता है।

अनियंत्रित रक्त शर्करा से उत्पन्न होने वाली इन जटिलताओं को देखते हुए शोधकर्ताओं ने विभिन्न पारंपरिक मशीन लर्निंग तकनीकों का उपयोग करके, रोग को वर्गीकृत करने के लिए कई शोध किए हैं, लेकिन रोग का पता लगाने की सटीकता (एक्यूरेसी) के स्तर के संदर्भ में एक तुलनात्मक अध्ययन हमेशा समय की आवश्यकता के रूप में महसूस किया जाता रहा है। इस प्रस्तुत पत्र में हमने पारंपरिक मशीन लर्निंग की तकनीकों के साथ आर्टिफिशियल न्यूरल नेटवर्क्स (ANN) तकनीक की डायबिटीज डिटेक्शन सटीकता से तुलना की है, क्योंकि डायबिटीज डिटेक्शन की सटीकता अधिक महत्वपूर्ण है। प्रस्तुत अध्ययन में वर्गीकरण तकनीकों की सटीकता जांचने के लिए हमने UCI रिपॉजिटरी द्वारा जारी किये गए पीमा इंडियन (PIMA INDIAN)नामक डाटा सेट [1] पर डाटा इंप्यूटेशन तकनीक का उपयोग किया है। उसके बाद तुलनात्मक अध्ययन के लिए इस डाटा सेट पर हमने मशीन लर्निंग की विभिन्न वर्गीकरण तकनीकों का उपयोग किया और उन वर्गीकरण तकनीकों की मधुमेह डिटेक्शन में सटीकता की तुलनात्मक जाँच की है।

UCI रिपॉजिटरी द्वारा जारी किये गए पीमा इंडियन (PIMA INDIAN)नामक डाटा सेट [1] में 8 एट्रिब्यूट, 768 इंस्टैंस और 1 बाइनरी क्लास एट्रिब्यूट शामिल हैं। विशेषताओं का विवरण तालिका –1 में नीचे प्रस्तुत किया गया है।

तालिका 1. पीमा भारतीय डेटासेट के एट्रिब्यूट

क्र.सं.	विशेषता का नाम	विवरण
1	गर्भावस्था	संख्यात्मक (Numeric)
2	प्लाज्मा ग्लूकोज कंसंट्रेशन	ग्लूकोज टॉलरेंस टेस्ट में 2 घंटे
3	डायस्टोलिक रक्तचाप	mm Hg
4	ट्राइसेप्स स्किन फोल्ड थिकनेस	Mm
5	2– एच सीरम इंसुलिन	mu U/ml
6	बॉडी मास इंडेक्स (BMI)	Kgm-2
7	मधुमेह वंशावली समारोह	संख्यात्मक (Numeric)
8	आयु	वर्ष

इस डेटासेट के चुनाव के पीछे का मुख्य कारण यह है कि इसमें गर्भावस्था और रक्त प्लाज्मा विवरणों के साथ उम्र, अधिक वजन, बीएमआई, 2 एच सीरम इंसुलिन, डायस्टोलिक रक्तचाप जैसी विशेषताएं भी शामिल हैं। दूसरी बात यह है कि इसमें डायबिटीज पेडिग्री फंक्शन (DPF) शामिल है, जो कि रिश्तेदारों में डायबिटीज मेलिटस के इतिहास और मरीज के रिश्तेदारों के आनुवांशिक संबंधों के आंकड़ों से संबंधित है।

इसके अलावा, यह देखा गया कि कई डेटासेटों की तरह, इस डेटासेट में भी कहीं कहीं पर कुछ आंकड़ों की अनुपस्थिति थी (Missing data values), जो कि मजबूत तकनीकी वर्गीकरण मॉडल तैयार करने के लिए शोधकर्ताओं द्वारा सबसे बड़ी चुनौती है। इस लिए, हमारा शोध सबसे पहले अनुपस्थित आंकड़ों को और ऑउटलायर्स के निरीक्षण के लिए है [3], निरीक्षण के उपरांत उन आंकड़ों को सही करने के पश्चात, हमने विभिन्न वर्गीकरण तकनीकों का डेटासेट पर प्रयोग किया और उन वर्गीकरण तकनीकों की मधुमेह डिटेक्शन में सटीकता की तुलनात्मक जाँच की है।

पेपर का शेष भाग निम्नानुसार आयोजित किया गया है: खण्ड (Section)2 में साहित्य समीक्षा, खण्ड (Section)3 में मेथड्स और डेटासेट शामिल हैं, खण्ड (Section) 4 में अनुसंधान के लिए अपनाई गई पद्धति का वर्णन किया है। इसके अलावा खण्ड (Section)5 में परिणाम और चर्चा प्रस्तुत की गयी है, और अंत में खण्ड (Section)6 में प्रस्तुत शोधपत्र का निष्कर्ष दिया गया है।

साहित्य समीक्षा :

शोधकर्ताओं ने विभिन्न पारंपरिक मशीन लर्निंग दृष्टिकोणों का उपयोग करके, रोग को वर्गीकृत करने के लिए कई शोध किए हैं, और इन शोधों के अध्ययन से यह पता चलता है कि टाइप 2 डायबिटीज मेलिटस (T2DM) का सटीक विश्लेषण करने के लिए एक मॉडल की जबरदस्त जरूरत है। नाहला एट. अल. (Nahla et al.) [5] ने, मधुमेह रोग के निदान के लिए सिक्वेनशियल कवरिंग अप्रोच (Sequential Covering Approach) लागू करके एक सपोर्ट वेक्टर मशीन (Support Vector Machine & SVM) आधारित मॉडल प्रस्तुत किया। उनके प्रस्तुत काम में डेटा की सब–सैम्पलिंग के लिए के–मीन्स क्लस्टरिंग तकनीक का उपयोग किया गया व उनके नियम रक्त शर्करा के स्तर (एफबीएस-FBS-Fasting Blood Sugar) और कमर परिधि (Waist Circumference) जैसे कारकों पर आधारित थे। इस प्रस्तावित मॉडल में, डायबिटीज की बीमारी को डायग्नोज करने के लिए, डेटा के बजाय मॉडल

से प्राप्त नियमों का उपयोग किया गया है। झेंग एट अल (Zhang et. al.) [5] ने 23,281 मधुमेह से संबंधित रोगियों में से 300 रोगियों के नमूनों पर एक अध्ययन किया, उनका प्रदर्शित अध्ययन कार्य तीन चीजों पर आधारित था जैसे मधुमेह की दवा (medication of diabetes), असामान्य प्रयोगशाला परीक्षण (Abnormal Laboratory Tests) और मधुमेह निदान (Diagnosis of Diabetes)। उन्होंने अपने एल्गोरिथ्म की तुलना, विभिन्न मशीन लर्निंग एल्गोरिथ्म (जैसे की सपोर्ट वेक्टर मशीन (Support Vector Machine-SVM) (एसवीएम), के–नियरेस्ट नेबर (K-Nearest Neighbour-KNN) (केएनएन) एल्गोरिथ्म, डिसिशन ट्री (Decision Tree), नैव बैस क्लासिफायर (Naïve Bayes Classifier), लोजिस्टिक रिग्रेशन (Logistic Regression) से, एक्यूरेसी स्पेसिफिसिटी और सेंस्टिविटी के आधार पर की थी।

शोधकर्ताओं ने मधुमेह के वर्गीकरण (Classification) और प्रीडिक्शन के लिए न्यूरल नेटवर्क (Neural Network) का भी अध्ययन किया, शोधकर्ताओं ने पाया कि, ट्रेनिंग डाटासेट से उत्पन्न समानताओं के आधार पर न्यूरल नेटवर्क में मधुमेह की सटीक प्रीडिक्शन/भविष्यवाणी करने की क्षमता है। दीजाना सेजदनोवी एट अल (Dijana Sejdinovi et al.) [6] ने टाइप–2 डायबिटीज मेल्लिटस (T2DM) के वर्गीकरण के लिए न्यूरल नेटवर्क्स का उपयोग किया था। तत्पश्चात मधुमेह को प्रेडिक्ट करने के लिए पारास्टो रहिमलो एट अल (Parastoo Rahimloo et al.) [7] ने आर्टिफिशियल न्यूरल नेटवर्क और लॉजिस्टिक रिग्रेशन का उपयोग करके एक हाइब्रिड न्यूरल नेटवर्क मॉडल प्रस्तुत किया।महमूद हैदरी एट अल (Mahmoud Heydari et al.) [8] [Classification] ने डायबिटीज डाटासेट पर विभिन्न क्लासिफिकेशन/वर्गीकरण के एल्गोरिथ्म का अध्ययन किया और आर्टिफिशियल न्यूरल नेटवर्क्स के उपयोग से उन्होंने आशाजनक परिणाम पाए, पर और बेहतर परिणामों के लिए उनके शोध अध्ययन में हाइपर पैरामीटर्स को ऑप्टिमाइज करने की गुंजाइश थी। पिछले शोध कार्यों की समीक्षा से निष्कर्ष निकलता है कि पारंपरिक टेक्निक्स [9] द्वारा हासिल की गई सटीकता पर सुधार करने की गुंजाइश है। और न्यूरल नेटवर्क्स, पारम्परिक मशीन लर्निंग क्लासिफायर्स [10] की तुलना में सटीक परिण ाम देने में सक्षम है।

पिछले शोध कार्यो के अध्ययन से पता चलता है कि क्लासिफायर्स की सटीकता और परफॉरमेंस को बेहतर बनाने के लिए, शोधकर्ताओं ने अपने द्व ारा बनाये गए प्रेडिक्शन मॉडल में विभिन्न कारको को शामिल किया था, जिस से उनके मॉडल की जटिलता बढ़ गयी, फलस्वरूप उनके मॉडल का कम्प्यूटेशन टाइम भी बढ़ गया।

इस समस्या को हल करने के लिए प्रस्तुत शोध पत्र में हमने प्रयास किया है कि मॉडल की जटिलता को बढ़ाये बिना, क्लासिफिकेशन परफॉरमेंस और कम्प्यूटेशन टाइम में सामंजस्य बिठाया जाए, साथ ही हमने एल्गोरिथ्म की एक्यूरेसी का तुलनात्मक अध्ययन भी किया है। आँकी गयी समस्या के निवारण के लिए हमने विभिन्न डाटा इम्प्युटेशन टेक्निक्स का प्रयोग किया है,जैसे कि अनुपस्थित आंकड़ों एवं डाटा ऑउटलायर्स का निरीक्षण तथा अनुपस्थित आंकड़ों को संजोना। अनुपस्थित डाटा को किस प्रकार संजोना है, यह समस्या एक शोधकर्ता के लिए जटिल समस्या होती है, और एक मजबूत डेटा वर्गीकरण मॉडल बनाने के लिए, इन अनुपस्थित आंकड़ों को संजोना अति आवश्यक भी है। इस लिए, हमारा शोध सबसे पहले अनुपस्थित आंकड़ों एवं ऑउटलायर्स के निरीक्षण के लिए है [3], निरीक्षण के उपरांत उन अनुपस्थित आंकड़ों को सही करने के पश्चात, हमने विभिन्न वर्गीकरण तकनीकों का नए डेटासेट पर प्रयोग किया। एएनएन (ANN) को इस्तेमाल करके, मधुमेह (डायबिटीज) डिटेक्शन के सटीकता स्तर को ज्ञात किया और फिर एएनएन

(ANN) के सटीकता स्तर की पारंपरिक क्लासिफायर के सटीकता स्तर से तुलना की गयी।

डेटासेट में दिए गए मानों में से असामान्य मानों को ऑउटलायर (Outlier) कहा जाता है, वे या तो बहुत ज्यादा या बहुत कम मान रखते हैं, और अपेक्षित सीमा में नहीं आते हैं। यह ऑउटलायर (Outlier) किसी भी माप में परिवर्तनशीलता या प्रयोगात्मक त्रुटि का परिणाम हो सकते है। इस लिए एक शोधकर्ता को ऑउटलायरस को बड़ी सावधानी से संभालना होता है, विशेष रूप से चिकित्सा डेटा में, जहां यह बीमारी की भविष्यवाणी की सटीकता के लिए बहुत महत्वपूर्ण है। सामान्य तौर पर शोधकर्ता अनुपस्थित आंकड़ों [9] [11] को नियंत्रित करने के लिए माध्य या माध्यिका की गणना करते हैं, और इस दृष्टिकोण को एक अन्य शोधकर्ता कांग एट अल (Kang et al.) [12] ने सुधारा, जो कि एनेस्थेसियोलॉजी और दर्द चिकित्सा विभाग में विभिन्न प्रकार के अनुपस्थित आंकड़ों का अध्ययन करने वाले एक शोधकर्ता है।

प्रस्तुत कार्य का उद्देश्य है कि डाटा सेट में डाटा से जुडी त्रुटियों को हटाकर एक सुदृढ़ डाटासेट का निर्माण करना, तत्पश्चात डायबिटीज के डायग्नोसिस की एक्यूरेसी को ज्ञात करने के लिए डाटासेट पर मशीन लर्निंग के विभिन्न क्लासिफिकेशन एल्गोरिथ्म का प्रयोग करना और अंतत: डायबिटीज डिटैक्शन में क्लासिफिकेशन एल्गोरिथ्म की एक्यूरेसी को आर्टिफिशियल न्यूरल नेटवर्क की एक्यूरेसी से तुलना करना है।

3. वर्गीकरण तकनीकें :

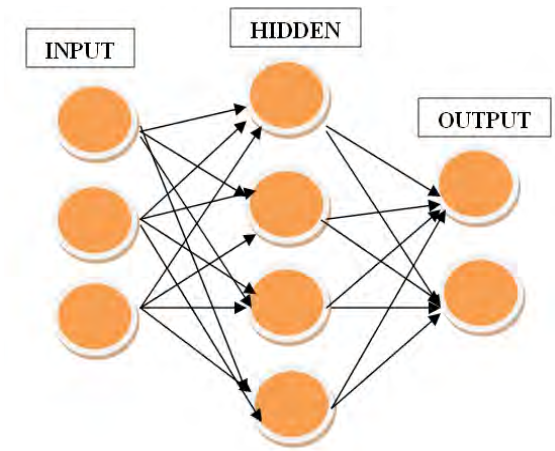
प्रस्तुत अध्ययन में वर्गीकरण तकनीकों की सटीकता जांचने के लिए हमने UCI रिपॉजिटरी द्वारा जारी किये गए पीमा इंडियन (PIMA INDIAN) नामक डाटा सेट [1] पर डाटा इंप्यूटेशन तकनीक (क्लास–वाइज मीन, तथा 2% विनसोराइजेशन) का उपयोग किया है। उसके बाद तुलनात्मक अध्ययन के लिए इस डाटा सेट पर हमने मशीन लर्निंग की विभिन्न वर्गीकरण तकनीकों का उपयोग किया और मधुमेह निदान की सटीकता जाँची है।प्रस्तुत शोध में प्रयोग की गयी तकनीक इस प्रकार है, सपोर्ट वेक्टर मशीन (Support Vector Machine), के–नियरेस्ट नेबर (K-Nearest Neighbour) एल्गोरिथ्म, डिसिशन ट्री (Decision Tree), आर्टिफिशियल न्यूरल नेटवर्क (Artificial Neural Network), नैव बैस क्लासिफायर (Naïve Bayes Classifier), इस्तेमाल की गयी इन तकनीकों की चर्चा नीचे की गयी है।

3.1. आर्टिफिशियल न्यूरल नेटवर्क (ANN)

आर्टिफिशियल न्यूरल नेटवर्क ऐसा डीप लर्निंग सिस्टम है जो वर्गीकरण (classification) और प्रतिगमन (Regression)विश्लेषण [13] के लिए उपयोग किया जाता है, इस सिस्टम की कार्यशैली मानव मस्तिष्क के समान है, जहां पर न्यूरॉन्स परस्पर जुड़े होते हैं और समस्याओं को हल करने के लिए सामूहिक रूप से काम करते हैं।

ANN में तीन लेयर होती हैं, इनपुट (Input), हिडन (Hidden) और आउटपुट (Output) लेयर होते हैं।

1. इनपुट (Input) लेयर (Layer): डेटा को प्राप्त करती है और उस डाटा को नेटवर्क के अगले भाग को सौंप देती है।
2. हिडन (Hidden) लेयर (Layer): इनपुट डाटा तथा इनपुट लेयर न्यूरॉन्स और हिडेन लेयर न्यूरॉन्स के बीच के कनेक्शन वेट्स, हिडन लेयर का कार्य निर्धारण करते है। इनपुट लेयर और हिडेन लेयर के बीच के कनेक्शन वेट्स, निर्धारित करते है कि कब हिडन लेयर न्यूरॉन सक्रिय (एक्टिव) होगा और कब वह निष्क्रिय (इनएक्टिव) होगा।
3. आउटपुट (Output) लेयर (Layer) हिडन लेयर के न्यूरॉन्स का आउटपुट तथा हिडन लेयर और आउटपुट लेयर के बीच के कनेक्शन वेटस, आउटपुट लेयर के न्यूरॉन्स का कार्य निर्धारण करते है।



चित्र 1. आर्टिफिशियल न्यूरल नेटवर्क (ANN) की संरचना।

किसी भी फीड फोरवर्ड न्यूरल नेटवर्क में हर लेयर अपने से पिछली लेयर के आउटपुट पर निर्भर करती है और सभी लेयर्स के वैट्स को समायोजित करके मशीन कुछ सीख पाती है।

इस संरचना में पहली परत यानी इनपुट लेयर को इस प्रकार लिखा जा सकता है:

$$h1 = g1\,(W1 * x + b1)$$

जहाँ, g1 एक्टिवेशन फंक्शन है, W1 का अर्थ है वेट्स (Weights) और b1 बायस टर्म है

इसी प्रकार, हिडन लेयर (दूसरी लेयर) को इस प्रकार लिखा जा सकता है:

$$h2 = g2\,(W2 * h1 + b2)$$

और थर्ड लेयर (आउटपुट लेयर) को इस प्रकार लिखा जा सकता है:

$$y = g3\,(W3 * h2 + b3)$$

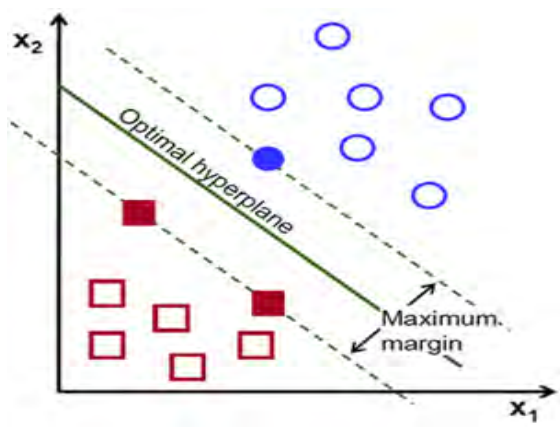
अब, h1 और h2 का प्रयोग करके हम लिख सकते है

$$y = g3\,(W3 * g2\,(W2 * g1\,(W1 * x)) + + [b + ---]$$

ऊपर दिए गए समीकरणों के आधार पर सभी लेयर्स के बीच त्रुटि (Error) का आंकलन किया जाता है और इसके आधार पर सभी हिडन नोड्स के कनेक्शन वेट्स पर करेक्शन (Correction) लगायी जाती है, और इस प्रकार इनपुट नोड्स प्रत्येक अवलोकन (ऑब्जरवेशन) के लिए बार–बार सेट किये जाते है।

3.2. सपोर्ट वेक्टर मशीन (SVM)

एक सपोर्ट वेक्टर मशीन (SVM) विभिन्न डेटा वर्गों को वर्गीकृत करने के लिए हाइपर–प्लेन का उपयोग करती है। एसवीएम का मॉडल ट्रेनिंग डाटा पर बनाया जाता है और टेस्ट डाटा से एक हाइपर प्लेन जेनरेट किया जाता है। एसवीएम मॉडल का उपयोग करके हम, डेटा मैट्रिक्स में उस स्थान का पता लगते है, जहां विभिन्न डेटा समूहों को व्यापक रूप से अलग किया जा सकता है, और इस स्थान को हाइपर–प्लेन निर्धारित करता है। नीचे दिए गए चित्र–2 में लाल और नीले रंग के दो ट्रेनिंग डाटा पॉइंट्स दिखाए गए हैं।



चित्र 2. सपोर्ट वेक्टर मशीन (SVM)

हाइपर–प्लेन को रैखिक रूप से परिभाषित करने के लिए हमें एक ऐसा प्लेन बनाना होता है, जो डाटा को अधिकतम मार्जिन से अलग कर सके, यह एक आदर्श हाइपर–प्लेन (Optimal hyper plane) कहलाता है। यह हाइपर प्लेन रेखिक भी हो सकता है और अरेखिक लीनियर भी।

3.3. के–नियेरेस्ट नेबर (KNN)

के नियेरेस्ट नेबर एल्गोरिदम सभी मशीन लर्निंग एल्गोरिदम में सबसे सरल हैं, इन एल्गोरिदम के पीछे मूल अवधारणा यह है कि, प्रशिक्षण सेट को याद रखना है और फिर अपने निकटतम पड़ोसियों के लेबल के आधार पर प्रशिक्षण सेट में हर नए उदाहरण बिंदु के लेबल का निर्धारण करना है। इस वर्गीकरण की तकनीक में डोमेन बिंदु लेबलिंग को परिभाषित करने के लिए उपयोग किए जाते हैं , क्योंकि हर डोमेन बिंदु के पास के अधिकांश बिंदुओं में एक ही लेबल होता है, इसलिए नियेरेस्ट नेबर एल्गोरिदम में यह माना जाता है कि वह डेटा बिंदु जो रिफरेन्स बिंदु से न्यूनतम दूरी पर स्थित होगा, डेटा बिंदु उसी वर्ग (बसें) का होगा।

के–नियेरेस्ट नेबर एल्गोरिथ्म में, डेटा बिंदुओं के बीच की दूरी को खोजने के लिए यूक्लिडियन डिस्टेंस का उपयोग किया जाता है, इस फंक्शन का फॉर्मूला नीचे दिया गया है:

$$d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 \ldots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

जहां, n आयामों की संख्या है, या हमारी मशीन सीखने की विशेषताएं हैं।

3.4. डिसिशन ट्री (Decision Tree)

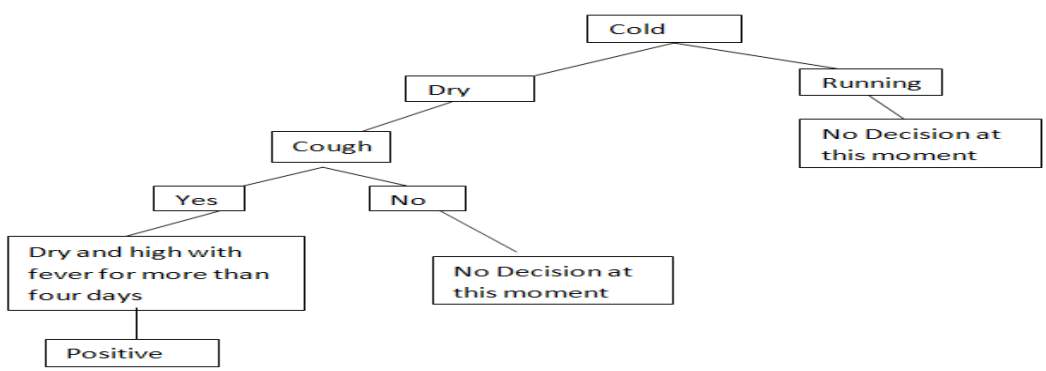
डिसिशन ट्री, मशीन लर्निंग की तकनीको में से, सुपरवाइसड लर्निंग की श्रेणी की एक तकनीक है। डिसिशन ट्री, एक पेड़ जैसा आकार होता है, जहां नोड एक फीचर पर टेस्ट का प्रतिनिधित्व करते है, क्लास लेबल, डिसिशन ट्री की लीफ नोड्स द्वारा दर्शाए जाते हैं और ब्रांच उन फीचर्स के कंजक्शन को दर्शाते हैं, जो उन क्लास लेबल्स को कनेक्ट करती है। डिसिशन ट्री की जड़ (रुट नोड–Root Node) से डिसिशन ट्री की पत्ती (लीफ नोड – leaf node) तक के रास्ते वर्गीकरण नियमों (Classification rules) को दर्शाते हैं। सभी फीचर्स की गणना के बाद वर्गीकरण पर निर्णय लिया जाता है। नीचे चित्र 3. में डिसिशन ट्री की संरचना के द्वारा वर्गीकरण करने के लिए एक उदहारण दिया गया है कि व्यक्ति कोरोना वायरस से संक्रमित है या नहीं।

व्यक्ति Covid19 से संक्रमित है या नहीं (ऊपर निर्णय वृक्ष का प्रदर्शन करने के लिए सिर्फ एक उदाहरण है और चिकित्सा दिशा निर्देशों के अधीन नहीं है)

3.5. नैव बैस क्लासिफायर

नैव बैस क्लासिफायर, प्रायकिता के आधार पर वर्गीकरण करने की एक तकनीक है, जो डेटा को वर्गीकृत करने के लिए बैस प्रमेय (Theorem) का उपयोग करती है।

बैस प्रमेय (Theorem) के हिसाब से घटना ए के होने की सम्भावना, जबकि घटना बी पहले ही घटित हो चुकी हो, कुछ नीचे दिए गए फॉर्मूले से दे जाती है



चित्र 3. डिसिशन ट्री ;(Decision Tree) की संरचना

यहाँ, B प्रमाण (Evidence) है और A परिकल्पना (हाइपोथिसिस– Hypothesis) है। यहाँ पूर्वधारणा यह है कि प्रीडिक्टर्स या फीचर्स मतलक स्वतंत्र हैं। कि एक प्रीडिक्टर्स या फीचर्स की उपस्थिति दूसरे को प्रभावित नहीं करती है। इसलिए इसे नैव (Naïve) कहा जाता है।

डेटासेट :

इस शोध पत्र में, इस्तेमाल किया गया डेटासेट PIMA भारतीय डेटासेट है।ऐसा डेटासेट यूसीआई रिपॉजिटरी [1] में सभी के लिए उपलब्ध है। इस डेटासेट के चुनाव के पीछे मुख्य कारण यह है इसमें गर्भावस्था और रक्त प्लाज्मा विवरणों के अलावा

$$P(A|B) = \frac{P(B \mid A)P(A)}{P(B)}$$

बुनियादी, मधुमेह से जुड़े आंकड़े उपस्थित हैं, जैसे कि उम्र, अधिक वजन, बीएमआई, सीरम, इंसुलिन, डायस्टोलिक रक्तचाप, और इसके अलावा इसमें डायबिटीज पेडिग्री फंक्शन (DPF) भी शामिल है, जो रिश्तेदारों में डायबिटीज मेल्लिटस के इतिहास तथा रोगी का उन रिश्तेदारों के आनुवंशिक संबंधों से संबंधित डेटा है।

इस डेटासेट में भी कहीं कहीं पर कुछ आंकड़ों की अनुपस्थिति थी (Missing data values), जो कि मजबूत तकनीकी वर्गीकरण मॉडल तैयार करने के लिए शोधकर्ताओं द्वारा सबसे बड़ी चुनौती है। इसलिए, अपनी शोध में हमने सबसे पहले अनुपस्थित आंकड़ों को और आउटलायर्स को संभालने के लिए प्रयास किये हैं। इस पत्र में, हमने, अनुपस्थित डाटा के आंकलन के लिए क्लास–वाइज मीन डाटा इम्पयुटेशन तकनीक का उपयोग किया है और ऑउटलायर को सँभालने के लिए 2% विनसोराइजेशन तकनीक का उपयोग किया है। हमने अपने पिछले अध्ययन [3] में पाया था कि क्लास–वाइज मीन तकनीक की सटीकता (accuracy), संवेदनशीलता (sensitivity)और विशिष्टता (Specificity) के आधार पर सर्वोत्तम है। अंतत: डाटा सेट को क्लास–वाइज मीन और विनसोराइजेशन तकनीक से सुनिश्चित करने के पश्चात (निरिक्षण के उपरांत), हमने विभिन्न वर्गीकरण तकनीकों का डेटासेट पर उपयोग किया और उन वर्गीकरण तकनीकों की मधुमेह (डायबिटीज) डिटेक्शन में सटीकता (एक्यूरेसी) का तुलनात्मक अध्ययन किया है।

4.1 परफॉरमेंस इवैल्यूएशन

प्रस्तुत कार्य में, क्लासिफिकेशन तकनीक की एक्यूरेसी को उनके परफॉरमेंस फैक्टर के रूप में माना गया है। क्लासिफिकेशन तकनीक सटीकता को नापने के लिए हमे वास्तविक सकारात्मक (ट्रू पॉजिटिव – TP) और वास्तविक नकारात्मक (ट्रू नेगेटिव – TN) वर्गों का उपयोग, मरीज के वर्गीकरण के लिए करना पड़ता है, कि वह व्यक्ति मधुमेह से पीड़ित है या नहीं। दूसरे शब्दों में क्लासिफिकेशन तकनीक/मेथड की एक्यूरेसी/सटीकता निकलने के लिए, मामलों को सही ढंग से वर्गीकृत करने की दर को निकाला जाता है। एक्यूरेसी (Accuracy) निकलने का फार्मूला नीचे दिया गया है

$$\text{accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (1)$$

$$sensitivity = \frac{TP}{TP+FN} \quad (2)$$

$$specificity = \frac{TN}{TN+FP} \quad (3)$$

$$\text{precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (4)$$

$$\text{recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (5)$$

$$\text{F} - \text{measure} = \frac{2*\text{precision}*\text{recall}}{\text{precision+recall}} \quad (6)$$

यहाँ,

TP: अर्थात वास्तविक सकारात्मक (ट्रू पॉजिटिव) जिसका अर्थ है सही ढंग से वर्गीकृत सकारात्मक मामले।

TN: अर्थात वास्तविक नकारात्मक (ट्रू नेगेटिव) जिसका अर्थ है सही ढंग से वर्गीकृत नकारात्मक मामले।

FP: अर्थात काल्पनिक सकारात्मक (फाल्स पॉजिटिव) जिसका अर्थ है गलत तरीके से वर्गीकृत नकारात्मक मामले।

FN: अर्थात काल्पनिक नकारात्मक (फाल्स नेगेटिव) जिसका अर्थ है गलत तरीके से वर्गीकृत सकारात्मक मामले।

हमने डाटा मॉडल को के– फोल्ड क्रॉस (K-Fold Cross) वेलिडेशन/मान्यीकरण (Validation) तकनीक का उपयोग करके मान्य किया है, जो डेटा मॉडल के प्रदर्शन को मान्य करने के लिए सबसे अधिक बार उपयोग की जाने वाली विधि है। इस शोध कार्य में, 10– फोल्ड क्रॉस (10-Fold Cross)– वेलिडेशन का उपयोग किया गया है, जिसमें प्रारंभिक डेटासेट को 10 उप–नमूनों में विभाजित किया जाता है, और एक भाग का उपयोग परीक्षण के उद्देश्य के लिए किया जाता है तथा शेष नौ भागों को प्रशिक्षण के लिए उपयोग किया जाता है। इस प्रक्रिया को दस बार दोहराया गया और औसत सटीकता को मॉडल की अंतिम सटीकता के रूप में लिया गया।

मेथोडोलॉजी :

इस पत्र में, हमने, अनुपस्थित डाटा के आंकलन के लिए क्लास–वाइज मीन डाटा इम्पयुटेशन तकनीक का उपयोग किया है और ऑउटलायर को सँभालने के लिए 2% विनसोराइजेशन तकनीक का उपयोग किया है। हमने अपने पिछले अध्ययन [3] में पाया था की क्लास–वाइज मीन तकनीकय सटीकता (accuracy), संवेदनशीलता (sensitivity) और विशिष्टता (Specificity) के आधार पर सर्वोत्तम है। अंतत: डाटा सेट को क्लास–वाइज मीन और विनसोराइजेशन तकनीक से सुनिश्चित करने के पश्चात (निरिक्षण के उपरांत), हमने विभिन्न वर्गीकरण तकनीकों का डेटासेट पर प्रयोग किया और उन वर्गीकरण तकनीकों की मधुमेह (डायबिटीज) डिटेक्शन में सटीकता (एक्यूरेसी) का तुलनात्मक अध्ययन किया है।

प्रयोग को दो चरणों में आयोजित किया गया था, पहले चरण/स्टेज में डाटा प्रीप्रोसेसिंग की गयी है, जहां डुप्लीकेट और अनुपस्थित डाटा के लिए डेटासेट की जाँच की गई, और डेटा इंप्यूटेशन (Data Imputation) के लिए, क्लास वाइज मीन(Class wise mean) तकनीक [3] का उपयोग किया गया। उसके बाद डेटा को दो भागों में विभाजित किया गया, पहले विभाजन में मिन–मैक्स स्केलर (Min-Max Scalar) को फीचर डाटा सेट पर अप्लाई कर के उनकी वैल्यूज को 0 से 1 के बीच सेट कर दिया गया, और दूसरी बार टेस्टिंग तथा ट्रेनिंग के लिए फीचर डाटा सेट को विभाजित किया गया।

दूसरे चरण में तीन परत के एएनएन (ANN) मॉडल का निर्माण किया गया था, जहां इनपुट (input) और अदृश्य (hidden) परत में न्यूरॉन्स और एक्टिवेशन फंक्शन थे और आउटपुट (output) परत/लेयर्स में सिग्मोइड (Sigmoid) फंक्शन का इस्तेमाल किया है। उसके बाद एएनएन (ANN) के आउटपुट को के–फोल्ड क्रॉस (K-fold cross) वेलिडेशन तकनीक से वैलिडेट किया गया। फिर पायथन साइकेट लाइब्रेरी का उपयोग करके, एएनएन (ANN) की सटीकता की तुलना अन्य क्लासिफायर (यानी SVM, KNN, डिसीजन ट्री और नैव बैस क्लासिफायर) की सटीकता के साथ की गयी है।

चरण 1 का सूडो कोड (Pseudo Code): डाटा प्रीप्रोसेसिंग स्टेज और चरण 2: एएनएन बनाने के चरणों का सूडो कोड (Pseudo Code) नीचे प्रस्तुत किया गया है:

चरण/स्टेज–1 :	चरण/स्टेज–2 :
डाटा प्रीप्रोसेसिंग स्टेज का सूडो कोड (Pseudo Code): इनपुट: बाइनरी लॉस फंक्शन और स्टोकॉस्टिक ग्रेडिएंट ऑप्टिमाइजर इनपुट: पिमा इंडियन डेटासेट .csv के रूप में डेटासेट की प्री–प्रोसेसिंग (Pre Processing): • डुप्लिकेट के लिए जाँच कि यदि मौजूद है तो हटा दिया गया है • अनुपस्थित आंकड़ों को संभालने के लिए क्लास वाइज डाटा इपुटेशन तकनीक का प्रयोग किया गया। • ऐरे/सरणी (Array) मैं उपस्थित डाटा से न्यूरल नेटवर्क का निर्माण • स्प्लिट डेटासेट डी(D): ☐ – डेटासेट में x और ल, जहां x स्वतंत्र फीचर है और ल निर्भर फीचर है। – मिन–मैक्स स्केलर विधि का प्रयोग:फीचरडेटासेट में 0 और 1 के बीच रखने केलिए। • दोबारा/फिर से डेटासेट(D) स्प्लिट: – प्रशिक्षण डेटा 80% – परीक्षण डेटा 20%	ANN मॉडल बनाने का सूडो कोड (Pseudo Code): इस मॉडल में इनपुट, आउटपुट और हिडन लेयर के रूप में तीन लेयर है। • इनपुट परत/लेयर में 15 न्यूरॉन्स और एक्टिवेशन फंक्शन हैं, • दूसरी परत/लेयर में 17 न्यूरॉन्स और एक्टिवेशन फंक्शन हैं और • आउटपुट एक्टिवेशन में 1 न्यूरॉन और सिग्मॉइड फंक्शन होते है। एएनएन(ANN) के आउटपुट को के–फोल्ड क्रॉस (K-fold cross) वेलिडेशन तकनीक से वैलिडेट किया गया। मॉडल में प्रशिक्षण डाटा पर बाइनरी लॉस फंक्शन और स्टोकॉस्टिक ग्रेडिएंट ऑप्टिमाइजर का उपयोग करके,उसकी सटीकता/एक्यूरेसी को मापने के लिए संकलित किया गया था और सटीकता को मैट्रिक्स के रूप में संजोया गया था।

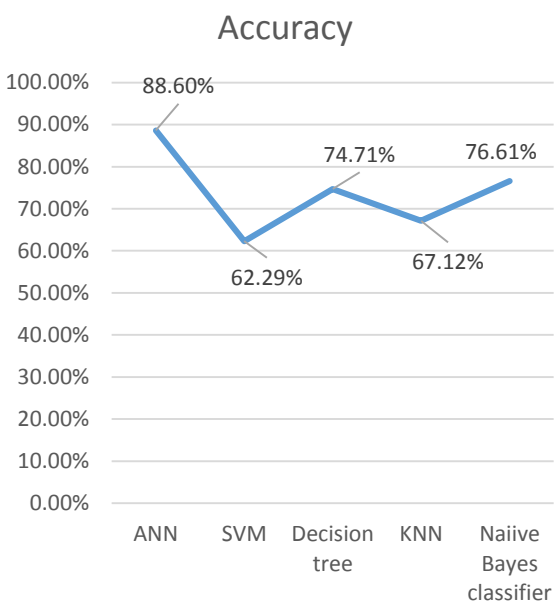
6. परिणाम और चर्चा

प्रत्येक क्लासिफायर की सटीकता/एक्यूरेसी तालिका–2 में नीचे दी गई है, परिणामों से यह स्पष्ट है कि आर्टिफिशियल न्यूरल नेटवर्क (एएनएन), चुने हुए डेटासेट पर 88.6% की सटीकता/एक्यूरेसी दर्शाते हुए सबसे अच्छा प्रदर्शन किया है

तालिका 2. : विभिन्न वर्गीकरण एल्गोरिदम की सटीकता (%)

वर्गीकरण तकनीक	सटीकता
आर्टिफिशियल न्यूरल नेटवर्क	88.6%
सपोर्ट वेक्टर मशीन	62.29%
डिसिशन ट्री	74.71%
के–नियेरेस्ट नेबर	67.12%
नैव बैस क्लासिफायर	76.61%

चूंकि आर्टिफिशियल न्यूरल नेटवर्क (ANN) केवल 0 और 1 के बीच संभाव्यता संख्या (Probability अंसनम) को मान देता है, इसलिए 0.5 की सीमा को डेटा को 1 और 0. के रूप में वर्गीकृत करने के लिए सेट किया गया था। यदि मान 0.5 से ऊपर था तो इसे 1 के रूप में वर्गीकृत किया गया था अन्यथा 0 के रूप में वर्गीकृत किया गया था आर्टिफिशियल न्यूरल नेटवर्क (ANN) पारंपरिक मॉडल की तुलना पारंपरिक मशीन लर्निंग क्लासिफायर के साथ की गई थी।



चित्र 4. एएनएन के साथ तुलना में विभिन्न क्लासिफायर का सटीकता

वास्तव में, बहुत मुश्किल है यह कह पाना की कब कौनसी वर्गीकरण तकनीक सबसे अच्छा प्रदर्शन करेगी, लेकिन विशेषज्ञों ने देखा कि डीप लर्निंग तकनीक में सर्वोत्तम संभव परिणाम प्राप्त करने की क्षमता है। भविष्य के कार्यों में, रक्त शर्करा के स्तर की भविष्यवाणी करने के लिए समय श्रृंखला (time series) डेटा को उपयोग करके, डीप लर्निंग मॉडल को विकसित किया जा सकता है। जो इंसुलिन के स्तर को बनाए रखने में महत्वपूर्ण योगदान हो सकता है। डीप लर्निंग की क्षमता इतनी जबरदस्त है कि यह भविष्य में एक बड़ी प्रगति ले सकती है।

7. निष्कर्ष :

इस पत्र में, पारंपरिक वर्गीकरण तकनीकों की तुलना आर्टिफिशियल न्यूरल नेटवर्क (ANN) के साथ की गयी है। तालिका –2 में दिए गए परिणामों के आधार पर, यह निष्कर्ष निकाला गया है कि आर्टिफिशियल न्यूरल नेटवर्क (ANN) में मधुमेह के वर्गीकरण के लिए उच्च स्तर की सटीकता (88.6%) पाई गई, जिसमें Pima Indian Dataset की सभी विशेषताएं/फीचर्स शामिल हैं। अन्य शोधों में, शोधकर्ताओं ने अधिक सटीकता पाई, लेकिन वे आमतौर पर उन्होंने अनुपस्थित आंकड़े/डेटा या आउटलायर्स को नजरअंदाज कर दिया था। प्रस्तुत शोध कार्य में क्लास वाइज मीन (class wise mean) [3] का उपयोग करके अनुपस्थित आंकड़े/डेटा और आउटलायर्स को संभालते हुए,और मॉडल की जटिलता को बिना बढ़ाये, हमने पाया की मधुमेह के वर्गीकरण के लिए आर्टिफिशियल न्यूरल नेटवर्क (ANN) की तकनीक, सर्वोत्तम है और सटीक रिजल्ट्स देती है।

8. प्रमुख शब्दों की तालिका :

Technical Terms (English)	तकनीकी शब्द (हिन्दी)
Classification Algorithm	वर्गीकरण तकनीक
Support Vector Machine - SVM	सपोर्ट वेक्टर मशीन
K&Nearest Neighbor - KNN	के–नियरेस्ट नेबर
Decision Tree	डिसिशन ट्री
Artificial Neural Network - ANN	आर्टिफिशियल न्यूरल नेटवर्क्स
Naïve Bayes Classifier	नैव बैस क्लासिफायर
Logistic Regression	लोजिस्टिक रिग्रेशन

Missing data values	आंकड़ों की अनुपस्थिती
Sequential Covering Approach	सिक्वेनशियल कवरिंग अप्रोच
Medication of Diabetes	मधुमेह की दवा
Abnormal Laboratory Tests	असामान्य प्रयोगशाला परीक्षण
Diagnosis of Diabetes	मधुमेह निदान
Classification	वर्गीकरण
Accuracy	सटीकता
Sensitivity	संवेदनशीलता
Specificity	विशिष्टता
Probability	सम्भावना
Data Imputation	डेटा इंप्यूटेशन
Time series	समय श्रृंखला

References:

[1] UCI repository, Pima Indian Dataset.

[2] Wenqian, Shuyu Chen, Hancui Zhang, and Tianshu Wu Chen, "A hybrid prediction model for type 2 diabetes using K-means and decision tree," in Software Engineering and Service Science (ICSESS), pp. 386-390,2017.

[3] Sofia Goel and Sudhansh Sharma," Advanced Data Imputation Techniques for Predicting Type 2 Diabetes using Machine Learning, "International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, vol.9,issue 2, December 2019.

[5] Nahla, Andrew P. Bradley, and Mohamed Nabil H. Barakat Barakat, "Intelligible support vector machines for diagnosis of diabetes mellitus," IEEE transactions on information technology in biomedicine, vol. 14, no. 4, pp. 1114-1120, July 2010.

[5] Xie W, Xu L, He X, Zhang Y, You M, Yang G, Chen Y Zheng T, "A machine learning-based framework to identify type 2 diabetes through electronic health records," International journal of medical informatics, vol. 97, pp. 120-127, Jan 2017.

[6] Dijana Sejdinovie et al.,"Classification Of Prediabetes And Type 2 Diabetes Using Artificial Neural Network", CMBEBIH Springer, Singapore. pp. 685-689, March 2017.

[7] Rahimloo, Parastoo, and and Ahmad Jafarian, "Prediction of Diabetes by Using Artificial Neural Network, Logistic Regression Statistical Model and Combination of Them",Bulletin de la Société Royale des Sciences de Liège, 85, pp.1148-1164.Jan 2016.

[8] Mahmoud, Mehdi Teimouri, Zainabolhoda Heshmati, and Seyed Mohammad Alavinia Heydari, "Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran," International Journal of Diabetes in Developing Countries, vol. 36, pp. 167-173, 2016.

[9] Longfei, Senlin Luo, Jianmin Yu, Limin Pan, and Songjing Chen Han, "Rule extraction from support vector machines using ensemble learning approach: an application for diagnosis of diabetes," IEEE journal of biomedical and health informatics, vol. 19, pp. 728-734., 2015.

[10] Ali, Tinna B. Aradóttir, Alexander R. Johansen, Henrik Bengtsson, Marco Fraccaro, and Morten Mørup Mohebbi, "A deep learning approach to adherence detection for type 2 diabetics," in Engineering in Medicine and Biology Society (EMBC, vol. 39, pp. 2896-2899, 2017.

[11] Xie W, Xu L, He X, Zhang Y, You M, Yang G, Chen Y Zheng T, "A machine learning-based framework to identify type 2 diabetes through electronic health records," International journal of medical informatics, vol. 97, pp. 120-127, Jan 2017.

[12] Hyun Kang, "The prevention and handling of the missing data.," Korean journal of anesthesiology, pp. 402-406, 2013.

[13] Aliza Ahmad, Aida Mustapha, Eliza Dianna Zahadi, Norhayati Masah, and Nur Yasmin Yahaya, "Comparison between Neural Networks against Decision Tree in Improving Prediction Accuracy for Diabetes Mellitus," in International Conference on Digital Information Processing and Communications, pp. 537-545, 2011.