ڈیٹا مائننگ ٹیکنیکس کے استعمال سے امراض قلب کے خدشات کا تشخیصی ماڈل

مقالہ

ڈگری برائے

ڈاکٹر آف فلاسفی

مقالہ نگار

سید امام ملا نصار اللہ

Enrolment Number (A160920)

زیر نگرانی

ڈاکٹر پردیپ کمار

ایسوشیٹ پروفیسر و سربراہ شعبہ کمپیوٹر سائنس و انفارمیشن ٹیکنالوجی

ستمبر۔ 2019

شعبہ کمپیوٹر سائنس و انفارمیشن ٹیکنالوجی

مولانا آزاد نیشنل اردو یونیورسٹی، گچی بولی، حیدر آباد (تلنگانہ)، انڈیا

मौलाना आज़ाद नेशनल उर्दू यूनिवर्सिटी مولانا آزاد نیشنل اردو یونیورسٹی
MAULANA AZAD NATIONAL URDU UNIVERSITY
(A Central University established by an Act of Parliament in 1998)

Department of Computer Science and Information Technology

CERTIFICATE

It is certified that research work presented in the thesis entitled **“Heart Disease Risk Evaluation Model Using Data Mining Techniques”** in partial fulfillment of the requirements for the award of the degree of **Doctor of Philosophy** in **Computer Science** has been carried out under my guidance and supervision. He has fulfilled all the requirements for submission of the thesis, which to the best of my knowledge has reached the requisite standard. This thesis presented by him, to the best of my knowledge and belief, did not form the basis for the award of any other degree earlier.

Dr. Pradeep Kumar
Supervisor
Associate Professor and Head
Department of Computer Science & Information Technology
Maulana Azad National Urdu University
Gachibowli, Hyderabad, INDIA

Gachibowli, Hyderabad – 500 032
www.manuu.ac.in
Tel: 040 23006612-15 Fax: 23006603

DECLARATION

I, **Syed Immamul Ansarullah**, solemnly declare that the thesis entitled "**Heart Disease Risk Evaluation Model Using Data Mining Techniques"** is my original work. The study has been conducted under the guidance of **Dr. Pradeep Kumar** with the Department of Computer Science and Information Technology with **Maulana Azad National Urdu University** (A Central University), Gachibowli, Hyderabad, India. It is further declared that to the best of my knowledge and belief, it has not been submitted earlier for the award of any other degree, by anyone.

Dated: ___/___/2019

Syed Immamul Ansarullah
Research Scholar
Department of Computer Science & Information Technology
School of Technology
Maulana Azad National Urdu University
Gachibowli, Hyderabad, INDIA

اظہار تشکر

سب سے پہلے میں اللہ تعالیٰ کا شکر ارا کرنا چاہتا ہوں کہ انہوں نے مجھے انمول احسانات اور نہ ختم ہونے والی نعمتوں سے نوازا اور مجھے غلط راستے سے الگ رہنے کی جرات، صبر، اور دانائی عطا کی۔ یہ تحقیق ان کی برکت کے بغیر ممکن نہیں ہوتی۔

میں اپنے سپروائزر ڈاکٹر پردیپ کمار، ایسو شیٹ پروفیسر و سربراہ شعبہ کمپیوٹر سائنس و انفارمیشن ٹیکنالوجی کی انتھک اور فوری مدد کی تعریف کرنا چاہتا ہوں جنھوں نے مجھے تحقیق میں اپنی سمتوں کی وضاحت اور اس کی کھوج کی مکمل آزادی کی اجازت دی۔ اگرچہ یہ کسی حد تک انتہائی مشکل ثابت ہوا لیکن اس میں ان کے انداز کی حکمت کی بھی تعریف کرتا ہوں۔ میں پروفیسر عبدل واحد، ڈاکٹر مدثر منظور کرمانی، ڈاکٹر سید محسن سیف اور زیما مشتاق کے مفید مباحثوں، مشوروں، اور ان کے قیمتی تبصروں کے لئے ان کا شکریہ ادا کرنا چاہتا ہوں۔ ان کے ساتھ ساتھ میں محبوب باشا کی تکنیکی مدد پر بھی ان کا شکریہ ادا کرنا چاہتا ہوں۔

میں تمام فیکلٹی ممبران اور دوسرے معاون عملہ ، شعبہ کمپیوٹر سائنس و انفارمیشن ٹیکنالوجی، اسکول آف ٹکنالوجی، مانو کا انتہائی شکریہ ادا کرتا ہوں۔ میں اپنے مطالعے کے دوران اخلاقی مدد فراہم کرنے پر محکمہ کمپیوٹر سائنس و انفارمیشن ٹیکنالوجی کے تمام ریسرچ اسکالرز اور دیگر افراد کا شکریہ ادا کرتا ہوں۔

میں اپنے پیارے والدین اور دیگر کنبہ کے ممبروں کا مخلصانہ شکریہ ادا کرتا ہوں جنہوں نے مجھے اپنے مطالعے کے پورے عرصے کے دوران لامحدود پیار، محبت اور مدد فراہم کی۔ ان کی مستقل حمایت نے مجھے اپنی زندگی میں کامیابی حاصل کرنے دی ہے اور مجھے اپنے تحقیقی مطالعہ کے مقصد کو سمجھنے میں مدد فراہم کی ہے۔

سید امام ملا نصار اللہ

تلخیص

امراض قلب دنیا بھر میں موت کی سب سے اہم وجہ کے طور پر ابھر رہی ہے اور یہ ایک انتہائی مہنگی ترین دائمی بیماری ہے۔ زبردست بہتری کے باوجود امراض قلب مریضوں اور میڈیکل سسٹم پر ایک بہت بڑا بوجھ ڈال رہی ہے۔ نقصان دہ پیچیدگیوں کے باوجود امراض قلب سب سے زیادہ قابل علاج اور قابل کنٹرول بیماری ہے لہذا اس کی پیشن گوئی کرنا اور وقت سے پہلے ہی اس کی تشخیصی کرنا ضروری ہے۔ اگرچہ جدید ترین تشخیصی اور بازآبادکاری جیسے الیکٹروکارڈیوگرام (ای سی جی)، کارڈیک کمپیوٹرائزڈ ٹوموگرافی (سی ٹی) اسکین، اور کارڈیک مگنٹک ریزونینسی امیجنگ (ایم آر آئی) اب دیکھ بھال کا معیار بن گئے ہیں جو امراض قلب کا براہ راست اور درست ثبوت فراہم کرتے ہیں تاہم یہ طرز عمل ناگوار ہے، کیونکہ ان میں خون کے نمونوں کے تحقیقات کی بھی ضرورت پڑتی ہے اور یہ نسبتا قیمتی اور عملی طور پر پیچیدہ ہیں جو دیہی علاقوں اور عوامی سطح پر اسکریننگ کی تشخیصی پر ان کے استعمال کو روکتی ہے۔

ادبیات اور تکنیکی رپوٹوں کا جائزہ لینے کے بعد یہ پتہ چلا ہے کہ امراض قلب سے معذوری اور اموات کی شرح بڑھ رہی ہے۔ اموات شرح کے بڑھتے ہوئے واقعات کی وجہ سے امراض قلب کے خطرے سے متعلق تشخیصی ماڈل تیار کیا گیا ہے تاکہ ابتدائی پیش گوئی اور تشخیصی میں معالجین کی مدد کی جاسکے۔ میڈیکل ڈومین کے ماہرین کے ذریعہ اہم نان انویزو خطرے (عمر، سسٹولک بی پی، ڈیاسٹولک بی پی، بی ایم آئی، ور ہرڈیٹری فیکٹر، تمباکو نوشی، الکوحل، اور فزیکل ان ایکٹوٹی) کی خصوصیات کی نشاندہی کی گئی ہے اور امراض قلب کی تشخیصی میں ان کی قابل اعتمادیت کی تفتیش مختلف ذریعے سے کی گی ہے۔ ڈیٹا مییننگ تکنیکسی جیساکہ ڈسیشن ٹری، کے نیریسٹ نیبر، رینڈم فارسٹ، نیوی بائیس اور سپورٹ ویکٹر مشین کو خطرے والے عوامل پر منحصر کرنے کے لئے مخصوص تفتیشی تکنیکوں کے اطلاق میں اضافے کا تجربہ کیا گیا ہے۔ مزید یہ کہ کم سے کم غلط تشخیصی کی شرح کے ساتھ امراض قلب کی زیادہ درست طریقے سے پیش گوئی کرنے کے لئے ہائپر پیرامیٹر کی اصلاح کی گئی ہے۔

ڈیولپڈ رسک ماڈل جو پیٹر نوٹ بک ویب ایپلی کیشن کا استعمال کرتے ہوئے تیار کیا گیا ہے اور اس کی کارکردگی کو نہ صرف میڈیکل ڈومین اقدامات جیسے sensitivity, specificity, accuracy, precision, misclassification rate and cross validation بلکہ کمپیوٹیشنل ماڈل اقدامات کے ذریعے بھی جانچا گیا ہے جیسے کہ model complexity and comprehensibility۔ تجرباتی نتائج سے پتہ چلا ہے کہ رینڈم فارسٹ ماڈل دوسرے امراض قلب ماڈلس سے ڈیفالٹ اور ہائپر پرامیٹرس کی اصلاح کی ترتیبات پر بہتر ہے۔ رینڈم فارسٹ تشخیصی ماڈل نے 85 فیصد کی sensitivity، 83 فیصد کی specificity، 84 فیصد کی accuracy، 85 فیصد کی precision، 15 فیصد کی misclassification rate، 85 فیصد AUROC اسکور اور کم وقت کی پیچیدگی کو حاصل کیا ہے۔ ڈیولپڈ رسک ماڈل کے نتائج سے پتہ چلتا ہے کہ وہ قابل احتمال پیش گوئی کی درستگی کے ساتھ موجودہ رسک تشخیصی ماڈلز کو آگے بڑھاتا ہے اور امراض قلب کی ابتدائی تشخیصی اور جانچ میں اس کی افادیت کو ثابت کرتا ہے۔ ماڈل کی استعداد کار کو بڑھانے اور غلط تشخیص کی شرح کو کم سے کم کرنے کے لئے امراض قلب کے خطرہ کے ماڈل کو hyperparameter optimization سے بہتر بنایا گیا ہے۔ ہائپر پرامیٹر کے نتائج سے پتہ چلتا ہے کہ رینڈم فارسٹ ماڈل دوسرے امراض قلب ماڈلس سے sensitivity میں 87 فیصد، specificity میں 84 فیصد، accuracy میں 86 فیصد، AUROC اسکور میں 86 فیصد اور misdiagnosis rate of 13 فیصد بہترین ہے۔ خطرے کی خصوصیت کے مرکب سبسیٹ [سسٹولک بی پی، ڈیاسٹولک بی پی، اتج، اور ہیریٹیٹی] نے decision tree کا استعمال کرتے ہوئے سب سے زیادہ اسکور دکھایا جو 78 فیصد کی sensitivity، 80 فیصد کی specificity اور 77 فیصد accuracy کے ساتھ ہے۔ تاہم رینڈم فارسٹ ماڈل نے اس امتزاج کے سبسیٹ میں ایک اور امراض قلب کے خطرے کی خصوصیت BMI کا اضافہ کیا اور 78.9 فیصد کی اعلی ترین درستگی حاصل کی۔ مجوزہ ماڈل کے نتائج سے پتہ چلتا ہے کہ وہ قابل احتمال پیش گوئی کی درستگی کے ساتھ موجودہ خطرے کی تشخیص کرنے والے ماڈلز کو آگے بڑھاتا ہے اور دل کی بیماری کی ابتدائی پیش گوئی اور امتحان میں اس کی افادیت کو ثابت کرتا ہے۔

ڈیولپڈ نان انویزورسک ماڈل امراض قلب کے خطرے سے وابستہ طبّی پریکٹیشنرز کی مدد کرے گا اور ساتھ ہی امراض قلب کے مریضوں کو بیماری کی ممکنہ موجودگی کے بارے میں انتباہ کرے گا اس سے پہلے کہ وہ کسی اسپتال میں داخلہ لئے۔ یہ ماڈل وہاں قابل استعمال ہے جہاں لوگوں کو جلد پتہ لگانے اور علاج کے لئے مربوط بنیادی صحت کی نگہداشت کی ٹیکنالوجیز کا فائدہ نہیں ہوتا ہے۔ ڈیولپڈ نان انویزورسک ماڈل کے ذریعہ پیدا ہونے والی دل کی بیماری کے قواعد مختلف میڈیکل ڈومین ماہرین کے ذریعہ جانچ اور توثیق کیے جاتے ہیں۔ امراض قلب کے نکالے جانے والے تشخیصی قاعدے تجویز کردہ ہیں لیکن حتمی نہیں ہیں کیونکہ یہ مخصوص نسل (کشمیر) پر مبنی ہیں۔ اگرچہ اس تحقیق نے ناول نان انویزورسک اوصاف کے امتزاج پر ڈیٹا میننگ تکنیکوں کا استعمال کرتے ہوئے کم لاگت سے امراض قلب کے خطرے کی تشخیصی کا ماڈل تیار کیا ہے تاہم بیماری کے بارے میں نئی شناخت کی گئی دریافتوں کو سمجھنے کے لئے اضافی تحقیق کی ضرورت ہے۔

فہرست مشمولات

مضمون	صفحہ نمبر
اظہار تشکر	I
تلخیص	II
فہرست مشمولات	V
فہرست مخففات	X
فہرست جدول	XII
فہرست ترسیم	XIII
فہرست مساوات	XV
تحقیقی اشاعتیں	XVI
باب 1. تعارف	**1-11**
1.1۔ پس منظر	1
1.2۔ امراض قلب کا جائزہ	1-2
1.2.1۔ امراض قلب کی شرح اموات	2-3
1.2.2۔ امراض قلب کا عالمی بوجھ	3-4
1.2.3۔ امراض قلب کی شناخت اور تشخیص	4-5
1.2.4۔ امراض قلب کے خطرے کا تخمینہ	5-6
1.3۔ ڈیٹا مائننگ کا عمومی جائزہ	6-9
1.3.1۔ کے ڈی ڈی میں ڈیٹا مائننگ ایک بنیادی اقدام کے طور پہ	7-8
1.3.2- ہیلتھ کیئر میں ڈیٹا مائننگ کا استعمال	8-9
1.4۔ مسئلے کا بیان	9
1.5۔ تحقیقی مقاصد	10
1.6۔ تحقیق کی تشکیل	10-11

باب2.ادب کا جائزہ **12-28**

2.1۔ مختلف ڈیٹا مائننگ ٹاسکس اور ٹکنیکس کا استعمال کرتے ہوئے امراض قلب کی پیشن گوئی 12-26

2.1.1۔ سوپروایزڈ ٹاسکس کے ذریعے امراض قلب کی پیشن گوئی 12-17

2.1.2۔ عن سوپروایزڈ ٹاسکس کے ذریعے امراض قلب کی پیشن گوئی 17-18

2.1.3۔ ہائبرڈ ڈیٹا مائننگ ٹکنیکس کا استعمال کرتے ہوئے امراض قلب کی پیشن گوئی 18-26

2.2۔ ریسرچ گیپس 26-27

2.3۔ باب کا خلاصہ 27-28

باب3.امراض قلب کی تشخیص میں ڈیٹا مائننگ ٹکنیکس کا اطلاق **29-46**

3.1۔ فیچر سلیکشن ٹکنیکس 29

3.1.1۔ اکسٹرا ٹری کلاسیفائر 29

3.1.2۔ گریڈینٹ بوسٹنگ کلاسیفائر 29-30

3.1.3۔ رینڈم فوریسٹ 30

3.1.4۔ ریکرسیو فیچر ایلمنیشن 30

3.1.5۔ ایکس جی بوسٹ کلاسیفائر 30

3.2۔ ڈیٹا مائننگ ٹاسکس 30-32

3.2.1۔ Predictive ڈیٹا میننگ ٹاسکس 31

3.2.1.1 Classification 31

3.2.1.2 Regression 31

3.2.2۔ Descriptive ڈیٹا میننگ ٹاسکس 32

3.2.2.1 کلسٹرنگ 32

3.2.2.2 ایسوسیشن 32

3.3۔ ڈیٹا مینگ ٹکنیکس 32-40

3.3.1۔ ڈسیشن ٹری 32-34

3.3.2- کے نیریسٹ نیبر 34-36
3.3.3۔ سپورٹ ویکٹر مشین 36-38
3.3.4۔ رینڈم فاریسٹ 38-39
3.3.5۔ نیوی بائیس 39-40
3.4۔ ماڈل پرفارمنس ٹکنیکس 40-43
3.4.1۔ کنفیوژن میٹرکس 41-42
3.4.2۔ کراس ویلڈیشن 42
3.4.3۔ AUROC 42-43
3.5۔ رسک ایوالویشن ماڈل ڈویلپمنٹ کے لئے ڈیٹا کلیکشن اور ریسرچ میتھوڈس کا استعمال 43-45
3.6۔ ہارٹ ڈیزیز میں نان انویسو رسک فیچرز کی اہمیت 45-46
3.7 ۔ باب کا خلاصہ 46
باب 4. ڈیٹا مائننگ ٹکنیکس کا استعمال کرکے ہارٹ ڈیزیز ڈیٹا میں علم کی دریافت 47-69
4.1 ۔ امراض قلب کی پیش گوئی کے لئے ڈیٹا مائننگ طریقہ کار 47-48
4.2 ۔ ہارٹ ڈیزیز رسک اوالویشن ماڈل کے لئے ریسرچ ڈیزائن 49-50
4.3 ۔ Exploratory ڈیٹا انالیسس پراسیس 50-54
4.3.1 ۔ ڈیٹاسیٹ میں کلاس عدم توازن اور ڈیٹا کی تقسیم کے مسائل کی جانچ پڑتال 51-52
4.3.2۔ مختلف ہارٹ ڈیزیز رسک فیچرز میں correlations کی تلاش 52-54
4.4۔ ہارٹ ڈیزیز رسک اوالویشن فیچر سلیکشن ٹکنیکس کا نمایاں انتخاب 54-57
4.5 ۔ تجویز کردہ ڈیٹا مائننگ تراکیب کے تجرباتی نتائج 57-66
4.5.1 ڈسیشن ٹری ماڈل کے تجرباتی نتائج 57-59
4.5.2 کے نیریسٹ نیبر ماڈل کے تجرباتی نتائج 59-61
4.5.3۔ سپورٹ ویکٹر مشین ماڈل کے تجرباتی نتائج 61-62
4.5.4۔ رینڈم فاریسٹ ماڈل کے تجرباتی نتائج 63-64
4.5.5۔ نیوی بائیس ماڈل کے تجرباتی نتائج 64-66

4.6 ۔ ڈیولپڈ ہارٹ ڈیزیز رسک اوالویشن ماڈلز کی کارکردگی کا موازنہ 66-67
4.7۔ امراض قلب کے لئے ڈیٹا مائننگ ماڈل کی درست تقویت 67-69
4.8۔ باب کا خلاصہ 69
باب5. ہارٹ ڈیزیز رسک اوالویشن ماڈلز میں ہائپر پرامیٹر آپٹیمائزیشن ٹکنیکس کا استعمال 70-96
5.1 ۔ ہائپر پرامیٹر آپٹیمائزیشن ٹکنیکس 70-74
5.1.1- گرڈ سرچ ہائپر پرامیٹر آپٹیمائزیشن 71-72
5.1.2- رینڈم سرچ ہائپر پرامیٹر آپٹیمائزیشن 72-73
5.1.3۔ Bayesian ہائپر پرامیٹر آپٹیمائزیشن 73-74
5.2 ۔ Optimizing ہارٹ ڈیزیز رسک اوالویشن ماڈلز 74-88
5.2.1 ۔ ڈیسیشن ٹری ہائپر پرامیٹر آپٹیمائزیشن ماڈل 75-78
5.2.2۔ کے نیریسٹ نیبر ہائپر پرامیٹر آپٹیمائزیشن ماڈل 78-81
5.2.3۔ سپورٹ ویکٹر مشین ہائپر پرامیٹر آپٹیمائزیشن ماڈل 81-84
5.2.4۔ رینڈم فارسٹ ہائپر پرامیٹر آپٹیمائزیشن ماڈل 85-88
5.3۔ ڈیولپڈ ہارٹ ڈیزیز رسک ماڈلز میں کارکردگی کا موازنہ 88-89
5.4۔ ڈیفالٹ اور optimized رسک اوالویشن ماڈلز کے مابین کارکردگی کا موازنہ 89-90
5.5۔ ابتدائی بیماری کی پیش گوئی اور شناخت کے لئے ہارٹ ڈیزیز رسک فیچرز کے مختلف مجموعے 90-93
5.6۔ ہارٹ ڈیزیز ایکسپرٹ سسٹم اوالویشن ماڈل کمپونٹس 93-94
5.7۔ ہارٹ ڈیزیز رسک اوالویشن ماڈل(HDREM) 94-96
5.8۔ باب کا خلاصہ 96

باب6. نتیجہ اور مستقبل کے کام کا منصوبہ 97-
6.1 ۔ تحقیقی کام کا خلاصہ 98-99

6.1.1 ہارٹ ڈیزیز رسک تشخیص میں اہم خصوصیات 99-100

6.1.2 ہارٹ ڈیزیز رسک اوالویشن میں نان انویسو رسک فیچرز کی اہمیت 100

6.1.3 ہارٹ ڈیزیز رسک اوالویشن ماڈل ڈیولپمنٹ (HDREM) 100

6.2 ۔ تحقیق کی محدودیات 101

6.3۔ مستقبل کے کام کا منصوبہ 101-102

حوالہ جات 103-117

فہرست مخففات

ABS	Australian Bureau of Statistics
AHA	American Heart Association
AIRS	Artificial Immune Recognition System
AUROC	Area Under the Receiver Operating Characteristics
BHF	British Heart Foundation
BMI	Body Mass Index
CAD	Coronary Artery Disease
CANFIS	Coactive Neuro-Fuzzy Inference System
CHD	Coronary Heart Disease
CHF	Coronary Heart Failure
CRISP_DM	Cross-Industry Standard Process for Data Mining
CT	Computerized Tomography
CVD	Cardiovascular Disease
DALY	Disability Adjusted Life Years
DMP	Default Model Parameter
DMX	Data Mining Extension
DSS	Decision Support System
ECG	Electrocardiogram
EDA	Exploratory Data Analysis
EF	Ejection Fraction
ESCAP	Economic and Social Commission of Asia and the Pacific
ETC	Extra Tree Classifier
GBC	Gradient Boosting Classifier
GNI	Gross National Income

HBP	High Blood Pressure
HDREM	Heart Disease Risk Evaluation Model
HPT	Hyper-Parameter Tuning
HVD	Heart Valve Disease
IHDPS	Intelligent Heart Disease Prediction System
KDD	Knowledge Discovery from Data
KNN	K Nearest Neighbor
LAD	Left Anterior Descending
LCX	Left Circumflex
LMICs	Low-and-Middle-Income-Countries
MLP	Multi-Layer Perceptron
MRI	Magnetic Resonance Imaging
OOB	Out-of-Bag
PCA	Principal Component Analysis
RCF	Right Coronary Artery
RFE	Recursive Feature Elimination
SAS	Statistical Analysis Software
SEMMA	Sample Explore Modify Model Assess
SVM	Support Vector Machine
TNR	True Negative Rate
TPR	True Positive Rate
WEKA	Waikato Environment for Knowledge Analysis
WHO	World Health Organization
XGB	Extreme Gradient Boosting
YLL	Years of Life Lost

فہرست جدول

جدول نمبر	جدول	صفحہ نمبر
3.1	Contingency Matrix for Two-Class Classification	41
3.2	Description of the Heart Disease Dataset	44
4.1	Feature Selection Techniques Providing Weight to Each Risk Attribute	55
4.2	Mean Ranking of Risk Attributes by Feature Selection Techniques	56
4.3	Performance Measures of Developed Heart Disease Models	67
5.1	Experimental Results of the Optimized Decision Tree Model	76
5.2	Experimental Results of the Optimized K NN Model	79
5.3	Experimental Results of the Optimized SVM Model	83
5.4	Experimental Results of the Optimized Random Forest Model	86
5.5	Performance Measures of Developed Optimized Heart Disease Models	88
5.6	Performance Comparison of Developed Heart Disease Risk Models	90
5.7	Integrating Different Non-Invasive Heart Disease Risk Factors	91

فہرست ترسیم

ترسیم نمبر	فہرست ترسیم	صفحہ نمبر
1.1	The Process of Knowledge Discovery in Data	8
3.1	Categorization of Data Mining Tasks	31
3.2	Decision Tree Model Working for Heart Disease Prediction	34
3.3	K Nearest Neighbour classification Example	35
3.4	Linear SVM Classifier for Two-Class Representation	37
3.5	Random Forest Algorithm Working	39
3.6	AUROC Representation	43
4.1	Heart Disease Risk Evaluation Model Methodology	48
4.2	Detailed Steps of Research Design	50
4.3	Heart Disease Distribution Based On Sex Attribute	51
4.4	Correlation in Risk Attributes Through Heatmap Representation	53
4.5	Risk Attribute Hierarchy by Feature Selection Techniques	56
4.6	Decision Tree Model Confusion Matrix	58
4.7	AUROC by the Decision Tree Model	59
4.8	K Nearest Neighbor Confusion Matrix on the Test Dataset	60
4.9	AUROC by K Nearest Neighbor Model	60
4.10	SVM Confusion Matrix on Test Dataset	62
4.11	AUROC by Support Vector Machine Model	62
4.12	Random Forest Model Confusion Matrix on Test Dataset	63
4.13	AUROC by Random Forest Model	64
4.14	Gaussian Naive Bayes Model Confusion Matrix on Test Dataset	65

66 AUROC Curve by Gaussian Naive Bayes Model 4.15

67 Combined AUROCs of the Developed Risk Evaluation Models 4.16

68 Bias and Variance Contributing to the Total Error 4.17

71 Hyperparameter Optimization Representation 5.1

72 Grid Search Layout 5.2

73 Random Search Layout 5.3

74 Single Cross-Validation Methodology for Hyperparameter Optimization 5.4

77 Confusion Matrix of the Optimized Decision Tree Model 5.5

78 AUROC of the Optimized Decision Tree Model 5.6

80 Confusion Matrix of the Optimized K NN Model 5.7

81 AUROC of the Optimized K NN Model 5.8

84 Confusion Matrix of the Optimized SVM Model 5.9

84 AUROC of the Optimized SVM Model 5.10

87 Confusion Matrix of the Optimized Random Forest Model 5.11

88 AUROC of the Optimized Random Forest Model 5.12

89 Combined AUROCs of the Optimized Risk Evaluation Models 5.13

93 Heart Disease Expert System Evaluation Tool Components 5.14

95 Heart Disease Risk Evaluation Model Interface 5.15

95 High-Risk Heart Disease Evaluation Example 5.16

96 Low-Risk Heart Disease Evaluation Example 5.17

فہرست مساوات

مساوات نمبر	مساوات	صفحہ نمبر
3.1	Information Gain for Decision Tree Construction	33
3.2	How to Calculate the Values of Information Gain	33
3.3	Calculating Information Gain Values	33
3.4	Calculating Euclidean Distance	35
3.5	Calculating Minimum-Maximum Normalization	36
3.6	SVM Discriminant Function f (T) for a Test Sample T	36
3.7	SVM Discriminant Function for Non-Linear Separation	37
3.8	Quadratic Problem on Maximizing the Lagrangian Dual Objective Function	38
3.9	Constraints of the Quadratic Problem Objective Function	38
3.10	Calculating the Prior Probability and the Conditional Probability Through Naive Bayes Algorithm	40
3.11	Calculating Sensitivity	41
3.12	Calculating Specificity	41
3.13	Calculating Accuracy	42
3.14	Calculating Precision	42
3.15	Calculating Error Rate	42
4.1	Calculating Total Bias-Variance Error	68
5.1	Representation of Hyperparameter Optimization	71

List of Publications

1. **Syed Immamul Ansarullah, Dr. Pradeep Kumar,** “Heart Disease Prediction and Diagnosis by Finding the Correlation and Significance Among Different Risk Attributes Through Machine Learning Techniques”, Journal of Advanced Research in Dynamical and Control Systems (JARDCS) , Vol. 10, 02-Special Issue, 2018 (**Scopus Indexed**) (ISSN Number: 1943023X)

2. **Syed Immamul Ansarullah, Dr. Pradeep Kumar,** ” Performance and evaluation of heart disease risk assessment model using machine learning classification techniques”, Journal of Advanced Research in Dynamical and Control Systems (JARDCS) , Vol. 10, 02-Special Issue, 2018 (**Scopus Indexed**) (ISSN Number: 1943023X)

3. **Syed Immamul Ansarullah, Dr. Pradeep Kumar,** ” A Systematic Literature Review On Cardiovascular Disorder Identification Using Knowledge Mining and Machine Learning Methods”, International Journal of Recent Technology and Engineering (IJRTE), Revised Manuscript Received on December 22, 2018,Vol. 10, 02-Special Issue, 2018 (**Scopus Indexed**) (ISSN Number: 2277-3878)

4. **Syed Immamul Ansarullah, Dr. Pradeep Kumar,** “Heart Disease Prognosis and Identification Using Different Machine Learning Techniques”, at 6th International Conference on Recent Challenges in Engineering and Technology held on 24th to 25th Nov, 2018 at Nagpur, Maharashtra, INDIA.

5. **Syed Immamul Ansarullah, Dr. Pradeep Kumar,** “Heart Disease Risk Assessment Model Development Using Machine Learning Techniques”, at 6th International Conference on Recent Challenges in Engineering and Technology held on 24th to 25th Nov, 2018 at Nagpur, Maharashtra, INDIA.

6. **Syed Immamul Ansarullah, Dr. Pradeep Kumar,** “Heart Disease Prediction using ensemble Classification Technique” at National conference on emerging trends and issues in information technology & communication, ETIITC-18, MANUU Hyderabad, INDIA.

7. **Syed Immamul Ansarullah, Pradeep Kumar, Abdul Wahid, Mudasir M Kirmani,** “Heart Disease Prediction System using Data Mining Techniques: A study” International Research Journal of Engineering and Technology (IRJET) (ISSN: 2395 -0056), Volume: 03 Issue: 08 | Aug-2016.

باب 1

تعارف

1.1 پس منظر

حالیہ برسوں میں اموات اور معذوری infectious بیماریوں سے کینسر، diabetes اور امراض قلب جیسی دائمی بیماری میں منتقل ہو گئی ہے۔ Infectious بیماریوں سے دائمی بیماری کی طرف جانے والے اس اقدام کو epidemiologic transition کہا جاتا ہے۔ ہر وقت دنیا کے مختلف ممالک یا یہاں تک کسی ملک کے اندر بھی مختلف خطے epidemiologic transition پر ہوتے ہے[1] ۔ گلوبل ہیلتھ کیئر ڈیٹا (2015) سے پتہ چلتا ہے کہ ہر سال 17.3 ملین اموات امراض قلب کے بذریعہ ہوتی ہے[2] اور سال 2030 تک اموات کا تناسب بڑھ کر 23.6 ملین ہو جائے گا[3]۔ عالمی ادارہ صحت نے بتایا ہے کہ مختلف ممالک میں دل اور سانس کی بیماریاں موت کے سب سے بڑے ذرائع ہیں[6][5][4]۔

اگرچہ امراض قلب وسیع و عریض بیماری میں سے ایک ہے جس کی وجہ سے پوری دنیا میں معذوری اور اموات بڑی فیصد میں ہوتی ہے لیکن اس کو انتہائی قابل علاج اور قابل کنڑول بیماری کے طور پر تسلیم کیا جاتا ہے [7]-کارڈیک ڈس آرڈر کے شکار افراد کی ابتدائی شناخت مریضوں کی صحت یاب ہونے اور اموات کے تناسب کو کم کرنے میں فائدہ مند ثابت ہو سکتی ہے[8]۔ ورلڈ ہیلتھ آرگنائزیشن (ڈبلیو ایچ او) نے رپورٹ کیا ہے کہ امراض قلب کی ابتدائی پیش گوئی اور اس کے علاج سے اہم حالات اور پیچیدگیوں میں پیشرفت کم ہوتی ہے[9]۔ شدید خطرے میں مریضوں کو پہچاننا اور ابتدائی تشخیص کے لئے علم کی فراہمی کے لئے درست منظم ماڈل کی ضرورت ہے[11][10]۔ مختلف محققین صحت کی دیکھ بھال کرنے والی صنعتوں میں ڈیٹا مائننگ techniques کا استعمال کرتے ہیں تاکہ ابتدائی مرحلے میں امراض قلب کی تشخیص میں ہیلتھ کیئر کے پیشہ ورافراد کی مدد کی جا سکے۔

1.2 امراض قلب کا جائزہ

بیماری کی وضاحت کرنے کی متعدد کوششیں ہوئی ہیں لیکن بیماری کی تسلی بخش وضاحت بیان کرنا حیرت انگیز طور پر مشکل ہے۔ The disease can loosely be defined “as a condition of the body or some part or organ of the body in which its functions are disrupted or deranged” [12]. تمام کم ، درمیانی اور اعلی آمدنی والے ممالک میں دل کی بیماری موت کی ایک اہم وجہ ہے اور مرد اور خواتین دونوں کے لئے موت کا سب سے اہم ذریعہ ہے[13][14]۔ امراض قلب ایک umbrella کی اصطلاح ہے جو کسی بھی عارضے کے لئے استعمال ہوتی ہے جو دل کو متاثر کرتی ہے۔ امراض قلب کے تحت ہونے والی بیماریوں میں Coronary Artery Disease (CAD), Arrhythmias, Congenital Heart Defects شامل ہیں اور امراض قلب کو اکثر Cardiovascular Diseases (CVDs) بھی کہا جاتا ہے[16] [15] ۔ امراض قلب اس وقت ہوتا ہے جب plaque کی تعمیر سے کورونری arteries تنگ ہو جاتی ہیں۔ ایتھروما (plaque) کولیسٹرول، چربی، اور دیگر مادوں کا جمع ہے جس کے نتیجے میں وقت کے ساتھ ساتھ دل کو خون کی فراہمی میں کمی واقع ہوتی ہے[18] [17]۔ خون کی رگوں کو اس طرح تنگ کرنے سے محسوس ہونے والا درد فالج، انجائنا یا دل کا دورہ پڑنے کا باعث بن سکتا ہے[20][19]۔ امراض قلب سے وابستہ علامات ایک دوسرے کے دل کے مرض سے مختلف ہو سکتے ہیں لیکن عام طور پر عام علامتوں میں جیسے بازوؤں، کہنیوں، بائیں کندھے، جبڑے یا کمر میں تکلیف اور سینے کے بیچ میں اضطراب یا پریشانی شامل ہیں [21]۔

اس باب میں امراض قلب اور اس کی اموات کی شرح کا خاکہ پیش کیا گیا ہے۔ امراض قلب کا عالمی بوجھ، اس کی پہچان اور تشخیصی اور اس کے خطرے کی تشخیصی کی بھی وضاحت کی گئی ہے۔ اس کے بعد، اس باب میں ڈیٹا مائننگ کا جائزہ، کے ڈی ڈی میں ڈیٹا مائننگ ایک بنیادی اقدام کے طور پہ ، صحت کی دیکھ بھال میں ڈیٹا مائننگ کی ایپلی کیشنز، مسئلے کا بیان، تحقیقی مقاصد اور آخر میں مقالہ خاکہ پر بھی تبادلہ خیال کیا گیا ہے۔

1.2.1 امراض قلب کی شرح اموات

امراض قلب کو ترقی پذیر اور ترقی یافتہ دونوں ملکوں کی عمومی آبادی میں موت کی بنیادی وجہ قرار دیا گیا ہے۔ WHO نے بتایا ہے کہ مردوں اور خواتین میں امراض قلب کے اموات کا تناسب نسبتا یکساں ہے جن میں 3.8 ملین اموات مرد اور 3.4 ملین اموات خواتین ہیں[2]۔ مختلف ممالک اور براعظموں میں صحت کی تنظیموں کے مطابق امراض قلب اموات کی سب سے اہم وجہ ہے۔ برٹش ہارٹ فاؤنڈیشن کا کہنا ہے کہ برطانیہ میں ہونے والی اموات میں 26% دل کی بیماری کا سبب بنتا ہے[22]۔ The Australian Bureau of Statistics کا کہنا ہے کہ آسٹریلیا میں اموات کی بنیادی وجہ دل اور گردش کے امراض ہیں جو کہ 33.7 فیصد اموات کا مجموعی

ہے[23][24][25][26] Economic and Social Commission of Asia and the Pacific (ESCAP 2010) کی رپورٹ کے مطابق ایشیائی ممالک کا 1/5th غیر سنجیدہ بیماریوں جیسے کینسر، امراض قلب اور سانس کی دائمی بیماریوں کا شکار ہے[27]۔

مختلف علاقوں کی صحت رپورٹس امراض قلب کے ذریعہ اموات کی شرح کو مختلف بتاتی ہیں۔ The East Asia and Pacific Region میں ہونے والی تمام اموات 35.2% ہے جو اسکیمک بیماری سے ہوا ہے [28]اور مشرق وسطی اور شمالی افریقہ کے علاقوں میں 47 فیصد اموات دل کی بیماری کی وجہ سے ہوتی ہیں[11]۔ اسی طرح جنوبی ایشیاء میں، اموات کی سب سے بڑی وجہ امراض قلب ہے جو رپورٹ ہونے والی اموات کا 10.6 فیصد ہے[29] اور مغربی افریقہ میں سب سہارن افریقہ نے رپورٹ دی ہے کہ تمام اموات میں سے 13% CVD کی وجہ سے ہوئی ہیں[31] [30]۔ اسی طرح مشرقی یورپ اور وسطی ایشیاء، لاطینی امریکہ اور کیریبین اور ایشیاء پیسیفک، آسٹرالاسیا، مغربی یورپ اور شمالی امریکہ نے یہ ظاہر کیا کہ دل کی بیماری گردش کی بیماریوں میں غلبہ رکھتی ہے[33] [32]۔ عالمی ادارہ صحت کے اعداد و شمار سے پتہ چلتا ہے کہ امراض قلب تمام ممالک میں موت کا باعث بنی ہے۔

1.2.2 امراض قلب کا عالمی بوجھ

امراض قلب ایک اہم عالمی مسئلہ ہے جس کے متعدد مراحل میں اس کے خاطر خواہ نتائج برآمد ہوتے ہیں: انفرادی اموات اور معذوری، خاندانی تکلیف اور حیرت انگیز معاشی اخراجات۔ امراض قلب کے بوجھ متنوع ہیں جن کی وضاحت مندرجہ ذیل ہے : برٹش ہارٹ فاؤنڈیشن کا تخمینہ ہے کہ برطانیہ میں دل کی بیماری کی قیمت ہر سال 9 بلین پاؤنڈ ہے اس معاشی لاگت میں قبل از وقت موت اور دل کی بیماریوں کی وجہ سے معذوری سے متعلق اخراجات شامل ہیں[24]۔ United States میں اسٹروک اور دل کی خرابی کے سالانہ اخراجات کا تخمینہ 312.6 بلین ڈالر ہے اور 2035 تک لاگت 1.1 ٹریلین ڈالر تک بڑھے گی[10]۔ چین میں دل کی بیماریوں کے سالانہ اخراجات 40 ارب سے زیادہ یا مجموعی قومی آمدنی کا تقریبا 4% ہیں[34]۔ جنوبی افریقہ 2 سے 3 فیصد Gross National Income (GNI) کے امراض قلب کے علاج پر خرچ کرتا ہے جو تقریبا جنوبی افریقہ کی بنیادی نگہداشت کے اخراجات کے چوتھائی کے برابر ہے[11]۔ عالمی سطح پر سال 2001 کے لئے امراض قلب کے صحت کی دیکھ بھال کے اخراجات کا تخمینہ 370 بلین امریکی ڈالر تھا جو اس سال کے لئے کل عالمی سطح پر صحت کی دیکھ بھال کے

اخراجات کا 10 فیصد نمائندگی کرتے ہیں [35] [32]۔ اسی طرح مشرقی یورپی خطے میں ہائی بلڈ پریشر کے اخراجات کا تخمینہ تقریبا 25% طبّی دیکھ بھال کے اخراجات میں لگایا گیا تھا [14] [13]۔

امریکن ہارٹ ایسوسی ایشن نے ہائی بلڈ پریشر، Coronary Artery Disease، اسٹروک، انجینا اور سی وی ڈی کی دیگر اقسام کے مستقبل کے اخراجات کی پیش گوئی کے لئے ایک طریقہ کار وضع کیا[35]۔ ڈیولپڈ طریقہ کار سے پتہ چلتا ہے کہ سال 2030 تک امریکی آبادی کے 40.8 % لوگوں کو کسی طرح کی بھی دل کی بیماری ہو سکتی ہے۔ سال 2013 اور 2030 کے درمیان امراض قلب کی دیکھ بھال کے کل اخراجات 320 سے لے کر 818 امریکی ارب بلین ڈالر تک بڑھنے کی پیش گوئی کی گئی ہے۔ ان رپورٹوں میں یہ واضح کیا گیا ہے کہ امراض قلب کے پھیلاؤ اور اخراجات میں کافی اضافہ ہونے کی پیش گوئی کی گئی ہے۔ ان رپورٹس سے یہ بھی ظاہر ہوتا ہے کہ امراض قلب کی وبا نے دنیا پر کافی اثر ڈالا ہے اور وہ دونوں انسانی مصائب اور اس سے ہونے والے نقصان کے لحاظ سے صحت اور ترقی کے سب سے بڑے چیلنجوں میں سے ایک ہے جو انہوں نے ممالک کی سماجی و اقتصادی بنیاد پر عائد کیا ہے[37] [36]۔ امراض قلب کے تباہ کن معاشرتی، معاشی اور عوامی صحت کے اثرات کو سمجھنے کے بعد ابتدائی مراحل میں اس کی تشخیصی ضروری ہے اور کارروائی کرنے میں تاخیر سے صورتحال خراب ہونے کا نتیجہ برآمد ہوگا[38]۔

1.2.3 امراض قلب کی شناخت اور تشخیص

امراض قلب کو ایک انتہائی روک تھام اور قابل قابو والا مرض سمجھا جاتا ہے۔ صحت مند غذا، باقاعدگی سے جسمانی سرگرمی اور سگریٹ نوشی کو روکنے کے ذریعہ کم از کم 80% ہارٹ ڈیزیز سے بچا جا سکتا ہے[40] [39]۔ امراض قلب سے مرنے والی آبادی میں خطرے کے کچھ عمومی عوامل ہوتے ہیں جو طرز زندگی سے متاثر ہوتے ہیں[41]۔ امراض قلب کی نشاندہی اور روک تھام کا بنیادی مقصد بیماری سے اسٹریٹجک فاصلہ برقرار رکھنا اور بیماری کی بہتری پر دخل اندازی کرنا ہے۔ امراض قلب کی روک تھام کی سرگرمیاں تین مختلف سطحوں پر انجام دی جاتی ہیں جیسے پرائمری روک تھام، ثانوی روک تھام اور tertiary روک تھام[42]۔ پرائمری روک تھام کا تعلق صحت مند لوگوں سے ہے جو اس خطرے کے عوامل کو کم کرنے کے طریقہ کار سے نمٹتے ہے جس سے بیماریوں کے واقعات پیدا ہو سکتے ہیں[35]۔ ثانوی روک تھام کا تعلق خطرے کے عوامل اور بیماری کے ابتدائی پتہ لگانے سے ہے جس میں کامیاب طبّی علاج کے امکانات میں اضافہ ہوتا ہے۔ Tertiary روک تھام کا تعلق بیماری کے طبّی علاج اور خطرے کے عوامل پر قابو پانے سے ہے[43] [17]۔

پرائمری اور ثانوی روک تھام امراض قلب کو قابو کرنے میں دو اہم عوامل ہیں۔ پرائمری روک تھام امراض قلب کے اثرات کو کنٹرول کرنے میں اہم کردار ادا کرتا ہے۔ دل کے عارضے کے شکار افراد کی غیر وقتی تشخیص سے مریضوں کی صحت یابی اور اموات کی شرح کو کم کرنے میں مدد مل سکتی ہے[17]۔ امریکن ہارٹ ایسوسی ایشن کی رپورٹ ہے کہ 30 سے 69 سال کی عمر کے افراد میں 11.4 ملین امراض قلب سے ہونے والی اموات اور 70 سال یا اس سے زیادہ عمر کے افراد کے درمیان 15.9 ملین امراض قلب سے ہونے والی اموات کو روکا جاسکتا ہے اگر تمبا کو اور الکحل کو ختم کرنا، نمک کی مقدار میں کمی، موٹاپا کو کنٹرول کرنا اور بلڈ پریشر کو کم کرنا جیسے مقاصد پورے ہو جاتے ہیں[10]۔ WHO نے اطلاع دی ہے کہ امراض قلب کی پیش کش کی شناخت اور علاج اہم اور مہنگی بیماریوں اور پریشانیوں میں پیشرفت کو کم کرنے کے لئے پرعزم ہے لہذا کارڈیک ڈس آرڈر کی ابتدائی پیش گوئی اور علاج کے اس مقصد کو حاصل کرنے کے لئے صحیح اور منظم ماڈل کی بنیادی ضرورت ہے جو ان متاثرین کی درجہ بندی کرے جو امراض قلب کے خطرے سے دوچار ہیں[39]۔

امراض قلب کے متعدد ٹیسٹ جیسے ای سی جی (الیکٹرو کارڈیو گرام)، کورونری انجیو گرافی اور Exercise Stress Test وغیرہ سے ثانوی سطح کی روک تھام کی نشاندہی کی جاسکتی ہے۔ تاہم یہ ٹیسٹ مہنگے ہیں اور انکے استعمال کے لئے پیچیدہ قسم کے آلات کی ضرورت پڑتی ہے۔ بدقسمتی سے یہ خیال کیا جارہا ہے کہ امراض قلب کی ہلاکتوں میں 82% آنے والے اضافے LMICs میں ہوگا[12]۔ LMICs کے معاشی حالات امراض قلب کی تشخیص کے تقاضے کو پورا کرنے کے لئے ضروری آلات اور طبّی سہولیات کی دستیابی کو محدود کرتے ہیں۔ ناکافی وسائل کے پیش نظر کم لاگت کے طریقوں کی نشاندہی کرنا ایک اولین ترجیح ہے۔ اعلی خطرے میں لوگوں کی شناخت کے لئے کمیونٹی سطح کی اسکریننگ ٹیسٹوں کا استعمال ایک ثانوی روک تھام کا طریقہ کار شامل ہے اور یہ LMICs میں مؤثر ثابت ہوگا[44]۔

1.2.4 امراض قلب کے خطرے کا تخمینہ

اگرچہ امراض قلب کا ایک بہت بڑا حصہ قابو پانے کے قابل ہے لیکن وہ آگے بڑھتے رہتے ہیں کیونکہ بچاؤ کے طریقے ناکافی ہیں۔ چونکہ عالمی سطح پر امراض قلب کی شدت میں تیزی آرہی ہے مریضوں کی صحت کی دیکھ بھال کو بڑھانے کے لئے فارمیسیوں میں اسکریننگ ٹیسٹ کی ضرورت ہے۔ عوامی سطح کے اسکریننگ ٹیسٹ غیر وقتی تخمینہ اور دل کی بیماری کی جانچ میں مدد کر سکتے ہیں[28]۔ اگرچہ جدید ٹیکنالوجیز کو دل کی حالت کی ابتدائی تشخیص کے لئے استعمال کیا جاتا ہے، تاہم یہ ٹیسٹ مہنگے پڑتے ہیں اور اسے کمیونٹی سطح کی اسکریننگ ٹیسٹ کے طور پر استعمال نہیں کیا جاسکتا

ہے۔ لہذا کمیونٹی سطح کی اسکریننگ ٹیسٹ کے طور پر سستے ٹیسٹ لینے کی ضرورت ہے[36]۔ اگر ہم امراض قلب کے بڑھتے ہوئے بوجھ کو کم کرنا چاہتے ہیں تو اس کے خطرناک عوامل کو تسلیم کرنا ضروری ہے جو دنیا کو ناگوار صورتحال کی طرف کھینچتے ہیں[45]۔ یہ بات بڑے پیمانے پر قبول کی گئی ہے کہ عمر، شراب نوشی، غیر صحت بخش غذا، سگریٹ نوشی، اور جسمانی عدم فعالیت جیسے عوامل امراض قلب کے اہم فیچرس ہیں [46] [47] اور ان فیچرس کی مسلسل نمائش کے نتیجے میں ہائی بلڈ پریشر [48]، Diabetes [49]، Dyslipidaemia [50]، موٹاپا [51] اور اسٹروک[52] میں اضافہ ہوتا ہے۔

ہارٹ ڈیزیز ڈیاگنوسس ٹولز بڑے پیمانے پر دستیاب ہیں جیسے رینالڈس رسک اسکور [53]، Framingham ہارٹ ڈیزیز رسک اوالویشن ٹول [55][54] اور Australian Absolute Cardiovascular Risk Calculator[56]، تاہم یہ تمام خطرہ تشخیصی ٹولز پیش گوئی کے نتیجے میں کم استعمال ہوتے ہے کیونکہ ان میں خون کے نمونے کی بھی ضرورت ہوتی ہے جو کہ ایک ناگوار اور نسبتا مہنگا عمل ہے جو طبّی ترتیبات کے علاوہ ان کے استعمال کو کم کر دیتا ہے۔ لہذا ان ٹیسٹوں کو آسان بنانے اور امراض قلب سے متاثرہ افراد کی جلد تشخیصی کے لئے صرف نان انویسیو ہارٹ ڈیزیز رسک اوالویشن فیچرز کا استعمال کرنے کی ضرورت ہے۔ اس سے قبل امراض قلب کے خطرے کے حساب سے مکمل طور پر نان انویسیو رسک فیچرز کے استعمال کی جانچ نہیں کی گئی ہے۔ مزید برآں اگر نان انویسیو رسک فیچرز Cardiac Disorder کی شکایت کے خطرات کی تشخیصی میں خاطر خواہ کامیابی کا مظاہرہ کرتی ہیں تو یہ تجزیہ امراض قلب کے ابتدائی تشخیصی کے لئے بے حد فائدہ مند ثابت ہوگا۔

1.3 ڈیٹا مائننگ کا عمومی جائزہ

متعدد محققین نے اس بات پر اتفاق کیا کہ ڈیٹا مائننگ کثیر الضابطہ field ہے اور اس کی وضاحت مختلف نقطہ نظر سے کی جاسکتی ہے۔ محقق[57] کے مطابق ” Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques.” اسطرح محققین[58] نے ڈیٹا مائننگ کی تعریف میں یہ اضافہ کیا۔ “data mining is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large databases.” اسی طرح محقق [59] نے ڈیٹا مائننگ کو بطور “a process of nontrivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database” بیان کیا ہے۔ محققین [60] ڈیٹا مائننگ کی تعریف کچھ اس طرح کرتے ہیں “as an attempt to discover hidden patterns where these patterns are difficult to detect with traditional statistical methods.” کرتے ہیں۔ یہ تعریفیں ہمیں یہ نتیجہ اخذ کرنے کی اجازت دیتی ہیں کہ ڈیٹا مائننگ گہری چھپی دلچسپ اور صحیح نمونوں کی شناخت کے لئے بھاری مقدار میں خام ڈیٹا سے مفید معلومات کا حصول ہے۔

1.3.1 کے ڈی ڈی میں ڈیٹا مائننگ ایک بنیادی اقدام کے طور پہ

محققین ڈیٹا مائننگ کو مختلف نقطہ نظر سے بیان کرتے ہیں کچھ محققین ڈیٹا مائننگ کو KDD (Knowledge Discovery from Data) کے طور پر بیان کرتے ہیں تاہم دیگر محققین ڈیٹا مائننگ کو کے ڈی ڈی میں محض ایک بنیادی اقدام قرار دیتے ہیں۔ ذیل میں دیئے گئے ترسیم 1.1 مندرجہ ذیل مراحل کی تکراری سیریز کے بطور ڈیٹا بیس میں علم کی دریافت کو ظاہر کرتا ہے [59]۔ Data Cleaning and Integration کے ڈی ڈی کا پہلا مرحلہ ہے جو noise کو دور کرنے اور ڈیٹا میں عدم مطابقت کو دور کرنے کے لئے لاگو ہوتا ہے جب کہ ڈیٹا انضمام متضاد ذرائع سے ڈیٹا کو مربوط ڈیٹا اسٹور میں جوڑتا ہے۔ دوسرا مرحلہ Data Selection and Transformation ہے جس میں تجزیہ مقصد کے لئے ڈیٹا اسٹور سے موزوں ڈیٹا نکالا جاتا ہے اور پھر خلاصہ اور اجتماعی تکنیک کا مظاہرہ کرکے مائننگ کے مقصد کے لئے موزوں شکلوں میں مستحکم اور تبدیل ہو جاتا ہے۔ تیسرا مرحلہ ڈیٹا مائننگ ہے جہاں بصیرت انگیز طریقوں کا اطلاق mined ڈیٹا کے نمونوں پر ہوتا ہے۔ چوتھا مرحلہ Pattern Evaluation ہے جو دلچسپی کے اقدامات پر مبنی علم کی نشاندہی کرنے والے انتہائی دلائل نمونوں کی نشاندہی کرتا ہے۔ کے ڈی ڈی عمل کا حتمی مرحلہ Knowledge Representation ہے جہاں صارفین کو مائننگ معلومات فراہم

کرنے کے لئے بصارت اور علم کی نمائندگی کے طریقوں کا استعمال کیا جاتا ہے[59]۔ اگلے دہائی میں ڈیٹا مائننگ کی جدید ترین پیشرفت کی پیش گوئی کی جا رہی ہے کیونکہ ایپلی کیشنز کی ایک بڑی حد میں یہ روزانہ زیادہ وسیع ہوتا جاتا ہے [61]۔

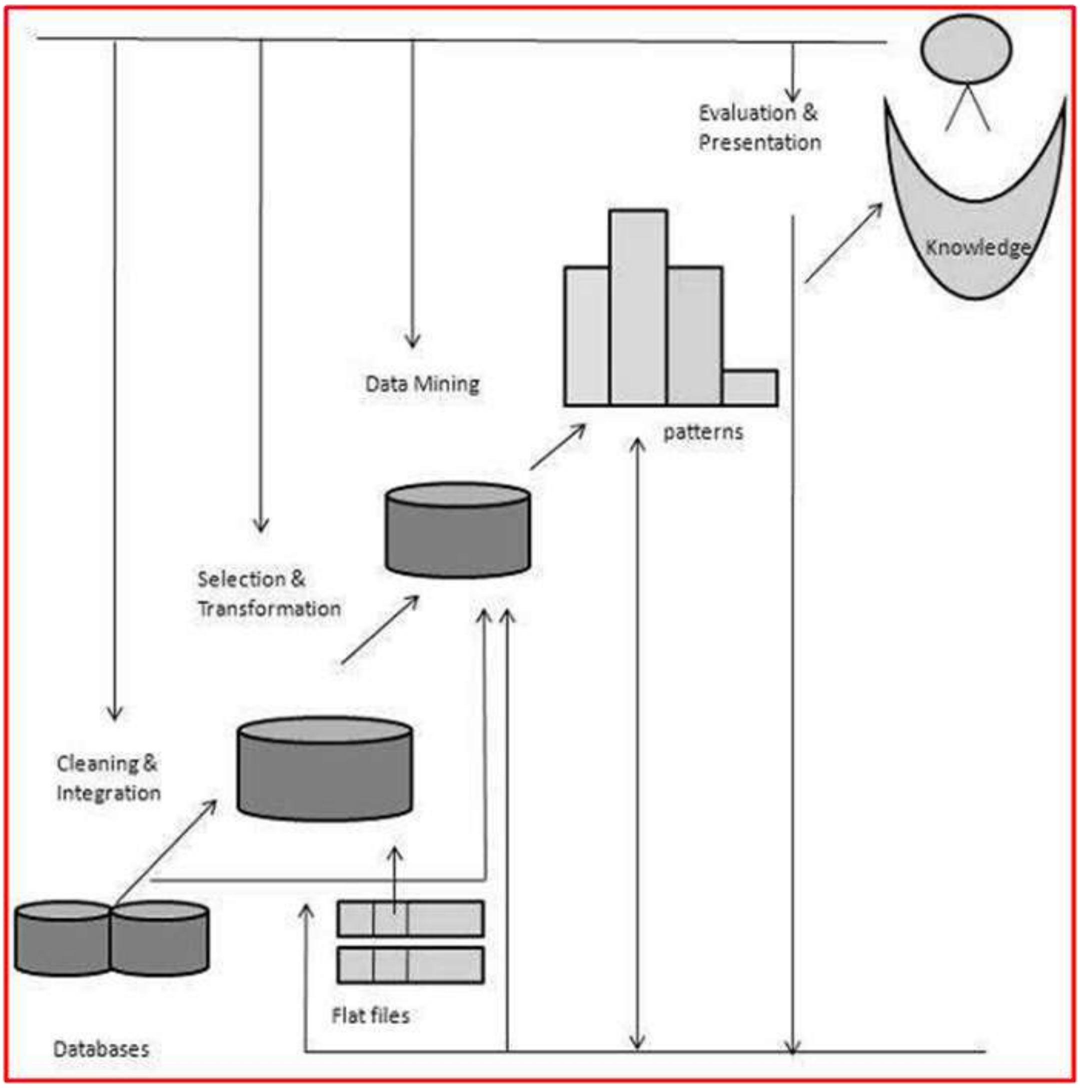

Evaluation &
Presentation
Knowledge
Data Mining
patterns
Selection &
Transformation
Cleaning &
Integration
Databases
Flat files

Figure 1.1 The process of Knowledge Discovery in Data [Source: 59]

1.3.2 ہیلتھ کیئر میں ڈیٹا مائننگ کا استعمال

انتہائی اطلاق پر مبنی مضمون کی وجہ سے ڈیٹا مائننگ نے مختلف شعبوں میں غیر معمولی کارنامے دیکھے ہیں۔ ویب مائننگ ، Business Intelligence ، Load Forecasting، Diagnosis، Marketing and Sales، Oil Refinement اور اسکریننگ کی تصاویر جیسے ایپلی کیشنز کے وسیع دائرہ کار میں ڈیٹا مائننگ کامیابی کے ساتھ بڑھ رہی ہے[64][63 [62]۔ پیش گوئی اور تخمینہ فراہم کرنے کے لئے صحت کی دیکھ بھال میں ڈیٹا مائننگ ایک اعلی field کی اہمیت کا حامل ہے[67] [66] [65]۔ ہیلتھ کیئر میں ڈیٹا

مائننگ کی اپیلی کیشنز صحت کی بہتر حکمت عملی وضع کرنے اور ہسپتال کی ناکامیوں سے بچنے، ابتدائی روک تھام اور بیماریوں کی نشاندہی اور اسپتال میں غیر ضروری اموات، ہیلتھ کیئر کی فراہمی میں رقم اور قیمت کی بچت۔ جعلی انشورنس دعووں کے بارے میں تحقیقات شامل ہیں[68] ۔

ہیلتھ کیئر ڈیٹا مائننگ میں میڈیکل ڈومین کے ڈیٹا سیٹس میں پوشیدہ نمونوں کی تفتیش کی غیر معمولی صلاحیت ہے۔ صحت اور طب کے ذریعہ پیش کیا جانے والا سب سے اہم چیلنج ایک ایسی ٹکنالوجی کی تشکیل ہے جس میں قابل اعتماد مفروضے پیش کیے جا سکیں جو ان اقدامات پر مبنی ہیں جن پر طبّی صحت کی تحقیق پر بھروسہ کیا جاسکتا ہے اور اسے کلینیکل ماحول میں لاگو کیا جاسکتا ہے[69] تربیتی ڈیٹا سیٹس سے مختلف نظریاتی مضمرات کی دریافت ایک مشکل عمل ہے یہاں تک کہ بہترین ماہرین جمع شدہ اعداد و شمار سے مغلوب ہیں۔ لہذا ڈیٹا مائننگ کو decision making میں صحت کے پیشہ ور افراد کو فائدہ پہنچانے کے لئے استعمال کیا جاتا ہے[70]۔

محققین متعدد بیماریوں کی تخمینہ میں ڈیٹا مائننگ کی تکنیکس استعمال کررہے ہیں جیسے Diabetes[71], Stroke[72], Cancer[73] and Heart Disease[74]. ہیلتھ کیئر کا مستقبل صحت کی دیکھ بھال کے اخراجات کو کم کرنے، علاج کی حکمت عملیوں اور بہترین طریقوں کا تعین کرنے، کارکردگی کی جانچ کرنے، جعلی انشورنس اور طبّی دعووں کا پتہ لگانے اور بالآخر مریضوں کی نگہداشت کے معیار کو بہتر بنانے کے لئے ڈیٹا مائننگ کے استعمال پر بھرپور انحصار کرسکتا ہے۔

1.4 مسئلے کا بیان

امراض قلب بہت ہی مشہور اور دائمی بیماری ہے جس کی وجہ سے پوری دنیا میں اموات کی زبردست شرح ہوتی ہے تاہم اس کی ابتدائی پیش گوئی مریضوں کی صحت کی بحالی اور اموات کی شرح کو کم کرنے میں فائدہ مند ثابت ہوسکتی ہے۔ اموات کی شرح، معذوری اور لاگت کی بنیاد پر امراض قلب کی تشخیص کے لئے ایک درست طریقہ کار کی لازمی ضرورت ہے۔ کم لاگت سے امراض قلب کے ابتدائی مرحلے میں تشخیص کرنے کے لئے اسکریننگ ماڈل استعمال کیے جاتے ہیں لیکن ان ماڈلز میں پہلے خون کے نمونے لینے کی ضرورت ہوتی ہے جو ایک ناگوار اور مہنگا عمل ہے۔ اس وجہ سے خطرہ کی خصوصیات کو ہموار کرنے اور ڈیٹا مائننگ ماڈل تیار کرنے کا مطالبہ کیا جارہا ہے جو عوام کی معیاری اسکریننگ کے ذریعہ امراض قلب کے شدید مریضوں کو پہچاننے اور ابتدائی مداخلت کی سہولت فراہم کرنے اور مریضوں کی صحت کو بڑھانے کے لئے علم پیدا کرنے کے لئے استعمال کیا جاسکتا ہے۔

1.5 تحقیقی مقاصد

اس تحقیق کے مقاصد حسب ذیل ہیں:

i. امراض قلب کی تشخیصی کے بارے میں ادب کا جائزہ لینے کے لئے ڈیٹا مائننگ تکنیکوں کا استعمال کرنا تاکہ شناخت شدہ تحقیقات کو پُر کیا جاسکے۔

ii. نان انویسیو ہارٹ ڈیزیز رسک فیچرز اور امراض قلب کی تشخیصی میں ان کی اہمیت کا مطالعہ اور تجزیہ کرنا۔

iii. امراض قلب کے متعلق تشخیصی ماڈل تیار کریں تاکہ امراض قلب مریضوں کو خطرے سے پہچان سکے اور جلد مداخلت کو قابل بنا سکے۔

iv. مختلف میٹرکس کا استعمال کرتے ہوئے امراض قلب ماڈل کی کارکردگی کا تجزیہ کرنا۔

v. ڈومین ماہرین کے علم سے، ادب میں موجودہ ماڈلز اور بینچ مارک امراض قلب کے ڈیٹا سیٹس کے ذریعہ امراض قلب کے خطرے سے متعلق تشخیصی ماڈل کی توثیق کرنا۔

1.6 تحقیق کی تشکیل

اس مقالہ کو چھ ابواب میں تقسیم کیا گیا ہے:

پہلا باب ایک تعارفی باب ہے جو تحقیقی کام کے پس منظر پر گفتگو کرتا ہے۔ اس کا آغاز امراض قلب، اس کی شرح اموات اور عالمی بوجھ کے جائزہ کے ساتھ ہوتا ہے۔ اس باب میں امراض قلب کی شناخت، امراض قلب کی تشخیصی اور خطرے کی تشخیصی پر تبادلہ خیال کیا گیا ہے۔ اس باب میں صحت سے متعلق نظاموں میں ڈیٹا مائننگ کا استعمال، مسئلے کا بیان اور تحقیقاتی کام کے مقاصد کی بھی وضاحت کی گئی ہے۔

باب 2 میں مختلف ڈیٹا مائننگ تراکیب کا استعمال کرتے ہوئے امراض قلب کی پیش گوئی اور پتہ لگانے کے بارے میں ایک تفصیلی منظم ادب کے جائزہ پر بات چیت کی گئی ہے۔

باب 3 میں فیچر سلیکشن تکنیکس پر تبادلہ خیال کیا گیا ہے جو امراض قلب کی ابتدائی پیش گوئی کے لئے خطرے کی صفات کے اہم غیر حملہ آور سبسیٹ کی تلاش کے لئے لگائے جاتے ہیں۔ اس باب میں ڈیٹا مائننگ کی مختلف تکنیکوں کا اطلاق ہوتا ہے جیسے رینڈم فارسٹ، نیوی بائیس، کے نیریسٹ نیبر، سپورٹ ویکٹر مشین اور ڈسیشن ٹری تاکہ یہ دیکھیں کہ آیا یہ تکنیکس بیماری کے ابتدائی پیش گوئی میں طبّی پیشہ ور افراد کی مدد

کریں گی یا نہیں جس کے نتیجے میں شدید اور مہنگی بیماری اور پیچیدگیوں میں کمی واقع ہوگی۔ ڈیولپڈ رسک ماڈل کی کارکردگی کا اندازہ کرنے کے لئے کنفیوژن میٹرکس، AUROC، ماڈل کی پیچیدگی وغیرہ جیسے اقدامات استعمال کیے گیے ہے۔ آخر میں باب غیر ناگوار امراض قلب کے خطرے کی خصوصیات کی اہمیت پر گفتگو کرتے ہوئے اختتام پزیر ہوتا ہے۔

باب 4 ڈیٹا مائننگ ٹاسکس اور تکنیکس کو بیان کرتا ہے۔ اس باب میں ڈیٹا مائننگ کی درجہ بندی تکنیکس جیسے رینڈم فارسٹ، نیوی بائیس، کے نیریسٹ نیبر، سپورٹ ویکٹر مشین اور ڈسیشن ٹری پر تبادلہ خیال کیا گیا ہے تاکہ ابتدائی مراحل میں امراض قلب کی پیش گوئی اور اس کا پتہ لگ سکے۔ اس باب میں اس بات کے طریقہ کار پر تبادلہ خیال کیا گیا ہے کہ جپٹر نوٹ بک ویب اپلی کیشن پر امراض قلب کے خطرے کی تشخیصی کا ماڈل کس طرح تیار کیا گیا ہے۔ تیار کردہ ماڈل کی کارکردگی کی پیائش کا اندازہ مختلف تشخیصی تکنیکوں کے ذریعے کیا جاتا ہے۔ مجوزہ تراکیب کے تجرباتی نتائج کی وضاحت کی گئی ہے اور ان کے موازنہ پر تبادلہ خیال کیا گیا ہے۔ تجرباتی نتائج سے ظاہر ہوتا ہے کہ رینڈم فاریسٹ امراض قلب ماڈل نے دوسرے ڈیولپڈ ماڈل سے بہتر کارکردگی کا مظاہرہ کیا۔ آخر میں اس باب کی پیش گوئی کرنے والے انداز کے bias اور variance کی غلطیوں پر تبادلہ خیال کر کے ختم کیا گیا ہے۔

باب 5 ہائپر پرامیٹر کی اصلاح اور اس کی تکنیکس کا تعارف پیش کرتا ہے۔ اس میں اس بات پر تبادلہ خیال کیا گیا ہے کہ ماڈل کی درستگی کو بہتر بنانے کے لئے کس طرح ڈیولپڈ ماڈل کو بہتر بنایا جائے۔ اس باب میں ہائپر پرامیٹر default and ماڈلز کے مابین موازنہ بھی انجام دیا گیا ہے اور پہلے سے طے شدہ اور ہائپر پرامیٹر کے ساتھ امراض قلب کے خطرے والے ماڈلوں کے مابین کارکردگی کی تشخیصی پر بھی تبادلہ خیال کیا گیا ہے۔ امراض قلب کے خطرے کی تشخیصی کے لئے پیدا کردہ قواعد کے زیادہ سے زیادہ سیٹ کی وضاحت کی گئی ہے۔ اس باب میں کارڈیک ڈس آرڈر کی تشخیصی کے لئے خطرے کی صفات کے مختلف امتزاج کی اہمیت بیان کی گئی ہے۔ آخر میں امراض قلب کے ماہر نظام کی تشخیصی کے اجزاء پر تبادلہ خیال کیا گیا ہے اور امراض قلب کی تشخیصی کا ماڈل پیش کیا گیا ہے۔

باب 6 تحقیق کے اختتام اور مستقبل کے کام کی وضاحت کرتا ہے۔ اس باب میں امراض قلب کے خطرے کی تشخیصی میں اہم خصوصیات جیسے قلبی بیماری کی تشخیصی کے لئے غیر ناگوار خطرہ خصوصیات کی اہمیت جیسے اہم سوالات کا خلاصہ بھی کیا گیا ہے۔ آخر میں تحقیقی حدود پر مستقبل کے کام کے ساتھ بھی تبادلہ خیال کیا گیا۔

باب 2

ادب کا جائزہ

اس باب میں مختلف محققین کی جانب سے ڈیٹا مائننگ کی مختلف تکنیکوں کا استعمال کرتے ہوئے امراض قلب رسک ماڈل تیار کرنے کے لئے دیئے گئے اہم شراکت کا خاکہ پیش کیا گیا ہے۔ اس باب میں ابتدائی تشخیصی اور امراض قلب کی شناخت کی اہمیت پر بھی روشنی ڈالی گئی ہے۔ آخر میں مروجہ ادب میں پائے جانے والے تحقیقی خلیوں پر تبادلہ خیال کیا گیا۔

2.1 مختلف ڈیٹا مائننگ ٹاسکس اور ٹکنیکس کا استعمال کرتے ہوئے امراض قلب کی پیشن گوئی

حالیہ برسوں میں محققین نے مختلف ڈیٹا مائننگ ٹاسکس اور ٹکنیکس کا استعمال کرتے ہوئے امراض قلب کی خرابی کی نشاندہی کرنے میں فیصلہ کن شراکت کی۔ پہلے سے امراض قلب کی پیش گوئی کے لئے بہت سارے محققین ڈائیورجنٹ ڈیٹا سیٹ، مختلف مشین لرننگ الگورتمس، مختلف ڈیٹا مائننگ اپروچ اور متعدد ٹولز کا استعمال کرتے ہوئے ماڈل تیار کیئے ہیں جن کی وضاحت مندرجہ ذیل ہے۔

2.1.1 سوپروایزڈ ٹاسکس کے ذریعے امراض قلب کی پیشن گوئی

Predictive Tasks کا مقصد یہ ہوتا ہے کہ کسی دوسرے وصف کی قدروں کی بنیاد پر کسی خاص وصف کی قیمت کی پیش گوئی کی جائے۔ پیش گوئی کی جانے والی وصف کو عام طور پر target یا dependent متغیر کے طور پر جانا جاتا ہے، جبکہ پیش گوئی کرنے کے لئے استعمال ہونے والی صفات کو exploratory یا independent متغیر کے نام سے جانا جاتا ہے[59] ۔ مندرجہ ذیل محققین کارڈیک عارضے کی پیش گوئی کے لئے زیر نگران ڈیٹا مائننگ کی تراکیب کا استعمال کیئے ہیں:

(Colombet et al. 2000) [75] نے 15444 مثالوں پر مشتمل ایک حقیقی ڈیٹاسیٹ سے امراض قلب کی تشخیصی کے لئے CART (Classification and Regression Tree) اور ملٹی لیئر پرسیپٹرن ڈیٹا مائننگ الگورتمس کا استعمال کیا ہے۔ ڈیٹاسیٹ کو تصادفی طور پر (10،296) مثالوں کے ٹریننگ سیٹ اور (5،148) مثالوں کے ٹیسٹ سیٹ میں تقسیم کیا گیا ہے۔ محققین نے ٹیسٹ ڈیٹاسیٹ کی بنیاد پر مختلف کارکردگی کی پیمائش کا استعمال کرتے ہوئے ڈیولپڈ رسک ماڈل کی کارکردگی کا جائزہ لیا ہیں اور ROC

Analysis کی بنیاد پر عمل درآمد کے معیار، وضاحتی معیار اور امتیازی کارکردگی کے معیار پر غور کیا ہے۔ محققین نے ROCKIT سافٹ ویئر کا استعمال کرتے ہوئے ROC Curve کا تجزیہ کیا۔ تجرباتی نتائج سے ظاہر ہوتا ہے کہ امراض قلب کی تشخیص میں ایم ایل پی کی پیش گوئی کرنے کی صلاحیت CART سے زیادہ ہے۔

(Yan et al. 2003) [76] نے ملٹی لیئر پرسیپٹرن نیورل نیٹ ورک کا استعمال کرتے ہوئے ایک پیش گو کارڈیک ڈس آرڈر رسک ماڈل کو ڈیزائن کیا ہے۔ ڈیولپڈ ہارٹ ڈیزیز رسک اوالویشن سسٹم کو Back-Propogation الگورتھم momentum term، the adaptive learning rate اور forgetting mechanics کی مدد سے استعمال کرتے ہوئے تربیت دی گی ہے۔ سسٹم کی تربیت اور جانچ کے لئے مجموعی طور پر 352 میڈیکل ریکارڈ استعمال ہوئے ہیں۔ محققین نے Cross-Validation، ہولڈ آؤٹ اور بوٹسٹریپنگ جیسے تین مختلف تشخیصی طریقوں کا استعمال کرتے ہوئے ڈیولپڈ رسک ماڈل کی کارکردگی کا جائزہ لیا ہیں۔ تجرباتی نتائج سے پتہ چلتا ہے کہ ایم ایل پی پر مبنی دل کی بیماریوں کے خطرہ تشخیصی ماڈل میں قوی پیش گوئی کرنے والے اندازے کے ساتھ پانچ مختلف کارڈیک ڈس آرڈر کے اقسام کی درجہ بندی کرنے کی بڑی صلاحیت ہے۔

(Noh et al. 2006)[77] نے ای سی جی سے Heart Rate Valve کا تجزیہ کر کے ای سی جی پیٹرن اور کلینیکل تفتیش کے فریم ورک کے تحت Coronary Artery Disease کی تشخیص کے لئے ڈیٹا مائننگ کی درجہ بندی کی تکنیک کا استعمال کیا گیا ہے۔ مجوزہ طریقہ efficient FP-growth method پر مبنی ایک associative classifier ہے جس کی pruning بے کار قواعد کے لئے ایک ہم آہنگی اقدام کو استعمال کرتی ہے۔ 670 مریضوں کا ڈیٹا سیٹ اسوسیٹیو درجہ بندی کے لئے ایک تجربہ کرنے کے لئے استعمال کیا جاتا ہے جو متعدد قواعد، pruning اور متعص بانہ اعتماد (یا ہم آہنگی کی پیمائش) کو استعمال کرتا ہے۔ میڈیکل ڈومین ماہرین علم کا استعمال کر کے محققین مریضوں کی درجہ بندی کرتے ہیں جیسا کہ سی اے ڈی سے متاثر ہوتا ہے یا عام طور پر لومینل ٹریونگ کے اسٹینوسس کی بنیاد پر۔ محققین نے ڈیٹاسیٹ پر stratified 10-fold cross validation کا استعمال کیا اور مختلف کارکردگی کی پیمائش جیسے precision، F-measure، Recall اور روٹ میین اسکوائر ایرر (آر ایم ایس ای) کا استعمال کرتے ہوئے ڈیولپڈ رسک ماڈل کی کارکردگی کا جائزہ لیا۔ نتائج سے ظاہر ہوتا ہے کہ مجوزہ درجہ بندی کرنے والے نے سی اے ڈی کی تشخیص کے لئے دوسرے درجہ بندی الگورتھم کو بہتر بتایا۔

(Palaniappan and Awang 2008)[78] نے ڈیسیشن ٹری ، نیورل نیٹ ورک اور نیوی بائیس ڈیٹا مائننگ ٹکنیکس کا استعمال کرتے ہوئے ہارٹ ڈیزیز رسک اوالویشن ماڈل تیار کیا۔ ڈیولپڈ رسک ماڈل قلبی عارضے سے وابستہ دلچسپ پوشیدہ نمونوں کو نکال سکتا ہے اور ایسے پیچیدہ سوالوں کے جواب دے سکتا ہے جو موجودہ رسک تشخیصی ٹولز میں ناکام رہتے ہیں۔ محققین کلیولینڈ ہارٹ ڈیزیز ڈیٹا بیس سے ایک چھوٹا سا ڈیٹا سیٹ حاصل کرتے ہیں جس میں کل 909 ریکارڈز شامل ہیں جن میں 15 میڈیکل رسک فیچیرس ہیں۔ .NET پلیٹ فارم میں رسک ماڈل تیار کیا گیا ہے اور اس کے ساتھ interaction کرنے کے لئے data mining extension query language استعمال کیے جاتے ہیں۔ ڈیولپڈ رسک ماڈل کی کارکردگی کی جانچ پڑتال کے لئے Lift Chart اور Classification میٹرکس کو استعمال کیا گیا ہے تاکہ ہم دیکھ سکیں کہ کس ماڈل نے امراض قلب کے مریضوں کی تشخیص کے لئے صحیح پیش گوئیوں کا زیادہ سے زیادہ فیصد دیا ہے۔ تجرباتی نتائج سے یہ پتا چلا ہے کہ نیوی بائیس رسک ماڈل نے نیورل نیٹ ورک اور ڈیسیشن ٹری ماڈلز کو مات دی ہے۔

(Shouman, Turner, and Stocker 2011)[79] نے ڈیسیشن ٹری ڈیٹا مائننگ تکنیک کا استعمال کرتے ہوئے امراض قلب کے مریضوں کی جلد پیش گوئی کے لئے ایک ناول کلاسیفکیشن ماڈل تیار کیا۔ while developing the risk ماڈل evaluation researchers integrate multiple classifiers voting technique with different multi-interval discretization methods like (equal frequency, chi-merge, equal width and entropy) جیسے مختلف ڈیسیشن ٹری کی مختلف variants استعمال کرنا like (Gini Index, Gain Ratio and Information Gain) امراض قلب کے فیصلے کے قواعد نکالے جاتے ہیں اور پھر error pruning کرنے والی ٹکنیکس کا استعمال کرکے قواعد کے موثر سیٹ کا انتخاب کیا جاتا ہے۔ ڈیولپڈ ماڈل نے voting بغیر equal width discretization information gain کے ڈیسیشن ٹری کی تشکیل پر سب سے زیادہ درستگی 79.1 فیصد حاصل کی۔ ووٹنگ کا اطلاق کرنے کے بعد ماڈل نے equal frequency discretization gain ratio decision tree کے ذریعہ 84.1 فیصد کی اعلی ترین درستگی حاصل کی۔

(Rani 2011) [80] نے نیورل نیٹ ورک الگورتھم کا استعمال کرتے ہوئے ایک روبوٹک اور قابل اعتماد امراض قلب رسک ماڈل تیار کیا۔ رسک ماڈل کلیولینڈ ہارٹ ڈیزیز ڈیٹاسیٹ سے قابل اعتماد درجہ بندی کے قواعد نکالتا ہے۔ اس ماڈل کو فیڈ فارورڈ نیورل نیٹ ورک اور کی Back-propogation learning algorithm with momentum and variable learning شرح کے ساتھ تربیت دی گی ہے۔ نیٹ ورک کی کارکردگی کا تجزیہ کرنے کے لئے نیٹ ورک میں ان پٹ کے بطور بہت بڑا ٹیسٹ ڈیٹا دیا گیا ۔ learning کے عمل کو تیز کرنے کے لئے all hidden and output layers میں ہر نیوران میں ہم آہنگی کا نفاذ کیا گیا ۔ تجرباتی نتائج نے ثابت کیا کہ نیورل نیٹ ورک کی تکنیک نے درجہ بندی کے کام کے لئے تسلی بخش نتائج مہیا کیے ہے۔

(Kumari and Godara 2011) [81] نے کلیولینڈ ہارٹ ڈیزیز ڈیٹاسیٹ پر ڈسیشن ٹری، نیورل نیٹ ورک، RIPPER اور سپورٹ ویکٹر مشین الگورتھم کا استعمال کیا ۔ محققین نے 14 صفات کے ساتھ 303 ریکارڈ والے ڈیٹاسیٹ کا استعمال کیا ہیں اور WEKA Tool پر اس کی جانچ کی ہیں۔ ڈیٹاسیٹ پر پہلے Pre-Processing سلیکشن الگورتھم کا اطلاق کیا گیا جس کے نتیجے میں ڈیٹا مائننگ کی درجہ بندی کی ٹکنیکس کے 296 ریکارڈ رہے ۔ Accuracy، sensitivity اور specificity میٹرکس کا حساب لگانے کے لئے ایک کنفیوژن میٹرکس کا استعمال کیا گیا ۔ نتائج سے پتہ چل ہے کہ SVM نے accuracy, sensitivity, specificity, اور misclassification rate کے تمام پیرامیٹرز میں دیگر درجہ بندی الگورتھم کو پیچھے چھوڑ دیا ہے۔ ROC اسپیس پوائنٹ پر بھی ایک ایس وی ایم ماڈل دوسرے ماڈلز کے مقابلے میں کامل پوائنٹ (0.1) کے قریب ہے، جو ایس وی ایم کو امراض قلب کا بہترین پیش گو ثابت کرتا ہے۔

(Shouman, Turner, and Stocker 2012) [82] نے کلیولینڈ ہارٹ ڈیزیز ڈیٹاسیٹ کا استعمال کرتے ہوئے کے نیریسٹ نیبر ہارٹ ڈیزیز رسک اوالویشن ماڈل تیار کیا تاکہ زیادہ سے زیادہ درستگی کے ساتھ پہلے سے ہی کارڈیک ڈس آرڈر کے مریضوں کا پتہ لگ سکے۔ ابتدائی طور پر K کی ویلیو 1 پر رکھی جاتی ہے اور پھر تکراری طور پر 13 کی بالائی حد تک بڑھادی جاتی ہے اور جب K کی ویلیو 7 پہنچ جاتی ہے تب accuracy اور specificity الترتیب 97.4% اور 99% حاصل کی جاتی ہے۔ اس کام میں، محققین نے دریافت کیا کہ ووٹنگ کو لاگو کرنے سے پیرامیٹر کی مختلف اقدار کا تخمینہ لگانے کے بعد بھی صحت سے متعلق کوئی پیشرفت نہیں ہو سکتی ہے۔

(Chaurasia 2013)[83] محقق نے ID3, CART and Decision Table جیسے مختلف درجہ بندی algorthims کا استعمال کرتے ہوئے ایک ناول کارڈیک ڈس آرڈر کا پتہ لگانے والا ماڈل تیار کیا۔ ڈیولپڈ رسک ماڈل کو معیاری کلیولینڈ ہارٹ ڈیزیز ڈیٹاسیٹ پر تجربہ کیا گیا ۔ ہارٹ ڈیزیز ڈیٹاسیٹ کا تجزیہ کرنے کے بعد اس کا تجربہ WEKA ٹول پر کیا گیا ہے۔ رسک تشخیصی ماڈل کی کارکردگی کو مختلف اقدامات سے جانچا گیا ہے۔ محققین نے bias کو کم کرنے اور کارکردگی کو بہتر بنانے کے لئے 10-fold cross validation کا استعمال کیا۔ امراض قلب کی پیش گوئی کے لئے ہر انفرادی ان پٹ متغیر کی اہمیت کو بہتر طور پر سمجھنے کے لئے Chi-square, information gain and gain ratio کے ٹیسٹ کروائے گئے۔ تجرباتی نتائج سے پتہ چلتا ہے کہ CART اعلی ترین درستگی اور کم سے کم غلطی کی شرح کے ساتھ بہتر کارکردگی کا مظاہرہ کرتا ہے تاہم ڈیولپڈ رسک ماڈل کے وقت پیچیدگی میں اضافہ ہوتا ہے.

(Al-Milli 2013) [84] محققین نے back propogation neural network الگورتھم کا استعمال کرتے ہوئے کارڈیک ڈس آرڈر رسک ڈیولپڈ ماڈل تیار کیا۔ محقق کلیولینڈ بینچ مارک ہارٹ ڈیزیز ڈیٹاسیٹ استعمال کرتا ہے جس میں 13 میڈیکل اوصاف شامل ہیں۔ ریسرچر نے MATLAB ٹول کا استعمال کرتے ہوئے رسک ماڈل تیار کیا ہے۔ پیرامیٹرز کی ترتیبات کے بعد تجربات 10,000 لوپس تک چلائے گئے تھے۔ میٹلیب ٹول میں رسک تشخیص ماڈل 11 مرتبہ عمل میں لایا گیا ہے تاہم ہر رن نے مختلف results دیے ہے۔ تجرباتی نتائج سے یہ پتا چلا ہے کہ جب ماڈل کو 10 ویں بار سرانجام دیا گیا تھا تو تربیت اور جانچ کے عمل میں سب سے زیادہ تغیر پایا گیا تھا۔ محقق نے ڈیٹا سیٹس کی تربیت اور جانچ کے معیار کی تقسیم کو واضح کرنے کے لئے باکس پلاٹ کی نمائندگی کا استعمال کیا۔ دونوں ہی صورتوں میں output data میں less error آیا ہے جو یہ ظاہر کرتا ہے کہ یہ ایک مضبوط الگورتھم ہے۔ کیئے گئے تجربات نے موجودہ ٹولز کے اسی طرح کے نقطہ نظر کے مقابلے میں زیادہ سے زیادہ کارکردگی دکھائی۔

(Masethe Hlaudi and Masethe Mosima 2014)[85] محققین نے J48 ، نیوی بائیس ، Simple CART، REPTREE، اور بائیس نیٹ ڈیٹا مائننگ algorithms استعمال کرکے دل کے دورے کی پیش گوئی کرنے کے لئے ایک ماڈل تیار کیا ہے۔ ہارٹ ڈیزیز ماڈل کی تعمیر کے لئے استعمال ہونے والا ڈیٹاسیٹ جنوبی افریقہ میں صحت کی دیکھ بھال کرنے والے پیشہ ور افراد سے جمع کیا جاتا ہے جس میں 490 ریکارڈس اور 11 اوصاف ہیں۔ وہ امراض قلب کی پیش گوئی کے لئے WEKA کا استعمال کرتے ہیں۔ محققین

نے غیر جانبدارانہ نتائج کا تخمینہ لگانے کے لئے ڈیٹاسیٹ پر 10 fold cross validation کا اطلاق کیا۔ تجرباتی نتائج سے یہ پایا گیا ہے کہ جب مختلف classification algorithms کا اطلاق ہوتا تھا تو نتائج امراض قلب کی پیشن گوئی میں کوئی قابل ذکر تفاوت فراہم نہیں کرتے ہیں۔

(Ngueilbaye, Lei, and Wang 2016)[86] محققین نے 315 مثالوں کا چھوٹا ڈیٹاسیٹ مختلف اسپتالوں کے databases سے جمع کیا اور کارڈیک ڈس آرڈر کے مریضوں کی ابتدائی پیش گوئی کے لئے نیوی بائیس اور سپورٹ ویکٹر مشین الگورتھم استعمال کیا۔ اطلاق شدہ درجہ بندی کی کارکردگی کو جانچنے کے لئے محققین نے مختلف اقدامات جیسے احتمال اور درجہ بندی کی درستگی کا استعمال کیا۔ تجرباتی نتائج سے پتہ چلا ہے کہ نیوی بائیس الگورتھم نے ایس وی ایم ماڈل کو مات دیدی ہے۔

2.1.2۔ عن سوپروائیزڈ ٹاسکس کے ذریعے امراض قلب کی پیشن گوئی

Descriptive Data mining ٹاسک کا مقصد یہ ہے کہ ڈیٹا کے بنیادی relationships کا خلاصہ کرنے والے نمونوں (correlations, trends, clusters, trajectories and anomilies) کو اخذ کرنا ہے۔ Descriptive data mining tasks are often exploratory in nature اور نتائج کی توثیق کرنے اور ان کی وضاحت کے لئے post-processing تکنیک کی کثرت سے ضرورت ہوتی ہے۔

(Nguyen and Davis 2007) [87] محققین نے قلبی بیماری کی ابتدائی پیش گوئی کے لئے KMIX الگورتھم propose کیا ۔ ہارٹ ڈیزیز ڈیٹاسیٹ پری پروسیسنگ ٹکنیکس کے ذریعے صاف کیا گیا ہے۔ اس کے بعد صاف شدہ ڈیٹاسیٹ جس میں 341 ریکارڈس اور 19 اہم خصوصیات ہیں ماڈل کو تیار کرنے کے لئے استعمال کی جاتی ہیں۔ Linear Transformation کا استعمال کرتے ہوئے مستقل اہم عددی صفات کو [0،1] میں تبدیل کیا جاتا ہے اور بولین ڈیٹا کو متنی نمبر کی شکل میں تبدیل کر دیا جاتا ہے۔ WEKA ٹول کا استعمال کرتے ہوئے الگورتھم کی کارکردگی کو جانچنے کے لئے sensitivity اور specificity کا استعمال کیا گیا ہے۔ تجرباتی نتائج سے پتہ چلتا ہے کہ KMIX نے 0.25 کی sensitivity اور 0.89 کی specificity کے ساتھ K-Means الگورتھم کو مات دی۔ لہذا یہ دیکھا جاسکتا ہے کہ CVD کے مریضوں کی تشخیصی کے لئے KMIX کلسٹرنگ الگورتھم کی کارکردگی مناسب ہے۔

(Patil and Kumaraswamy 2009)[88] محققین نے K-Means کلسٹرنگ الگورتھم کا استعمال کرتے ہوئے کارڈیک ڈس آرڈر پیشن گوئی کے لئے ایک موثر نقطہ نظر تیار کیا۔ امراض قلب ماڈل Java میں ہارٹ ڈیزیز ڈیٹاسیٹ پر تیار کیا گیا ہے جو یو سی آئی مشین لرننگ ریپوزٹری سے حاصل کیا گیا ہے۔ محققین نے متواتر نمونوں کو حاصل کرنے کے لئے Maximal Frequent Itemset الگورتھم کا استعمال کیا جو امراض قلب کے لئے انتہائی موزوں ہیں۔ امراض قلب کے متواتر خطرے کے نمونوں کو اخذ کرنے کے بعد پیٹرن کو تفویض کردہ وزن کا حساب لگایا جاتا ہے اور ایک اہم وضاحتی حد سے زیادہ اہم وزن والے نمونوں کو امراض قلب کے مریضوں کی جلد شناخت کے لئے استعمال کیا جاتا ہے۔ میڈیکل ڈومین کے ماہرین امراض قلب کے خطرے کے اہم نمونوں کی تصدیق کرتے ہیں۔

(Shouman, Turner, and Stocker 2012a) [89] محققین نے کارڈیک ڈس آرڈر کے مریضوں کی ابتدائی پیش گوئی کے لئے نیوی بائیس کو بہتر بنانے میں K-Means کلسٹرنگ تکنیک کی کارکردگی کا مظاہرہ کیا۔ لگاتار صفات سے نمٹنے کے لئے نیوی بائیس الگورتھم کی انبلٹ محدودیت کی وجہ سے، مساوی تعدد صوابدیدی طریقہ کار ان کو مجرد افراد میں تبدیل کرنے کے لئے استعمال کیا جاتا ہے۔ غیر جانبدارانہ نتائج کو حاصل کرنے کے لئے محققین نے ڈیٹاسیٹ کو تربیت اور جانچ کے سیٹوں میں تقسیم کیا جس کا استعمال 10-fold cross validation ہے۔ محققین کلیو لینڈ ہارٹ ڈیزیز ڈیٹاسیٹ کا استعمال کرتے ہیں جو 297 ریکارڈس پر مشتمل ہے جس میں 13 طبّی خطرے کی صفت ہیں۔ ابتدائی سینٹروڈ سلیکشن کے مختلف طریقے جیسے حد، انیلر، آؤٹلیئر، رینڈم اٹریبیوٹ ویلیوز، اور بے ترتیب قطار کے طریقوں کا اطلاق امراض قلب کے مریضوں کے لئے کیا جاتا ہے۔ تجرباتی نتائج سے یہ ظاہر ہوتا ہے کہ مختلف ابتدائی سینٹروڈ سلیکشن کا استعمال کرتے ہوئے K میونز کلسٹرنگ کو نیوی بائیس کے ساتھ مربوط کرنے سے امراض قلب کے مریضوں کی پیش گوئی کرنے میں نیوی بائیس کی درستگی میں اضافہ ہوا- نتائج سے پتہ چلتا ہے کہ رینڈم انتساب اور رینڈم قطار طریقوں نے دو کلسٹروں والے انیلر، آؤٹلیر اور حد کے طریقوں سے زیادہ درستگی حاصل کی۔ حاصل کردہ بہترین درستگی دو کلسٹرز رینڈم رو ابتدائی سینٹروڈ کے انتخاب کے طریقہ کار کے ذریعہ ہے۔ تاہم رینڈم صفات اور رینڈم صف ابتدائی سینٹروڈ کے انتخاب کے طریقوں کے جھرمٹ کی تعداد میں اضافہ ہونے سے امراض قلب کے مریضوں کی تشخیصی میں ان کی درستگی میں اضافہ نہیں ہوا۔

2.1.3۔ ہائبرڈ ڈیٹا مائننگ ٹکنیکس کا استعمال کرتے ہوئے امراض قلب کی پیشن گوئی

ہائبرڈ سسٹم ایک ہی نظام کے ڈیزائن میں دو یا دو سے زیادہ طریق کار کا مجموعہ ہے۔ ہائبرڈ سسٹم تمام طریق کار میں سے بہترین ہیں اور بیماری کی پیش گوئی کے لئے ایک بہترین حل فراہم کرتے ہیں۔ مندرجہ ذیل محققین امراض قلب کی ابتدائی مرحلے میں پیش گوئی کرنے کیلئے ہائبرڈ ڈیٹامائننگ کی مختلف ٹکنیکس استعمال کرتے ہیں:

(Parthiban and Subramanian 2007) [90] محققین نے CoActive Neuro Fuzzy System (CANFIS) کا استعمال کرتے ہوئے کلیولینڈ ہارٹ ڈیزیز ڈیٹاسیٹ پر امراض قلب کے ایک پیش گوئی ماڈل تیار کیا۔ CANFIS Proposed ماڈل ایک ماڈیولر نیورل نیٹ ورک کے ساتھ موافقت پذیر فجی ان پٹ کو تیز اور درست انداز میں پیچیدہ افعال کے لئے مربوط کرتا ہے۔ کینفس ماڈل لرننگ کو بہتر بنانے کے لئے genetic algorithm کو ہر ان پٹ اور کنٹرول پیرامیٹرز کی اصلاح کے کے لئے ممبر فنکشن کی بہترین تعداد تلاش کرنے کے لئے استعمال کیا جاتا ہے۔ Genetic الگورتھم سلیکشن آپریٹر، کراس اوور آپریٹر اور mutation آپریٹر کو یکجا کر کے مسئلے کا بہترین حل تلاش کرتے ہیں جب تک کہ مخصوص معیار کو پورا نہیں کیا جاتا ہے۔ ڈیولپڈ رسک ماڈل کا استعمال کرتے ہوئے یہ محسوس کیا گیا کہ ماڈل کا Mean Square Error بہت کم ہے۔

(Polat, Sahan, and Gunes 2007) [91] ان محققین نے fuzzy resource allocation mechanism کے ساتھ artificial immune recognition system algorithm کا استعمال کرتے ہوئے کارڈیک ڈس آرڈر کی ابتدائی پیش گوئی کے لئے ایک ناول سسٹم تجویز کیا۔ محققین نے سب سے پہلے K NN پر مبنی وزن کے عمل کو کارڈیک ڈس آرڈر ڈیٹاسیٹ پر لگایا اور (0 اور 1) کی حد میں وزن کی پیمائش کی۔ ایک بار جب pre-processing مکمل ہو جاتا ہے تو Fuzzy-AIRS الگورتھم weight کارڈیک ڈس آرڈر ڈیٹاسیٹ پر لگایا جاتا ہے۔ محققین یو سی آئی مشین لرننگ ڈیٹا بیس سے کارڈیک ڈس آرڈر ڈیٹاسیٹ حاصل کرتے ہیں جس میں 13 صفات اور 270 ریکارڈس ہیں۔ سسٹم اپلی کیشنز میں کارڈیک ڈس آرڈر ڈیٹاسیٹ کو k کی مختلف ویلیوز جیسا کہ 10، 15 اور 20 کے لئے درجہ بند کیا گیا ہے جو K-NN قبل از پروسیسنگ مرحلہ کے طور پر استعمال ہوتا ہے۔ ہر بار درجہ بندی کے لئے دوسرے پیرامیٹرز میں کوئی تبدیلی نہیں کی گئی but except to k value۔ درجہ بندی کی اعلی ترین درستگی اس وقت پہنچ جاتی ہے جب k کی قیمت 15 ہو۔ مجوزہ نظام کی درجہ بندی کی درستگی کا نتیجہ 87% ہے اور یہ اس مسئلے کے لئے ادب میں دیگر درجہ بندی کی

درخواستوں کے سلسلے میں بہت امید افزا ہے۔ نتائج مشورہ دیتے ہیں کہ کے weighted K NN preprocessing اور fuzzy resource mechanism with AIRS کارڈیک ڈس آرڈر arrhythmia میں مدد کر سکتا ہے۔

(Tsipouras et al. 2008) [92] ان محققین نے فوزی رول پر مبنی ماڈل تیار کر کے CAD کی پیش گوئی کی ہے۔ محققین نے ہارٹ ڈیزیز ڈیٹاسیٹ استعمال کیا جس میں آبادی، تاریخی اور لیبارٹری ڈیٹا کے 199 ریکارڈز اور 19 صفات ہیں۔ غیر جانبدارانہ نتائج حاصل کرنے کے لئے بے ترتیب اسٹریٹیفائیڈ 10-fold cross validation کا اطلاق اس ڈیٹاسیٹ پر کیا ہے۔ ڈیولپڈ ماڈل کی کارکردگی کو مختلف اقدامات سے جانچا گیا ہے۔ تجرباتی نتائج ڈسیشن ٹری کے اصولوں پر 62% اور 54% کی specificity اور sensitivity کو ظاہر کرتے ہیں۔ جب افزائش اور اصلاح کے مرحلے استعمال کیے جاتے ہیں تو اوسط حساسیت اور خاصیت بالترتیب 80% اور 65% تک بڑھ جاتی ہے۔

(Tu, Shin, and Shin 2009) [93] ان محققین نے Bagging with Naive Bayes, C4.5 اور C4.5 with classification الگورتھمس کا استعمال کرتے ہوئے ایک predictive کارڈیک ڈس آرڈر رسک ماڈل تیار کیا۔ محققین امراض قلب کے مریضوں سے براہ راست جمع کردہ ڈیٹاسیٹ استعمال کیے ہیں۔ بیگنگ الگورتھم ایک دیئے گئے ٹریننگ سیٹ کا استعمال کرتے ہوئے عمل کی تقلید کرتے ہوئے لرننگ تکنیک کی عدم استحکام کو بے اثر کرنے کی کوشش کرتا ہے۔ ہر بار نیا ٹریننگ ڈیٹاسیٹ نمونے لینے کے بجائے اصل ٹریننگ ڈیٹا میں کچھ مثالوں کو حذف کر کے اور دوسروں کو نقل کر کے ترمیم کی جاتی ہے۔ محققین نے WEKA Tool پر تین مختلف تجربات کیے۔ تجربہ 1 نے ڈسیشن ٹری الگورتھم کا استعمال کیا، تجربہ 2 میں Bagging with Decision Tree with reduce error pruning operation کا استعمال کیا گیا اور تجربے 3 نے Bagging with Naive Bayes الگورتھم کا استعمال کیا۔ ہر تجربے کے لئے 10-fold cross validation ٹریننگ اور ٹسٹنگ ڈیٹا کے بے ترتیب نمونے لینے سے تیار کردہ bias کو کم کرنے کے لئے استعمال کیا گیا ہے۔ تجرباتی نتائج سے پتہ چلتا ہے کہ Naive Bayes نے بہترین نتائج دکھا ے-

(Das, Turkoglu, and Sengur 2009) [94] ان محققین نے جلد از جلد امراض قلب کی پیشن گوئی کرنے کے لئے ایک موثر Neural Network Ensemble ماڈل تیار کیا۔ محققین نے پیشگی احتمال کے متعدد ماڈلوں سے posterior probability کو ملا کر نئے ماڈل بنانے کے لئے جوڑنے والے جزو کا استعمال کیا۔ اس کے بعد یہ نیا تیار کیا ہوا رسک ماڈل unseen ڈیٹا کو اسکور کرنے کے لئے استعمال ہو ہے۔ محققین امراض قلب کی ابتدائی پیش گوئی کے لئے نیورل نیٹ ورک کو جوڑنے پر مبنی طریقہ کار تخلیق کرنے کے لئے SAS Enterprise Miner 5.2 کا استعمال کیے ہیں۔ کلیولینڈ کارڈیک ڈس آرڈر ڈیٹا بیس جس میں 297 ریکارڈس اور 14 اوصاف ہیں 89.01 % درجہ بندی کی درستگی تجربات سے حاصل کی گئی ہے۔ محققین ان پٹ متغیر کی حالت کو مسترد کرتے ہوئے تشکیل کر کے ان پٹ کی مقدار کو کم کرنے کے لئے متغیر انتخاب کے جزو کا استعمال کرتے ہیں۔

(Anbarasi, Anupriya, and Iyengar 2010) [95] ان محققین نے امراض قلب کے کم وجود میں صفات کی موجودگی کے بارے میں صحیح طور پر پیش گوئی کرنے کے لئے ایک ماڈل تیار کیا۔ Genetic Algorithm کو ان خصوصیات کا تعین کرنے کے لئے شامل کیا گیا ہے جو امراض قلب کی تشخیصی اور علاج میں زیادہ اہم کردار ادا کرتے ہیں۔ دل کے عارضہ کے مریضوں کی تشخیصی کے لئے تین درجہ بندیوں جیسے Naive Bayes, Classification by Clustering and Decision Tree کا استعمال کیا گیا ہے۔ مشاہدات کی نمائش ہے کہ ڈسیشن ٹری ڈیٹا میننگ الگورتھم اعلی ماڈل کی تعمیر کے وقت کے ساتھ فیچر سبسیٹ سلیکشن کو شامل کرنے کے بعد دیگر ڈیٹا میننگ techniques کو پیچھے چھوڑ دیتی ہے۔ WEKA ٹول کے ساتھ 909 ریکارڈز کے ڈیٹاسیٹ پر تجربات کیے گئے تھے۔ صفات کے ایک زیادہ سے زیادہ مجموعہ کے لئے genetic search صفر صفات، ایک ابتدائی آبادی اور تصادفی طور پر پیدا کردہ قواعد سے شروع ہوتی ہے۔ New generation اس وقت تک جاری رہتی ہے جب تک کہ اس آبادی کا ارتقا نہ ہو جہاں ہر اصول آبادی سے مطمئن ہوتا ہو۔ تمام صفات کو الگ الگ بنایا گیا ہے اور اس کی تضادات سادگی کے لئے حل کر دیئے گئے ہیں۔ cross over 0.6 probability اور mutation corss over probability 0.033 کے ساتھ generic search کے نتیجے میں 6 اوصاف پیدا ہوئے جو کارڈیک بیماری کی تشخیصی میں زیادہ اہم کردار ادا کرتے ہیں۔

(Adeli and Neshat 2010) [96] ان ریسرچرز نے بینچ مارک کلیولینڈ ہارٹ ڈیزیز ڈیٹاسیٹ پر ایک fuzzy ماڈل تیار کیا جو 303 ریکارڈز اور 12 خصوصیات پر مشتمل ہے۔ تمام 11 ان پٹ متغیرات اور 1 آؤٹ پٹ متغیر کی رکنیت کا کام انفرنس میکانزم کا استعمال کرتے ہوئے تیار کیا گیا ہے۔ محققین Mamdane Fuzzification approach استعمال کرتے ہیں

اور Defuzzification کے عمل کے لئے سینٹروڈ طریقہ شامل کیا گیا تھا۔ فجی انفینس سسٹم میں نتائج کا معیار فجی قوانین پر منحصر ہوتا ہے۔ مجوزہ نظام نے 44 قواعد پیدا کیے اور یہ دوسرے اصولوں کے نتائج کے مقابلے میں بہترین ہے۔ ہر قاعدہ کے لئے موزوں ڈگری تیار کی جاتی ہے اور قواعد کو اکٹھا کرنے کے لئے زیادہ سے زیادہ درستگی ڈگری کا حساب K=(1, 2, ... 44) سے کیا جاتا ہے۔ تشخیصی سے نمٹنے والا مبہم ماہر نظام نافذ کیا گیا ہے اور تجرباتی نتائج سے معلوم ہوا ہے کہ اس نظام نے غیر ماہروں کے مقابلے میں کافی بہتر کارکردگی کا مظاہرہ کیا۔

(Aqueel and Hannan 2012) [97] ان ریسرچرز نے سپورٹ ویکٹر مشین، جینیٹک الگورتھم، سیٹ تھیوری، ایسوسی ایشن رولز اور نیورل نیٹ ورک الگورتھم کا استعمال کرتے ہوئے ہارٹ ڈیزیز رسک ماڈل بنائیں۔ محققین نے ڈیٹا سیٹ پر متعدد تجربات کیے جن میں 909 ریکارڈ اور 13 اوصاف شامل ہیں۔ ان الگورتھم کی کارکردگی کو جانچنے کے لئے محققین نے WEKA ٹول پر تجربات کیے۔ تجرباتی نتائج سے پتہ چلتا ہے کہ فیچر سلیکشن تراکیب کا اطلاق کرنے کے بعد developed decision tree کے رسک ماڈل نے دیگر تکنیکوں کے مقابلے میں اعلی پیش گوئی کرنے کی صلاحیت ظاہر کی تاہم رسک ماڈل میں وقت کی پیچیدگی بڑھ جاتی ہے۔

(Alizadehsani et al. 2013) [98] ان ریسرچرز نے Coronary Artery Disease کی پیش گوئی کرنے کے لئے classification ڈیٹا مائننگ الگورتھم استعمال کیا ہے۔ موجودہ مطالعے میں لیبارٹری اور ای سی جی کے ڈیٹا کی تحقیقات کے لئے C4.5 درجہ بندی اور بیگنگ درجہ بندی کرنے والوں Left Circumflex, Left Anterior Descending and Right Coronary Artery کی الگ الگ شناخت کی جاسکتی ہے۔ ان محققین نے 303 بے ترتیب زائرین Rajaie Cardiovascular, Medical and Research Center, Tehran, Iran. سے ڈیٹا جمع کیا۔ ایل ای ڈی اسٹینوسس کی پیش گوئی کرنے کی درستگی کو فیچر سلیکشن کے ذریعے حاصل کیا گیا۔ اس تحقیق میں ریپڈ مائنر ٹول کی ڈیفالٹ سیٹنگ کا استعمال کیا گیا تھا اور الگورتھم کی درستگی، حساسیت اور مخصوصیت حاصل کی گئی تھی۔ انتہائی اہم خصوصیات کو منتخب کرنے کے لئے گنی انڈیکس اور معلوماتی حصول استعمال کیا گیا۔ مزید برآں معلومات کے حصول کی بنیاد پر منتخب کردہ خصوصیات کے استعمال نے ایل اے ڈی کی علامت تشخیص کی درستگی کو بڑھا کر 79.54 فیصد کردیا۔ نتائج سے پتا چلتا ہے کہ تمام شریانوں کی stenosis پر 10 انتہائی موثر خصوصیات میں ejection fraction, age, lymph اور ایچ ٹی این شامل تھے۔

(Jabbar, Deekshatulu, and Chandra 2015) [99] نے کارڈیک ڈس آرڈر متاثرین کی پیش گوئی کے لئے K Nearest Neighbor and Genetic classifier ملا کر ایک نیا نقطہ نظر پیش کیا۔ Genetic search is applied as a goodness measure to prune redundant and irrelevant attributes جو درجہ بندی کی طرف زیادہ شراکت کرتے ہیں۔ کم درجہ بندی کی خصوصیات کو خارج کر دیا گیا ہے اور درجہ بند خصوصیات طبقاتی خصوصیات کی بنیاد پر ڈیزائن کیا گیا ہے۔ تیار کردہ طریقہ کار کی کارکردگی کی تصدیق 6 میڈیکل اور 1 نان میڈیکل ڈیٹاسیٹ سے کی جاتی ہے۔ امراض قلب کے ان سات ڈیٹاسیٹوں میں سے، ایک ڈیٹاسیٹ آندھرا پردیش، انڈیا کے مختلف اسپتالوں سے جمع کیا جاتا ہے اور بقیہ ڈیٹاسیٹ یو سی آئی مشین لرننگ ریپوزٹری سے حاصل کیے جاتے ہیں۔ Experimental results show that the classifiers increase the efficiency of the model.

(Amin, Agarwal, and Beg 2013)[100] ان محققین نے Neural Network and Genetic algorithm کا استعمال کرتے ہوئے نمایاں خطرے کی خصوصیات پر مبنی اپنی ابتدائی پیش گوئی کے لئے ہائبرڈ ہارٹ ڈیزیز تشخیصی ماڈل تیار کیا۔ امریکن ہارٹ ایسوسی ایشن کے ذریعہ جمع کردہ سروے کے اعداد و شمار کا استعمال کیا گیا ہے جو خطرے کے 12 اہم عوامل اور 50 واقعات پر مشتمل ہے۔ ڈیٹا کی Pre-Processing کے بعد، Neural Network Weights were initlized with MATLAB Tool configure function in MATLAB. میں تیار کیا ہوا ensembled ماڈل نے 12 epochs کے بعد Minimum Mean Square Error 0.034683 کے ساتھ ساتھ توثیق ڈیٹاسیٹ پر ایک زیادہ سے زیادہ درستگی حاصل کر لی۔

(Chaurasia and Pal 2014) [101] نے زیادہ سے زیادہ درستگی کے ساتھ دل کی بیماری کی پیش گوئی کرنے کے لئے نیوی بیئس، Decision Tree اور بیگنگ الگورتھم استعمال کرتے ہوئے ہنگری کے ڈیٹاسیٹ پر ہارٹ ڈیزیز ماڈل تیار کیا۔ پیش گوئی کرنے والے ماڈلز کے غیر جانبدارانہ تخمینے کی پیمائش کرنے کے لئے محققین نے 10-fold cross validation کا استعمال کیا۔ WEKA ٹول پر ہارٹ ڈیزیز ماڈل کی تربیت اور جانچ کی جاتی ہے۔ تجرباتی نتائج سے پتہ چلتا ہے کہ بیگنگ الگورتھم Naive Bayes اور J48 الگورتھم سے بہتر کارکردگی کا مظاہرہ کرتا ہے جس کی اعلی درستگی 85.03% ہے۔

(Srinivas, Rao, and Govardhan 2014)[102] ان محققین نے امراض قلب کی تشخیصی کے لئے فجی سیٹ کے ساتھ rough set theory کو ملا کر ایک نئی درجہ بندی کرنے کی تجویز پیش کی۔ Fuzzy base rules کسی نہ کسی طرح سیٹ تھیوری کا استعمال کرتے ہوئے بنائے جاتے ہیں اور یہ پیش گوئی فجی درجہ بند کرنے والے کے ذریعہ کی جاتی ہے۔ درست فجی قوانین حاصل کرنے کے لئے بنیادی تجزیہ غیر ضروری سیٹ میٹرکس کی تشکیل کے بعد کسی نہ کسی حد تک نظریہ سے متعلقہ صفات کی نشاندہی کرنے کے لئے کیا گیا ہے۔ پھر فجی سسٹم کو فجی قواعد اور ممبر شپ کے افعال کی مدد سے ڈیزائن کیا گیا ہے تاکہ ڈیزائن کی گئی اس فجی سسٹم میں پیش گوئی کی جاسکے۔ مجوزہ نظام MATLAB 7.11 کا استعمال کرتے ہوئے نافذ کیا گیا ہے اور ڈیٹا کو system میں داخل کرکے امراض قلب کی موجودگی کی نشاندہی کی جاتی ہے۔ درجہ بندی کرنے والے تجربات تین بڑے پیمانے پر لگائے گئے ڈیٹا سیٹس یعنی کلیولینڈ، ہنگری اور سوئٹزرلینڈ کے ساتھ جو UCI مشین لرننگ ریپوزٹری سائٹ سے ڈاؤن لوڈ ہوتے ہیں۔ نتائج سے محققین اس بات کو یقینی بناتے ہیں کہ developed rough set fuzzy کلاسفائیر نے سوئٹزرلینڈ امراض قلب ڈیٹاسیٹ پر 80% اور ہنگری کے امراض قلب کے ڈیٹاسیٹ پر 42% کی درستگی حاصل کرکے گذشتہ طریقوں کو بہتر بنایا۔

(Dewan and Sharma 2015) [103] ان محققین نے Back Propogation کے ذریعہ Genetic algorithm کا استعمال کرکے ایک قابل ensembled ہارٹ ڈیزیز ماڈل تیار کیا۔ تیار کردہ ماڈل کو ڈیٹاسیٹ پر عمل میں لایا جاتا ہے جس میں 303 ریکارڈ شامل ہیں البتہ جانچ کے مقصد کے لئے صرف 270 استعمال ہوا ہے۔ پری پروسیسنگ میں WEKA tool کی سب سے عام تکنیک یعنی missing value filter کو تبدیل کیا گیا۔ Local minima میں پھنس جانے کی خرابی کو حل کرنے کے لئے بہترین optimizers یعنی genetic algorithm جو مختلف نسلوں سے زیادہ تغیر اور کراس اوور کے مظاہر کو استعمال کرتا ہے وہ ماڈل میں سرایت کرتا ہے۔ وزن کی واپسی کے لئے استعمال ہونے والے وزن کو پہلے بہتر بنایا جاتا ہے اور پھر بہتر نتائج حاصل کرنے کے لئے نیٹ ورک کو ان پٹ کے بطور دیا جاتا ہے۔ تجرباتی نتائج سے پتہ چلتا ہے کہ non linear data کی پیش گوئی کرنے یا درجہ بندی کرنے کے لئے classification کی تمام ٹکنیکس میں نیورل نیٹ ورک بہترین ہے۔

(Sumana and Santhanam 2015) [104] ان ریسرچرز نے امراض قلب کی ابتدائی تشخیصی کے لئے ہائبرڈ ماڈل تجویز کیا۔ ڈیٹا صاف کرنے کے بعد امراض قلب سے متعلقہ خصوصیات کو حاصل کرنے کے لئے سب سے پہلے best first search اور

feature selection techniques کو cascading انداز میں شامل کیا گیا۔ Testing ڈیٹاسیٹ K Means الگورتھم کا استعمال کرتے ہوئے کلسٹر کیا جاتا ہے اور صحیح طور پر کلسٹرڈ نمونے 12 مخصوص درجہ بندی کے ساتھ تربیت دیئے جاتے ہیں تاکہ تیار شدہ 10 fold cross validation کا استعمال کرتے ہوئے حتمی ماڈل تیار کیا جاسکے۔ ڈیولپڈ ماڈل کی درجہ بندی کی درستگی اور وقت کی پیچیدگی کی جانچ کرنے کے لئے UCI مشین لرننگ ریپوزٹری سے جمع کردہ 5 دیگر بائنری کلاس میڈیکل ڈیٹا سیٹس پر WEKA ٹول کا استعمال کرتے ہوئے اندازہ کیا جاتا ہے۔ مشاہدہ شدہ نتائج یہ ظاہر کرتے ہیں کہ ڈیٹا سیٹس اور الگورتھم سے قطع نظر ڈولوپڈ ہارٹ ڈیزیز رسک ماڈل نے تمام 12 طبقاتی طبقوں کے ساتھ پانچ مختلف طبّی ڈیٹا سیٹس پر درستگی کو بڑھایا۔

(Beena, Rajinikanth, and Viswanadha 2016) [105] ان ریسرچرز نے کارڈیک ڈس آرڈر کی تشخیصی کے لئے پیش گوئی کی درستگی اور فیصلہ سازی کو بڑھانے کے لئے کمپیوٹرائزڈ فیچر سلیکشن طریقوں اور میڈیکل خصوصیات کے امتزاج سے امراض قلب کی نمایاں خصوصیات کا انتخاب کیا۔ کلیولینڈ ہارٹ ڈیزیز ڈیٹاسیٹ کے پہلے سے طے شدہ ملٹی کلاس درجہ بندی کے وضع کو بائنری درجہ بندی کی شکل میں تبدیل کر دیا گیا ہے۔ محققین نے Matlab tool کا استعمال کرتے ہوئے رسک ماڈل تیار کرنے کے لئے feature selection الگورتھم کا استعمال کیا۔ تجرباتی نتائج سے یہ پایا گیا ہے کہ خصوصیت کے انتخاب کے طریقہ کار کی درستگی مجرد خصوصیات کو کنٹرول کر کے بڑھتی ہے تاہم ماڈل کی پیچیدگی میں اضافہ ہوتا ہے۔

(Bialy et al. 2016) [106] ان محققین نے نیوی بایس، بائیسین نیٹ، ملٹی لیئر پرسیپٹرن (ایم ایل پی)، sequential minimal optimization, C4.5 and Decision Tree کا استعمال کرتے ہوئے امراض قلب کے خطرے کی پیش گوئی کا ماڈل پیش کیا۔ یہ system ہر ایک algorithm کے پیش گوئی کے نتائج کو weight کے ذریعے ensemble model میں جوڑتا ہے۔ اس system میں دو مختلف بیماریوں کے ڈیٹا سیٹس ہے جیسے 920 records پر مشتمل CAD ڈیٹاسیٹ جس میں 14 اوصاف ہیں اور 103 records پر مشتمل Heart Disease Valve ڈیٹاسیٹ پر مشتمل ہے The outliers and extreme values were detected and removed by using the inter-quartile range technique. - تجرباتی نتائج سے پتہ چلتا ہے کہ کسی ایک درجہ بندی کے لئے اعلی ترین درستگی کا نتیجہ دونوں ڈیٹا سیٹس پر نیوی بایس الگورتھم

کے ذریعہ حاصل کیاجاتا ہے۔ حاصل کردہ نتائج کو WEKA Tool کا استعمال کرتے ہوئے تجزیہ کیا گیا اور درجہ بندی کی کارکردگی کو 10-fold cross validation کے ذریعہ ماپا گیا۔

(Arabasadi et al. 2017) [107] ان محققین نے ناگوار طریقوں کی ضرورت کے بغیر کلینکل ڈیٹا پر مبنی امراض قلب کی تشخیصی کے لئے ہائبرڈ ماڈل تجویز کیا۔ محققین نیٹ ورک کی تربیت کرنے اور وزن میں ترمیم کرنے کے لئے جینی انڈیکس، ایس وی ایم کے ذریعہ وزن ، انفارمیشن گیین اور Pricnipal Component Analysis فیچر سلیکشن تراکیب کا استعمال کیا ہیں تاکہ minimum error کو حاصل کیا جاسکے۔ وہ مصنوعی نیورل نیٹ ورک میں back propogation error الگورتھم کا استعمال کیا ہیں جس میں ایم ایل پی ڈھانچہ اور sigmoid exponential فنکشن ہوتا ہے جس سے وہ امراض قلب کے ماڈل کو تشکیل دے سکتے ہیں۔ ڈیولپڈ رسک ماڈل جینیٹک الگورتھم کا استعمال کرتے ہوئے ابتدائی وزن میں اضافہ کرکے نیورل نیٹ ورک کی کارکردگی کو بڑھاتا ہے۔ ماڈل Z-Alizadeh Sani ڈیٹاسیٹ پر ایک زیادہ سے زیادہ درستگی، حساسیت اور خصوصیت حاصل کرتا ہے جو موجودہ نظاموں کے مقابلے میں زیادہ ہے۔

2.2۔ ریسرچ گیپس

امراض قلب کی تشخیص کئی طریقوں کے ذریعے کی جاسکتی ہے تاہم انتہائی سستے اور قابل اعتماد طریقے کار غیر ناگوار خطرے کی صفات کی تشخیص پر مبنی ہے۔ مختلف محققین نے ڈیٹا مائننگ کی مختلف تکنیکوں کا استعمال کرتے ہوئے خطرے کے عوامل پر امراض قلب کی پیش گوئی کی ہے لیکن ادب کو قریب سے دیکھنے سے متعدد کوتاہیوں کا انکشاف ہوتا ہے جن کو ذیل میں بیان کیا گیا ہے:

i. امراض قلب کے زیادہ سے زیادہ ڈیولپڈ رسک ماڈل میں عمومی صلاحیت کی کمی ہوتی ہے۔

ii. امراض قلب سے متعلق ڈیٹا سے اخذ کردہ رسک اصول پیچیدہ اور فطرت میں بڑے ہوتے ہیں جو نظام کو سست بناتے ہیں اور غلط فیصلوں کا باعث بنتے ہیں۔

iii. مختلف ٹولز (WEKA, RappidMiner, Orange and KNIME) تجربات اور تعلیمی مقاصد کے لئے استعمال کیے جاتے ہیں البتہ ہر ٹول کے ساتھ پیچیدگیاں ہوتی ہیں۔ پیشن گوئی کے لئے بہت سارے میکانزم موجود ہیں لیکن سب کی محدودیتیں ہیں جیسے جی یو آئی کے لئے دستاویزات محدود ہیں، اسکیلنگ ایک مسئلہ ہے، بگ ڈیٹا کو سنبھالا نہیں جاسکتا وغیرہ۔

iv. طبّی ڈومین کی کارکردگی کے اقدامات جیسے حساسیت، صراحت، درستگی، صحت سے متعلق وغیرہ کا استعمال کیا جاتا ہے تاہم محققین کے ذریعہ کمپیوٹیشنل پیچیدگی اور فہمیت جیسے ماڈل اقدامات استعمال نہیں کیے جاتے ہیں۔

v. عددی طبّی متغیر پر ڈسیشن ٹری الگورتھم کی خودکار طریقے سے الگ ہونے کی حالت طبّی پیشہ ورافراد کے لئے غلط تشخیص کا باعث بنتی ہے۔ میڈیکل سوسائٹی میں معیاری تقسیم کے معیارات ہیں جو عالمی سطح پر تسلیم کیے جاتے ہیں (ہائی بلڈ پریشر، ہائی کولیسٹرول وغیرہ) لہذا ڈسیشن ٹری الگورتھم کو صحیح پیش گوئیاں حاصل کرنے اور غلط مفروضوں سے روکنے کے لئے میڈیکل ڈیٹاسیٹ پر اطلاق کرنے سے پہلے اس کو کٹ آف اقدار پر تربیت دی جانی چاہئے۔

vi. ڈیولپڈ رسک ٹولز امراض قلب کے خطرے سے متاثرہ افراد کی درجہ بندی کرنے میں مدد کرتے ہیں تاہم ان کی کارکردگی واضح طور پر معلوم نہیں ہے۔

vii. بیشتر محققین نے اپنی شراکت میں کلینیکل اوصاف کا استعمال کیا جس سے طبّی ترتیبات کے علاوہ ان کو کم استعمال کیا جاتا ہے۔ تاہم کوئی بھی ہارٹ ڈیزیز ڈیولپڈ رسک ماڈل غیر ناگوار خطرے کی خصوصیات پر مبنی نہیں ہے۔

viii. زیادہ تر محققین اہم خصوصیات کو حاصل کرنے کے لئے صرف ایک خصوصیت کے انتخاب کی تکنیک کا استعمال کرتے ہیں تاہم ابتدائی امراض قلب کے خطرے کی تشخیصی کے لئے ان کی اہم اقدار کے ساتھ نمایاں غیر ناگوار صفات حاصل کرنے کے لئے متعدد خصوصیت کے انتخاب کی ٹکنیکس کے استعمال کی کوئی تحقیقات نہیں ہیں۔

ان تحقیقی حدود پر قابو پانے کے لئے ایک غیر مؤثر خطرہ کی خصوصیت کا استعمال کرتے ہوئے کم لاگت سے امراض قلب کے خطرے سے متعلق تشخیصی ماڈل تیار کیا گیا ہے۔ اس کے بعد کے ابواب میں رسک ماڈل کے ترقیاتی طریقہ کار پر تبادلہ خیال کیا گیا ہے۔

2.3۔ باب کا خلاصہ

اس باب میں امراض قلب کی پیش گوئی اور ڈیٹا مائننگ کی مختلف تکنیکوں کا استعمال کرتے ہوئے تشخیصی کا تفصیلی جائزہ پیش کیا گیا ہے۔ محققین نے دنیا کو اس مہلک بیماری کی طرف کھینچنے والے عوامل کا پتہ لگانے کے لئے ڈیٹا مائننگ کے طریقوں کا استعمال کرتے ہوئے کارڈیک ڈس آرڈر کی شناخت میں فیصلہ کن شراکت کی۔ انہوں نے پایا کہ طرز عمل کے خطرے کے عوامل دل کی بیماریوں کی بنیادی وجوہات ہیں۔ بہت سے محققین

ڈائیورجنٹ ڈیٹاسیٹ، مختلف مشین لرننگ الگورتھم، مختلف ڈیٹامائننگ کے طریقہ کار اور متعدد ٹولز کا استعمال کرتے ہوئے رسک ماڈل تیار کرتے ہیں۔ محققین نے محسوس کیا کہ یہاں ایک بھی الگورتھم موجود نہیں ہے جو ہر ڈیٹاسیٹ کے لئے بہترین نتائج پیدا کرتا ہے تاہم ہائبرڈائزیشن اور ensemble methods زیادہ سے زیادہ نتائج دکھاتے ہیں۔ محققین نے مختلف ٹولوں کے استعمال سے تجربات اور similation مقاصد کے لئے cross validation اور error rates کا استعمال کیا لیکن ہر ٹول میں پیچیدگیاں ہیں۔ ہیلتھ کیئر کے نظام کو بہتر بنانے اور موثر ماڈل بنانے کے لئے ہمیں درست تشخیص کو پیشگی بہتر طریقے سے حاصل کرنے کے لئے ریئل ٹائم ڈیٹا سیٹس پر مطالبہ کرنے والے تکنیک کا مطالبہ کرنے کے لئے انتہائی مناسب اور ناول ڈیٹا استعمال کرنے کی ضرورت ہے۔ جلد سے جلد ڈیٹا مائننگ کی تراکیب کے ذریعے دل کی بیماری کی پیش گوئی اور اس کا پتہ لگانے کے لئے اس کے بارے میں نئی دریافتوں کو سمجھنے کے لئے اضافی تحقیق کا مطالبہ کرتا ہے۔

باب 3

امراض قلب کی تشخیص میں ڈیٹا مائننگ ٹکنیکس کا اطلاق

میڈیکل انڈسٹری databases نامکمل اور raw ڈیٹا سے مغلوب ہیں اور ان بڑے ڈیٹا اسٹورس سے کسی واضح ڈھانچے میں چھپی ہوئی معلومات کو نکالنے کے لئے ڈیٹا مائننگ ٹکنیکس کا اطلاق ہوتا ہے۔ صحت کی دیکھ بھال میں ڈیٹا مائننگ ٹکنیکس کو متعارف کروانے کی وجہ ماہرین کو برطرف کرنا نہیں ہے بلکہ جہاں وہ مصائب سے دوچار ہوتے ہو وہاں ان کی مدد کرنا ہے۔ اس باب میں فیچر سلیکشن ٹکنیکس کی وضاحت کی گئی ہے جو امراض قلب کی ابتدائی پیش گوئی کے لئے خطرے کی صفات کے اہم غیر حملہ آور سبسیٹ کو تلاش کرنے کے لئے استعمال ہوتے ہیں۔ اس باب میں ڈیٹا مائننگ تکنیکوں پر تبادلہ خیال کیا گیا ہے جو ہارٹ ڈیزیز رسک ماڈل کو تیار کرنے کے لئے استعمال ہوتے ہیں۔ اس باب میں رسک ماڈل کی کارکردگی کا اندازہ مختلف اقدامات کے ذریعے کیا جاتا ہے اور آخر میں کارڈیک مریضوں کی ابتدائی شناخت اور علاج کے لئے نان انویسو ہارٹ ڈیزیز رسک فیچرز کی اہمیت بھی پیش کی گئی ہے۔

3.1 فیچر سلیکشن ٹکنیکس

فیچر سلیکشن ٹکنیکس کا استعمال فیچرز کے ذیلی ذخیرے کو دریافت کرنے کے لئے کیا جاتا ہے جو صحیح اور کومپیکٹ پیش گوئی ماڈل تیار کرتی ہیں [134]۔ اس تحقیقی کام میں فلٹر، ریپر اور ایمبیڈڈ feature selection methods کا امتزاج استعمال کیا گیا ہے تاکہ ڈیولپڈ ماڈل کی پیچیدگی کو کم کیا جاسکے اور اور امراض قلب کی پیشن گوئی کے لئے رسک فیچرز کا سب سے بڑا غیر منحرف سبسیٹ حاصل کیا جاسکے۔

3.1.1۔ اکسٹرا ٹری کلاسیفائر

اکسٹرا ٹری کلاسیفائر کو extremely randomized trees کے طور پر بھی جانا جاتا ہے جو بغیر کسی متبادل کے متعدد trees بناتا ہے۔ اکسٹرا ٹری کلاسیفائر کے نوڈس رینڈم splitting کی بنیاد پر تقسیم ہوتے ہیں جس کی وجہ سے درستگی میں اضافہ ہوتا ہے اور سٹینڈرڈ ٹریس اور رینڈم فوریسٹس میں زیادہ سے زیادہ کٹ پوائنٹس کے عزم سے منسلک کمپیوٹیشنل load میں بڑے پیمانے پر کمی واقع ہوتی ہے [108]-

3.1.2۔ گریڈینٹ بوسٹنگ کلاسیفائر

گریڈینٹ بوسٹنگ کلاسیفائر کا استعمال classification اور Regression دونوں کے لئے کیا جاتا ہے۔ اس میں loss function شامل ہوتا ہیں جو ڈیسیشن ٹریز کا استعمال کرتے ہوئے اپنی مرضی کے مطابق greedy انداز میں ڈیسیشن ٹریز بنا دیتا ہے اور آخر میں ان ٹریز کو ایک وقت میں ایک ساتھ شامل کیا جاتا ہے تاکہ نقصان کی تقریب کو کم سے کم کیا جا سکے [109]۔

3.1.3۔ رینڈم فوریسٹ

رینڈم فارسٹ ڈیسیشن ٹریز کی پیشن گوئوں سے بنائے گئے ہیں جن کو classification اور regression کے کاموں کے لئے استعمال کیا جاتا ہے۔ رینڈم فوریسٹس تصادفی طور پر منتخب کردہ تربیتی سیٹ سے متعدد ڈیسیشن ٹریز کا استعمال کرتے ہوئے تخلیق کیے گے ہے تاکہ انفرادی ڈیسیشن ٹریز کی حد سے تجاوز کرنے والے مسئلے کو عبور کر سکیں [110]۔ رینڈم فارسٹ کلاسیفائر کو ڈیٹا مائننگ تکنیکس سیکشن میں مکمل طور پر سمجھایا جائے گا۔

3.1.4۔ ریکرسیو فیچر ایلمنیشن

ریکرسیو فیچر ایلمنیشن ایک greedy optimization طریقہ ہے جو بہترین کارکردگی کا مظاہرہ کرنے والی خصوصیت کا سبسیٹ حاصل کرنے کی کوشش کرتا ہے۔ یہ بار بار ماڈل بناتا ہے اور ہر اعادہ پر بہترین یا کم ترین کارکردگی پیش کرنے والی فیچرز کو الگ رکھتا ہے۔ یہ باقی فیچرز کا استعمال کرتے ہوئے اگلے ماڈل کی تشکیل کرتی ہے جب تک کہ تمام خصوصیات ختم نہ ہو جائیں اور پھر ان کے خاتمے کے درجے کی بنیاد پر خصوصیات کو پوزیشن میں رکھتا ہے [111] ۔

3.1.5۔ ایکس.جی بوسٹ کلاسیفائر

ایکسٹریم گریڈینٹ بوسٹنگ ایک ensemble الگورتھم ہے جو متوازی پروسیسنگ، tree pruning، گمشدہ اقدار کو سنبھالنے اور باقاعدگی کے ذریعہ overfitting اور bias سے بچنے کے لئے ایک اصلاحی میلان کو فروغ دینے والے الگورتھم کا استعمال کرتا ہے۔ ایکس.جی بوسٹ کی درجہ بندی کرنے والوں کی توسیع پزیرائی کی وجہ سے یہ تیزی سے سیکھتا ہے اور memory کا موثر استعمال حاصل کرتا ہے[112]۔

3.2۔ ڈیٹا مائننگ ٹاسکس

ڈیٹا مائننگ کا بنیادی مقصد ڈیٹا سے سیکھنا ہے۔ ڈیٹا مائننگ کے عمل میں پائے جانے والے نمونوں کی قسم کا تعین کرنے کے لئے ڈیٹا مائننگ ٹاسکس کا استعمال کیا جاتا ہے۔ ڈیٹا مائننگ ٹاسکس کو عام طور پر دو بڑی اقسام میں تقسیم کیا جاتا ہے: Predictive Tasks اور Descriptive

Tasks جیسا کہ ذیل میں دیئے گئے (3.1) ترسیم میں دکھایا گیا ہے [113] [114]۔ Predictive Tasks میں مقصد آزاد (تلاشی) وصفوں کی اقدار پر مبنی انحصار (ہدف) کے وصف کی قدر کی پیش گوئی کرنا ہے۔ Descriptive Tasks کا مقصد یہ ہے کہ ڈیٹا میں بنیادی relationships کو بیان کرنے والے patterns کو نکالا جائے۔ Descriptive Tasks اکثر exploratory in nature ہوتے ہیں اور نتائج کی وضاحت اور توثیق کے لئے اکثر پری پروسیسنگ کے بعد کے طریقوں کی ضرورت ہوتی ہے[59]۔

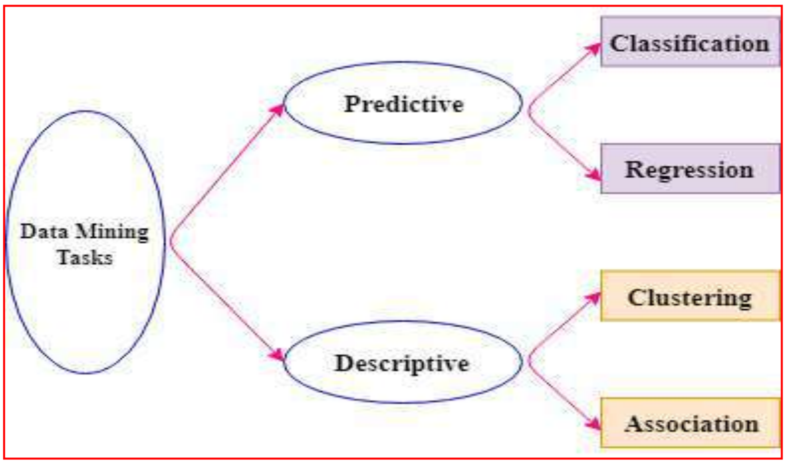


Figure3.1 Categorization of Data Mining Tasks

3.2.1۔ Predictive ڈیٹا مینگ ٹاسکس

Predictive ڈیٹا مینگ ماڈلنگ کا مطلب ہدف متغیر کے لئے وضاحتی متغیر کے فنکشن کے طور پر ایک ماڈل بنانا ہے۔ Predictive data mining ماڈلنگ کے کام دو طرح کے ہوتے ہیں[59]۔

3.2.1.1۔ Classification: Classification ڈیٹا مائننگ ایپلی کیشنز predictive ڈیٹا مینگ کے کاموں کو بتاتی ہیں جو ہدف کی خصوصیت کی متضاد اقدار کی پیش گوئی کرتی ہیں۔ اگر ہدف کی خصوصیت والی اقدار دو اقدار ہیں (جیسے ہاں اور نہیں)، تو اسے Binary Classification کہا جاتا ہے۔ لیکن، اگر ہدف کے وصف میں کسی بھی بیماری کے علاج کے لئے multiple متعدد ممکنہ اقدار (مثلا دوائی A، B اور C) ہو تو اس کو Multiple Classification کہا جاتا ہے [59] ۔

3.2.1.2۔ Regression: Regression تشریحی متغیر کے فعل کے طور پر مستقل ہدف متغیرات کے لئے ایک پیش گوئی ماڈل بناتی ہے۔ مثال کے طور پر مستقبل میں اسٹاک کی لاگت کا تین ماہ کی پیشن گوئی کرنا، کیونکہ لاگت ایک مستقل قیمت والی صفت ہے۔ دونوں کاموں کا مقصد ایک ایسا ماڈل سیکھنا ہے جو ہدف متغیر کی پیش گوئی اور صحیح قدروں کے مابین غلطی کو کم کردے[59] ۔

3.2.2۔ Descriptive ڈیٹامیننگ ٹاسکس

Descriptive ماڈلنگ سے وہ patterns اخذ ہوتے ہے جو ڈیٹامیں بنیادی relationships کا خلاصہ کرتے ہیں۔ Descriptive ماڈلنگ کے دو طرح کے کام ہیں: کلسٹرنگ اور ایسو سیشن ۔

3.2.2.1۔ کلسٹرنگ: کلسٹرنگ کا کام مضبوطی سے interrelated instances کو دریافت کرنا ہے تاکہ ایک ہی کلسٹر میں پائے جانے والے واقعات الگ الگ کلسٹر سے تعلق رکھنے والے واقعات کے علاوہ ایک دوسرے سے مضبوطی سے وابستہ ہوں[59]۔

3.2.2.2۔ ایسو سیشن: ایسو سیشن کے تجزیے کا اطلاق ایسے patterns کی نشاندہی کرنے کے لئے کیا جاتا ہے جو ڈیٹامیں مضبوطی سے وابستہ خصوصیات کو نمایاں کرتے ہیں۔ دریافت شدہ نمونوں کا استعمال کثرت سے مضمر قواعد یا خصوصیات سبسٹ کی شکل میں کیا جاتا ہے [60]

3.3۔ ڈیٹامیننگ ٹکنیکس

دل کی حالت کو مختلف علامات سے پیش گوئی کرنا ایک مسخ شدہ مسئلہ ہے جو غلط مفروضوں کا پابند ہے اور اس کے متاثر کن اثرات مرتب ہوتے ہیں۔ امراض قلب سے متعلق ڈیٹا سے معلومات حاصل کرنے کے لئے ڈیٹا مینگ کے مختلف طریقوں پر عمل کیا جاتا ہے۔ ہیلتھ کیئر میں ڈیٹا مائننگ کے طریقوں کو apply کرنے کا مقصد ماہرین یا معاونین کو سنبھالنا نہیں ہے، بلکہ جہاں وہ جدوجہد کرتے ہیں وہاں مدد فراہم کرنا ہے[-115] [113] ڈیٹا مائننگ کی بہت سی بنیادی techniques دستیاب ہیں لیکن اس تحقیق کا مرکز classification کی تکنیک کا اطلاق ہے جو طبّی پیشہ ورافراد کو امراض قلب کے خطرے کی شناخت کرنے میں مدد فراہم کرتا ہے۔

3.3.1۔ ڈسیشن ٹری

ڈسیشن ٹری ایک غیر پیرامیٹرک ٹکنیک ہے جو اکثر classification کے لئے استعمال ہوتا ہے۔ تاہم یہ regression کے کاموں کے لئے بھی استعمال کیا جاسکتا ہے[116]۔ ڈسیشن ٹری greedy نقطہ نظر کو اپناتے ہیں اور اوپر سے نیچے بار بار چلنے والے تقسیم اور فتح کے انداز میں تعمیر کیے جاتے ہیں۔ الگورتھم tuples اور ان سے وابستہ طبقاتی لیبلوں کے تربیتی سیٹ سے شروع ہوتا ہے۔ ٹریننگ سیٹ کو بار بار چھوٹے چھوٹے حصوں میں تقسیم کیا جاتا ہے تا کی ٹری بنایا جاے[59]۔ جب ڈسیشن ٹری تعمیر کیے جاتے ہیں تو بہت سے شاخیں ٹریننگ ڈیٹا

میں noise کرنے والوں کی عکاسی کر سکتی ہیں۔ Tree Pruning اس طرح کی شاخوں کی نشاندہی کرنے اور ان کو ختم کرنے کی کوشش کرتی ہے جس کا مقصد unseen ڈیٹا پر classification کی درستگی کو بہتر بنانا ہے۔ ڈسیشن ٹری میں ہر اندرونی نوڈ ایک وصف پر ایک ٹیسٹ کی نشاندہی کرتا ہے، ہر branch امتحان کے نتائج کی نمائندگی کرتی ہے، اور ہر پتی کے نوڈ میں کلاس کا لیبل ہوتا ہے [117]۔

ریسرچ میں مختلف قسم کے ڈسیشن ٹری دستیاب ہیں اور ان میں فرق صرف mathematical ماڈل کا ہے جو ٹری کے فیصلے کے اصولوں کو نکالنے میں الگ کرنے والے وصف کو منتخب کرنے کے لئے استعمال ہوتا ہے۔ وابستگی کے انتخاب کے مقبول اقدامات انفارمیشن گیین، گیین ریشو اور جینی انڈیکس ہیں [118]۔ انفارمیشن گیین کے انتخاب کے پیمائش کا استعمال اس وصف کو منتخب کرنے کے لئے کیا جاتا ہے جو مختلف طبقات کو الگ الگ کلاس میں تقسیم کرتا ہے۔ انفارمیشن گیین اپروچ الگ ہونے والی وصف کو منتخب کرتی ہے جو انٹروپی کی قدر کو کم سے کم کرتی ہے، اس طرح انفارمیشن گیین کو زیادہ سے زیادہ کرتے ہیں۔ ہر ایک وصف کے لئے انفارمیشن گیین کا حساب کتاب مساوات (3.1) کے ذریعے کیا جاتا ہے۔

$$Gain(A) = Info(D) - Info_A(D) \qquad (3.1)$$

جہاں Info(D) معلومات میں اوسط رقم ہوتی ہے جس میں D tuple کے کلاس لیبل کی شناخت ہوتی ہے اور مساوات (3.2) کا استعمال کر کے اس کا اندازہ کیا جاتا ہے۔ InfoA(D) کی طرف سے تقسیم کی بنیاد پر ڈی سے ٹپل کو درجہ بندی کرنے کے لئے مطلوبہ معلومات کی ضرورت ہے اور مساوات (3.3) میں اس کا حساب کیا جاتا ہے۔

$$Info(D) = -\sum_{i=1}^{m} pi \ \log_2 \ (pi) \qquad (3.2)$$

یہاں pi غیر صفر امکان ہے کہ D میں arbitrary tuple کا تعلق کلاس Ci سے ہے اور اس کا تخمینہ لگایا جاتا ہے |Ci, D|/|D| 2base پر log function استعمال کیا جاتا ہے کیونکہ معلومات کو بٹس میں انکوڈ کیا جاتا ہے۔

$$Info_A(D) = -\sum_{j=1}^{v} \frac{|Dj|}{|D|} * Info(D_j) \qquad (3.3)$$

اصطلاح $|D_j| / |D|$ jth تقسیم کے وزن کے طور پر کام کرتا ہے [59]۔

ذیل میں دیئے گئے ترسیم 3.2 سے پتہ چلتا ہے کہ ڈسیشن ٹری ہارٹ ڈیزیز رسک ماڈل کو بنانے کے لئے کس طرح کام کرتا ہے۔ ڈسیشن ٹری کو استعمال کرنے کی بنیادی وجہ یہ ہے کہ وہ ایک رسک اسسمنٹ ماڈل تیار کریں جو ٹریننگ ڈیٹا سیٹ سے فیصلے کے قواعد سیکھ کر امراض قلب سے متاثرہ افراد کی پیشن گوئی کر سکے۔ ڈسیشن ٹری سے حاصل کردہ تجرباتی نتائج کی وضاحت باب 4 میں کی گئی ہے۔

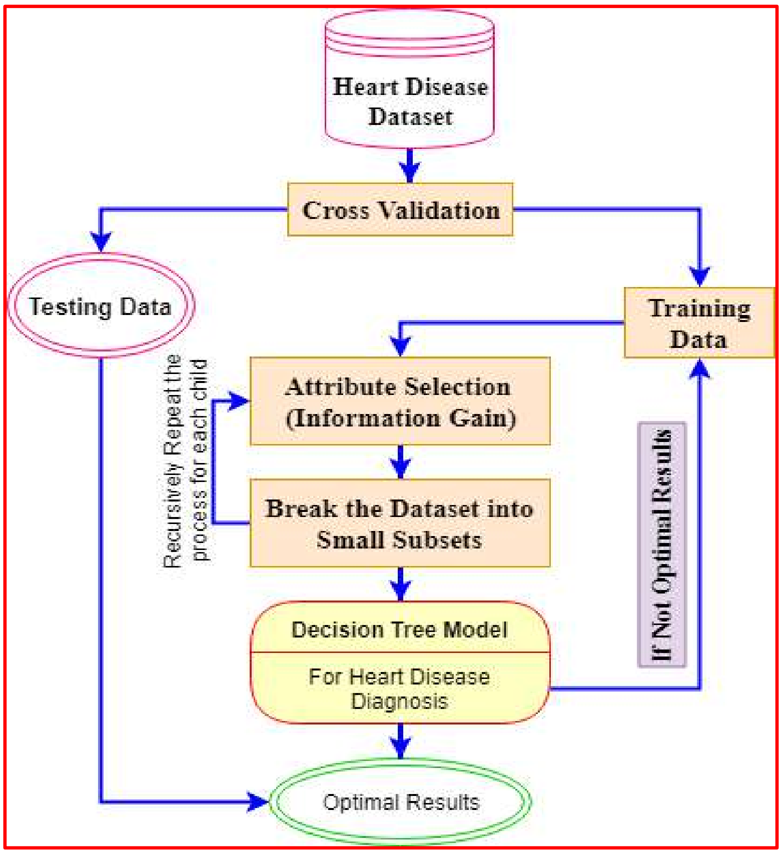

Heart Disease Dataset
Cross Validation
Testing Data
Training Data
Attribute Selection (Information Gain)
Break the Dataset into Small Subsets
Recursively Repeat the process for each child
If Not Optimal Results
Decision Tree Model
For Heart Disease Diagnosis
Optimal Results

Figure 3.2 Decision Tree Model Working for Heart Disease Prediction

3.3.2- کے نیریسٹ نیبر (کے این این)

کے نیریسٹ نیبر (کے این این) بنیادی، غیر پیرامیٹرک اور instance پر مبنی ڈیٹا میننگ تکنیک ہے [119]۔ کے این این مشابہت کے ذریعہ سیکھنے کا استعمال کرتا ہے، جو distance میٹرک کا استعمال کرتے ہوئے موجودہ ریکارڈوں کے ساتھ نئے غیر طبقاتی ریکارڈ کا موازنہ کرتا ہے۔ قریب ترین موجودہ ریکارڈ کلاس کو نئے غیر طبقاتی ریکارڈ میں تفویض کرنے کے لئے استعمال کیا جاتا ہے [59]۔ ذیل میں دیئے گئے اعداد 3.3 میں کے این این کی درجہ بندی کی مثال دکھائی گئی ہے۔

k کی value تجربات کے ذریعے چیک ہوتی ہے, k کی ویلیو 1 سے set کیا جاسکتا ہے اور پھر انکریمنٹ کر کے زیادہ نئے neighbors چیک ہوتے ہے۔ Minimum Error کی شرح دینے والے کے قدر کو منتخب کیا گیا ہے۔ درجہ بندی کی غلطی کی شرح

کا اندازہ لگانے کے لئے ٹیسٹ سیٹ کا استعمال کیا جاتا ہے۔ کے این این الگورتھم میں ایک نئی مثال neighbors کی قربت کے لحاظ سے درجہ بندی کی جاتی ہے جس کی تعریف فاصلہ افعال کے لحاظ سے کی جاتی ہے۔ بہت سارے distance measures ہیں جن کو استعمال کیا جا سکتا ہے، جیسے (یوکلیڈین، مینہٹن اور منکووسکی) لیکن اس تحقیق میں یکلیڈین اقدام کو امراض قلب کے ڈیٹا کی خصوصیات کی وجہ سے استعمال کیا ہے [60] [120]۔

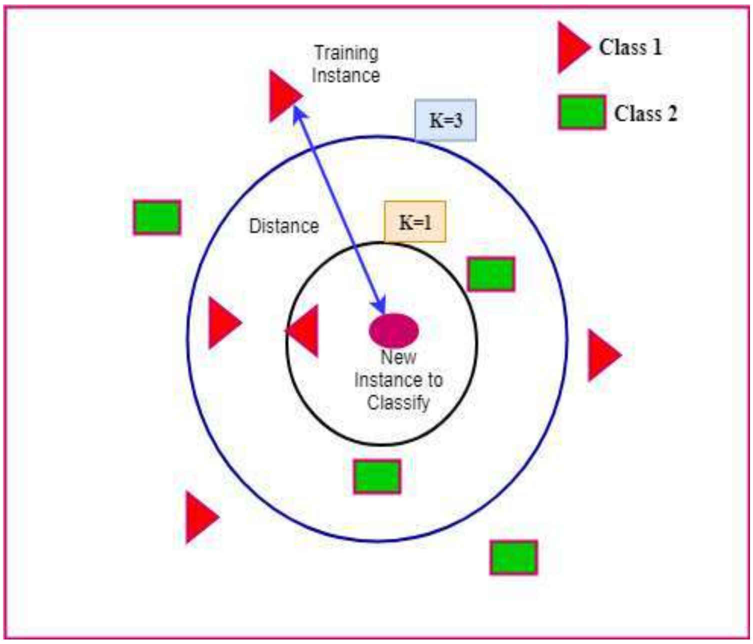


Figure 3.3 K-Nearest Neighbour classification Example

Euclidean distance دو پوائنٹس کے درمیان کو جوڑنے والے direct shortest راستے کی لمبائی کو کہتے ہے۔ یکلیڈین distance کا حساب کتاب i اور j کو تمام P input attributes کے درمیان مربع اختلافات کے مجموعی کے مربع جڑ کے طور پر شمار کیا جاتا ہے۔

$$d(i, j) = \sqrt{(xi1\text{-}xj1)^2+(xi2\text{-}xj2)^2+...+(xip\text{-}xjp)^2} \qquad (3.4)$$

Euclidean distance کی پیمائش کو استعمال کرنے سے پہلے اوصاف کی قدر کو معمول بنایا جاتا ہے تاکہ higher value والے attributes سب سے کم value والے attributes کو نظر انداز نہ کریں۔ اس تحقیقی کام میں Min-Max normalization technique کا استعمال کیا گیا ہے تاکہ P کی عددی صفت کی قدر/ P میں تبدیل کمپیوٹنگ کے ذریعہ [0،1] کی حد میں ہو ۔

$$P^{\backslash} = \frac{P - \min Z}{\max Z - \min Z} \qquad (3.5)$$

جہاں min_Z اور max_Z کم سے کم اور خصوصیت Z کی زیادہ سے زیادہ اقدار ہیں۔

اس تحقیق میں کے این این تکنیک کو امراض قلب کے مریضوں کو جلد سے جلد پیش گوئی کرنے کے لئے استعمال کیا جاتا ہے۔ Classification سے حاصل کیے گئے تجرباتی نتائج پر باب چہارم میں تبادلہ خیال کیا گیا ہے۔

3.3.3 سپورٹ ویکٹر مشین (ایس وی ایم)

سپورٹ ویکٹر مشین (ایس وی ایم) ایک supervised ڈیٹا مائننگ تکنیک ہے جو classification اور regression دونوں مقاصد کے لئے استعمال کی جاتی ہے۔ ایس وی ایم تربیت ڈیٹا کو ایک اعلی جہت میں تبدیل کرنے کے لئے non linear mapping استعمال کرتا ہے۔ اس نئی جہت کے اندر یہ زیادہ سے زیادہ الگ کرنے والے hyperplane کو تلاش کرتا ہے۔ ایس وی ایم کے لئے زیادہ سے زیادہ hyperplane کا مطلب دو کلاسوں کے درمیان سب سے بڑا مارجن ہوتا ہے۔ ایس وی ایم کو یہ hyperplane سپورٹ ویکٹرس اور مارجن کا استعمال کرتے ہوئے ملتا ہے [60] [121]۔ سپورٹ ویکٹر وہ ڈیٹا پوائنٹ ہیں جو الگ کرنے والے hyperplane کے قریب ہوتا ہے اور ڈیٹاسیٹ کے اہم عنصر سمجھے جاتے ہیں۔ اور مارجن کا مطلب hyperplane کے متوازی سلیب کی زیادہ سے زیادہ چوڑائی ہے جس کے اندرونی ڈیٹا کا کوئی پوائنٹ نہیں ہے۔

$$f(T) = \sum_{i} \alpha_i y_i \ (X_i \ . T) + b \qquad (3.6)$$

جہاں ویکٹر X_i معاون ویکٹر ہیں، Yi کلاس لیبل ہیں Xi کے جو ویکٹر ٹی ٹیسٹ کے نمونے کی نمائندگی کرتا ہے۔ (Xi.T) معاون Xi کے ساتھ ٹیسٹ نمونے T کی ڈاٹ پروڈکٹ ہے۔ اور α_i and b پیرامیٹرز ہیں جو learning الگورتھم کے ذریعہ طے کیے جاتے ہیں۔

درج ذیل ترسیم 3.4 linear سپورٹ ویکٹر مشین کی وضاحت کرتا ہے جہاں ہلکے سبز حلقے X1 کے ڈیٹا پوائنٹس اور x2 کے سرخ اشارے والے ڈیٹا پوائنٹس کی نمائندگی کرتے ہیں۔ ایس وی ایم کا مقصد hyperplane اور ٹریننگ سیٹ کے ساتھ کسی بھی ڈیٹا پوائنٹ کے مابین سب

سے زیادہ ممکنہ مارجن کے ساتھ ایک ہائپرپلان کا انتخاب کرنا ہے، جس سے نئے ڈیٹا کو صحیح درجہ میں classify کرنے کا زیادہ سے زیادہ امکان مل جاتا ہے۔ تاہم، اگر کوئی واضح ہائپرپائلن نہیں ہے تو ضروری ہے کہ higher dismensions والے نقطہ نظر میں منتقل ہونا ضروری ہے جسے ایس وی ایم میں کرنلنگ کہا جاتا ہے۔ خیال یہ ہے کہ اس ڈیٹا کو اعلی اور اعلی طول و عرض میں نقش کیا جاتا رہے گا جب تک کہ اس کو الگ کرنے کے لئے ایک ہائپرپلیٹ تشکیل نہیں دیا جاسکتا[61]۔

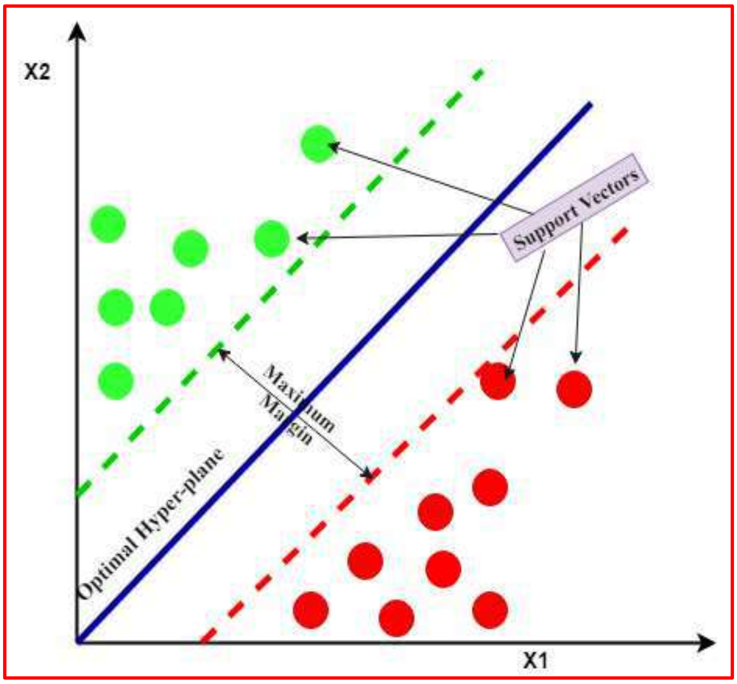


Figure 3.4 Linear SVM Classifier for Two-Class Representation

لہذا non-linear علیحدگی کی صورت میں تربیت کے ڈیٹا کو ایک higher dimensional space H میں بنایا جائے گا اور وہاں ایک زیادہ سے زیادہ hyperplane تعمیر کی جائے گی۔ The mapping kernel function K کے ذریعہ انجام دیا جاتا ہے جو H میں اندرونی مصنوع کی وضاحت کرتا ہے۔ مختلف میپنگ مختلف SVM تشکیل دیتے ہیں۔ جب mapping ہوتی ہے تو امتیازی سلوک اس طرح دیا جاتا ہے:

$$f(T) = \sum_{i} \partial_i y_i K(x_i, T) + b^{|} \qquad (3.7)$$

ایس وی ایم بڑے پیمانے پر اس کے kernel function کے انتخاب کی خصوصیت سے نمایاں ہے مثلا polynomial, Gaussian and radial basis kernel function ۔ تاہم ان فنکشن کے افعال کے علاوہ kernel فنکشن کے اور بھی

دوسرے کام ہیں۔ پیرامیٹرز ∂ and y کا تعین کرنے اور مذکورہ مساوات میں امتیازی تقریب کی تعمیر آخرکار Lagrangian dual objective function کو زیادہ سے زیادہ کرنے میں ایک مسئلہ بن جاتی ہے۔

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(X_i, X_j) \qquad (3.8)$$

Under constraints

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \alpha_i >= 0, (i = 1, 2, \ldots, n) \qquad (3.9)$$

یہاں training data میں samples کی تعداد n ہے۔ تاہم، مساوات میں کواڈریٹک پروگرامنگ کا مسئلہ (3.8) آسانی سے معیاری تکنیک کے ذریعہ حل نہیں کیا جاسکتا ہے کیونکہ اس میں ایک matrix شامل ہوتا ہے جس میں متعدد عناصر شامل ہوتے ہیں جن کی تربیت کے نمونے کے مربع کے برابر ہوتے ہیں[60]۔ ایس وی ایم تکنیک امراض قلبی پیش گوئی میں استعمال کی جاتی ہے کیونکہ حیاتیات اور ڈاکٹروں کی مدد کے لئے جامع اور صحیح قواعد انتہائی اہم ہیں۔ امراض قلبی پیشن گوئی کے نتائج ایس وی ایم ماڈل سے حاصل کیے گئے باب چہارم میں زیر بحث آئے گے۔

3.3.4 رینڈم فاریسٹ

رینڈم فاریسٹ is an ensemble of simple decision trees جو classification اور regression کی دشواری دونوں کے لئے استعمال ہوتے ہیں۔ رینڈم فاریسٹ الگورتھم انفرادی ڈسیشن ٹری کی زیادہ سے زیادہ overfitting مسئلہ پر قابو پانے کے مقصد کے ساتھ، تصادفی طور پر منتخب کردہ training dataset سے متعدد decision trees کے ساتھ فاریسٹ تخلیق کرتا ہے۔ Random forest classification میں ہر ڈسیشن ٹری کے ووٹ اور مجموعی ووٹ ٹیسٹ object کی آخری کلاسوں کا فیصلہ کرتے ہیں تاہم regression میں، انفرادی trees کی اسباب کی پیش گوئی یا regression کا حساب لگایا جاتا ہے[122]۔ رینڈم فاریسٹ. بیگنگ کی کافی حد تک تغیرات ہیں جو de-correlated trees کا ایک بہت بڑا مجموعہ بناتے ہیں اور پھر ان کا اوسط کرتے ہیں۔ ترسیم 3.5 رینڈم فاریسٹ الگورتھم کے کام کو ظاہر کرتا ہے جس میں ہر ڈسیشن ٹری اصل ڈیٹا کے ایک مختلف نمونے پر grow ہو جاتا

ہے اور ٹیسٹ سیٹ غلطی کا غیر جانبدارانہ تخمینہ حاصل کرنے کے لئے cross validation یا علیحدہ ٹیسٹ سیٹ کی ضرورت نہیں ہے کیونکہ، ہر iteration میں، تقریبا 1/3 نمونے بوٹسٹریپ کی نئی تربیت سے باہر رہ جاتے ہیں اور ڈسیشن ٹری کی تعمیر میں استعمال نہیں ہوتے ہیں۔ اس طرح تعمیر شدہ ڈسیشن ٹری میں سے ایک تہائی ٹریز میں ہر نمونے کے لئے ایک ٹیسٹ سیٹ کی درجہ بندی حاصل کی جاتی ہے۔ نمونے کی آخری درجہ بندی وہ کلاس ہے جس میں فاریسٹ میں ٹریز سے زیادہ سے زیادہ ووٹ ہوتے ہیں [60][61]۔ رینڈم فاریسٹ الگورتھم دل کی بیماری کی پیش گوئی اور تشخیصی میں مستعمل ہے۔ رینڈم فاریسٹ ماڈل سے حاصل کردہ امراض قلب کی پیش گوئی کے نتائج پر چہارم باب میں تبادلہ خیال کیا گیا ہے۔

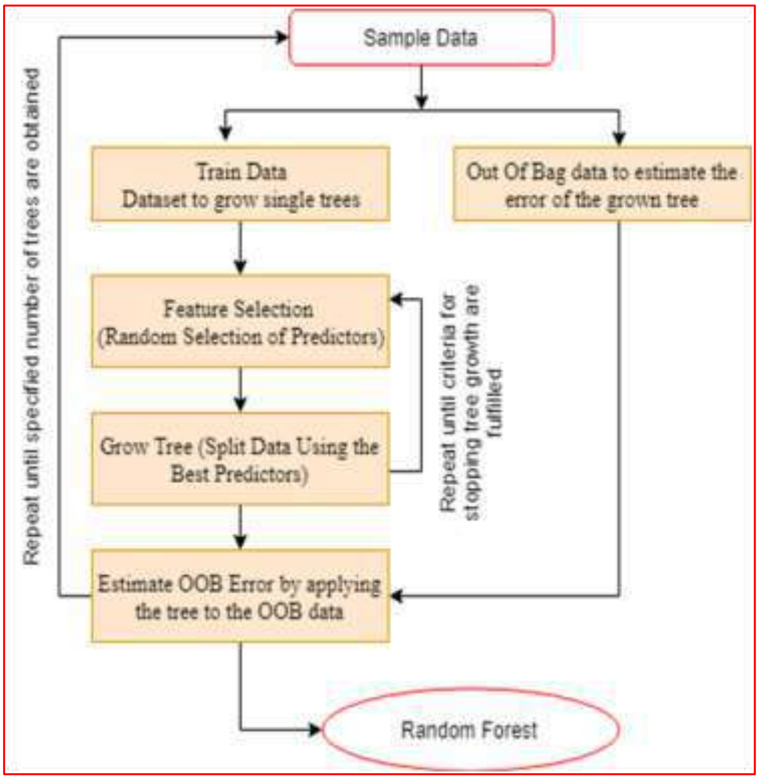

Sample Data
Train Data
Dataset to grow single trees
Out Of Bag data to estimate the error of the grown tree
Feature Selection
(Random Selection of Predictors)
Grow Tree (Split Data Using the Best Predictors)
Repeat until criteria for stopping tree growth are fulfilled
Estimate OOB Error by applying the tree to the OOB data
Repeat until specified number of trees are obtained
Random Forest

Figure 3.5 Random Forest Algorithm Working

3.3.5۔ نیوی بائیس الگورتھم

نیوی بائیس ڈیٹا کی امکانی صلاحیتوں کی بنیاد پر طبقاتی رکنیت کی پیش گوئی کر رہے ہیں۔ نیوی بائیس کی درجہ بندی Bayesian theorem پر مبنی ہے اور خاص طور پر موزوں ہے جب ان پٹ کی جہت زیادہ ہو۔ الگورتھم نے یہ یقین کیا ہے کہ کسی خاص طبقے پر کسی خاصیت کی قیمت کا نتیجہ دوسری صفات کی اقدار سے آزاد ہے۔ اس مفروضے کو class conditional independence کہا جاتا ہے۔ یہ مفروضہ اس میں شامل کمپیوٹیشن کو آسان بنانے کے لئے کیا گیا ہے اور اس لحاظ سے اسے نیوی سمجھا جاتا ہے [59]۔

نیوی بائیس classification کرنے والا متعدد آزاد متغیرات کو سنبھال سکتا ہے چاہے وہ continuous ہو یا categorical۔ نیوی بائیس کی درجہ بندی ہدف کے اوصاف کی prior probability اور باقی صفات کے conditional proability کا حساب لگاتی ہے۔ Training dataset کے لئے پیشگی اور مشروط امکان کا حساب لگایا جاتا ہے۔ ٹیسٹنگ ڈیٹاسیٹ میں ہر ٹیسٹنگ مثال کے لئے احتمال کا اندازہ ہدف کے انتہاہی اقدار میں سے ہر ایک کے ساتھ کیا جاتا ہے اور پھر سب سے بڑے امکان والے ہدف کی قیمت کو منتخب کیا جاتا ہے[123] [124]۔ اہداف کی خصوصیت کی قیمت کے لئے جانچ کی مثال کے امکانات ذیل میں دی گئی مساوات کا استعمال کرکے حساب کیا جاتا ہے۔

$$P(H / X) = \frac{P(X/H)P(H)}{P(X)} \qquad (3.10)$$

یہاں P (H /X) is the posterior probability of H conditioned on X اور P(X/H) is the posterior probability of X conditioned on H P(X/H) اور اسطرح P (X) is the prior probability of X. اور P(H) is the prior probability of H.

نیوی بائیس یہ یقین کرکے احتمالات کا حساب کتاب آسان کرتا ہے کہ ایک مخصوص طبقے کی قیمت سے متعلق ہر صفات کا احتمال دیگر تمام صفات سے آزاد ہے۔ یہ ایک مضبوط مفروضہ ہے اور اس کا نتیجہ تیز اور موثر طریقہ میں نکلتا ہے[60] [61]۔ اس تحقیق میں نیوی بائیس الگورتھم غیر حملہ آور خطرہ کی خصوصیات سے جلد از جلد امراض قلب کی پیش گوئی اور تشخیص کرنے کے لئے استعمال کیا جاتا ہے۔ نیوی بائیس ماڈل سے حاصل ہونے والی امراض قلب کی پیش گوئی کے نتائج کو چوتھا باب میں زیر بحث لایا گیا ہے۔

3.4 ماڈل پرفارمنس تکنیکس

پرفارمنس پیمائش ڈیولپڈ ماڈل کی کارکردگی کو جانچنے کے لئے استعمال کی جاتی ہے۔ دیئے گئے ڈیٹاسیٹ میں structured patterns کو سمجھنے کے بے شمار طریقے ہیں تاہم، یہ معلوم کرنے کے لئے کسی خاص مسئلے پر کس طریقہ کا اطلاق کیا جائے ہمیں اندازہ لگانے کے لئے منظم طریقے کی ضرورت ہے۔ classification کے مسائل میں ایک الگورتھم کی کارکردگی کو کنفیوژن میٹرکس، cross validation، error rate، sensitivity، specificity، accuracy، precision اور AUROC کے لحاظ سے ماپا جاتا ہے، جس پر مندرجہ ذیل گفتگو کی جاتی ہے[59] [125].

3.4.1۔ کنفیوژن میٹرکس

کنفیوژن میٹرکس درجہ بندی کے مسائل میں کارکردگی کی پیمائش کا ایک بنیادی ذریعہ ہے۔ دیئے گئے جدول 3.1 کے نیچے two-class confusion میٹرکس دکھایا گیا ہے جو درجہ بندی کرنے والے کے ذریعہ ہونے والی غلطیوں کی اقسام کے بارے میں بصیرت فراہم کرتا ہے۔

Table 3.1 Contingency Matrix for Two-Class Classification

		Predicted Cases	
		Positive	**Negative**
Actual Cases	**Positive**	True Positive	False Positive
	Negative	False Negative	True Negative

True Positive ان مثبت والے tuples کا حوالہ دیتے ہیں جو درجہ بندی کرنے والے کے ذریعہ صحیح طور پر لیبل لگائے جاتے ہیں، True Negative اس منفی tuples کا حوالہ دیتے ہیں جس کو درجہ بند کرنے والے نے صحیح طور پر لیبل لگایا ہے، False Positives (جسے Type I بھی کہا جاتا ہے) وہ منفی ٹیوپلس ہیں جن کو غلط طور پر لیبل لگا ہوا ہے۔ False Negative (جسے Type II ایرر بھی کہا جاتا ہے) وہ مثبت ٹیوپلس ہیں جو منفی کے طور پر غلط لیبل لگائے جاتے ہیں۔

i. Sensitivity (جس کو True Positive Rate یا Recognition یا Recall بھی کہا جاتا ہے) مثبت tuples کا تناسب ہے جنہیں مثبت کے طور پر درجہ بند کیا گیا ہے [126]۔

$$Sensitivity = \frac{TruePositive}{Positive} \tag{3.11}$$

ii. Specificity (True Negative Rate کے نام سے بھی جانا جاتا ہے) منفی tuples کا تناسب ہے جو منفی کے طور پر درجہ بندی کی گئی ہے [126]۔

$$Specificity = \frac{TrueNegative}{Negative} \tag{3.12}$$

iii. Accuracy ان معاملات کی کل فیصد ہے جو الگورتھم کے ذریعہ صحیح درجہ میں درجہ بندی کی گئی ہیں [127] ۔

$$Accuracy = \frac{True\ Positive + True\ Negative}{TP + FP + TN + FN} \qquad (3.13)$$

iv. Precision قطعیت کا ایک پیمانہ ہے (یعنی مثبت درجہ بندی کرنے والی کتنی فیصد حقیقت میں مثبت ہیں) [127] ۔

$$Precision = \frac{True\ Positive}{TruePositive + FalsePositive} \qquad (3.14)$$

vi. Missclassification Rate ایک پوری سیٹ میں غلطیوں کا تناسب ہے۔ Error Rate تربیت اور generalizaiton error کا مجموعہ ہے۔ تربیت کی غلطیاں تربیت کے اعداد و شمار پر عائد غلط فہمی کی غلطیوں کی تعداد ہیں، جبکہ عام طور پر غلطی پچھلے unseen ریکارڈوں پر ماڈل کی متوقع غلطی ہے۔ بہترین درجہ بندی ماڈل میں کم training اور error generalizaiton کرنے میں خرابی ہے۔ [59] [60] [61] ۔

$$Error\ Rate = \frac{False\ Positive + False\ Negative}{Positive + Negative} \qquad (3.15)$$

3.4.2۔ کراس ویلڈیشن

Cross validation technique measures the error rate of a learning model on a specific dataset۔ Cross Validaton میں مکمل ڈیٹا سیٹ تصادفی طور پر تقریبا برابر سائز کے باہمی خصوصی ذیلی حصوں میں تقسیم کیا جاتا ہے اور ہر ریکارڈ کو ایک ہی تعداد میں ٹریننگ کے لئے اور عین جانچ کے لئے ایک بار استعمال کیا جاتا ہے۔ Training ڈیٹا سیٹ ڈیٹا میننگ کی تکنیک کو اس ڈیٹا سے سیکھنے کی اجازت دیتا ہے۔ ٹیسٹنگ ڈیٹا سیٹ کو ٹریننگ ڈیٹا سیٹ سے جو سیکھا گیا ہے اس کے سلسلے میں ڈیٹا مائننگ تکنیک کی کارکردگی کی جانچ کرنے کے لئے استعمال کیا جاتا ہے[61][60][59]۔

3.4.3۔ اے یوآراوسی (AUROC)

AUROC ایک کارکردگی کا پیمانہ گراف ہے جو مختلف حدوں کی ترتیبات پر درجہ بندی ماڈل کی کارکردگی کو ظاہر کرتا ہے۔ AUROC ماڈل کلاسوں میں فرق کرنے میں ماہر ہے۔ ROC Curve کو ایکس محور پر Rate False Positive کے مقابلے میں Y-axis پر True Positive Rate کے ساتھ منصوبہ بنایا گیا ہے جیسا کہ نیچے دیئے گئے 3.6 ترسیم میں دکھایا گیا ہے[61] [60] ۔

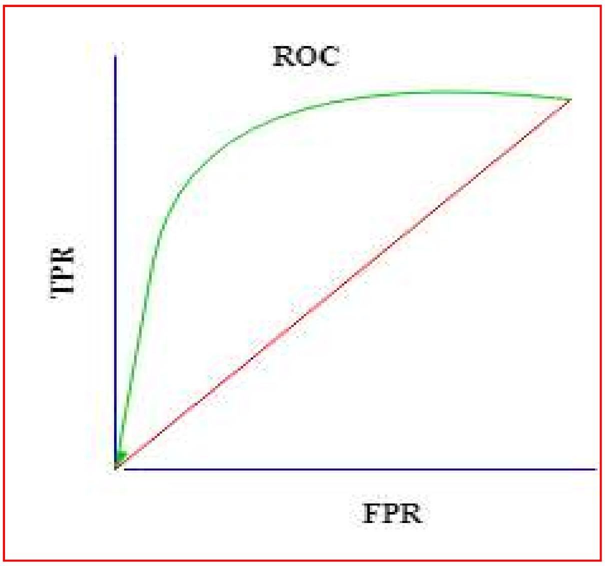


Figure 3.6 AUROC Curve Representation

اگر ایک ماڈل میں AUROC ویلیو 1 کے قریب ہے تو اس کا مطلب ہے کہ ماڈل میں علیحدگی کا عمدہ پیمانہ ہے اور وہ outstanding ہے ۔ ایک poor ماڈل میں AUROC ویلیو مساوی یا 0 کے قریب ہوتا ہے جس کا مطلب ہے کہ وہ نتائج کو reciprocate کرتا ہے اور 0 کو 1 اور 1 کو 0 کی طرح پیش گوئی کرتا ہے۔ جب AUROC value تقریبا 0.5 ہوتی ہے تو پھر ماڈل میں مثبت اور منفی کلاسوں میں فرق کرنے کی صلاحیت نہیں ہوتی۔

3.5۔ رسک ایوالویشن ماڈل ڈویلپمنٹ کے لئے ڈیٹا کلیکشن اور ریسرچ میتھوڈس کا استعمال

ڈیٹاسیٹ جمع کرنے کے لئے قابل ذکر چیلنج معیار اور متعلقہ ڈیٹا کو حاصل کرنا ہے۔ بنیادی ڈیٹا کو متعدد اعداد و شمار جمع کرنے کے طریقوں (انٹرویوز) کے ذریعے مختلف اضلاع سے متعلق اعداد و شمار کے ذرائع جیسے ضلعی اسپتالوں اور کشمیر کے نجی کلینک (انڈیا) سے جمع کیا جاتا ہے۔ اس دل کی بیماری کے ڈیٹاسیٹ میں 5776 مریضوں کے ریکارڈ شامل ہیں اور اس کے ساتھ چودہ (14) غیر ناگوار خطرے کی خصوصیات بھی ہیں جیسا کہ ذیل میں دیئے گئے جدول 3.2 میں بیان کیا گیا ہے۔ Risk Evaluation ماڈل تیار کرنے کے لئے وضاحتی تحقیقاتی طریقہ کار کی پیروی کی گئی ہے۔ رسک ایولویشن ماڈل python پروگرامنگ language میں ڈولوپڈ کیا گیا ہے اور ڈیٹا کی صفائی ، ڈیٹا ماڈلنگ ، ڈیٹا visualization، اور machine learning operations سب Jupyter ویب اپلی کیشن میں تیار کیا گیا ہے۔

Table3.2 Description of the Heart Disease Dataset

Features	**Data Types**	**Features with Subsequent values and Explanation**
Age	Numeric	Represents the age of a patient and is calculated in the number of years
Sex	Nominal	Represents the sex of a patient where 0= Female and 1= Male
Height	Numeric	Represents the height of the patient and is measured in Centimeters
Weight	Numeric	Represents the bodyweight of the patient and is measured in Kilograms
Systolic BP	Numeric	Represents the systolic blood pressure of the patient and is measured in mmHg
Diastolic BP	Numeric	Diastolic BP of a patient is measured in mmHg
Hereditary	Nominal	Whether the patient had inherited heart disease. It is represented in the form of 1 and 0, where 1=Yes, and 0= No
Healthy Diet	Nominal	Does the patient take a nutritious diet? It is represented as 0=Following, 1=occasionally and 2= Not Following
Physical Activity	Nominal	Whether the patient is exercising or not. Represented as 0= No Exercise, 1= Regular Exercise and 2= occasionally
Alcohol Consumption	Nominal	How often the patient drinks alcohol and it is represented as 0= Non-Alcoholic, 1= occasionally and 2= Alcoholic
Smoking	Nominal	Whether the patient is smoking or not 0= Non-Smoker, 1= Regular and 2= occasionally smoking
Socio-Economic Level	Nominal	Represents the economic level of the patient where 0=Poor, 1= Middle Class and 2= High Class
Diagnosis	Nominal	0= No and 1= Yes

ڈیٹا سے مختلف بصیرت حاصل کرنے کے لئے کشمیر ہارٹ ڈیزیز ڈیٹاسیٹ پر exploratory data analysis کیا گیا ہے۔ یہ پایا گیا ہے کہ ہارٹ ڈیزیز ڈیٹاسیٹ noisy ہے اور اس میں (؟) نشان کے ذریعہ متعدد missing values والے اقدار شامل ہیں۔ اعداد و شمار کا انقباض انجام دیا جاتا ہے (اعداد و شمار کی گمشدگی سے منسوب خاصیتوں کا مطلب یہ ہے کہ اعداد و شمار کو صاف کرنے کی تکنیک کا استعمال missing ہونے والی صفت اقدار کو بھرنے کے لئے کیا جاتا ہے اور واضح صفات کے لئے غائب شدہ اقدار کو بھرنے کے لئے mode

کا طریقہ استعمال کیا جاتا ہے)۔ ہارٹ ڈیزیز ڈیٹاسیٹ اوصاف کو دو اقسام برائے نام اور ہندسوں میں منظم کیا جاتا ہے۔ مثال کے طور پر، nominal and numeric اقدار جیسے "مرد" اور "خواتین" برائے nominal attribute کی نمائندگی کرتے ہیں اور "عمر" وصف کی اقدار 70 سال کی حیثیت سے ایک numeric attribute کی نمائندگی کرتی ہیں۔ مزید برآں، برائے نام اعداد و شمار بائنری اور عام متغیر کے مطابق ہوتے ہیں اور اعداد و شمار کے اوصاف عدد، وقفے سے پیمانے اور تناسب سے منسلک متغیر کے مساوی ہوتے ہیں۔

3.6 ۔ہارٹ ڈیزیز میں نان انویسو رسک فیچرز کی اہمیت

متعدد ٹیسٹوں کا استعمال کرکے کارڈیک ڈس آرڈر کی شناخت کی جا سکتی ہے۔ البتہ یہ ٹیسٹ مہنگے ہیں اور عوامی سطح کے اسکریننگ ٹیسٹوں کے بطور استعمال نہیں ہو سکتے ہیں۔ رسک Evaluation کی موجودہ تکنیکس امراض قلب کے خطرناک خصوصیات کو استعمال کرتی ہیں جن کو استعمال کرنے سے پہلے خون کے مختلف ٹیسٹوں سے معلومات کی ضرورت ہوتی ہے۔ اس مسئلے پر قابو پانے کے لئے کارڈیک ڈس آرڈر کے خطرے کی خصوصیات کو اس مقصد کے ساتھ ہموار کرنے کی ضرورت ہے کہ مناسب خطرے کی شناخت کی حکمت عملیوں کو حقیقت میں لایا جا سکے [11]۔ عمر، اونچائی، وزن، تمباکو نوشی کی عادات، جنسی تعلقات اور بلڈ پریشر جیسے ناگوار دل کی بیماری کے خطرے کی خصوصیات کو بغیر کسی پیچیدہ مشینوں اور آلات کی ضرورت کے بغیر آسانی سے تسلیم کیا جاتا ہے [129] [130]۔ اس حقیقت کے باوجود کہ جسمانی وزن اور بلڈ پریشر کو پیمائش کے کچھ آلات کی ضرورت ہے تاہم یہ آلات گھر یا دوائیوں کی دکان میں قابل رسائی ہو سکتے ہیں اور جسمانی نمونے لینے کے لئے اسپتال کی ضرورت نہیں ہوتی ہے۔

باب 4 اور 5 نان انویسو رسک فیچرز کی نشاندہی کرنے کے لئے مختلف الگورتھم کا اطلاق کرتے ہیں جو دل کی بیماری کے مریضوں کی تشخیص میں بہترین کارکردگی کا مظاہرہ کر سکتے ہیں۔ Non-Invasive صفات فائدہ مند ہیں کیونکہ وہ کم لاگت والے اوصاف ہیں (کم لاگت سے وابستہ ہونے کا مطلب ہے کہ مریض کی تشخیص کرتے وقت اس کی وابستگی کی قدر طے کرنے میں کچھ خرچ نہیں آتا ہے)۔ کشمیر ڈیٹا سیٹ سے non-invasive صفات کے مختلف مجموعے استعمال کرنے کے اثرات کی جانچ پڑتال کے لئے کارکردگی کی تشخیص کے لئے رینڈم فارسٹ، نیوی بائیس، کے نیریسٹ نیبر، سپورٹ ویکٹر مشین اور ڈیسیشن ٹری data mining classification کے ان پٹ کی حیثیت سے کی جاتی ہے۔

متعدد Non-invasive attributes کا استعمال کرتے ہوئے مساوات تیار کر کے جانچ پڑتال کی جاتی ہیں تاکہ معلوم کیا جاسکے کہ وہ امراض قلب کے مریضوں کی پیش گوئی کرنے میں کارکردگی کو بہتر بناتے ہیں یا نہیں۔

موجودہ رسک کیلکولیٹر امراض قلب کے خطرے سے متعلق صفات جیسے total cholestrol، ایچ ڈی ایل کولیسٹرول اور diabetes (عام طور پر بلڈ شوگر یا انسولین کی سطح کے ذریعے ماپا جاتا ہے) کے ساتھ ساتھ عمر، تمباکو نوشی، جنسی تعلقات، اور مریضوں کو دریافت کرنے کے لئے بلڈ پریشر کو آرام دینے جیسے ناگوار خصوصیات کا استعمال کرتے ہیں[131] ۔ امراض قلب کا خطرہ تاہم، صرف غیر ناگوار خصوصیات کا استعمال کرتے ہوئے کارکردگی کو ابھی تک ناپا نہیں جاسکا۔ اس تحقیق میں ڈیٹا مینگ تکنیکس (صرف Non-invasive attributes کے استعمال سے) درجہ بند مریضوں کی جانچ کی جاسکتی ہیں۔ یہاں کامیابی عوامی سطح پر اسکریننگ ٹیسٹوں کے لئے درخواست دینے کے لئے ایک غیر معمولی موقع فراہم کرے گی، اس طرح دل کی بیماری کے زیادہ خطرہ والے مریضوں میں ابتدائی شمولیت کو قابل بنائے گی اور ان مریضوں کے علاج معالجے کا مناسب فیصلہ کریں گے۔

3.7 باب کا خلاصہ

یہ تحقیق مختلف ڈیٹا مائننگ تکنیکوں کا استعمال کرتے ہوئے ہارٹ ڈیزیز رسک ماڈل تیار کرنے کی تحقیقات کرتی ہے۔ بنیادی ڈیٹا quantitative data collection کے طریقوں کے ذریعے heterogeneous data sources کے ذرائع سے جمع کیا جاتا ہے۔ یہ تحقیقی کام وضاحتی تحقیقی طریقہ کار کی پیروی کرتا ہے اور امراض قلب کے خطرے سے متعلق تشخیصی نمونہ تشکیل دینے کے لئے python programming language اور جو پیٹر ویب اپلی کیشن کا استعمال کرتا ہے۔ اس تحقیقی کام میں امراض قلب کی ابتدائی پیش گوئی کے لئے غیر حملہ آور خطرہ کی خصوصیات کے نمایاں ذیلی سیٹ کو منتخب کرنے کے لئے کشمیر امراض قلب کے ڈیٹاسیٹ پر مختلف خصوصیت کے انتخاب کی تکنیک کا استعمال کیا جاتا ہے۔ یہ تحقیقی کام ڈیٹا مائننگ کی مختلف تکنیکوں جیس ڈسیشن ٹری، سپورٹ ویکٹر مشین، رینڈم فاریسٹ، کے نیریسٹ نیبر اور نیوی بائیس پر عمل درآمد کرتا ہے تاکہ یہ معلوم کیا جاسکے کہ کیا یہ تکنیک طبّی پیشہ ور افراد کو ابتدائی پیش گوئی میں مددگار ثابت ہو گی جس کے نتیجے میں شدید اور پیچیدگیوں والی بیماریوں میں کمی واقع ہو گی۔ ڈیولپڈ رسک ماڈل کی کارکردگی کو جانچنے کے لئے مختلف ماڈل کی جانچ کی تکنیک استعمال کی جاتی ہے۔ آخر میں، باب غیر ناگوار خطرے کی خصوصیات کی اہمیت پر گفتگو کر کے اختتام پذیر ہوتا ہے۔

باب 4

ڈیٹا مائننگ ٹکنیکس کا استعمال کر کے ہارٹ ڈیزیز ڈیٹا میں علم کی دریافت

کشمیر ہارٹ ڈیزیز ڈیٹا سیٹ ابتدائی پیش گوئی، شناخت اور علم حاصل کرنے کے لئے تیار کیا گیا ہے۔ اس باب میں ڈیوس ڈیٹا مائننگ تکنیک کے طریقہ کار کی وضاحت کی گئی ہے جو امراض قلب کی ترقی کے لئے اس تحقیق میں عمل پیرا ہے۔ اس باب میں تحقیقی سر گرمیوں کو آسان بنانے اور مقاصد کے بیان سے تحقیق کو نتیجہ خیز بنانے کے لئے تحقیقی ڈیزائن تیار کیا گیا ہے۔ نمایاں significant نان انویسو ہارٹ ڈیزیز رسک attributes منتخب کرنے کے لئے مختلف feature selection تکنیکوں کا استعمال کیا گیا ہے۔ یہ تحقیق دل کے عارضے میں مبتلا مریضوں کی ابتدائی تشخیص کے لئے ڈیٹا مائننگ کی تراکیب کے لئے رسک attributes کے اہم ذیلی سیٹ کی نشاندہی کرتی ہے۔ آخر کار تحقیق میں دل کی بیماری کے خطرے سے متعلق تشخیصی نمونہ تیار کیا گیا تاکہ طبّی ماہرین کی مدد سے کارڈیک بیماری کے high risk میں متاثرین کی درجہ بندی کی جا سکے۔

4.1 امراض قلب کی پیش گوئی کے لئے ڈیٹا مائننگ طریقہ کار

ڈیٹا مائننگ طریقہ کار متبادل طریقوں کو استعمال کرنے کے لئے ایک تکنیک ہے جو raw ڈیٹا کو تبدیل شدہ ڈیٹا سیٹ میں لے جاتا ہے تاکہ صارفین کے لئے علم پیدا ہو سکے۔ "Knowledge Discovery from Data" کے عمل کے لئے ڈیٹا مائننگ کے دو مشہور طریقہ کار موجود ہیں: CRISP_DM[133][132] اور SEMMA [134] CRISP-DM کو صنعت کے زیر قیادت Cross Industry Standard Process Model کے طور پر تیار کیا ہے ۔ اور SEMMA (Simple Explore Modify Model Asssess) اعداد و شمار کی ڈیٹا مائننگ طریقہ ہے جو شماریاتی تجزیہ سافٹ ویئر انسٹی ٹیوٹ (SAS، 2008) سے اخذ کیا گیا ہے۔ یہ دونوں طریقے ہمارے تحقیقی کام کے لئے موزوں نہیں ہیں کیونکہ یہ بہت بڑے اور پیچیدہ ہیں۔ لہذا اس تحقیقی کام کے لئے ڈیوس ڈیٹا مائننگ طریقہ کار پر عمل کیا گیا ہے جیسا کہ ترسیم 4.1 میں دکھایا گیا ہے[135]۔ اس مخصوص طریقہ کار کو استعمال کرنے کی وجہ یہ ہے کہ وہ ہمارے تحقیقی مقاصد کو ظاہر کرتا ہے۔ اس طریقہ کار میں درج ذیل مراحل شامل ہیں:

i. Data Selection: اس مرحلے میں مختلف متضاد ذرائع سے امراض قلب کے متعلقہ اعداد و شمار کو منتخب کیا جاتا ہے اور پھر اسے معیاری ڈیٹابیس میں محفوظ کیا جاتا ہے۔

ii. Data Preparation: اس مرحلے میں امراض قلب کے ڈیٹاسیٹ کا تجزیہ کیا جاتا ہے اور ڈیٹا مائننگ الگورتھم کے لئے ایک مناسب شکل میں تیار کیا جاتا ہے تاکہ اس سے useful insights اور زیادہ سے زیادہ آؤٹ پٹ حاصل لئے۔

iii. Data Task Filter: اس مرحلے میں بعد میں آنے والے اقدامات میں دل کی بیماریوں کی تشخیص کے متوقع نتائج کا تعین کرنے کے لئے Heuristic decision rules استعمال کیے جاتے ہیں۔ اس کے بعد منتخب ڈیٹاسیٹ کو "Data Mining Task Warehouse" میں محفوظ کیا جاتا ہے۔

iv. Data Mining Techniques: اس مرحلہ میں ایک مناسب الگورتھم کا انتخاب ایک مناسب ڈیٹاسیٹ کے ساتھ کیا گیا ہے for the task requested in step 3۔

v. Comparison and Evaluation: اس مرحلے میں درجہ بند نتائج are contrasted and estimated based on different ڈیٹا مائننگ evaluation measures۔

vi. Building New Models: اس مرحلے میں next prediction problems کے لئے ماڈل کو data mine warehouse میں محفوظ کیا جاتا ہے۔ new prediction taks کے لئے عمل کو مرحلہ 3 سے لے کر مرحلہ 5 تک دہرایا جاتا ہے۔

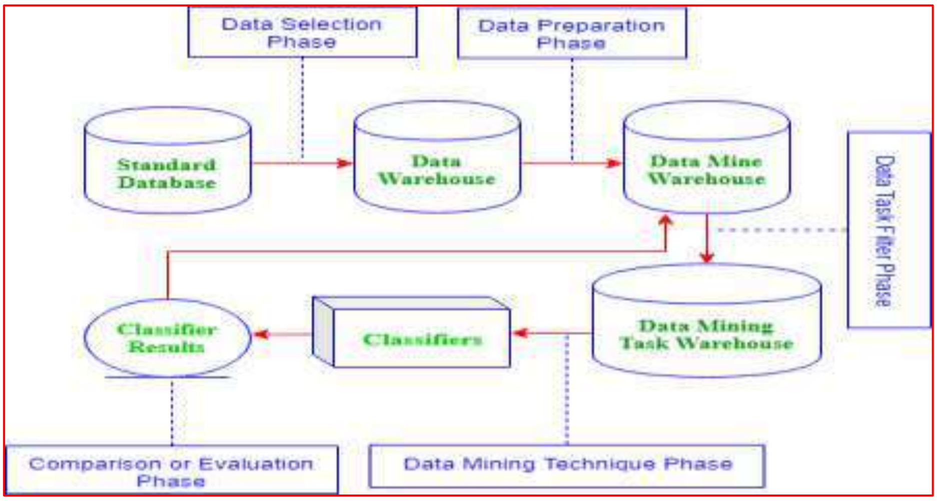

Data Selection Phase
Data Preparation Phase
Standard Database
Data Warehouse
Data Mine Warehouse
Data Task Filter Phase
Classifier Results
Classifiers
Data Mining Task Warehouse
Comparison or Evaluation Phase
Data Mining Technique Phase

Figure 4.1 Heart Disease Risk Evaluation Model Methodology

4.2 ہارٹ ڈیزیز رسک اوالویشن ماڈل کے لئے ریسرچ ڈیزائن

تحقیقی سرگرمیوں کو آسان بنانے اور مقاصد کے بیان سے تحقیق کو نتیجہ خیز بنانے کے لئے تحقیقی ڈیزائن کی پیروی کی جاتی ہے۔ اس میں مطلوبہ اعداد و شمار کے آدانوں، تجزیے کے طریق کار اور تحقیق کے مسئلے کو حل کرنے کے لئے مطالعہ کے مقاصد کا بیان فراہم کیا گیا ہے۔ مجوزہ تحقیقی ڈیزائن (جیسا کہ ذیل میں دیئے گئے ترسیم 4.2 میں دکھایا گیا ہے) ہارٹ ڈیزیز رسک ماڈل کی مرحلہ وار عمل کے تعمیر کے لئے استعمال کیا گیا ہے۔ یہ تحقیقی ڈیزائن تین اہم مراحل پر مشتمل ہے اور اس میں آٹھ اسٹیپس ہیں۔ اس ڈیزائن کے مراحل کی وضاحت مندرجہ ذیل ہے۔

i. Data Phase: ڈیٹا فیز میں ڈیٹا اکٹھا کرنے سے لے کر فیچر انجینئرنگ تک کا سارا عمل شامل ہوتا ہے۔ اس مرحلے میں کوالٹیٹو ڈیٹا اکٹھا کرنے کا مرحلہ، پری پروسیسنگ سب سسٹم مرحلہ، cleaning ڈیٹا سیٹ اسٹوریج مرحلہ اور آخر میں خصوصیت کے انتخاب کا مرحلہ شامل ہے۔

ii. Data Mining Phase: ڈیٹا مائننگ اسٹوریج مرحلے میں درجہ بندی کرنے والے رینڈم فارسٹ، نیوی بائیس، کے نیریسٹ نیبر، سپورٹ ویکٹر مشین اور ڈسیشن ٹری شامل ہیں جو امراض قلب کے خطرے سے متعلق تشخیصی ماڈل تیار کرنے کے لئے کام کریں گے۔

iii. Model Evaluation and Validation Phase: یہ مرحلہ مختلف ڈیٹا مائننگ تکنیکوں کا استعمال کرتے ہوئے ہارٹ ڈیزیز رسک ماڈل کا حساب اور توثیق کرتا ہے۔ ہارٹ ڈیزیز رسک ماڈل کی جانچ پڑتال ٹرین ٹیسٹ split اور 10-fold Cross Validation کے معیارات کا استعمال کرتے ہوئے کیا جائے گا. اس کے بعد ماڈل کو موجودہ کارکردگی اور مختلف ماڈل میٹرکس کے ساتھ نتائج کا موازنہ کرکے کارکردگی کے مختلف اقدامات (Sensitivity, Specificity, Accuracy, Error Rate, and ROC Curve Score) کا استعمال کرتے ہوئے توثیق کی جائے گی۔

iv. Knowledge Base Phase: نالج بیس مرحلے میں امراض قلب کے بارے میں معلومات کو ذخیرہ کرنے اور retreive کرنے کے اقدامات شامل ہیں۔ دل کی بیماریوں سے پیدا ہونے والے خطرے کی تشخیص کے قواعد کو knowledge-base میں محفوظ کیا جائے گا۔ The genearated heart disease قواعد کو میڈیکل ہدایات کے مطابق اور ڈومین مہارت کے ذریعہ جانچ پڑتال کی جائے گی۔

v. اس باب میں پہلے مرحلے (qualitative data collection, pre-processing subsystem and feature selection) پر تبادلہ خیال کیا گیا ہے اور باقی مرحلوں پر بحث کی جائے گی جب بھی تحقیق میں مطالبہ کیا جائے گا۔

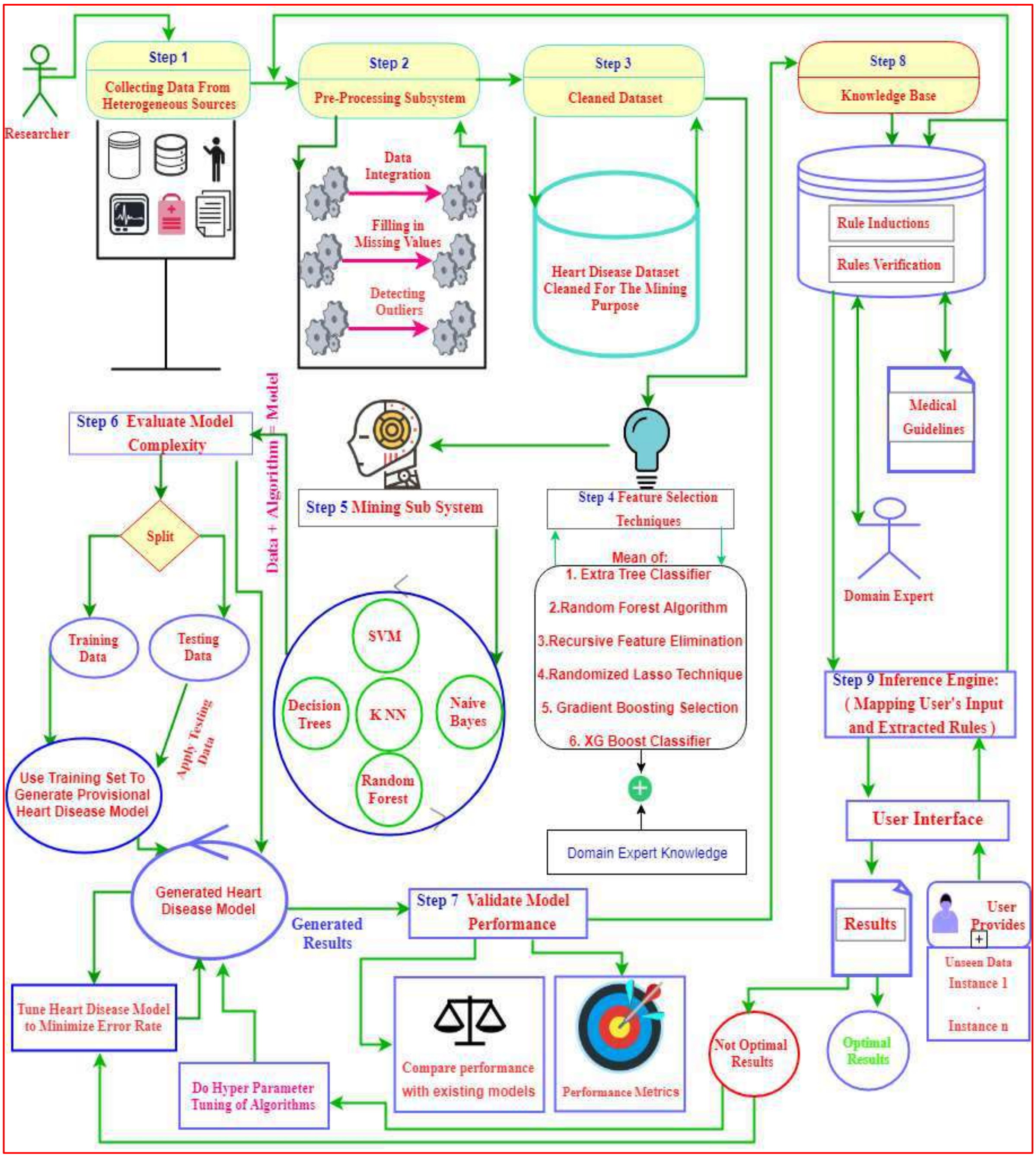

Researcher
Step 1
Collecting Data From Heterogeneous Sources
Step 2
Pre-Processing Subsystem
Data Integration
Filling in Missing Values
Detecting Outliers
Step 3
Cleaned Dataset
Heart Disease Dataset Cleaned For The Mining Purpose
Step 8
Knowledge Base
Rule Inductions
Rules Verification
Medical Guidelines
Domain Expert
Step 6 Evaluate Model Complexity
Data + Algorithm = Model
Split
Training Data
Testing Data
Apply Testing Data
Use Training Set To Generate Provisional Heart Disease Model
Generated Heart Disease Model
Step 5 Mining Sub System
SVM
Decision Trees
K NN
Naive Bayes
Random Forest
Step 4 Feature Selection Techniques
Mean of:
1. Extra Tree Classifier
2.Random Forest Algorithm
3.Recursive Feature Elimination
4.Randomized Lasso Technique
5. Gradient Boosting Selection
6. XG Boost Classifier
Domain Expert Knowledge
Step 9 Inference Engine: (Mapping User's Input and Extracted Rules)
User Interface
Results
User Provides
Unseen Data Instance 1
.
Instance n
Optimal Results
Not Optimal Results
Generated Results
Step 7 Validate Model Performance
Compare performance with existing models
Performance Metrics
Tune Heart Disease Model to Minimize Error Rate
Do Hyper Parameter Tuning of Algorithms

Figure 4.2 Detailed Steps of Research Design

4.3 Exploratory ڈیٹا انالیسس پراسیس

بنیادی اعداد و شمار کی تفصیل کشمیر ہارٹ ڈیزیز ڈیٹاسیٹ کی ہر ایک خصوصیت کی قیمت کے بارے میں جاننے کے لئے انجام دی جاتی ہے۔ ہر ایک وصف کے بارے میں اس طرح کے بنیادی ڈیٹا کو جاننے سے سموتھ نویسے ڈیٹا ، سپاٹ اوٹلیرس اور fill in the missing values میں مدد ملتی ہے۔ ہارٹ ڈیزیز ڈیٹاسیٹ numeric اور categorical ڈیٹا کے امتزاج پر مشتمل ہے۔ Numeric missing values کو simple mean imputation طریقہ کار کے ذریعہ ختم کیا جاتا ہے اور categorical missing values کو mode imputation technique سے بھر دیا جاتا ہے [137] [136]۔

4.3.1۔ ڈیٹاسیٹ میں کلاس عدم توازن اور ڈیٹا کی تقسیم کے مسائل کی جانچ پڑتال

ہارٹ ڈیزیز ڈیٹاسیٹ پر کسی بھی قسم کی کارروائی کرنے سے پہلے طبقاتی توازن کی جانچ کی ضرورت ہوتی ہے کیونکہ انتہائی عدم توازن والے ڈیٹا مشین لرننگ الگورتھم کو biased بناتی ہے۔ کلاس توازن کی جانچ پڑتال کے لئے، اعداد و شمار پر skewness اور kurtosis آپریشن کئے جاتے ہیں [139][138]۔ اسکیونیس اس توازن کا اندازہ لگاتا ہے کہ آیا ڈیٹا کی تقسیم سینٹر پوائنٹ کے بائیں اور دائیں کے برابر ہے۔ اور kurtosis اقدام کرتا ہے کہ whether data is light-tailed or heavy-tailed to normal distribution. -اسکیونیس اور کرٹوسس کے ڈیٹا پیمائش کے ٹیسٹ کے بعد یہ پتہ چلا ہے کہ ہارٹ ڈیزیز ڈیٹاسیٹ متوازن ہے جس میں اسکیونیس (-0.03065287) اور کرٹوسس (-2.000136)) ہے جس کا مطلب ہے کہ کشمیر ہارٹ ڈیزیز ڈیٹا سیٹ is normally distributed اور ڈیٹاسیٹ میں 5776 ریکارڈ موجود ہیں جن میں 2760 خواتین اور 3016 مرد ہیں۔ ان 5776 instances میں، 2745 (47.5%) مرض قلب ہیں اور 3031 (52.5%) صحت مند ہیں [141] [140] ۔

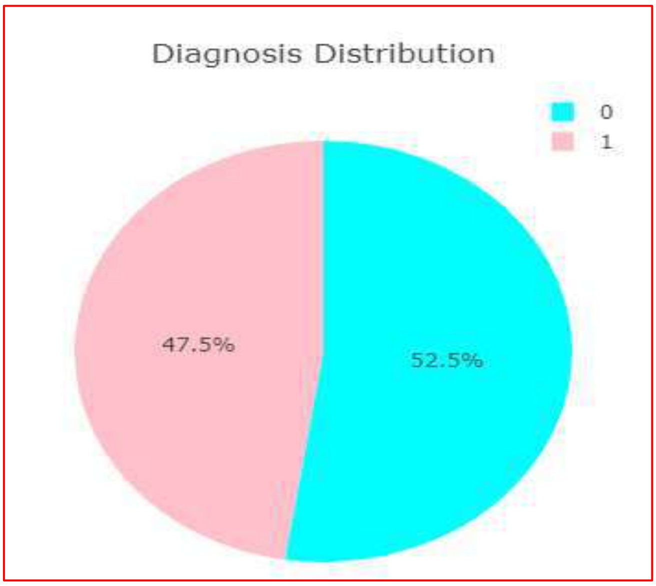


Figure 4.3 Heart Disease Distributions Based On Sex Attribute

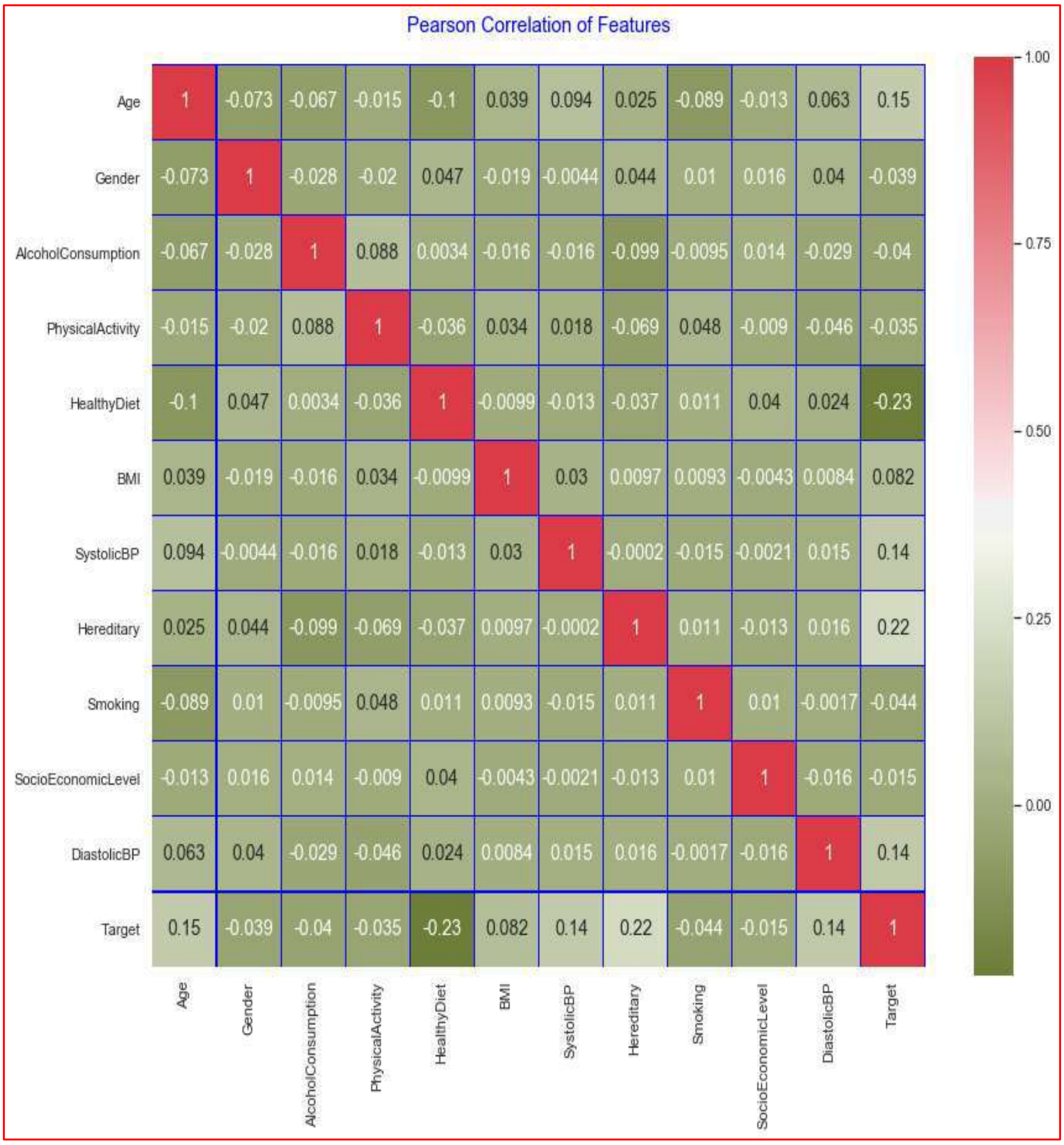
ذیل میں دیئے گئے ترسیم 4.3 ہارٹ ڈیزیز ڈیٹا کی تقسیم کی پائی چارٹ کی نمائندگی کو ظاہر کرتی ہے۔ امراض قلب مردوں اور عورتوں کو تقریباایک ہی تناسب سے متاثر کرتی ہے جس میں موت اور معذوری کی کافی شرح ہوتی ہے۔ امراض قلب کی درست پیش گوئی کرنااور ان کا پتہ لگانا کئی بنیادی وجوہات کی تشکیل کرتا ہے جیسے سماجی، تجارتی اور ثقافتی منتقلی۔ ان خطرات سے وابستہ افراد کے بارے میں طویل مدتی انکشاف کرنا سب سے مشکل پر اثر انداز ہوتا ہے اور اس کی موت تک ہوتی ہے۔ صحت کی اطلاعات سے پتہ چلتا ہے کہ اگر behavioural risk factors کو تبدیل نہ کیا گیا تو یہ بیماری بڑھتی رہے گی اور اس سے انسانی اور معاشی نقصان ہوگا۔

4.3.2۔ مختلف ہارٹ ڈیزیز رسک فیچرز میں correlations کی تلاش

کسی بھی ڈیٹاسیٹ میں متغیر کے درمیان کثیر جہتی اور عجیب وغریب relationships ہو سکتے ہیں لہٰذا یہ ضروری ہے کہ ڈیٹاسیٹ میں ایک دوسرے سے وابستہ ہونے والی ڈگری کا تعین اور پیمائش کیا جاے۔ ڈیٹاسیٹ اوصاف کے مابین تعلقات کی ڈگری تلاش کرنے کے اس عمل کو correlation کے نام سے جانا جاتا ہے۔ صفات کے مابین correlation کا علم اعداد وشمار کو بہتر طریقے سے تیار کرنے میں مدد کرتا ہے تاکہ مشین لرننگ الگورتھم کی توقعات کو پورا کیا جاسکے۔ کشمیر ہارٹ ڈیزیز ڈیٹاسیٹ کے درمیان باہمی تعلقات کی جانچ پڑتال کرنے کے لئے pearson’s correlation کو استعمال کیا گیا۔ ایک correlation مثبت ہو سکتا ہے (جس کا مطلب ہے کہ تمام متعلقہ صفات ایک ہی سمت میں حرکت پذیر ہوتی ہیں)، منفی (جس کا مطلب ہے کہ تمام متعلقہ صفات مخالف سمتوں میں حرکت پذیر ہوتی ہیں) یا neutral (جس کا مطلب یہ ہے کہ اوصاف ایک دوسرے سے متعلق نہیں ہیں) ہو سکتا ہے۔ دل کی بیماری کے متغیرات کے درمیان pearson’s coefficient correlation کا نتیجہ ہیٹ میپ کی نمائندگی کی شکل میں نیچے دیئے گئے ترسیم 4.4 میں دکھایا گیا ہے۔

ہیٹ میپ گرڈ دل کی بیماری کے اوصاف کے مابین corresponding correlation کی نمائندگی کرتا ہے۔ ہیٹ میپ سمیٹرک میٹرکس تمام خصوصیات کی نمائندگی across the top and down the side کرتا ہے جس میں تمام pair of attributes کے درمیان correlation شامل ہیں۔ The diagonal line across the matrix from the bottom-right corner to the top-left ہر ایک خصوصیت کا اپنے ساتھ کامل ارتباط کی نمائندگی کرتی ہے۔

value 1 کا مطلب صفات کے مابین کامل مثبت correlation ہے اور ویلیو -1 کا مطلب دل کی بیماری کے اوصاف کے مابین کامل منفی تعلق ہے۔ صفر کے قریب value دل کی بیماری کے اوصاف کے درمیان کمزور انحصار کی نشاندہی کرتا ہے۔



	Age	Gender	AlcoholConsumption	PhysicalActivity	HealthyDiet	BMI	SystolicBP	Hereditary	Smoking	SocioEconomicLevel	DiastolicBP	Target
Age	1	-0.073	-0.067	-0.015	-0.1	0.039	0.094	0.025	-0.089	-0.013	0.063	0.15
Gender	-0.073	1	-0.028	-0.02	0.047	-0.019	-0.0044	0.044	0.01	0.016	0.04	-0.039
AlcoholConsumption	-0.067	-0.028	1	0.088	0.0034	-0.016	-0.016	-0.099	-0.0095	0.014	-0.029	-0.04
PhysicalActivity	-0.015	-0.02	0.088	1	-0.036	0.034	0.018	-0.069	0.048	-0.009	-0.046	-0.035
HealthyDiet	-0.1	0.047	0.0034	-0.036	1	-0.0099	-0.013	-0.037	0.011	0.04	0.024	-0.23
BMI	0.039	-0.019	-0.016	0.034	-0.0099	1	0.03	0.0097	0.0093	-0.0043	0.0084	0.082
SystolicBP	0.094	-0.0044	-0.016	0.018	-0.013	0.03	1	-0.0002	-0.015	-0.0021	0.015	0.14
Hereditary	0.025	0.044	-0.099	-0.069	-0.037	0.0097	-0.0002	1	0.011	-0.013	0.016	0.22
Smoking	-0.089	0.01	-0.0095	0.048	0.011	0.0093	-0.015	0.011	1	0.01	-0.0017	-0.044
SocioEconomicLevel	-0.013	0.016	0.014	-0.009	0.04	-0.0043	-0.0021	-0.013	0.01	1	-0.016	-0.015
DiastolicBP	0.063	0.04	-0.029	-0.046	0.024	0.0084	0.015	0.016	-0.0017	-0.016	1	0.14
Target	0.15	-0.039	-0.04	-0.035	-0.23	0.082	0.14	0.22	-0.044	-0.015	0.14	1

Figure 4.4 Correlation in Risk Attributes Through Heatmap Representation

ہیٹ میپ correlation کے نتائج کا تجزیہ کرنے کے بعد یہ پتہ چلا ہے کہ ہارٹ ڈیزیز ڈیٹا سیٹ کی آزاد صفات ایک دوسرے کے ساتھ ڈھیلی پڑتی ہیں۔ امراض قلب کی آزاد خوبیوں کے درمیان یہ ڈھیلے باہمی تعلق ماڈل کی کارکردگی کو بہتر بنانے کے لئے ایک اچھی علامت ہے۔ تاہم اگر کسی ڈیٹا سیٹ میں صفات کو مضبوطی سے باہمی ربط کیا جاتا ہے (جسے Multi-colinearity کہا جاتا ہے) تو پھر ایک متغیر میں تبدیلی سے دوسرے متغیر میں تبدیلی آسکتی ہے جو الگورتھم کی کارکردگی کو خراب کر سکتی ہے۔ صفات کے مابین correlation کا مطلب یہ نہیں ہے

کہ اس وجہ سے صفات کے مابین مضبوط رشتہ کا نمایاں جائزہ لیا جانا چاہئے۔ کچھ نظر انداز عوامل کی وجہ سے اوصاف کے مابین مضبوط تعلقات میں قوی وابستگی نظر آسکتی ہے۔

4.4۔ ہارٹ ڈیزیز رسک اوالویشن کے لئے فیچر سلیکشن ٹیکنیکس کا نمایاں انتخاب

اگرچہ امراض قلب نے وبا کی حد کو حاصل کر لیا ہے لیکن اس کے بعد بھی خطرے کے عوامل کو کم کر کے قابو پایا جاسکتا ہے۔ وقت سے پہلے ہارٹ ڈیزیز رسک اوالویشن کرنا غلط مسائل سے پاک نہیں ہے اور غیر یقینی عوامل کی تعداد تشکیل دیتا ہے۔ فیچر سلیکشن کے طریقوں کا استعمال مختلف ہارٹ ڈیزیز رسک فیچرز کے اہم اور انتہائی مناسب سب سیٹ کو منتخب کرنے کے لئے کیا جاتا ہے۔ خصوصیت کا انتخاب نامناسب اور بیکار اوصاف کو کم کرنے میں مدد کرتا ہے، جو اکثر درجہ بندی کرنے والوں کی کارکردگی کو کم کرتا ہے۔ اس تحقیق میں ہارٹ ڈیزیز رسک اوالویشن کے لئے مناسب خصوصیت کا سبسیٹ حاصل کرنے کے لئے تین مختلف قسم کی فیچر سلیکشن تکنیکس (فلٹر، ریپر اور ایمبیڈڈ) کا اطلاق کیا گیا ہے۔

پانچ مختلف فیچر سلیکشن تکنیکس (ایکسٹرا ٹری کلاسیفائر، گراڈینٹ بوسٹنگ کلاسیفائر، رینڈم فاریسٹ، ریکورسیف فیچر Elimination، اور ایکس جی بوسٹ کلاسفائیر) بہترین نان انویسو رسک فیچرز سبسیٹ حاصل کرنے کے لئے ہارٹ ڈیزیز ڈیٹا سیٹ (جیسا کہ جدول 4.1 میں بتایا گیا ہے) پر لاگو کیا جاتا ہے۔ بیماریوں کی تشخیص میں ان کے کردار کے مطابق ہر خطرے کی خصوصیت ان فیچر سلیکشن تکنیکس سے weight کی جاتی ہے۔ خصوصیت کے انتخاب کی یہ تکنیک دل کی بیماریوں کے ہر خطرے سے منسوب ہونے کے لئے 0 اور 1 کے پیمانے کے درمیان weight فراہم کرتی ہیں۔ علیحدہ خصوصیت کے انتخاب کی تکنیک کے ذریعہ ہر خطرے سے associated weight کو تفویض کرنے کے بعد، ان خصوصیت کے انتخاب کی تکنیک کے ذریعہ ہر خاصیت کے لئے لگائے جانے والے تمام weight کا مجموعی مطلب آخری weight سمجھا جاتا ہے۔ دل کی بیماری کی تشخیص کے لئے قریب قریب سب سے زیادہ قیمت والی وابستگی کو اہم قرار دیا جاتا ہے اور ان خطرات سے وابستہ افراد جن کی وابستہ قیمتیں صفر (0) کے قریب ہیں دل کی بیماری کی پیش گوئی کرنے میں کم اہم سمجھی جاتی ہیں۔ ذیل میں دیئے گئے جدول 4.2 میں مختلف خصوصیت کے انتخاب کی تکنیک کے ذریعہ تفویض کردہ ان کے متعلقہ weight کے ساتھ دل کی بیماری کی مختلف خصوصیات سے پتہ چلتا ہے اور ٹیبل میں آخری کالم تمام تکنیکوں کا مجموعی مطلب ظاہر کرتا ہے۔ ان خطرات کی خصوصیات پیشہ ورانہ امراض قلب کے ماہرین صوفیہ ایئر ہارٹ (اسسٹنٹ پروفیسر)، اور انڈیا کے مختلف اسپتالوں میں محکمہ برائے امراض قلب میں کام کرنے والے بہت سے دوسرے عمومی معالجین

کی نشاندہی کرتے ہیں۔ دل کی بیماریوں کے ہر خطرے سے منسوب weight مختلف طبّی ماہرین کی توثیق اور منظوری کے ذریعہ دیتے ہیں جیسے ڈاکٹر وی جے سادھنا (ایم ڈی۔ ، فاہا میڈیکل ڈائریکٹر)، ڈاکٹر سید اعجاز ناصر (ایمز میں امراض قلب) اور ڈاکٹر محمد خطیب الدین انصاری (MANUU میں میڈیکل آفیسر) دریں اثنا، ان میڈیکل ڈومین کے ماہرین نے دل کی بیماری کی پیشگوئی اور شناخت کے لئے کچھ اہم خصوصیات (سینے میں درد، اور دمہ) کی طرح شامل کرنے کے لئے اپنی اپنی رائے دی۔

Table 4.1 Feature Selection Techniques with their Mean Value on Risk Attributes

Attributes	Feature Selection Techniques with Their Results and Mean Values					
	ETC	GBC	RF	RFE	XGB	MEAN
Age	0.92	0.92	0.87	0.25	0.92	0.78
Sex	0.0	0.0	0.11	0.83	0.0	0.19
Alcohol Consumption	0.09	0.09	0.09	0.75	0.09	0.22
Physical Activity	0.25	0.25	0.08	0.67	0.25	0.30
Healthy Diet	0.71	0.71	0.52	1.0	0.71	0.73
BMI	0.74	0.74	0.79	0.0	0.74	0.60
Hereditary	0.38	0.38	0.4	0.92	0.38	0.49
Smoking	0.17	0.17	0.09	0.5	0.17	0.22
Systolic BP	1.0	1.0	1.0	0.08	1.0	0.82
Diastolic BP	0.88	0.88	0.78	0.33	0.88	0.75
Socio Economic Level	0.17	0.17	0.11	0.42	0.17	0.21

نتائج کا تجزیہ کرنے کے بعد یہ نکلا ہے کہ صفات [سسٹولک بی پی، ڈیاسٹولک بی پی، ایج، بی ایم آئی، موروثی، صحت بخش غذا اور جسمانی سر گرمی] امراض قلب کی ابتدائی پیش گوئی کے لئے سب سے اہم خصوصیات ہیں کیونکہ ان کی متعلقہ عددی اقدار زیادہ ہیں اور میڈیکل ڈومین کے ماہرین کے ذریعہ ان کی توثیق اور منظوری بھی دی جاتی ہے۔ ان کے متعلقہ وزن کے ساتھ وابستگی کے درجات کی صیغی نمایندگی ذیل میں دیئے گئے figure 4.5 میں دکھائی گئی ہے۔

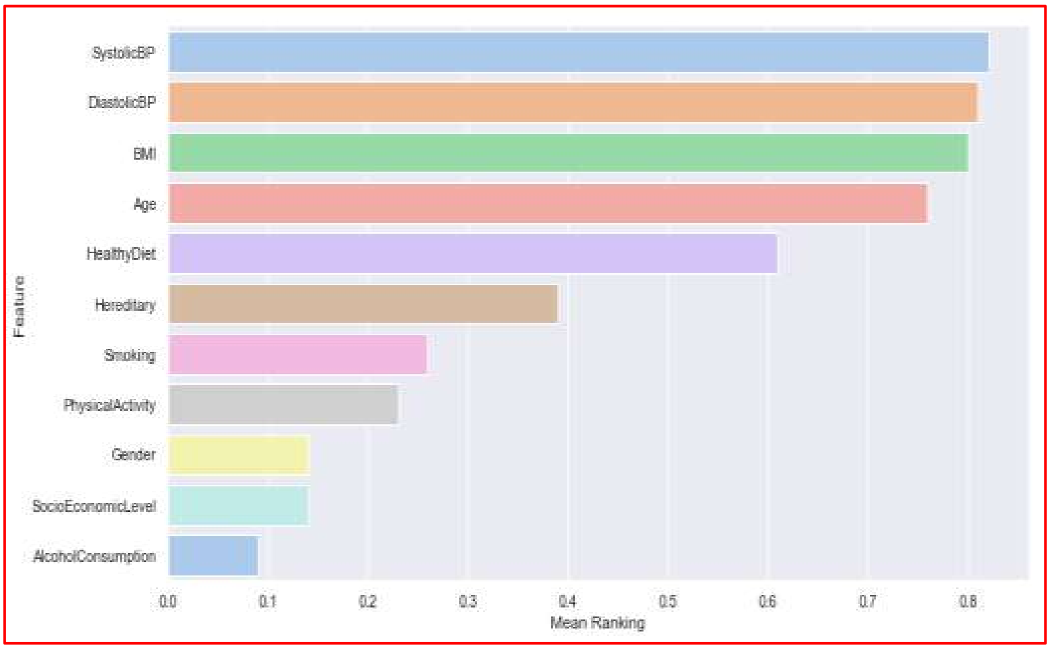


Figure 4.5 Risk Attribute Hierarchy by Feature Selection Techniques

Table 4.2 Mean Ranking of Risk Attributes using Feature Selection Techniques

Sr. No	Attributes	Mean Ranking of Attributes
1	Systolic BP	0.82
2	Diastolic BP	0.80
3	BMI	0.78
4	Age	0.76
5	Healthy Diet	0.54
6	Hereditary	0.42
7	Smoking	0.28
8	Physical Activity	0.24
9	Socio-Economic Level	0.16
10	Sex	0.14
11	Alcohol Consumption	0.12

دیئے گئے جدول 4.2 میں پانچ مختلف خصوصیت کے انتخاب کی تکنیکوں کے ذریعہ تفویض کردہ ان کی بنیادی قدروں کے مطابق دل کی بیماری کے اوصاف کا نزول ترتیب ظاہر ہوتا ہے۔ سب سے زیادہ وزن والی صفات انتہائی اہم ہیں اور جدول کے نیچے دیئے گئے صفات دل کی بیماری کی جلد سے

جلد پیش گوئی کرنے میں کم اہمیت رکھتے ہیں۔ خطرے کی خصوصیات کا انتہائی وزن والا اہم ذیلی سیٹ ہارٹ ڈیزیز رسک اوالویشن ماڈل تیار کرنے کے لئے استعمال ہوتا ہے۔

4.5 تجویز کردہ ڈیٹا مائننگ تراکیب کے تجرباتی نتائج

یہ پایا گیا ہے کہ موجودہ ہارٹ ڈیزیز رسک اوالویشن ماڈل مسائل سے آزاد نہیں ہیں کیونکہ وہ مختلف ڈیٹا سیٹس میں مختلف نتائج دکھاتے ہیں جو نظام کی تاثیر کو بہت کم کرتے ہیں. اس تحقیق میں ہارٹ ڈیزیز ڈیٹا سیٹ کا رینڈم فارسٹ، نیوی بائیس، کے نیریسٹ نیبر، سپورٹ ویکٹر مشین اور ڈیسیشن ٹری تکنیک کے ذریعہ استعمال کیا جاتا ہے اور 10-fold cross validation کے ذریعے غیر جانبدارانہ نتائج حاصل کرنے کے لئے تیار کیا گیا ہے۔ زیادہ سے زیادہ اور درست نتائج حاصل کرنے کے لے میڈیکل ڈومین کی کارکردگی کے مختلف اقدامات جیسے Sensitivity, Specificity, Accuracy ، AUROC اسکور اور Misclassification کی شرح اور کمپیوٹیشنل پیچیدگی اور فہمیت جیسے ماڈل اقدامات کا حساب لگایا جاتا ہے۔ ذیل میں دیئے گئے ذیلی حصوں میں امراض قلب کے مختلف خطرہ تشخیصی ماڈلز کے ذریعے حاصل کردہ تجرباتی نتائج کی وضاحت کی گئی ہے۔

4.5.1 ڈیسیشن ٹری ماڈل کے تجرباتی نتائج

ڈیسیشن ٹری کو استعمال کرنے کا مقصد ہارٹ ڈیزیز رسک اوالویشن ماڈل تیار کرنا ہے جو training data سے pruned اصولوں کو سیکھ کر کسی طبقے کی پیش گوئی کر سکتا ہے۔ تربیت ڈیٹا سیٹ پر cross validation غیر جانبدارانہ نتائج حاصل کرنے کے لئے استعمال کی جاتی ہے۔ ڈیسیشن ٹری ماڈل کی کارکردگی کے نتائج کنفیوژن میٹرکس کے ترسیم 4.6 میں دکھائے گئے ہیں۔ ترسیم 4.6 سے ہم True Positive Rate, True Negative Rate, Precision, Accuracy, Error Rate and AUROC اخذ کرتے ہیں جس کی تفصیل ذیل میں دی گئی ہے۔

کنفیوژن میٹرکس 4.6 میں مساوات (3.11) کا استعمال کرتے ہوئے ہمیں ڈیسیشن ٹری ماڈل کی sensitivity 0.826% حاصل ہوئی ہے ۔ اسی طرح ، کنفیوژن میٹرکس 4.5 میں مساوات (3.12) کا استعمال کرتے ہوئے ہمیں ڈیسیشن ٹری ماڈل کی True Negative Rate 0.80 فیصد حاصل ہوئی ہے جس کا مطلب ہے کہ ڈیولپڈ ڈیسیشن ٹری ماڈل صحت مندون کو% 80 کی درستگی

کے ساتھ پہچان سکتا ہے۔ ڈسیشن ٹری ماڈل کی Accuracy کنفیوژن میٹرکس 4.6 میں مساوات (3.13) کا استعمال کرکے حاصل کی گئی ہے جو حساب کے بعد 0.8185% کے برابر ہے اس میں یہ بتایا گیا ہے کہ مرض اور صحت مند دونوں معاملات کی تشخیصی میں ڈسیشن ٹری ماڈل کی مجموعی کارکردگی 81% ہے۔ اسی طرح ڈسیشن ٹری ماڈل کی precision حاصل کرنے کے لئے مساوات (3.14) استعمال ہوا ہے جو کہ 0.84 % ہے۔ ڈیولپڈ ڈسیشن ٹری ماڈل کی error rate مساوات (3.15) کا استعمال کرکے حاصل ہویے جو 0.18 % کے برابر ہے۔

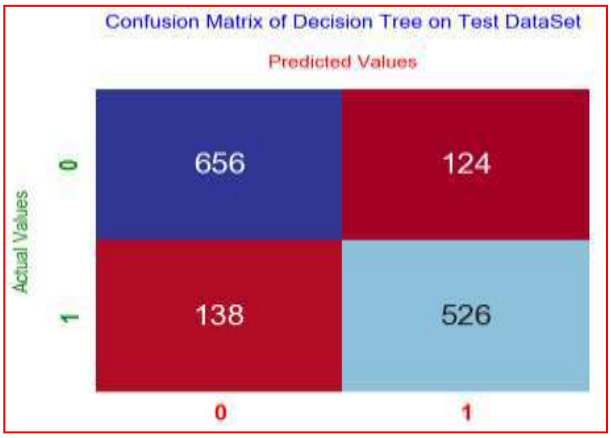


Figure 4.6 Decision Tree Model Confusion Matrix

AUROC کارکردگی کی پیمائش کا استعمال یہ دیکھنے کے لئے کیا جاتا ہے کہ ماڈل مریض اور صحت مندوں میں کتنا اچھا فرق کر سکتا ہے۔ بہتر ماڈل بیمار اور غیر مریض متاثرین میں بالکل مختلف ہے تاہم ناقص ماڈلز کو دونوں کے درمیان فرق کرنے میں دشواری کا سامنا کرنا پڑتا ہے۔ ذیل میں دیئے گئے ترسیم 4.7 کے تحت AUROC ڈسیشن ٹری ماڈل سے حاصل کیا گیا ہے جس کی AUROC اسکور 0.817% ہے۔ False Positive کے خلاف True Positive پلاٹ سازی منحنی خطوط کے تحت کا علاقہ ان ماڈلز کے لئے زیادہ بہتر ہے جو مثبت اور منفی معاملات کی صحیح شناخت کرنے کے قابل ہیں۔

ہم مروجہ تحقیق کے ساتھ مطلوبہ ہارٹ ڈیزیز ڈسیشن ٹری ماڈل کے کامیاب تجرباتی نتائج کی نقالی کرتے ہیں۔ ہمارے مشاہدے کے متبِک حاصل کردہ نتائج ادب میں شائع شدہ نتائج سے کہیں زیادہ بہترین ہیں۔ لہذا ہم امراض قلب کے مریضوں کی پیش گوئی کے لئے ڈیولپڈ ڈسیشن ٹری ماڈل کا استعمال کرتے ہیں البتہ ماڈل کی کارکردگی میں مزید بہتری کی ضرورت ہے۔ تاہم امراض قلب کے ڈیٹا سے اخذ کردہ ڈسیشن ٹری rules پیچیدہ ہیں جو ہارٹ ڈیزیز رسک اوالویشن ماڈل کے وقت کی پیچیدگی میں اضافہ کرتے ہیں اور ماڈل کو سست بناتے ہیں۔

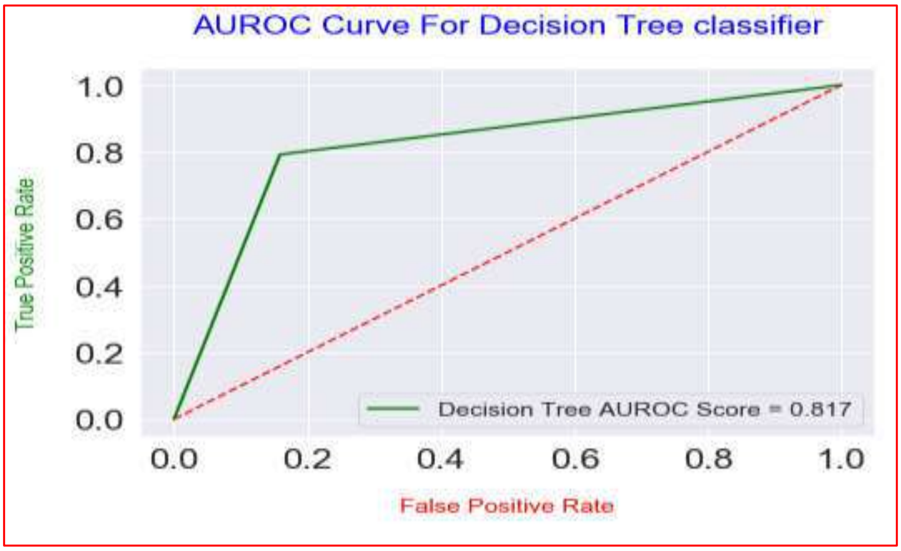


Figure 4.7 AUROC by the Decision Tree Model

4.5.2 کے نیریسٹ نیبر ماڈل کے تجرباتی نتائج

ہم نے کارڈیک ڈس آرڈر کے مریضوں کی ابتدائی پیش گوئی اور شناخت کے لئے کے نیریسٹ نیبر کلاسفائر کا اطلاق کیا۔ ماڈل کی کارکردگی کے نتائج کنفیوژن میٹرکس 4.8 ترسیم میں دکھائے گئے ہیں۔

کنفیوژن میٹرکس 4.8 میں مساوات (3.11) کا استعمال کرتے ہوئے ہمیں کے نیریسٹ نیبر ماڈل کی sensitivity 0.738% حاصل ہوئی ہے اور جس کا مطلب ہے کہ ڈیولپڈ کے نیریسٹ نیبر ماڈل 73% کی درستگی کے ساتھ دل کے امراض کے مثبت واقعات کو پہچان سکتا ہے۔ اسی طرح کنفیوژن میٹرکس 4.8 میں مساوات (3.12) کا استعمال کرتے ہوئے ہمیں کے نیریسٹ نیبر ماڈل کی specificity 0.66 فیصد کے برابر حاصل ہوئی ہے اور جس کا مطلب ہے کہ ڈیولپڈ کے نیریسٹ نیبر ماڈل صحت مندون کو 66% کی درستگی کے ساتھ پہچان سکتا ہے۔ کے نیریسٹ نیبر ماڈل کی Accuracy کنفیوژن میٹرکس 4.8 میں مساوات (3.13) کا استعمال کر کے حاصل کی گئی ہے جو حساب کے بعد 0.6980% کے برابر ہے اس میں یہ بتایا گیا ہے کہ مرض اور صحت مند دونوں معاملات کی تشخیصی میں کے نیریسٹ نیبر ماڈل کی مجموعی کارکردگی 69% ہے۔ اسی طرح کے نیریسٹ نیبر ماڈل کی precision حاصل کرنے کے لئے مساوات (3.14) استعمال کی جاتی ہے جو کہ تشخیصی کے بعد % 0.69 ہے اس کا مطلب یہ ہے کہ کے نیریسٹ نیبر ماڈل میں low false positive rateہے۔ ڈیولپڈ کے نیریسٹ نیبر ماڈل کی misclassification rate مساوات (3.15) کا استعمال کر کے حاصل کی جاتی ہے جو % 0.15 کے برابر ہے۔

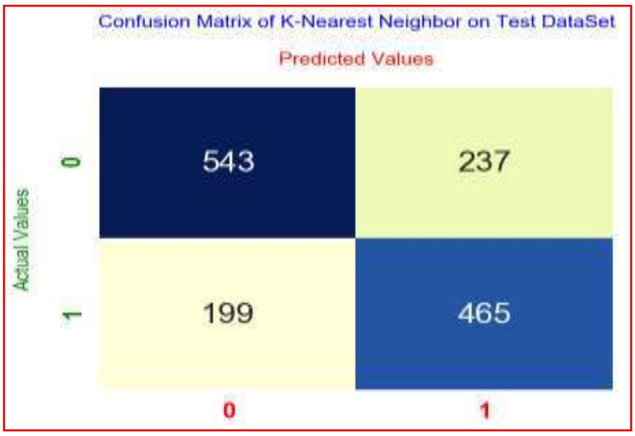


Figure 4.8 K Nearest Neighbor Confusion Matrix on Test Dataset

AUROC کارکردگی کی پیمائش کا استعمال یہ دیکھنے کے لئے کیا جاتا ہے کہ ڈیولپڈ ماڈل مریض اور صحت مندوں میں کتنا اچھا فرق کر سکتا ہے۔ بہتر ماڈل بیمار اور غیر مریض متاثرین میں بالکل مختلف ہے تاہم ناقص ماڈلز کو دونوں کے درمیان فرق کرنے میں دشواری کا سامنا کرنا پڑتا ہے۔ ذیل میں دیئے گئے ترسیم 4.9 کے تحت AUROC ڈیولپڈ کے نیریسٹ نیبر ماڈل سے حاصل کیا گیا ہے جس کی AUROC اسکور 0.70% ہے۔

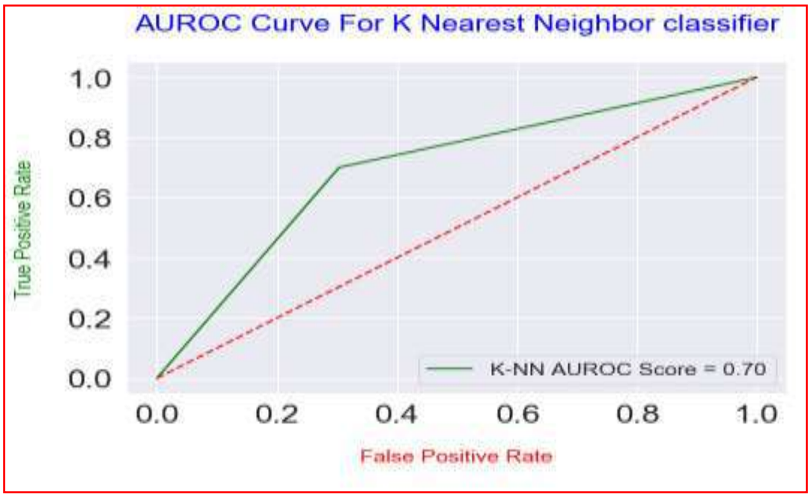


Figure 4.9 AUROC K Nearest Neighbor Model

ہم مروجہ تحقیق کے ساتھ کے این این ہارٹ ڈیزیز رسک اوالویشن ماڈل کے کامیاب تجرباتی نتائج کی نقالی کرتے ہیں۔ نتائج سے پتہ چلتا ہے کہ کے این این ماڈل ہارٹ ڈیزیز کی پیشن گوئی کے لئے زیادہ مناسب نہیں ہے کیوں کہ غلط درجہ بندی کی شرح زیادہ ہے۔ میڈیکل ڈومین

پرفارمنس کے علاوہ ماڈل کی کارکردگی کے اقدامات کا حساب لگا کر یہ پتا چلا کی وہ زیادہ ہے۔ غلط درجہ بندی کی شرح اور ماڈل کی پیچیدگی کے عوامل کی اعلی اقدار اس کے اطلاق کو روکتی ہیں کیونکہ میڈیکل پیشن گوئی کے ماڈل کو زیادہ سے زیادہ پیشن گوئی کی درستگی کو پورا کرنا ہوتا ہے اور ایک بھی غلط تشخیصی موت جیسے سنگین نتائج کا باعث بن سکتا ہے۔

4.5.3 سپورٹ ویکٹر مشین ماڈل کے تجرباتی نتائج

سپورٹ ویکٹر مشین ہائپر پلین کی بنیاد پر ڈیٹا کو کلاسوں میں جدا کرتی ہے جس میں زیادہ سے زیادہ مارجن ہوتا ہے اور سپورٹ ویکٹر بنانے کے لئے کلاسوں کے مابین زیادہ سے زیادہ ہائپر پلین تلاش کیا جاتا ہے۔ اس تحقیق میں ایس وی ایم کا استعمال ایک رسک ماڈل تیار کرنے کے لئے کیا جاتا ہے جو ابتدائی مرحلے میں ہی ہارٹ ڈیزیز کی پیش گوئی کر سکتا ہے۔ ہارٹ ڈیزیز ڈیٹاسیٹ پر سپورٹ ویکٹر مشین ماڈل کی کارکردگی کے نتائج کنفیوژن میٹرکس کے ترسیم 4.10 میں دکھائے گئے ہیں اور اس سے sensitivity، specificity، AUROC,accuracy، precision اور misclassification اخذ کی گئی ہیں جس کی تفصیل ذیل میں دی گئی ہے۔ ہم نے کارڈیک ڈس آرڈر کے مریضوں کی ابتدائی پیش گوئی اور شناخت کے لئے سپورٹ ویکٹر مشین کلاسفائر کا اطلاق کیا۔ ماڈل کی کارکردگی کے نتائج کنفیوژن میٹرکس 4.10 کے ترسیم میں دکھائے گئے ہیں۔

کنفیوژن میٹرکس 4.10 میں مساوات (3.11) کا استعمال کرتے ہوئے ہمیں سپورٹ ویکٹر مشین ماڈل کی sensitivity 0.825% حاصل ہوئی ہے اور جس کا مطلب ہے کہ ڈیولپڈ سپورٹ ویکٹر مشین 82 % کی درستگی کے ساتھ دل کے امراض کے مثبت واقعات کو پہچان سکتا ہے۔ اسی طرح کنفیوژن میٹرکس 4.10 میں مساوات (3.12) کا استعمال کرتے ہوئے ہمیں سپورٹ ویکٹر مشین ماڈل کی 0.815 specificity فیصد حاصل ہوئی ہے اور جس کا مطلب ہے کہ ڈیولپڈ سپورٹ ویکٹر مشین ماڈل صحت مندون کو 81% کی درستگی کے ساتھ پہچان سکتا ہے۔ سپورٹ ویکٹر مشین ماڈل کی Accuracy کنفیوژن میٹرکس 4.10 میں مساوات (3.13) کا استعمال کر کے حاصل کی گئی ہے جو 0.82% کے برابر ہے اس میں یہ بتایا گیا ہے کہ مرض اور صحت مند دونوں معاملات کی تشخیص میں سپورٹ ویکٹر مشین ماڈل کی مجموعی کارکردگی 82% ہے۔ اسی طرح سپورٹ ویکٹر مشین ماڈل کی precision حاصل کرنے کے لئے مساوات (3.14)

استعمال کی جاتی ہے جو کہ تشخیص کے بعد % 0.84 ہے۔ ڈیولپڈ سپورٹ ویکٹر مشین ماڈل کی misclassification rate مساوات (3.15) کا استعمال کرکے حاصل کی جاتی ہے جو% 0.17 کے برابر ہے۔

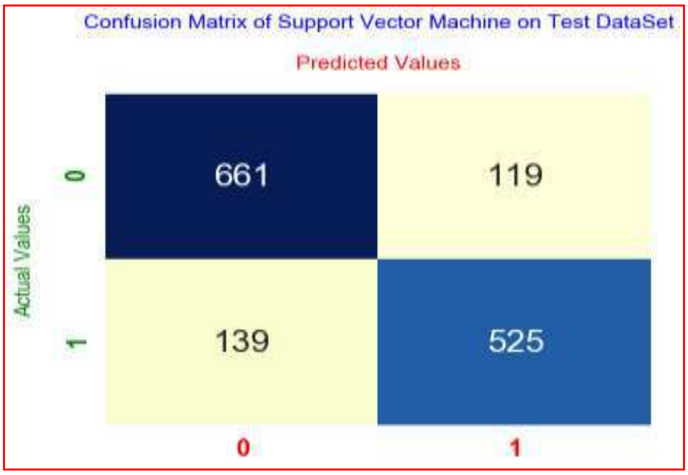


Figure 4.10 SVM Confusion Matrix on Test Dataset

AUROC کی کارکردگی کی پیمائش کا استعمال یہ دیکھنے کے لئے کیا جاتا ہے کہ ڈیولپڈ ماڈل مریض اور صحت مندوں میں کتنا اچھا فرق کر سکتا ہے۔ بہتر ماڈل بیمار اور غیر مریض متاثرین میں بالکل مختلف ہے تاہم ناقص ماڈلز کو دونوں کے درمیان فرق کرنے میں دشواری کا سامنا کرنا پڑتا ہے۔ اوپر دیئے گئے ترسیم 4.11 کے تحت AUROC سپورٹ ویکٹر مشین ماڈل سے حاصل کیا گیا ہے جس کی AUROC اسکور 0.82% ہے۔ ڈیولپڈ سپورٹ ویکٹر مشین ماڈل کے کامیاب تجرباتی نتائج کو موجودہ تحقیق کے ساتھ تیار کیا گیا ہے۔ حاصل کردہ نتائج ادب میں شائع شدہ نتائج سے کہیں زیادہ ہمارے علم میں ہیں لہذا ڈیولپڈ ایس وی ایم ماڈل کو اس کے عملی نفاذ کے لئے استعمال کیا جاتا ہے۔

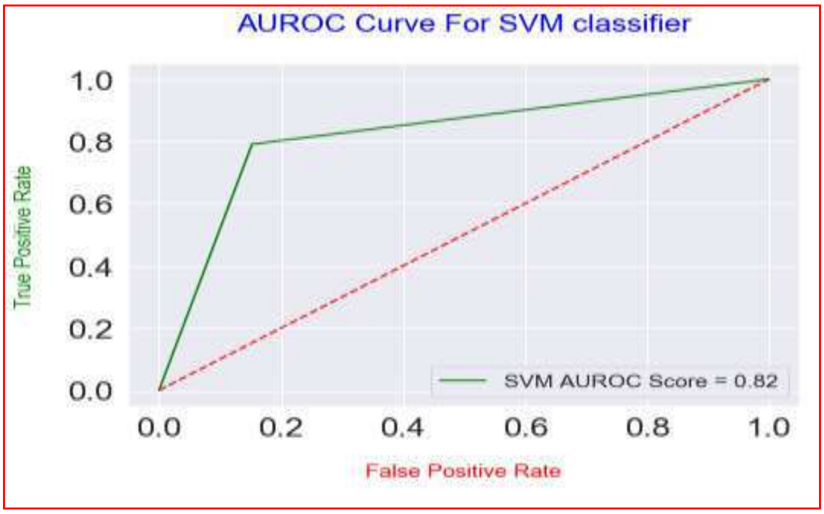


Figure 4.11 AUROC by Support Vector Machine Model

4.5.4 رینڈم فاریسٹ ماڈل کے تجرباتی نتائج

امراض قلب ڈیٹاسیٹ پر رینڈم فاریسٹ ماڈل کے پیش گوئی نتائج کنفیوژن میٹرکس کے ترسیم 4.12میں دکھائے گئے ہیں۔ ترسیم4.12سے ہم True Positive Rate, True Negative Rate, Precision, Accuracy,Error Rate and AUROCاخذ کرتے ہیں جس کی تفصیل ذیل میں دی گئی ہے۔

کنفیوژن میٹرکس 4.12 میں مساوات (3.11) کا استعمال کرتے ہوئے ہمیں رینڈم فاریسٹ ماڈل کی sensitivity 0.8567% حاصل ہوئی ہے اور جس کا مطلب ہے کہ ڈیولپڈ رینڈم فاریسٹ 85% کی درستگی کے ساتھ دل کے امراض کے مثبت واقعات کو پہچان سکتا ہے۔ اسی طرح کنفیوژن میٹرکس 4.12 میں مساوات (3.12) کا استعمال کرتے ہوئے ہمیں رینڈم فاریسٹ ماڈل کی specificity 0.83 فیصد حاصل ہوئی جس کا مطلب ہے کہ ڈیولپڈ رینڈم فاریسٹ ماڈل صحت مندون کو 83% کی درستگی کے ساتھ پہچان سکتا ہے۔ رینڈم فاریسٹ ماڈل کی Accuracy کنفیوژن میٹرکس4.12 میں مساوات(3.13)کااستعمال کرکے حاصل کی گئی ہے جو0.84% کے برابر ہے اس میں یہ بتایا گیا ہے کہ مرض اور صحت مند دونوں معاملات کی تشخیصی میں رینڈم فاریسٹ ماڈل کی مجموعی کارکردگی84% ہے۔ اسی طرح رینڈم فاریسٹ ماڈل کی precisionحاصل کرنے کے لئے مساوات(3.14)استعمال کی جاتی ہے جو کہ تشخیصی کے بعد% 0.85 ہے اس کا مطلب یہ ہے کہ رینڈم فاریسٹ ماڈل میں low false positive rateہے۔ ڈیولپڈ رینڈم فاریسٹ ماڈل کی misclassification rateمساوات(3.15)کااستعمال کرکے حاصل کی ہے جو0.15%کے برابر ہے۔ ماڈل کی غلط درجہ بندی کی شرح میں جو تناسب کم ہے وہ مریض اور صحت مند معاملات کی نشاندہی کرنے میں ماڈل کی اتنی ہی درست ہے۔

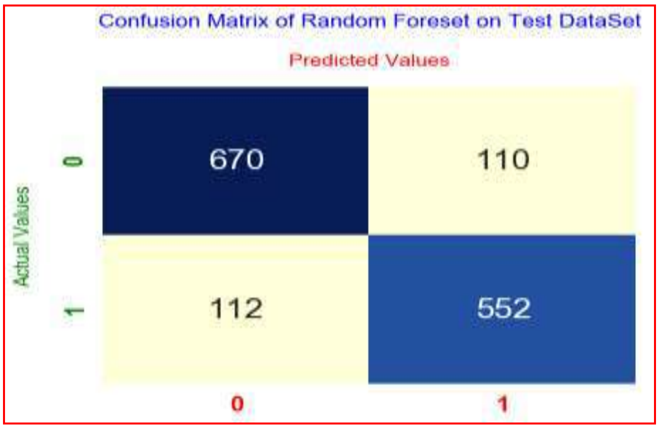


Figure 4.12 Random Forest Model Confusion Matrix on Test Dataset

AUROC کارکردگی کی پیمائش کا استعمال یہ دیکھنے کے لئے کیا جاتا ہے کہ ماڈل مریض اور صحتمندوں میں کتنا اچھا فرق کر سکتا ہے۔ بہتر ماڈل بیمار اور غیر مریض متاثرین میں بالکل مختلف ہے تاہم ناقص ماڈلز کو دونوں کے درمیان فرق کرنے میں دشواری کا سامنا کرنا پڑتا ہے۔ ذیل میں دیئے گئے ترسیم 4.13 کے تحت AUROC ہائپر پیمٹرائزڈ رینڈم فاریسٹ ماڈل سے حاصل کیا گیا ہے جس کی AUROC اسکور 0.85% ہے۔ ہم مروجہ تحقیق کے ساتھ مطلوبہ رینڈم فاریسٹ ماڈل کے کامیاب تجرباتی نتائج کی نقالی کرتے ہیں۔ ہمارے مشاہدے کے متابِک حاصل کردہ نتائج ادب میں شائع شدہ نتائج سے کہیں زیادہ بہترین ہیں۔ لہذا ہم امراض قلب کے مریضوں کی پیش گوئی کے لئے ڈیولپڈ رینڈم فاریسٹ ماڈل کا استعمال کرتے ہیں البتہ ماڈل کی کارکردگی میں مزید بہتری کی ضرورت ہے۔

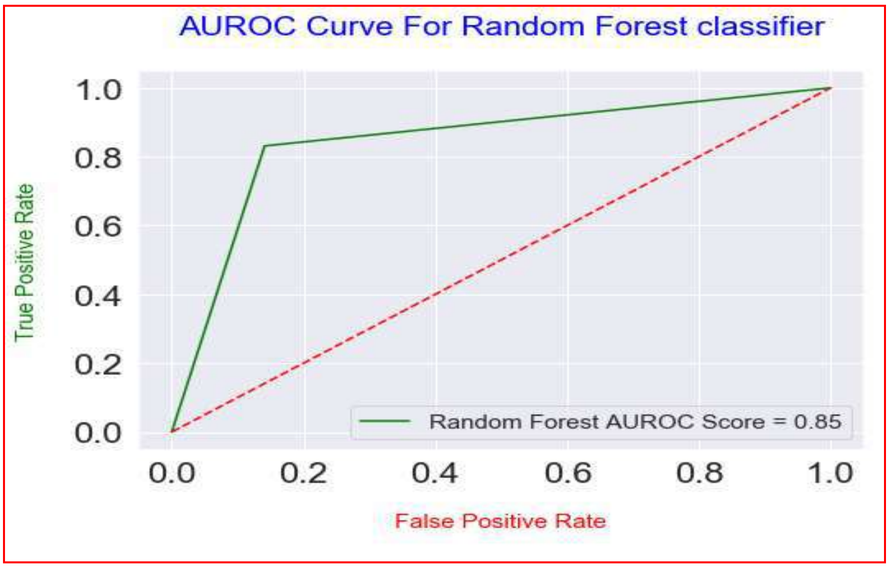
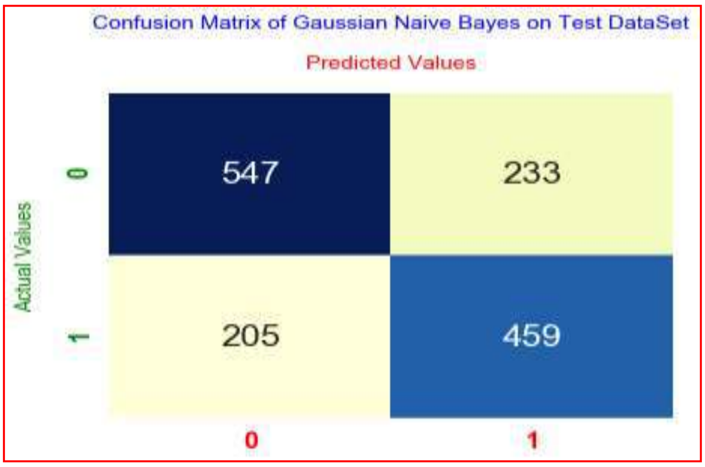


Figure 4.13 AUROC by Random Forest Model

4.5.5 نیوی بائیس ماڈل کے تجرباتی نتائج

نیوی بائیسا لگورتھم basic propabilistic الگورتھم کا ایک گروپ ہے جو بائیس کے نظریے پر مبنی ہے جو features کے مابین مضبوط آزاد مفروضے رکھتے ہیں۔ زیادہ سے زیادہ درستگی اور غیر جانبدارانہ نتائج کے حصول کے لئے ہارٹ ڈیزیز ڈیٹاسیٹ پر 10 fold cross validation کا استعمال کیا گیا ہے۔ نیوی بائیس الگورتھم کے کارکردگی کے نتائج کنفیوژن میٹرکس کے ترسیم 4.14 میں دکھائے گئے ہیں اور AUROC , precision, accuracy, sensitivity, specificity اور misclassification rate کا حساب ذیل میں لگایا گیا ہے-



	0	1
0	547	233
1	205	459

Figure 4.14 Gaussian Naive Bayes Model Confusion Matrix on Test Dataset

کنفیوژن میٹرکس 4.14 میں مساوات (3.11) کا استعمال کرتے ہوئے ہمیں نیوی بائیس ماڈل کی 0.72% True Positive Rate حاصل ہوئی ہے۔ اسی طرح کنفیوژن میٹرکس 4.14 میں مساوات (3.12) کا استعمال کرتے ہوئے ہمیں نیوی بائیس ماڈل کی True 0.66 Negative Rate حاصل ہوئی ہے۔ نیوی بائیس ماڈل کی Accuracy کنفیوژن میٹرکس 4.14 میں مساوات (3.13) کا استعمال کرکے حاصل کی گئی ہے جو حساب کے بعد 0.69% کے برابر ہے اس میں یہ بتایا گیا ہے کہ مرض اور صحت مند دونوں معاملات کی تشخیص میں نیوی بائیس ماڈل کی مجموعی کارکردگی 69% ہے۔ اسی طرح نیوی بائیس ماڈل کی precision حاصل کرنے کے لئے مساوات (3.14) استعمال کی جاتی ہے جو کہ 0.70 % ہے۔ ڈیولپڈ نیوی بائیس ماڈل کی misclassification rate مساوات (3.15) کا استعمال کرکے حاصل کی جاتی ہے جو 0.30% کے برابر ہے۔ ماڈل کی غلط درجہ بندی کی شرح میں جو تناسب کم ہے وہ مریض اور صحت مند معاملات کی نشاندہی کرنے میں ماڈل کی اتنی ہی درست ہے۔

AUROC کی کارکردگی پیمائش کا استعمال یہ دیکھنے کے لئے کیا جاتا ہے کہ ماڈل مریض اور صحت مندوں میں کتنا اچھا فرق کر سکتا ہے۔ بہتر ماڈل بیمار اور غیر مریض متاثرین میں بالکل مختلف ہے تاہم ناقص ماڈلز کو دونوں کے درمیان فرق کرنے میں دشواری کا سامنا کرنا پڑتا ہے دیئے گئے ترسیم 4.15 کے تحت AUROC نیوی بائیس ماڈل سے حاصل کیا گیا ہے جس کی AUROC اسکور 0.70% ہے۔ ہمارے علم کے مطابق نوئی بائیس ماڈل کے یہ تجرباتی نتائج امراض قلب کی تشخیصی کے لئے بہترین نہیں ہیں کیونکہ ادب میں موجود مجوزہ ماڈلز سے misclassification rate زیادہ ہیں۔ میڈیکل ڈومین کی کارکردگی کے علاوہ ماڈل کی complexity and comprehensibility زیادہ ہے۔ غلط درجہ بندی کی شرح اور ماڈل کی پیچیدگی کے عوامل کی اعلی اقدار اس کے اطلاق کو روکتی ہیں کیونکہ میڈیکل پیشن گوئی کے ماڈل کو زیادہ سے زیادہ پیشن گوئی کی درستگی کو پورا کرنا ہوتا ہے اور ایک بھی غلط تشخیصی سنگین نتائج کا باعث بن سکتا ہے۔

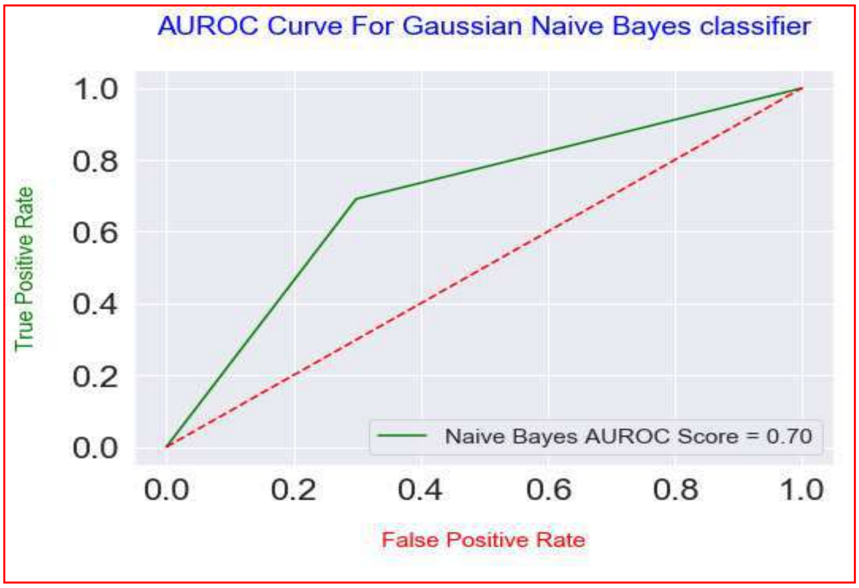


Figure 4.15 AUROC Curve by Gaussian Naive Bayes Model

4.6۔ ڈیولپڈ ہارٹ ڈیزیز رسک اوالویشن ماڈلز کی کارکردگی کا موازنہ

یہ سیکشن رینڈم فارسٹ، نیوی بائیس، کے نیریسٹ نیبر، سپورٹ ویکٹر مشین اور ڈسیشن ٹری ہارٹ ڈیزیز ماڈلز کی کارکردگی اور موازنہ پیش کرتا ہے جیسا کہ مندرجہ ذیل جدول 4.3 میں بیان کیا گیا ہے۔ تجرباتی نتائج سے ظاہر ہوتا ہے کہ رینڈم فارسٹ ماڈل دوسرے رسک ماڈلز کے مقابلے میں بہترین کارکردگی کا مظاہرہ کرتا ہے۔

ڈیولپڈ ہارٹ ڈیزیز رسک اوالویشن ماڈل کی کارکردگی کو موجودہ رسک ٹولز کے ساتھ آزمایا جاتا ہے جس سے یہ ظاہر ہوتا ہے کہ نتائج غیر معمولی پیش گوئی کی درستگی کے ساتھ حوصلہ افزائی کر رہے ہیں۔ تجرباتی نتائج کا بنیادی جائزہ لینے کے بعد ڈیٹا کو محتاط انداز میں جانچنا ضروری ہے تاکہ اہم معلومات کو نکالا جا سکے، بہترین نمونے تیار ہوں اور زیادہ سے زیادہ عنصر کی ترتیبات کا تعین ہو سکے۔ نتائج سے پتہ چلتا ہے کہ رینڈم فارسٹ ماڈل دوسرے ڈیولپڈ ہارٹ ڈیزیز رسک ماڈلز سے زیادہ بہتر کارکردگی کا مظاہرہ کرتا ہے جسے accuracy of 85%,specificity of 83%, sensitivity of 85%, precision of 85%، اور AUROC سکور 85 فیصد کے ساتھ بہتر ہے۔ رینڈم فارسٹ ماڈل سے حاصل ہونے والی درستگی امراض قلب تشخیصی کے لئے سب سے زیادہ ہے جو پچھلی مطالعات سے حاصل نہیں ہوتی ہے۔

Table 4.3 Performance Metrics of Different Proposed Heart Disease Models

Models	Performance Measures					
	Sensitivity	*Specificity*	*Accuracy*	*Precision*	*Error Rate*	*AUROC*
Decision Tree	82%	80%	81%	84%	18%	81%
K Nearest Neighbor	73%	66%	70%	69%	30%	70%
Support Vector Machine	82%	81%	82%	84%	17%	82%
Random Forest	85%	83%	84%	85%	15%	85%
Naive Bayes	72%	66%	69%	70%	30%	70%

نیچے دی گئی ترسیم 4.16 امراض قلب کے مختلف خطرہ تشخیصی ماڈل کے مشترکہ AUROC منحنی خطوط کو ظاہر کرتا ہے۔

ترسیم 4.16 سے یہ بات واضح ہے کہ رینڈم فارسٹ ماڈل امراض قلب کے خطرے سے متعلق تشخیصی ماڈل میں 85% کا سب سے زیادہ اسکور ہے جس کا مطلب ہے کہ ماڈل مریض اور غیر مریضوں کی تشخیص میں انتہائی ہنر مند ہے۔

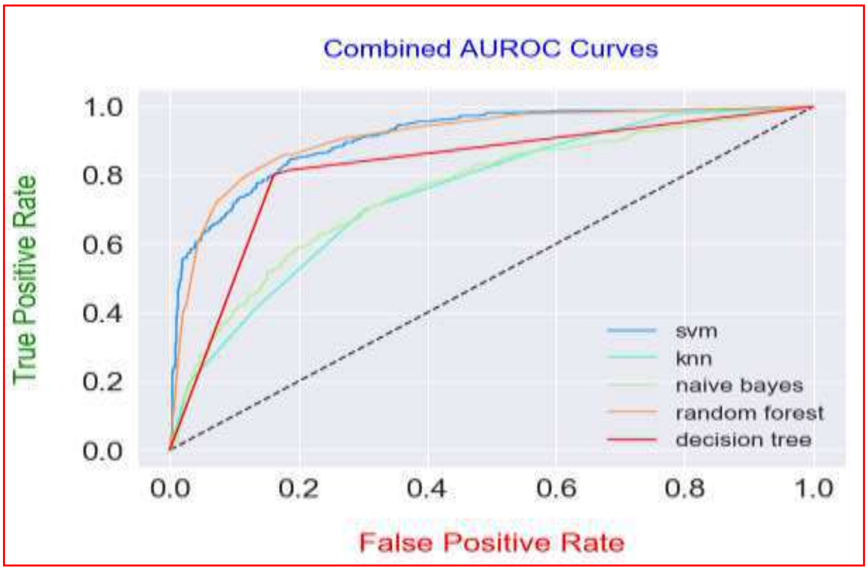


Figure 4.16 Combined AUROCs of the Developed Risk Evaluation Models

4.7 امراض قلب کے لئے ڈیٹا مائننگ ماڈل کا درست تقویت

پریڈکتیو ڈیٹامینگ ماڈلز کی ڈویلوپمنٹ کے دوران، ٹیسٹ سیٹ کی غلطی کو مؤثر طریقے سے کم کرنے کے لئے bias اور variance کا صحیح توازن تلاش کرنا ضروری ہے۔ Bias اور Variance کو ایڈجسٹ کرنے کے لئے ضروری ہے کہ ہم underfitting اور overfitting کا

شکار نہ ہوں۔ اگر ہم bias اور variance کو کم کرنے میں کامیاب ہو گئے تو درست ماڈل بنایا جاسکتا ہے۔ ”Bias is an error from faulty assumptions in the learning algorithm” اور جب bias high ہوتا ہے تو ماڈل صفات اور اہداف کے نتائج کے مابین تعلقات کی درست وضاحت نہیں کر سکے گا۔ however,”variance is an error resulting from fluctuations in the training dataset” اور جب اس کی value زیادہ ہو تو ماڈل نئے ڈیٹا پوائنٹس پر generalize نہیں کر سکے گا۔

$$\text{Total Error} = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error} \qquad (4.1)$$

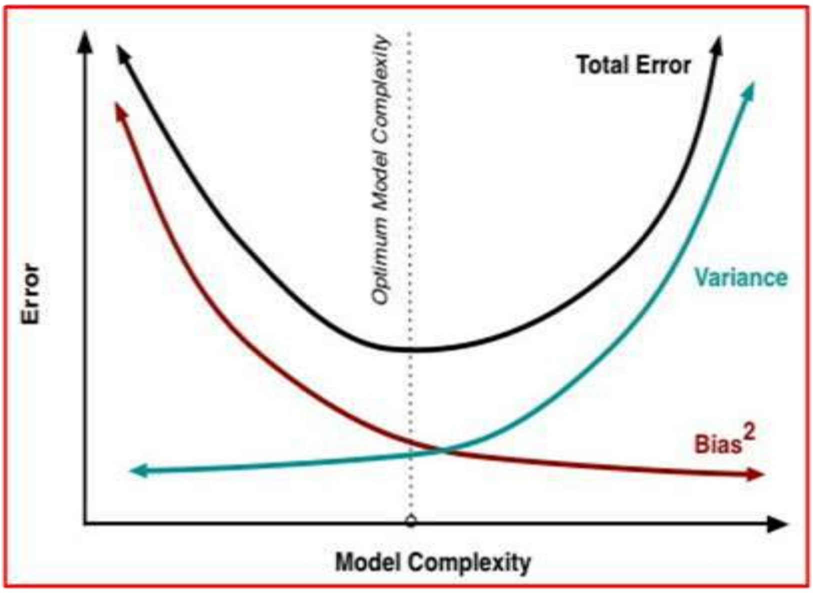


Figure 4.17 Bias and Variance Contributing to the Total Error

غلطی کو کم کرنے کا طریقہ کار bias اور variance کو کم کرنا ہے۔ لیکن ان کو بیک وقت کم کرنا آسان نہیں ہے، کیونکہ ایک اصطلاح کو کم کرنا دوسری اصطلاح میں اضافے کا باعث بنتا ہے۔ لہذا یہ ضروری ہے کہ bias اور variance کے مابین توازن تلاش کریں تاکہ total error کو کم کیا جاسکے اور بہترین فٹ ماڈل مل سکے۔ عملی طور پر، اس جگہ کو تلاش کرنے کے لئے تجزیاتی طریقہ کار موجود نہیں ہے، اس کے بجائے، ہم ہائپر پیرمیٹر آپٹیمائزیشن (باب V میں تبادلہ خیال) کا استعمال کرتے ہیں جو کارکردگی کو بہتر بنانے اور bias اور varaince کے مابین صحیح توازن تلاش کرنے پر مشین لرننگ کلاسیفائیر کے رویے کو منظم کرنے میں مدد کرتا ہے۔ ذیل میں دیئے گئے ترسیم 4.17 کے تحت

supervised learning ماڈل کی غلطی کی کل نمایندگی دکھائی گئی ہے۔ مساوات 4.1 کا استعمال کرتے ہوئے total error کا حساب لگایا جاتا ہے۔

4.8۔ باب کا خلاصہ

اس تحقیقی کام میں ڈیوس ڈیٹا مائننگ طریقہ کار پر عمل کیا گیا ہے تاکہ وہ امراض قلب کے خطرے سے متعلق تشخیصی نمونہ تیار کر سکیں۔ اس باب میں تحقیقی سرگرمیوں کو آسان بنانے کے لئے تحقیقی ڈیزائن وضع کیا گیا ہے اور مقاصد کے بیان سے تحقیق کو نتیجہ خیز بنایا گیا ہے۔ ہارٹ ڈیزیز ڈیٹاسیٹ کو امراض قلب کے خطرے سے متعلق صفات کے انتخاب کے لئے مختلف فیچر سلیکشن تکنیکوں کا استعمال کیا گیا ہے۔ خطرے کی یہ اہم خصوصیات دل کے مریضوں کی ابتدائی پیش گوئی کے لئے ڈیٹا مائننگ کی تراکیب کے لئے استعمال کی جاتی ہیں۔ ڈیٹا مائننگ تکنیک جیسے ڈسیشن ٹری، ایس وی ایم، کے این این، رینڈم فاریسٹ اور نیوی بائیس دل کے مریضوں کی جلد پیش گوئی اور ان کی نشاندہی کے لئے استعمال کیا جاسکتا ہیں۔ تجرباتی نتائج سے ظاہر ہوتا ہے کہ رینڈم فاریسٹ ماڈل سب سے زیادہ درستگی، کم غلط درجہ بندی کی شرح اور کم وقت کی پیچیدگی کے ساتھ دوسرے ماڈل سے بہتر بنا ہے۔ امراض قلب کا خطرہ تشخیصی ماڈل طبّی پیشہ ور افراد کو ابتدائی پیش گوئی اور تشخیصی میں مدد فراہم کرے گا لہذا ترقی کو شدید پیچیدگیوں میں کم کردے گا۔

باب 5

ہارٹ ڈیزیز رسک اوالویشن ماڈلز میں ہائپر پرامیٹر آپٹیمائزیشن کا استعمال

امرض قلب کی پیشن گوئی اور تشخیص اپنی بنیادی پیچیدگیوں کی وجہ سے ایک دشوار کام ہے۔ اس بیماری کی کم سے کم misdiagnose کے ساتھ اور زیادہ سے زیادہ درست اور موثر اندازہ لگانے کے لئے ہم ڈیولپڈ ماڈلز کے ہائپر پرامیٹرز کو بہتر بناتے ہیں۔ Hyperparameter ایک learning ماڈل کے لئے زیادہ سے زیادہ ہائپر پرامیٹرز کی معطلی کا عمل ہے[155]۔ بنیادی طور پر اس سے مراد ممکن عنصری اتصال کا سیٹ ہے جو مطلوبہ اعداد و شمار انعام کی مستحق خوبی کے optimize کے لئے ہیپر پآرامیٹر کی ایک large universe کے ذریعے تلاش کر رہا ہے [156]۔ ماڈل پیرامیٹرز تربیتی مرحلے کے دوران سیکھتے ہیں اور ہیپر پآرامیٹر براہ راست learning ڈیٹا سے نہیں سیکھا رہے ہیں but are tunned independently [157] [158]۔ اس باب میں بتایا گیا ہے کہ باب چہارم میں تیار شدہ ماڈلز کے ہائپر پرامیٹرز کو کس طرح بہتر بنایا جائے۔ اس باب میں، ہم ہائپر پرامیٹر آپٹیمائزیشن تکنیک کی مختلف اقسام اور ہائپر پر میٹر کے ساتھ اور اس کے بغیر امراض قلب کے خطرے کی جانچ کی ماڈلوں کا موازنہ بھی بیان کرتے ہیں۔ آخر میں، ہم کارڈیک ڈس آرڈر کی پیشن گوئی پیدا ہونے والی دل کی بیماری کے قواعد امراض قلب کے ماہر سسٹم کی تشخیصی ماڈل کے اجزاء، اور امراض قلب کے خطرے سے متعلق تشخیصی ماڈل کے مختلف امتزاج پر بھی تبادلہ خیال کرتے ہیں اور باب کے خلاصے اور اختتام کو ختم کرتے ہیں۔

5.1۔ ہائپر پرامیٹر آپٹیمائزیشن تکنیکس

ہائپر پرامیٹرس are the handles and levels that we draw and turn جب مشین learning ماڈل تعمیر کرتے ہے۔ ہائپر پرامیٹرس مختلف ترتیبات کی تلاش کر کے بہتر بناتے ہیں تاکہ یہ معلوم کیا جاسکے کہ کون سی اقدار زیادہ سے زیادہ accuracy دیتی ہیں [159]۔ ماڈل کوڈ میں مسلسل ترمیم کرنے کی وجہ سے ماڈل کو بہتر بنانے کے نفاذ میں ایک سب سے مشکل چیلنج ہے تاکہ جانچ کی غلطی کو کم کیا جاسکے[160][161]۔ ہائپر پرامیٹر کی اصلاح کی نمایاں نمائندگی نیچے دیئے گئے 5.1 ترسیم میں دکھائی گئی ہے۔ ہائپر پرامیٹر ٹیوننگ ماڈل

کی کارکردگی کو بہتر بنانے کا ایک فن ہے اور مناسب ہائپر پر میٹر کا انتخاب انتہائی درست نتائج پیدا کرے گا اور ڈیٹا میں انتہائی قیمتی بصیرت کا باعث بنے گا۔ ہائپرپرامیٹرس مشین ماڈل کے لئے کارکردگی دیکھتے اور bias-variance کے درمیان عین مطابق استحکام کی تلاش سیکھنے کا طرز عمل کی ہدایت کرنے میں مدد کرتے ہیں۔

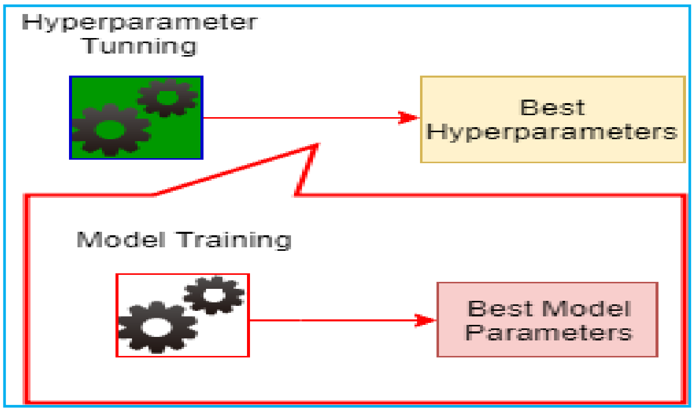


Figure 5.1 Hyperparameter Optimization Representation

مساوات کی شکل میں ہائپرپرامیٹرس کی اصلاح کی نمائندگی اس طرح ہوتی ہے:

$$x^{*} = \arg\min_{x \varepsilon X} f(x) \qquad (5.1)$$

یہاں f(x) توثیق سیٹ پر تشخیصی شدہ غلط طبقاتی تناسب کو کم سے کم کرنے کے لئے ایک مقصد کے نمونے کی نمائندگی کرتا ہے۔ X* ہائپرپرامیٹرس کا سیٹ ہے جو اسکور کی سب سے کم قیمت حاصل کرتا ہے، اور x ڈومین X میں کسی بھی قیمت کو حاصل کر سکتا ہے۔ آسان الفاظ میں ہم ہائپرپرامیٹرس تلاش کرنا چاہتے ہیں جو مختلف hyperparameteric تکنیکوں کا استعمال کرتے ہوئے validation سیٹ میٹرک پر بہترین اسکور حاصل کرتے ہیں۔

5.1.1۔ گرڈ سرچ ہائپر پر میٹر آپٹیمائزیشن

ہائپر پر میٹر آپٹیمائزیشن کے روایتی طریقہ کار کو گرڈ سرچ کیا گیا ہے۔ گرڈ سرچ defined search space میں تمام ممکنہ values سے زیادہ candidate values کے پیرامیٹر ویلیوز کی ایک جامع تلاش ہے۔ ماڈل کے تصوراتی پیرامیٹر کے امتزاجوں کی جانچ پڑتال کے بعد بہترین مجموعہ برقرار رکھا جائے گا۔ گرڈ سرچ ہائپر پر میٹرز کے دو سیٹوں کا استعمال کر کے ہر ایک کی اجازت کے لئے الگورتھم کی تربیت کرتا ہے اور

cross validation تکنیک کا استعمال کر کے درستگی کا تعین کرتا ہے [162]۔ اس validation تکنیک سے یہ یقین دہانی ہوتی ہے کہ trained ماڈل ڈیٹاسیٹ سے زیادہ سے زیادہ نمونوں کو حاصل کرتا ہے۔ ذیل میں دیئے گئے ترسیم 5.2 میں grid search کے طریقہ کار کی ترتیب کو ظاہر کیا گیا ہے۔ Grid Search Method استعمال کرنے کے لئے ایک آسان طریقہ ہے لیکن یہ ایک مہنگا طریقہ اختیار کرتے ہیں اور high dimensional data space میں کم موثر نتائج دیتا ہے [163]۔

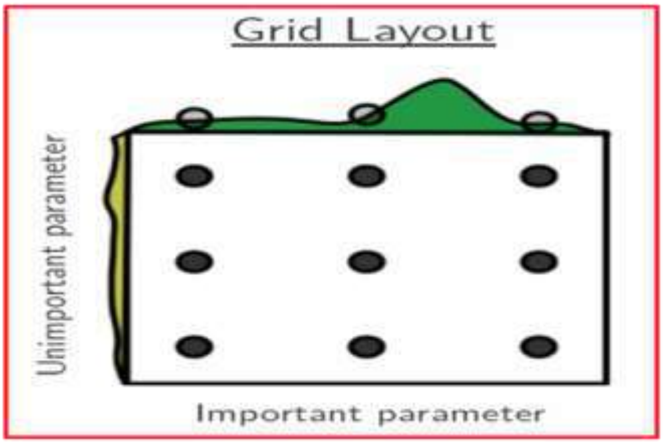


Figure 5.2 Grid Search Layout

5.1.2 رینڈم سرچ ہائپرپرمیٹر آپٹیمائزیشن

رینڈم سرچ تکنیک میں ہیپرپارامیٹر values تصادفی طور پر وضاحتی تلاش خلا سے منتخب کردہ ہیں۔ Random Sampling متفرد اور مسلسل hyperparameters شامل کرنے کے لئے search space کی اجازت دیتا ہے [164]۔ رینڈم سرچ ہائپرپرامیٹر آپٹیمائزیشن متوازی ہے اور سابقہ علم کی شمولیت کی اجازت سے تقسیم کو نمونہ کا تعین کرنے کے کی طرف کرتہ ہے۔ نیچے دیئے گئے ترسیم 5.3 رینڈم سرچ ہائپرپرامیٹر آپٹیمائزیشن کی اصلاح کی ترتیب کو ظاہر کرتا ہے۔

گرڈ اور رینڈم سرچ ہائپرپرامیٹر کی اصلاحات ماضی کی تشخیصوں سے مکمل طور پر ناواقف ہیں اور اس کے نتیجے میں اکثر "خراب" ہائپرپرامیٹرز کا اندازہ کرنے میں کافی وقت خرچ کرتے ہیں [166]۔ گرڈ اور رینڈم search کی اصلاح کی تراکیب تجربہ کار machine learning کے تجربہ کاروں پر مبنی ہیں اور یہ heuristic تکنیک ہیں۔ ان تراکیب میں اعلی جہتی ڈیٹا کو غلط انداز میں ڈھالنے کی وجہ سے انسانی مہارت کو ہائپرپرامیٹرز کی قریب ترین زیادہ سے زیادہ ترتیب نہیں مل پائے گی اور ایک سے زیادہ ہائپرپرامیٹرز کو ٹیون کرنے کی کوشش کرتے وقت آسانی

سے غلط تشریح کی جاسکتی ہے[167]۔ گرڈ اور رینڈم سرچ ہائپر پرامیٹر آپٹیمائزیشن کی تکنیک سے وابستہ ان خرابیوں کی وجہ سے، ہم دل کی بیماریوں کے خطرے کی تشخیصی ماڈل کی نشو نما کے لئے bayesian ہائپر پر میٹر آپٹیمائزیشن کو استعمال کرنے کو ترجیح دیتے ہیں۔

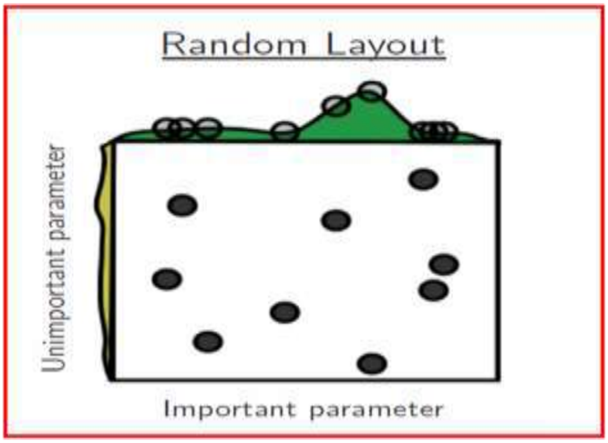


Figure5.3 Random Search Layout

5.1.3 Bayesian ہائپرپرا میٹر آپٹیمائزیشن

گرڈ اور رینڈم سرچ ہیپر پارامیٹر آپٹیمائزیشن میں ماڈل پیرامیٹرز تصادفی طور پر بنائے جاتے ہیں اور نئے پیرامیٹر کا امتزاج پچھلے ٹریلز سے آزاد ہے۔ تاہم Bayesian hyperparameter optimization خود بخود ہائپر پرامیٹر کی اصلاح کی تکنیک بہترین انتخاب کرنے کیلئے پچھلے ٹریل علم کا استعمال کرتی ہے۔ Bayesian hyperparameter optimization سب سے پہلے مختلف کنفیگریشنوں پر کارکردگی کو اکٹھا کرتی ہے پھر کچھ نتیجہ اخذ کرتی ہے اور فیصلہ کرتی ہے کہ آئندہ کون سی کنفیگریشن کی کوشش کی جائے تاکہ چک کی تعداد کم ہو۔ Bayesian Optimization کے لئے اس تحقیقی کام میں ترسیم 5.4 میں دکھائے گئے single cross validaton (S-CV) کے طریقہ کار کو ہر تکنیک کے ساتھ استعمال کیا گیا ہے۔ Hyperparameters کو بہتر بنانے اور اس پر عمل درآمد کرنے پر ہارٹ ڈیزیز ڈیٹاسیٹ k stratifiedfolds میں تقسیم کیا گیا ہے۔ رینڈم فارسٹ، نیوی بائیس، کے نیریسٹ نیبر، سپورٹ ویکٹر مشین اور ڈیسیشن ٹری الگورتھم کے ذریعہ پائے جانے والے ہر ایک کے لئے training dataset سے تربیت دی جاتی ہے۔ ڈیٹاسیٹ کا ایک حصہ ماڈل کی توثیق کے لئے استعمال ہوتا ہے اور ڈیٹاسیٹ کے بقیہ حصے جانچ کے مقصد کے لئے استعمال ہوتے ہیں۔ توثیق اور ٹیسٹ کی پرفارمنس کو ٹریننگ ڈیٹاسیٹ اور اصلاح کی تکنیک کے ذریعہ پائے جانے والے ہائپر پرمیٹر کی اقدار کے ساتھ آمادہ ماڈل کے ذریعے ماپا جاتا ہے۔ یہ عمل

single cross validation میں تمام K مرکبوں کے لئے دہرایا گیا ہے۔ اس کے بعد اوسط توثیق کی درستگی کو فٹنس ویلیو کے طور پر استعمال کیا جاتا ہے جو search کے عمل کو direct کرتا ہے۔ آخر میں زیادہ سے زیادہ توثیق کی درستگی والا individual واپس ہو جاتا ہے (اس کی ہائپر پیرامیٹر اقدار کے ساتھ) اور تکنیکی کارکردگی individual کی اوسط ٹیسٹ کی درستگی سمجھی جاتی ہے۔

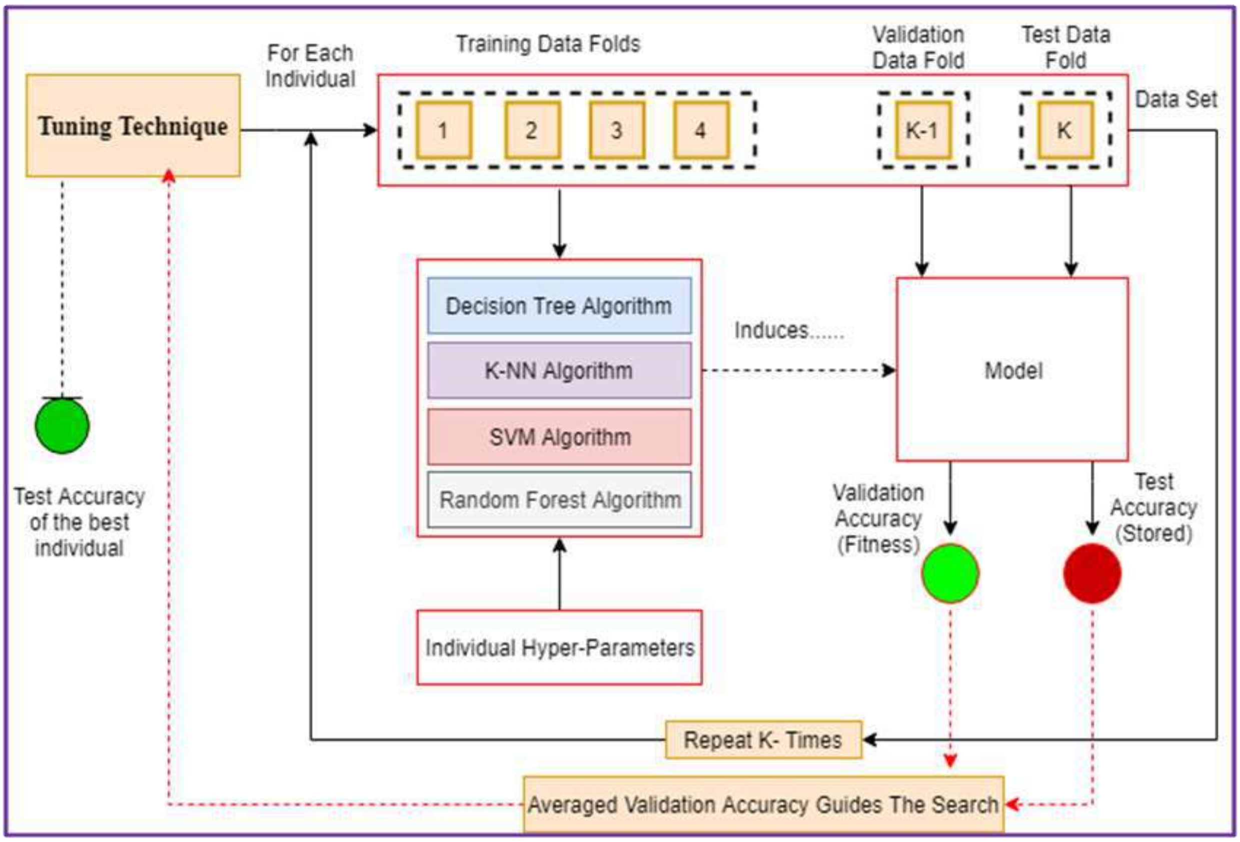

Tuning Technique
For Each Individual
Training Data Folds
Validation Data Fold
Test Data Fold
Data Set
1
2
3
4
K-1
K
Decision Tree Algorithm
K-NN Algorithm
SVM Algorithm
Random Forest Algorithm
Induces......
Model
Individual Hyper-Parameters
Test Accuracy of the best individual
Validation Accuracy (Fitness)
Test Accuracy (Stored)
Repeat K- Times
Averaged Validation Accuracy Guides The Search

Figure 5.4 Single Cross Validation Methodology for Hyperparameter Optimization

5.2 Optimizing ہارٹ ڈیزیز رسک اوالویشن ماڈلز

ڈیٹا مائننگ algorithms میں مختلف ہائپر پرامیٹرز شامل ہیں جن کی اقدار پیچیدہ طریقوں سے ڈیولپڈ ہارٹ ڈیزیز رسک اوالویشن ماڈل کی پیش گوئی کی کارکردگی کو متاثر کرتی ہیں۔ ہائپر پرامیٹر کی ترتیبات کے بہت زیادہ امکانات کی وجہ سے ہمارے پاس اس ضمن میں بصیرت کا فقدان ہے کہ تشکیل کی اس بے پناہ جگہ کو پیشہ ورانہ طور پر کیسے تلاش کیا جائے۔ ہر ماڈل پر روشنی ڈالنے کے لئے بہت سارے ممکنہ پیرامیٹرز موجود ہیں اور تمام پیرامیٹرز قابل قدر ہیں لیکن پیرامیٹرز کے اہم سبسیٹ کو منتخب کرنے کی ضرورت ہے۔ ذیل میں دیئے گئے sub sections میں یہ بیان کیا گیا ہے کہ ہائپر پیرمیٹرائزڈ ہارٹ ڈیزیز رسک اوالویشن ماڈل کو کس طرح بہتر بنایا گیا ہے۔

5.2.1- ڈیسیشن ٹری ہائپرپرامیٹر آپٹیمائزیشن ماڈل

زیادہ سے زیادہ درستگی کے لئے ڈیسیشن ٹری ماڈل کے سب سے اہم ہائپر پر میٹرز ترتیب دیئے گئے ہیں تاہم ہمیں محتاط رہنا چاہئے کہ ان کو validation dataset کے فولڈ پر overfitting کرنے سے بچنے۔ ڈیسیشن ٹری ماڈل کی کارکردگی کو بہتر بنانے والے اہم ہائپرپرامیٹرز مندرجہ ذیل ہیں۔

i. Max Depth: یہ ہائپرپرامیٹر اس بات کی نمائندگی کرتا ہے کہ ٹری کتنا گہرا ہو سکتا ہے۔ زیادہ سے زیادہ گہرائی والا ڈیسیشن ٹری ڈیٹا سے مزید معلومات حاصل کرتا ہے۔ اس تحقیقی کام میں ہم ایک ڈیسیشن ٹری الگورتھم کو 100 max depth کی گہرائی کے ساتھ فٹ کرتے ہیں اور training and testing AUROC اسکور کی بھی منصوبہ بندی کرتے ہیں۔ یہ پایا گیا ہے کہ اگر max depth attribute کی value high رکھا جائے تو ڈیولپڈ ماڈل overfit ہوتا ہے۔

ii. Min Samples Split: ڈیسیشن ٹری الگورتھم کا یہ ہائپر پارامیٹر اندرونی نوڈ کو تقسیم کرنے کے لئے درکار نمونوں کی کم از کم مقدار کی نمائندگی کرتا ہے۔ ہر نوڈ کے نمونے تقسیم کرنے میں ہر نوڈ کے ایک نمونے سے مختلف ہوتے ہیں۔. اگر min samples split پیرامیٹر کی قدر میں اضافہ کیا جائے تو ٹری زیادہ constrained ہو جاتا ہے کیونکہ اسے ہر نوڈ پر مزید نمونوں پر غور کرنا پڑتا ہے۔ تاہم اگر ہر نوڈ کے تمام نمونوں پر غور کیا جاتا ہے تو پھر ماڈل underfit ہو جاتا ہے.

iii. Min Samples Leaf: یہ ہائپر پارامیٹر leaf کے نوڈ پر ہونے والے نمونے کی کم از کم تعداد کی نشاندہی کرتا ہے۔ اگر min samples leaf کی پیرامیٹرک ویلیو بڑھ جاتی ہے تو پھر اس سے underfitting کا مسئلہ ہوتا ہے۔

iv. Max Features: Max Features ہیپرپآرامیٹر ڈیسیشن ٹری الگورتھم خصوصیات کی تلاش میں شامل کرنے کے لئے features کی تعداد کی نشاندہی کرتی ہیں تاہم اگر زیادہ سے زیادہ max features کو منتخب کیا گیا ہے تو پھر اس سے overfitting مسئلہ درپیش آتا ہے۔

v. Criteria : یہ ہیپرپآرامیٹر فیصلہ کرتی ہے کہ کس طرح impurity کو ماپا جائے گا. Criteria ہائپر پارامیٹر کی default value "Gini" ہے تاہم اسے "Entropy" کے طور پر مقرر کیا جاسکتا ہے۔

ڈسیشن ٹری ماڈل کے ہائپر پرامیٹرز کو ٹیوننگ کرنے کے بعد ہم نتائج حاصل کرتے ہیں جیسا کہ نیچے دیئے گئے جدول 5.1 میں دکھایا گیا ہے۔ ہائپر پیرامیٹرائزڈ ڈسیشن ٹری ماڈل کے تقویت اور امتزاج مختلف نتائج دکھاتے ہیں تاہم ہم صرف ان امتزاج کو بیان کرتے ہیں جو اعلی ترین اقدار فراہم کرتے ہیں۔

Table 5.1 Hyperparameter Optimization Results of the Decision Tree Model

Max Depth	***Min Samples Split***	***Min Samples Leaf***	***Max Features***	**Criterion**	**Accuracy**
10	18	15	Auto	Entropy	81%
15	25	12	Auto	Gini	82%
18	11	10	Sqrt	Gini	72%
20	15	20	Sqrt	Gini	83%
25	10	50	Auto	Entropy	71%
30	12	30	Auto	Entropy	74%
35	22	25	Sqrt	Entropy	75%
40	8	14	Sqrt	Gini	78%
45	5	16	Auto	Entropy	73%
50	14	Not Used	Auto	Gini	78%
70	17	18	Auto	Entropy	75%
80	13	Not used	Auto	Entropy	84%
100	20	Not used	Sqrt	Entropy	84%

ہم نے کارڈیک ڈس آرڈر کے مریضوں کی ابتدائی پیش گوئی اور شناخت کے لئے ہائپر پر میٹرائزڈ ڈسیشن ٹری ماڈل کا اطلاق کیا۔ ماڈل کی کارکردگی کے نتائج کنفیوژن میٹرکس کے ترسیم 5.5 میں دکھائے گئے ہیں۔ ترسیم 5.5 سے ہم True Positive Rate, True Negative Rate, Precision, Accuracy,Error Rate and AUROC، اخذ کرتے ہیں جس کی تفصیل ذیل میں دی گئی ہے۔

کنفیوژن میٹرکس 5.5 میں مساوات (3.11) کا استعمال کرتے ہوئے ہمیں ہائپر پر میٹرائزڈ ڈسیشن ٹری ماڈل کی **True Positive**

Rate حاصل ہوئی ہے جو حل کرنے کے بعد 0.833% کے برابر ہے اور جس کا مطلب ہے کہ ڈیولپڈ ہائپرپیمیٹرائزڈ ڈسیشن ٹری % 83 کی درستگی کے ساتھ دل کے امراض کے مثبت واقعات کو پہچان سکتا ہے۔ اسی طرح کنفیوژن میٹرکس 5.5 میں مساوات (3.12) کا استعمال کرتے ہوئے ہمیں ہائپرپرمیٹرائزڈ ڈسیشن ٹری ماڈل کی True Negative Rate حاصل ہوئی ہے جو کہ 0.80 فیصد کے برابر ہے اور جس کا مطلب ہے کہ ڈیولپڈ ہائپرپیمیٹرائزڈ ڈسیشن ٹری ماڈل صحت مندوں کو % 80 کی درستگی کے ساتھ پہچان سکتا ہے۔ ڈسیشن ٹری ماڈل کی Accuracy کنفیوژن میٹرکس 5.5 میں مساوات (3.13) کا استعمال کرکے حاصل کی گئی ہے جو 0.8185% کے برابر ہے اس میں یہ بتایا گیا ہے کہ مرض اور صحت مند دونوں معاملات کی تشخیص میں ڈسیشن ٹری ماڈل کی مجموعی کارکردگی 81% ہے۔ اسی طرح ہائپرپرمیٹرائزڈ ڈسیشن ٹری ماڈل کی precision حاصل کرنے کے لئے مساوات (3.14) کا استعمال ہوا ہے جو کہ % 0.8294 ہے اس کا مطلب یہ ہے کہ ہائپرپرمیٹرائزڈ ڈسیشن ٹری ماڈل میں low false positive rate ہے۔ ڈیولپڈ ہائپرپرمیٹرائزڈ ڈسیشن ٹری ماڈل کی misclassification rate مساوات (3.15) کا استعمال کرکے حاصل کی جاتی ہے جو % 0.18 ہے۔

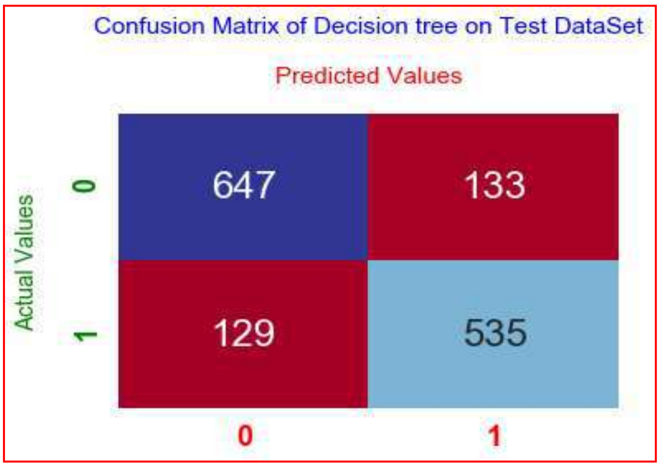


Figure 5.5 Confusion Matrix of the Optimized Decision Tree

AUROC کارکردگی کی پیائش کا استعمال یہ دیکھنے کے لئے کیا جاتا ہے کہ ماڈل مریض اور صحتمندوں میں کتنا اچھا فرق کر سکتا ہے۔ بہتر ماڈل بیمار اور غیر مریض متاثرین میں بالکل مختلف ہے تاہم ناقص ماڈلز کو دونوں کے درمیان فرق کرنے میں دشواری کا سامنا کرنا پڑتا ہے۔ ذیل میں دیئے گئے ترسیم 5.6 کے تحت AUROC ہائپرپرمیٹرائزڈ ڈسیشن ٹری ماڈل سے حاصل کیا گیا ہے جس کی AUROC اسکور 0.82% ہے۔ ہم مروجہ تحقیق کے ساتھ مطلوبہ ہارٹ ڈیزیز ڈسیشن ٹری ماڈل کے کامیاب تجرباتی نتائج کی نقالی کرتے ہیں۔ ہمارے مشاہدے کے

متِبک حاصل کردہ نتائج ادب میں شائع شدہ نتائج سے کہیں زیادہ بہترین ہیں۔ لہذا ہم امراض قلب کے مریضوں کی پیش گوئی کے لئے مجوزہ ہائپرپر میٹرائزڈ ڈسیشن ٹری ماڈل کا استعمال کرتے ہیں البتہ ماڈل کی کارکردگی میں مزید بہتری کی ضرورت ہے۔

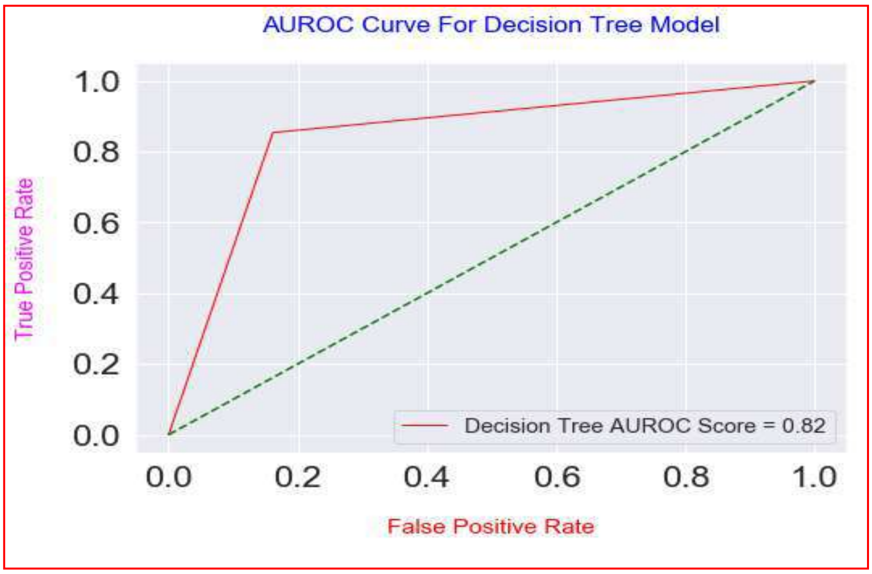


Figure 5.6 AUROC of Optimized Decision Tree Model

5.2.2۔ کے نیریسٹ نیبرہائپرپرامیٹرآپٹیمائزیشن ماڈل

کے این این اکثریتی ووٹوں کی بنیاد پر ایک نامعلوم نیبر کی درجہ بندی کرتا ہے۔ ہر نیبر کو یا تو مساوی وزن دیا جاسکتا ہے یا ووٹ distance پر مبنی ہو سکتا ہے۔ کے این این الگورتھم کا ہائپرپرامیٹر اصلاح بہت بہترین نمونہ تلاش کرنے کے لئے انجام دیا گیا ہے جو ٹیسٹ ڈیٹا سیٹ پر زیادہ سے زیادہ accuracy اور کم ترین error کو ظاہر کرتا ہے۔ کے این این درجہ بندی کے انتہائی اہم ہائپرپرامیٹرز کی جانچ پڑتال کرتا ہے تاکہ check کیا جاسکے کہ وہ ماڈل کو کس حد تک overfitting and underfitting کے لحاظ سے متاثر کرتے ہیں۔ کے این این ماڈل کے ابتدائی ہائپرپرامیٹرز جو tune ہو سکتے ہے وہ حسب ذیل بیان کئے گئے ہیں:

1۔ K Nearest Neighbors کی تعداد۔

2۔ similarity function یا distance میٹرک۔

کے این این الگورتھم کے یہ دو ائپرپرامیٹرز درستگی کو نمایاں طور پر متاثر کرتے ہیں۔ کے این این الگورتھم میں بہترین K وہی ہے جو lowest test error rate فراہم کرتا ہے لہذا K کی دیگر values کے لئے test error کو بار بار چیک کیا جاتا ہے بذریعہ fitting process تربیت کے اعداد و شمار کا سبسیٹ محفوظ کر کے test error rate کی پیمائش کی جاتی ہے۔ یہ سبسیٹ (validation set) الگورتھم کے flexibility کی صحیح سطح کا انتخاب کرنے کے لئے استعمال ہوتا ہے۔ Optimized کے این این ماڈل کے تجرباتی نتائج ذیل میں دیئے گئے جدول 5.2 میں دکھائے گئے ہیں۔ ہم maximum accuracy حاصل کرنے کے لئے کے این این ماڈل کے مختلف ترتیب اور مرکب استعمال کرتے ہیں۔ پیرامیٹر کے امتزاج جس کے نتیجے میں سب سے زیادہ درستیاں آتی ہیں درج ذیل table 5.2 میں بیان کی گئی ہے۔

Table 5.2 Hyperparameter Optimization Results by the K NN Model

Leaf Size	**Metric**	**Neighbors**	**Weights**	**Accuracy**
5	Euclidean	5	Distance	82%
10	Minkowski	11	Uniform	67%
30	City Block	13	Distance	85%
25	Euclidean	9	Distance	70%
15	Minkowski	7	Uniform	72%
20	City Block	11	Uniform	68%
12	Euclidean	15	Distance	75%
16	Minkowski	13	Uniform	77%
18	Minkowski	7	Uniform	80%
28	Euclidean	9	Distance	82%

ٹیبلر نتائج سے ہم دیکھ سکتے ہیں کہ جب میٹرک وصف کو "منکووسکی" کے طور پر تشکیل دیا جاتا ہے اور weight attribute کو "uniform" کے طور پر تشکیل دیا جاتا ہے تو کے این این ماڈل کی کارکردگی گھٹ کر 67% ہو جاتی ہے۔ ماڈل کی درستگی کو جانچنے کے لئے ‘best score’ فنکشن کا استعمال کیا جاتا ہے کیونکہ ‘best score’ cross validataion کے ذریعہ حاصل کردہ اسکور کی اوسطا درستگی کی نشاندہی کرتا ہے۔ جب k کی value کم رکھی جاتی ہے تو low Bias اور High Variance پایا جاتا ہے لیکن جب K کی value بڑی رکھی جاتی ہے تو high bias اور less variance حاصل کیا جاتا ہے۔ لہذا ہم balance position کے پیرامیٹر کی قدر کو تشکیل دیتے ہیں۔ Optimization search کے ذریعہ 85% تک ماڈل کی درستگی کو بہتر بنایا گیا ہے۔

تجرباتی نتائج سے پتہ چلتا ہے کہ جب ہائپر پیرامیٹر کے مجموعے [لیف سائز=30، میٹرک=سٹی بلاک، **weight** = 13] پر سیٹ کیے جاتے ہیں تو پھر **85%** کی **maximum accuracy** حاصل ہو جاتی ہے۔ ڈیولپڈ **optimized** کے این این ماڈل کی کارکردگی کے نتائج کنفیوژن میٹرکس کے ترسیم 7. 5 میں دکھائے گئے ہیں۔

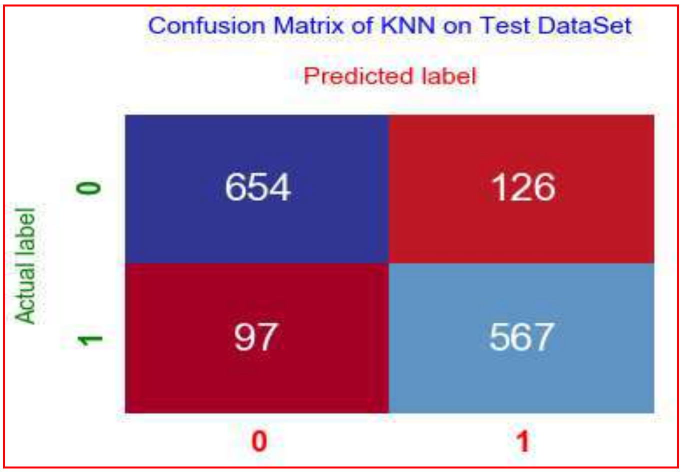


Figure 5.7 Confusion Matrix of the Optimized K NN Model

ہم نے کارڈیک ڈس آرڈر کے مریضوں کی ابتدائی پیش گوئی اور شناخت کے لئے optimized کے این این ماڈل کا اطلاق کیا۔ ماڈل کی کارکردگی کے نتائج ترسیم کنفیوژن میٹرکس 5.7 میں دکھائے گئے ہیں۔

کنفیوژن میٹرکس 5.7 میں مساوات (3.11) کا استعمال کرتے ہوئے ہمیں optimized کے این این ماڈل کی True Positive Rate حاصل ہوئی ہے جو 0.87% کے برابر۔ اسی طرح کنفیوژن میٹرکس 5.7 میں مساوات (3.12) کا استعمال کرتے ہوئے ہمیں ڈیولپڈ optimized کے نیریسٹ نیبر ماڈل کی True Negative Rate حاصل ہوئی ہے جو کہ 0.81 فیصد کے برابر ہے۔ optimized کے نیریسٹ نیبر ماڈل کی Accuracy کنفیوژن میٹرکس 5.7 میں مساوات (3.13) کا استعمال کر کے حاصل کی گئی ہے جو 0.84% کے برابر ہے۔ اسی طرح optimized کے نیریسٹ نیبر ماڈل کی precision % حاصل ہوئی 0.83 ہے۔ ڈیولپڈ optimized کے نیریسٹ نیبر ماڈل کی misclassification rate مساوات (3.15) کا استعمال کر کے حاصل ہوئی ہے جو% 0.15 کے برابر ہے۔

AUROC کارکردگی کی پیمائش کا استعمال یہ دیکھنے کے لئے کیا جاتا ہے کہ ڈیولپڈ ماڈل مریض اور صحتمندوں میں کتنا اچھا فرق کر سکتا ہے۔ بہتر ماڈل بیمار اور غیر مریض متاثرین میں بالکل مختلف ہے تاہم ناقص ماڈلز کو دونوں کے درمیان فرق کرنے میں دشواری کا سامنا کرنا پڑتا ہے۔ ذیل

میں دیئے گئے اعداد و شمار 5.8 کے تحت optimized AUROC کے نیریسٹ نیبر ماڈل سے حاصل کیا گیا ہے جس کی AUROC اسکور 0.85% ہے۔ ہم مروجہ تحقیق کے ساتھ مطلوبہ ڈیولپڈ کے نیریسٹ نیبر ماڈل کے کامیاب تجرباتی نتائج کی نقالی کرتے ہیں۔ ہمارے مشاہدے کے متبِک حاصل کردہ نتائج ادب میں شائع شدہ نتائج سے کہیں زیادہ بہترین ہیں۔ لہذا ہم امراض قلب کے مریضوں کی پیش گوئی کے لئے ڈیولپڈ optimized کے نیریسٹ نیبر ماڈل کا استعمال کرتے ہیں البتہ ماڈل کی کارکردگی میں مزید بہتری کی ضرورت ہے۔

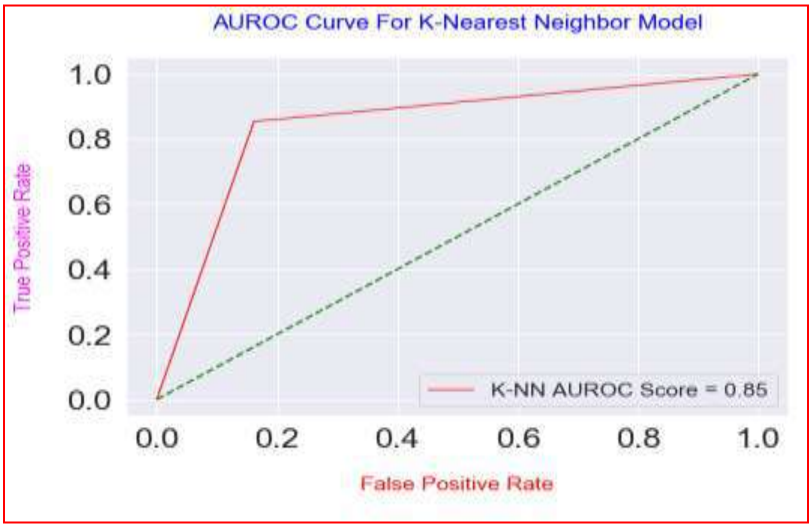


Figure 5.8 AUROC of Optimized K NN Model

5.2.3 سپورٹ ویکٹر مشین ہائپر پرامیٹر آپٹیمائزیشن ماڈل

امراض قلب کی ابتدائی شناخت کے لئے highest accuracy کے ساتھ ایس وی ایم ماڈل کے انتہائی اہم ہائپر پرامیٹرز ترتیب دیئے گئے ہیں تاہم ہمیں ٹیسٹ ڈیٹاسیٹ پر ان کی توثیق کرنے میں محتاط رہنا چاہئے۔ ایس وی ایم ماڈل کے ہائپر پرامیٹرز جو بہتر ہیں وہ مندرجہ ذیل ہیں:

i. Kernel: کرنل ہائپر پرامیٹر ڈیٹا کو علیحدہ کرنے کے لئے hyperplane کا انتخاب کرتی ہے۔ کرنل کرنل ہائپر پرامیٹر ڈیٹا کو مطلوبہ شکل میں تبدیل کرتا ہے۔ ایس وی ایم میں مختلف قسم کے kernel دستیاب ہیں جیسے Radial Basis Function (RBF), Polynomial, Linear and Sigmoid۔ اس تحقیق میں ہر طرح کے kernel استعمال کیے جاتے ہیں تاہم rbf کرنل نمایاں نتائج فراہم کرتا ہے۔

ii. Regularization: یہ ہائپرپرامیٹر penalty پیرامیٹر ہے ، جو misclassification کی نمائندگی کرتا ہے۔ Misclassification کا مطلب ہے how much error is bearable۔ اگر ریگولرائزیشن ہائپرپرامیٹر small value کے ساتھ مرتب کیا گیا ہے تو وہ ایک small مارجن ہائپرپلین تشکیل دیتا ہے اور اگر یہ large value کے ساتھ سیٹ کیا گیا ہے تو یہ ایک بڑے مارجن ہائپرپلین کو تشکیل دیتا ہے۔

iii. Gamma: SVM گاما ہائپرپرامیٹر non linear hyperplane کے لئے استعمال کیا جاتا ہے۔ اگر گاما ہائپرپرامیٹر small value کے ساتھ سیٹ کیا گیا ہے تو وہ تربیت کے ڈیٹاسیٹ کو loosly fit بیٹھتا ہے، تاہم higher value ٹریننگ ڈیٹاسیٹ کے ساتھ بالکل فٹ بیٹھتی ہے، جس کی وجہ سے زیادہ overfitting ہوتی ہے۔

ایس وی ایم ہارٹ ڈیزیز رسک اوالویشن ماڈل کا طرز عمل گاما ہائپر پیرامیٹر کے لئے انتہائی حساس ہے۔ ذیل میں دیئے گئے جدول 5.3 میں ایس وی ایم ماڈل کے مختلف ہائپرپرامیٹرز کو پورا کرنے کے بعد حاصل ہونے والی مختلف accuracies دکھائی گئی ہیں۔ زیادہ سے زیادہ accuracy حاصل کرنے کے لئے ایس وی ایم ماڈل کے "kernel" اور "regularization" ہائپرپرامیٹر بہترین permutations and combinations سیٹ کے ساتھ ترتیب دیئے گئے ہیں۔

یہ پایا گیا ہے کہ جب kernel ہائپرپرامیٹر کی (linear, sigmoid or Sqrt) values پر سیٹ کیے جاتے ہیں تو ہارٹ ڈیزیز رسک اوالویشن ماڈل کی time complexity بڑھ جاتی ہے۔ تجرباتی نتائج سے پتہ چلتا ہے کہ جب ایس وی ایم ہائپرپرامیٹر کے امتزاج [کرنل=آر بی ایف، گاما=0.1، ریگولرائزیشن = 1.0] پر سیٹ کیے جاتے ہیں تو 81% کی accuracy حاصل ہو جاتی ہے۔ ماڈل کی کارکردگی کے نتائج کنفیوژن میٹرکس ترسیم 5.9 میں دکھائے گئے ہیں۔

ہم نے کارڈیک ڈس آرڈر کے مریضوں کی ابتدائی پیش گوئی اور شناخت کے لئے ہائپرپرامیٹرائزڈ سپورٹ ویکٹر مشین ماڈل کا اطلاق کیا۔ ماڈل کی کارکردگی کے نتائج کنفیوژن میٹرکس ترسیم 5.9 میں دکھائے گئے ہیں۔

Table 5.3 SVM Hyperparameters Optimization with their Accuracies

Kernel	***Gamma***	***Regularization***	**Accuracy**
Linear	0.001	0.11	71%
Sigmoid	0.1	1.0	70%
Sqrt	0.00001	0.001	68%
rbf	0.1	1.0	81%
Linear	0.001	0.001	72%
rbf	0.0001	0.1	80%
Linear	0.01	0.10	73%
rbf	0.0011	0.0001	78%
Sqrt	0.0001	0.010	75%
Sqrt	0.1	0.11	76%
Sigmoid	0.01	1.0	74%
Linear	0.0001	1.0	71%
Sigmoid	0.010	0.11	77%
Rbf	0.11	0.0001	69%
Sqrt	0.10	0.001	73%

کنفیوژن میٹرکس 5.9 میں مساوات (3.11) کا استعمال کرتے ہوئے ہمیں Optimized سپورٹ ویکٹر مشین ماڈل کی True Positive Rate حاصل ہوئی ہے جو حل کرنے کے بعد 0.80% کے برابر ہے۔ اسی طرح کنفیوژن میٹرکس 5.9 میں مساوات (3.12) کا استعمال کرتے ہوئے ہمیں ڈیولپڈ سپورٹ ویکٹر مشین ماڈل کی True Negative Rate حاصل ہوئی ہے جو کہ 0.82 فیصد کے برابر ہے۔ optimized سپورٹ ویکٹر مشین ماڈل کی Accuracy کنفیوژن میٹرکس 5.9 میں مساوات (3.13) کا استعمال کرکے حاصل کی گئی ہے جو 0.81% ہے۔ اسی طرح optimized سپورٹ ویکٹر مشین ماڈل کی precision حاصل کرنے کے لئے مساوات (3.14) کا استعمال ہوا ہے جو کہ تشخیص کے بعد 0.86 % ہے۔ ڈیولپڈ optimized سپورٹ ویکٹر مشین ماڈل کی misclassification rate مساوات (3.15) کا استعمال کرکے حاصل کی جاتی ہے جو 0.18 % کے برابر ہے۔

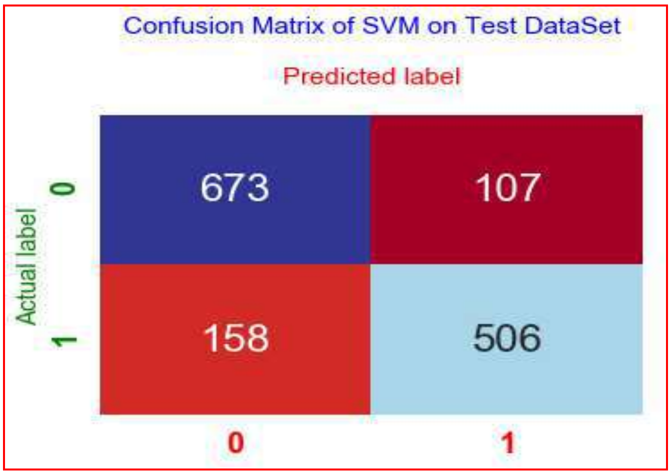


Figure 5.9 Confusion Matrix of the Optimized SVM Model

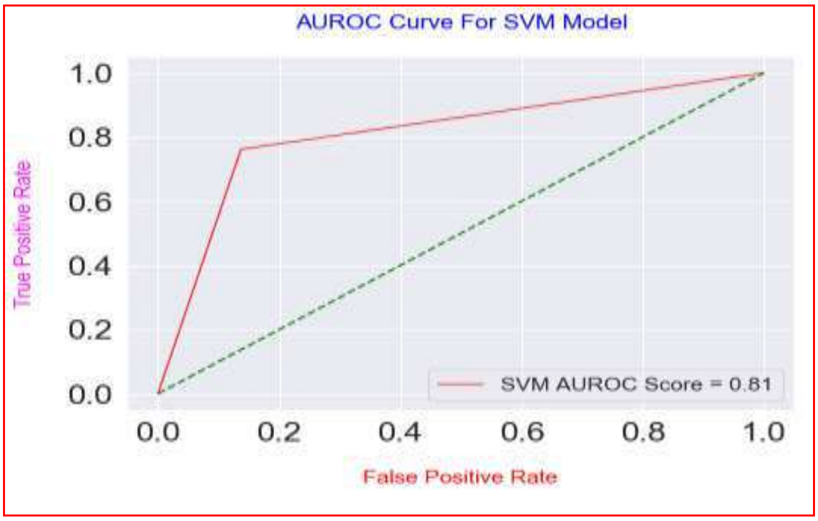


Figure 5.10 AUROC of the Optimized SVM Model

AUROC کارکردگی کی پیمائش کا استعمال یہ دیکھنے کے لئے کیا جاتا ہے کہ ڈیولپڈ optimized سپورٹ ویکٹر مشین ماڈل مریض اور صحتمندوں میں کتنا اچھا فرق کرسکتا ہے۔ بہتر ماڈل بیمار اور غیر مریض متاثرین میں بالکل مختلف ہے تاہم ناقص ماڈلز کو دونوں کے درمیان فرق کرنے میں دشواری کا سامنا کرنا پڑتا ہے۔ دیئے گئے ترسیم 5.10 کے تحت AUROC optimized سپورٹ ویکٹر مشین ماڈل سے حاصل کیا گیا ہے جس کی AUROC اسکور 0.81% ہے۔ ڈیولپڈ optimized سپورٹ ویکٹر مشین ماڈل کے کامیاب تجرباتی نتائج کو موجودہ تحقیق کے ساتھ تیار کیا گیا ہے۔ ہارٹ ڈیزیز رسک اوالویشن کے لئے سپورٹ ویکٹر مشین ماڈل کے تجرباتی نتائج بہترین نہیں ہیں۔ عملی تطبیق کے لئے اس کے استعمال کو روکنا ہے کیونکہ ڈیولپڈ رسک اوالویشن ماڈل کی time complexity زیادہ ہے اور اس سے overfitting مسئلہ درپیش ہے جو امراض قلب کی غلط تشخیص کا باعث بنتا ہے۔

5.2.4 رینڈم فارسٹ ہائپر پرامیٹر آپٹیمائزیشن ماڈل

رینڈم فاریسٹ الگورتھم ڈیٹاسیٹ کے ذیلی نمونوں کی مختلف اقسام پر متعدد ڈسیشن ٹری کو فٹ بیٹھتا ہے۔ رینڈم فاریسٹ الگورتھم کی پیش گوئی کی درستگی کو بہتر بنانے اور زیادہ مناسب مسئلہ کو کنٹرول کرنے کے لئے اوسطا استعمال کرتا ہے۔ اس ریسرچ میں رسک ماڈل ڈویلپمنٹ کے لئے رینڈم فاریسٹ کے درجہ بند کرنے والے انتہائی متاثر کن ہائپر پرمیٹرز کی کھوج اور ترتیب دی جاتی ہے جس پر ذیل میں تبادلہ خیال کیا گیا ہے:

i. N Estimators: رینڈم فاریسٹ الگورتھم کا یہ ہائپر پرامیٹر فاریسٹ میں ٹریز کی کل تعداد کی نمائندگی کرتا ہے۔ اگر رینڈم فاریسٹ میں بڑی تعداد میں ٹریز موجود ہیں تو الگورتھم ڈیٹا سے صحیح طور پر سیکھ سکتے ہیں۔ اس تحقیقی کام میں execution process کو 32 ٹریز پر روکا گیا ہے کیونکہ زیادہ تعداد میں ٹریز کا اضافہ ماڈل کی کارکردگی کو کم کرتا ہے۔

ii. Max Depth : Max Depth ہائپر پرامیٹر اس بات کی نمائندگی کرتا ہے کہ فاریسٹ میں ہر ٹری کتنا deep ہو سکتا ہے۔ Max Depth والا ٹری ڈیٹا سے مزید معلومات حاصل کرتا ہے۔ اس تحقیقی کام میں ہم ہر ایک ٹری کو max depth ہائپر پرامیٹر ویلیو کے ساتھ فٹ کرتے ہیں جس کی ویلیو 1 سے 32 کے درمیان ہے اور پھر ہم training and testing غلطیاں draw کرتے ہیں تاہم یہ پایا گیا ہے کہ اگر max depth attribute کی value high رکھا جائے تو ڈیولپڈ ماڈل overfit ہوتا ہے۔

iii. Min Samples Split : رینڈم فاریسٹ الگورتھم کا یہ ہائپر پرامیٹر اندرونی نوڈ کو تقسیم کرنے کے لئے مطلوبہ نمونے کی کم از کم تعداد کی نمائندگی کرتا ہے۔ ہر نوڈ کے نمونے تقسیم کرنے میں دوسرے نمونے سے مختلف ہوتے ہیں ۔ اگر min samples split پیرامیٹر کی value میں اضافہ کیا جائے تو ٹری زیادہ constrained ہو جاتا ہے کیونکہ اس کے بعد اسے ہر نوڈ پر مزید نمونے شامل کرنا پڑتے ہیں۔ تجرباتی نتائج سے پتہ چلتا ہے کہ جب ہر نوڈ کے تمام نمونے استعمال کیے جاتے ہیں تو ماڈل ڈیٹا سے نہیں سیکھتا ہے اور underfitting کا سبب بنتا ہے۔

iv. Min Samples Leaf: رینڈم فاریسٹ کا یہ ہائپر پرامیٹر leaf کے نوڈ پر ہونے والے نمونے کی کم از کم تعداد کی نشاندہی کرتا ہے۔ اگر اس ہائپر پرامیٹر کی پیرامیٹرک ویلیوز زیادہ مقرر کی گئی ہے تو پھر یہ underfitting کا باعث بنتی ہے۔

v. Max Features: یہ ہائپر پرامیٹر بہترین اسپلٹ کی تلاش کرتے وقت شامل صفات کی تعداد کی نمائندگی کرتا ہے اور تاہم زیادہ سے زیادہ خصوصیات کی value کو مقرر کرنا ایک overfitting مسئلہ کا سبب بنتا ہے۔

Table5.4 Experimental Results of the Optimized Random Forest Model

Criterion	**Max Depth**	**Max Features**	***N Estimators***	**Min Samples Leaf**	**Accuracy**
Gini	70	Not Used	Not Used	Not Used	85%
Entropy	60	Auto	Not Used	Not Used	86%
Gini	50	Auto	100	Not Used	87%
Entropy	80	Auto	100	100	73%
Gini	100	Auto	100	50	76%
Entropy	30	Not Used	80	60	80%
Gini	40	Not Used	90	40	78%
Gini	25	Auto	70	30	75%
Entropy	20	Auto	40	25	82%
Entropy	35	Auto	30	20	81%
Gini	45	Not Used	60	35	80%

رینڈم فاریسٹ ماڈل کے ہائپرپرامیٹرز کو ٹیوننگ کرنے کے بعد ہم کارکردگی کے نتائج حاصل کرتے ہیں جس کی تفصیل ذیل میں دیئے گئے جدول 5.4 میں بیان کی گئی ہے۔ ہائپر پیرامیٹرائزڈ رینڈم فاریسٹ ماڈل کے تقویت اور امتزاج مختلف نتائج دکھاتے ہیں مگر ہم صرف ان ہی پیرامیٹرک امتزاج کو بیان کرتے ہیں جو سب سے زیادہ accuracies فراہم کرتے ہیں۔ تجرباتی نتائج سے پتہ چلتا ہے کہ جب ہائپر پر میٹر کے امتزاج کو [Criterion= Gini, Max Depth= 50, Max Features = Auto, N Estimators=100] کے بطور تشکیل دیا گیا ہے تو 87% کی maximum accuracy حاصل ہوجاتی ہے۔ Optimized رینڈم فاریسٹ ماڈل امراض قلب کی ابتدائی پیش گوئی اور شناخت کے لئے لاگو ہوتا ہے۔ ماڈل کی کارکردگی کے نتائج کنفیوژن میٹرکس ترسیم 5.11 میں دکھائے گئے ہیں۔ ترسیم 5.11 سے ہم True Positive Rate, True Negative

Rate, Precision, Accuracy, Error Rate and AUROC، اخذ کرتے ہیں جس کی تفصیل ذیل میں دی گئی ہے۔

کنفیوژن میٹرکس 5.11 میں مساوات (3.11) کا استعمال کرتے ہوئے ہمیں optimized رینڈم فاریسٹ ماڈل کی sensitivity 0.87% حاصل ہوئی ہے۔ اسی طرح کنفیوژن میٹرکس 5.11 میں مساوات (3.12) کا استعمال کرتے ہوئے ہمیں ہائپرپرامیٹرائزڈ رینڈم فاریسٹ ماڈل کی specificity حاصل ہوئی ہے جو کہ 0.84 فیصد کے برابر ہے۔ رینڈم فاریسٹ ماڈل کی Accuracy کنفیوژن میٹرکس 5.11 میں مساوات (3.13) کا استعمال کرکے حاصل کی گئی ہے جو حساب کے بعد 0.86% کے برابر ہے۔ اسی طرح ہائپرپر میٹرائزڈ رینڈم فاریسٹ ماڈل کی precision 0.86 % ہے اس کا مطلب یہ ہے کہ ہائپرپر میٹرائزڈ رینڈم فاریسٹ ماڈل میں low false positive rate ہے۔ ڈیولپڈ optimized رینڈم فارسٹ ماڈل کی misclassification rate مساوات (3.15) کا استعمال کرکے حاصل کی جاتی ہے جو 0.13% کے برابر ہے۔

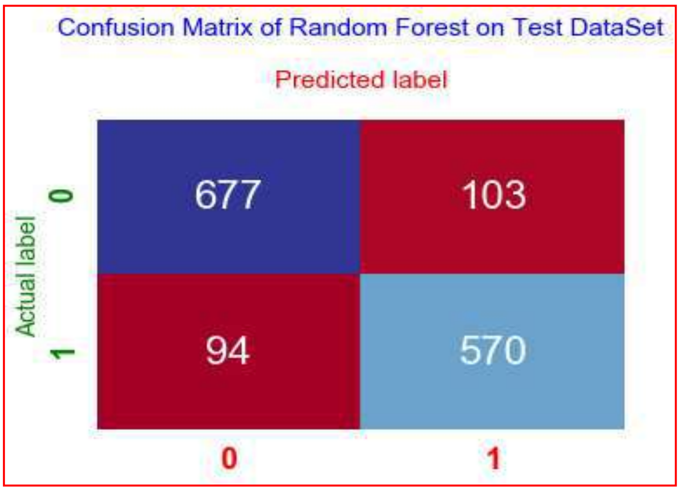


Figure 5.11 Confusion Matrix of the Optimized Random Forest Model

AUROC کارکردگی کی پیمائش کا استعمال یہ دیکھنے کے لئے کیا جاتا ہے کہ ماڈل مریض اور صحتمندوں میں کتنا اچھا فرق کر سکتا ہے۔ بہتر ماڈل بیمار اور غیر مریض متاثرین میں بالکل مختلف ہے تاہم ناقص ماڈلز کو دونوں کے درمیان فرق کرنے میں دشواری کا سامنا کرنا پڑتا ہے۔ ذیل میں دیئے گئے ترسیم 5.12 کے تحت AUROC ہائپرپر میٹرائزڈ رینڈم فاریسٹ ماڈل سے حاصل کیا گیا ہے جس کی AUROC اسکور 0.86% ہے۔

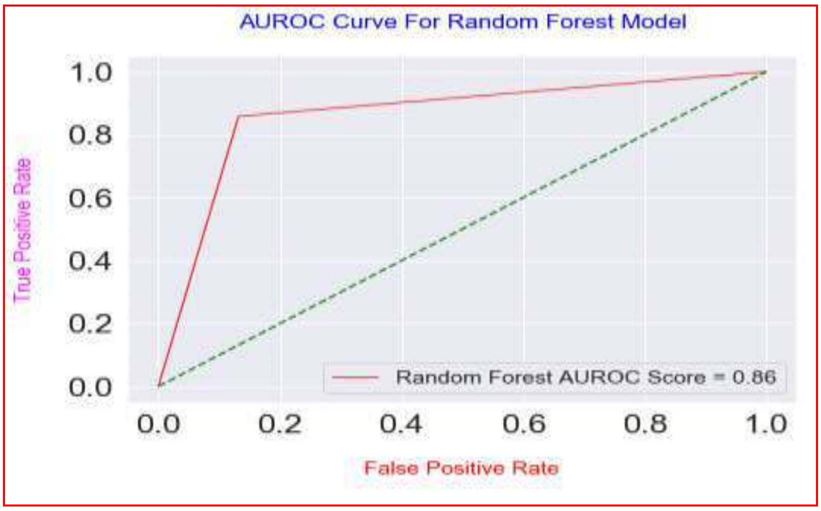


Figure5.12 AUROC of Optimized Random Forest Model

ہم مروجہ تحقیق کے ساتھ مطلوبہ رینڈم فاریسٹ ماڈل کے کامیاب تجرباتی نتائج کی نقالی کرتے ہیں۔ ہمارے مشاہدے کے متنِبک حاصل کردہ نتائج ادب میں شائع شدہ نتائج سے کہیں زیادہ بہترین ہیں۔ لہذا ہم امراض قلب کے مریضوں کی پیش گوئی کے لئے ڈیولپڈ optimized رینڈم فارسٹ ماڈل کا استعمال کرتے ہیں البتہ ماڈل کی کارکردگی میں مزید بہتری کی ضرورت ہے۔ لہذا ہم امراض قلب کے متاثرین کی ابتدائی پیش گوئی کے لئے optimized رینڈم فارسٹ ماڈل کا استعمال کرتے ہیں۔

Table 5.5 Performance Measures of Different Optimized Heart Disease Models

	Performance Measures of the Models					
Models	**TPR**	**TNR**	**Accuracy**	**Precision**	**Error Rate**	**AUROC**
Decision Tree	0.83%	0.80%	0.82%	0.82%	0.5%	0.82%
K Nearest Neighbor	0.87%	0.81%	0.84%	0.83%	0.15%	0.85%
Support Vector Machine	0.80%	0.82%	0.82%	0.86%	0.18%	0.82%
Random Forest	0.87%	0.84%	0.87%	0.86%	0.13%	0.86%

5.3۔ ڈیولپڈ ہارٹ ڈیزیز رسک ماڈلز میں کارکردگی کا موازنہ

اس سیکشن میں ڈیولپڈ ہارٹ ڈیزیز رسک ماڈلز کی تشخیص اور ان کا موازنہ بیان کیا گیا ہے اور ان ماڈلز کی کارکردگی کو جانچنے کے لئے مختلف

اقدامات استعمال کیے گئے ہیں جن کو ذیل میں دیئے گئے جدول 5.5 میں بیان کیا گیا ہے۔

Performance results یہ ظاہر کرتے ہے کہ ڈیولپڈ رینڈم فارسٹ ماڈل دوسرے رسک ماڈلز سے بہتر ہے۔ ڈیولپڈ ماڈل کی کارکردگی کی تصدیق موجودہ ڈیزائنوں کے ساتھ کی گئی ہے جس سے یہ ظاہر ہوتا ہے کہ نتائج شاندار پیش گوئی کرنے والے فعل کے ساتھ بالکل وابستہ ہیں۔ تجرباتی نتائج کے غیر یقینی جانچ کے بعد ہم سمجھتے ہیں کہ قیمتی معلومات کو حاصل کرنے اور ماڈلز کی نشوونما کے لئے اعداد و شمار کی درست جانچ پڑتال اور حساب کتاب کرنا ضروری ہے۔ نتائج سے پتہ چلتا ہے کہ رینڈم فارسٹ ماڈل کی کم سے کم غلط شرح بندی صرف 0.13% ہے اور درستگی زیادہ سے زیادہ 87% ہے۔

دیئے گئے ترسیم 5.13 مختلف **optimized** ہارٹ ڈیزیز رسک اوالویشن ماڈلکے مشترکہ **AUROC** منحنی خطوط ظاہر کرتی ہے۔ نتائج سے یہ واضح ہے کہ رینڈم فارسٹ ماڈل دل کی بیماریوں کے خطرے سے متعلق تشخیصی ماڈل میں سب سے زیادہ **AUROC score 0.87%** حاصل کیا ہے جس کا مطلب یہ ہے کہ ماڈل مریض اور غیر مریضوں میں فرق کرنے کی بہترین صلاحیت رکھتا ہے تاہم اس میں مزید بہتری کی ضرورت ہے۔

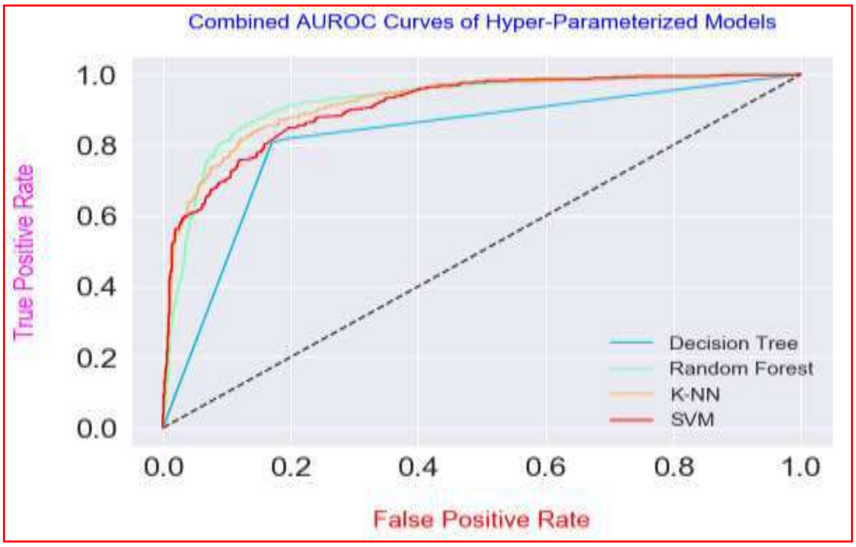


Figure 5.13 Combined AUROCs of the Optimized Risk Evaluation Models

5.4۔ ڈیفالٹ اور optimizedرسک اوالویشن ماڈلز کے مابین کارکردگی کا موازنہ

اگرچہ ہارٹ ڈیزیز ڈس آرڈر پوری دنیا میں اموات کا بنیادی سبب ہے تاہم اس کی نشاندہی سب سے زیادہ روک تھام اور قابل کنٹرول بیماریوں میں ہوتا ہے۔ صحت مند غذا، باقاعدگی سے جسمانی ورزش اور تمباکو کی مصنوعات سے گریز امراض قلب سے کم سے کم 80% بچا جا سکتا ہے۔ امراض

قلب کا جلد پتہ لگانے کا مقصدر اس مہنگی بیماری سے بڑھنے میں کمی لانا ہے۔ ان چیزوں کو دھیان میں رکھ کر ہارٹ ڈیزیزرسک ماڈل اعلی پیش گوئی کرنے والی طاقت کے ساتھ بیماری کے ابتدائی تشخیص کے لئے مختلف ڈیٹا مائننگ کے طریقوں کا استعمال کرتے ہوئے تیار کیا جاتا ہے۔ کارکردگی کو بہتر بنانے اور ہارٹ ڈیزیزرسک ماڈل کی غلط درجہ بندی کی شرح کو کم سے کم کرنے کے لئے ہم ماڈلز کے ہائپرپرامیٹرز کو بہتر بناتے ہیں۔ ذیل میں دیئے گئے جدول 5.6 میں پہلے سے طے شدہ اور ڈیولپڈ optimized ہارٹ ڈیزیزرسک اوالویشن ماڈل کے مابین موازنہ بیان کیا گیا ہے۔

Table 5.6 Performance Comparison of Proposed Heart Disease Models

Performance Measures	Comparison among Different Heart Disease Risk Evaluation Models								
	Decision Tree		**K NN**		**SVM**		**Random Forest**		**Naive Bayes**
	DMP	***HPT***	***DMP***	***HPT***	***DMP***	***HPT***	***DMP***	***HPT***	***DMP***
TPR	0.82	0.83	0.73	0.87	0.82	0.80	0.85	0.87	0.72
TNR	0.80	0.80	0.66	0.81	0.81	0.82	0.83	0.84	0.66
Accuracy	0.81	0.82	0.70	0.84	0.82	0.82	0.84	0.87	0.69
Precision	0.84	0.82	0.69	0.83	0.84	0.86	0.85	0.86	0.70
Error Rate	0.18	0.05	0.30	0.15	0.17	0.18	0.5	0.13	0.30
AUROC	0.81	0.82	0.70	0.85	0.82	0.82	0.85	0.86	0.70

TPR means (True Positive Rate), TNR (True Negative Rate), DMP means (Default Model Parameters) and HPT means (Hyperparameter Tuning).

تجرباتی نتائج وضاحت کرتی ہے کہ رینڈم فارسٹ ماڈل دوسرے ماڈلز کو default اور optimized کی ترتیبات پر بہتر بنا دیتا ہے۔ ڈیولپڈ کارڈیک ڈس آرڈر رسک اوالویشن ماڈل کی کارکردگی کی تصدیق موجودہ ڈیزائنوں سے کی گئی ہے جس سے یہ ظاہر ہوتا ہے کہ نتائج شاندار پیش گوئی کرنے والے فعل کے ساتھ بالکل وابستہ ہیں۔ تجرباتی نتائج کے غیر یقینی جانچ کے بعد ہم سمجھتے ہیں کہ قیمتی جانکاری حاصل کرنے اور ماڈل کی نشوونما کے لئے اعداد و شمار کی درست جانچ پڑتال اور حساب کتاب کرنا ضروری ہے۔

5.5۔ ابتدائی بیماری کی پیش گوئی اور شناخت کے لئے ہارٹ ڈیزیزرسک فیچرز کے مختلف مجموعے

اس سیکشن میں ہارٹ ڈیزیز prognosis and identification کے لئے نان انویسو رسک فیچرز میں سے مختلف ترتیب کے نتائج پیش کی جاتی ہیں. ذیل میں دیئے گئے جدول 5.7 ہارٹ ڈیزیز کی پیش گوئی میں نان انویسو رسک فیچرز کے مختلف مجموعوں کی کارکردگی کو ظاہر کرتا ہے۔ سسٹولک بی پی، Diastolic BP ، heredity and age کے مجموعے کی طرف سے ڈیسیشن ٹری نے 77.3% کی بہترین درستگی حاصل کی ہے. ہم تمام اوصاف کے امتزاج کی sensitivity اور specificity کی پیائش بھی کیے ہے۔ یہاں sensitivity مناسب دیکھ بھال فراہم کرنے کے لئے بیمار مقدمات کی تشخیص میں سب سے زیادہ مؤثر ہے۔ [عمر، سسٹولک بی پی، ڈیاسٹولک بی پی، اور ہیریڈٹی] کے امتزاج کے ساتھ BMI (اونچائی اور وزن) کی خصوصیت شامل کرنے سے رینڈم فارسٹ ماڈل کے ذریعہ ایک درستگی میں 78.9 فیصد تک اضافہ ہوا ہے۔

Table5.7 Integrating Different Non-Invasive Heart Disease Risk Factors

Techniques	Risk Attributes	Sensitivity	Specificity	Accuracy
Decision Tree	Systolic BP, Diastolic BP, Age, Heredity	78%	80%	77.3%
	Systolic BP, Diastolic BP, Age, BMI	72%	70%	70.9%
	Age, Healthy Diet, BMI	68%	61%	63.3%
	Systolic BP, Diastolic BP, Age, Physical Activity	53%	60%	58.6%
	Healthy Diet, BMI, Physical Activity, Age	58%	41%	50.9%
	Healthy Diet, Physical Activity, Age, Systolic BP, Diastolic BP	45%	43%	42.5
	Physical Activity, Age, Healthy Diet, BMI, Systolic BP, Diastolic BP	38%	30%	38.2%
	Age, Physical Activity, Smoking, Systolic BP, Diastolic BP, Healthy Diet, Alcohol Consumption, BMI	30%	28%	42.7%
K Nearest Neighbor (KNN)	Age, Healthy Diet, Alcohol Consumption, Smoking	42%	45%	38.2%
	Age, BMI, Healthy Diet	70%	60%	67.9%
	Age, BMI, Alcohol Consumption, Smoking, Sex	52%	50%	48.9%

	BMI, Systolic BP, Diastolic BP, Age, Physical Activity	38%	35%	42.7%
	BMI, Systolic BP, Diastolic BP, Age	68%	74%	72.5%
	Age, Systolic BP, BMI, Diastolic BP, Heredity	68%	70%	72.8%
Random Forest	Systolic BP, Diastolic BP, Age, Healthy Diet, Smoking	51%	48%	45.4%
	BMI, Age, Systolic BP, Diastolic BP, Heredity	72%	78%	78.9%
	Alcohol Consumption, Physical Activity, Age, Systolic BP, Diastolic BP, BMI, Smoking, Healthy Diet	35%	45%	58.7%
	Age, Sex, Physical Activity, BMI,	32%	34%	40.8%
	Age, Sex, Physical Activity, BMI, Systolic BP, Diastolic BP	39%	45%	42.6%
Support Vector Machine (SVM)	Systolic BP, Diastolic BP, Age	72%	62%	76.1%
	Systolic BP, Diastolic BP, Age, BMI, Heredity	70%	78%	75.2%
	Healthy Diet, Age, BMI	41%	53%	50.9%
	Systolic BP, Diastolic BP, Age, BMI, Physical Activity	50%	44%	51.6%
	BMI, Physical Activity, Alcohol Consumption, Age	49%	50%	52.4%
	Age, Alcohol Consumption, BMI, Healthy Diet	41%	59%	52.2%
Naive Bayes	Systolic BP, Diastolic BP, Age	74%	78%	75.1%
	Age, Alcohol Consumption, Healthy Diet, Sex, BMI	40%	44%	48.8%
	Systolic BP, Diastolic BP, Age, BMI, Heredity	68%	75%	77.2%
	Systolic BP, Diastolic BP, Alcohol Consumption, Heredity, Age, BMI, Smoking, Healthy Diet, Sex, Physical Activity,	46%	51%	50.6%

پیش گوئی کرنے والے خطرے کے قواعد کا ایک زیادہ سے زیادہ مجموعہ مذکورہ بالا صفت انتساب کے امتزاج کا استعمال کرتے ہوئے تیار کیا گیا ہے جو امراض قلب سے متاثرہ افراد کی ابتدائی تشخیصی اور شناخت میں مدد کرتے ہیں۔ امراض قلب کے خطرے سے متعلق تشخیصی قواعد کو مختلف میڈیکل ڈومین ماہرین کے ذریعہ evaluate and validate کیا گیا ہیں تاہم ان کا استعمال ممنوع ہے کیونکہ نکالا ہوا قاعدہ مخصوص نسلی ہارٹ ڈیزیز ڈیٹا سیٹ پر مبنی ہیں۔

5.6۔ ہارٹ ڈیزیز ایکسپرٹ سسٹم اوالویشن ماڈل کمپوننٹس

ہم نے امراض قلب کے خطرے سے متعلق تشخیصی ماڈل تیار کیا ہوا ہے کیونکہ یہ دل کے مریضوں کے خطرے کی ڈگری کی نشاندہی کرتا ہے جو صرف نان انویسو رسک فیچرز کا استعمال کرتا ہے، اس طرح اس کی پبلک اسکریننگ جانچ کے طور پر اس کی درخواست کی حمایت ہوتی ہے۔ For simplicity ہم نے اس ماڈل کو HDREM ماڈل نام رکھا ہے۔ ترسیم 5.14 میں HDREM اور ان کے کام کرنے کے تین اہم اجزاء دکھائے گئے ہیں: Knowledge Base, Inference Engine andUser Interface.

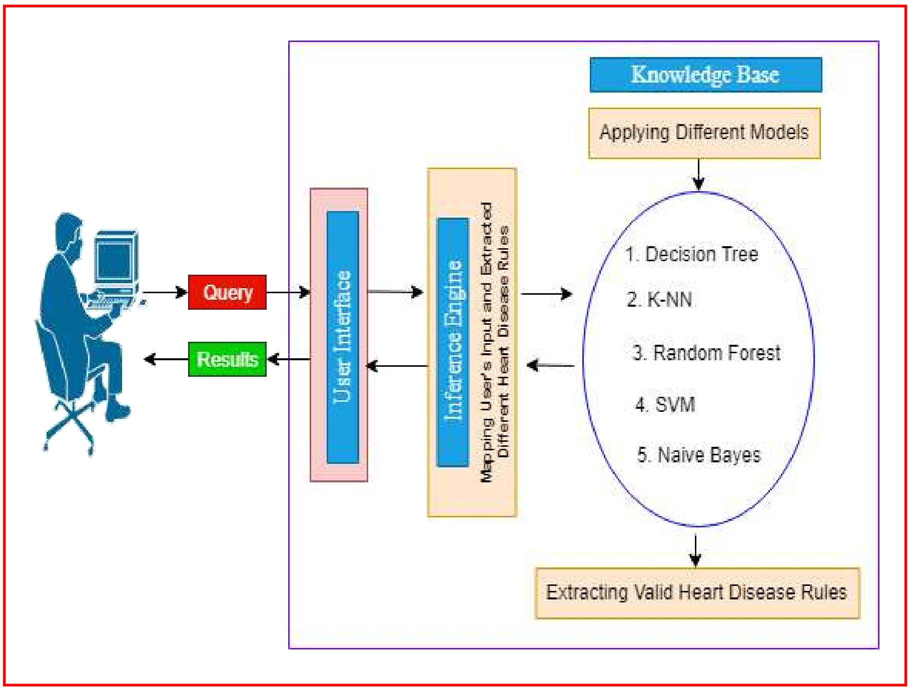

Knowledge Base
Applying Different Models
Query
Results
User Interface
Inference Engine
Mapping User's Input and Extracted Different Heart Disease Rules
1. Decision Tree
2. K-NN
3. Random Forest
4. SVM
5. Naive Bayes
Extracting Valid Heart Disease Rules

Figure 5.14 Heart Disease Expert System Evaluation Tool Components

ماہر نظام کے قواعد کو نکالنے کے لئے Knowledge Base غیر ناگوار امراض قلب کے ڈیٹا اوصاف پر تجویز کردہ ماڈل کا اطلاق کرتا ہے۔

Inference engine نکلے ہوئے اصولوں اور صارفین کے ان پٹ کا استعمال کرتے ہیں تاکہ علم کی بنیاد سے نتائج اخذ کریں اور انٹرفیس کے

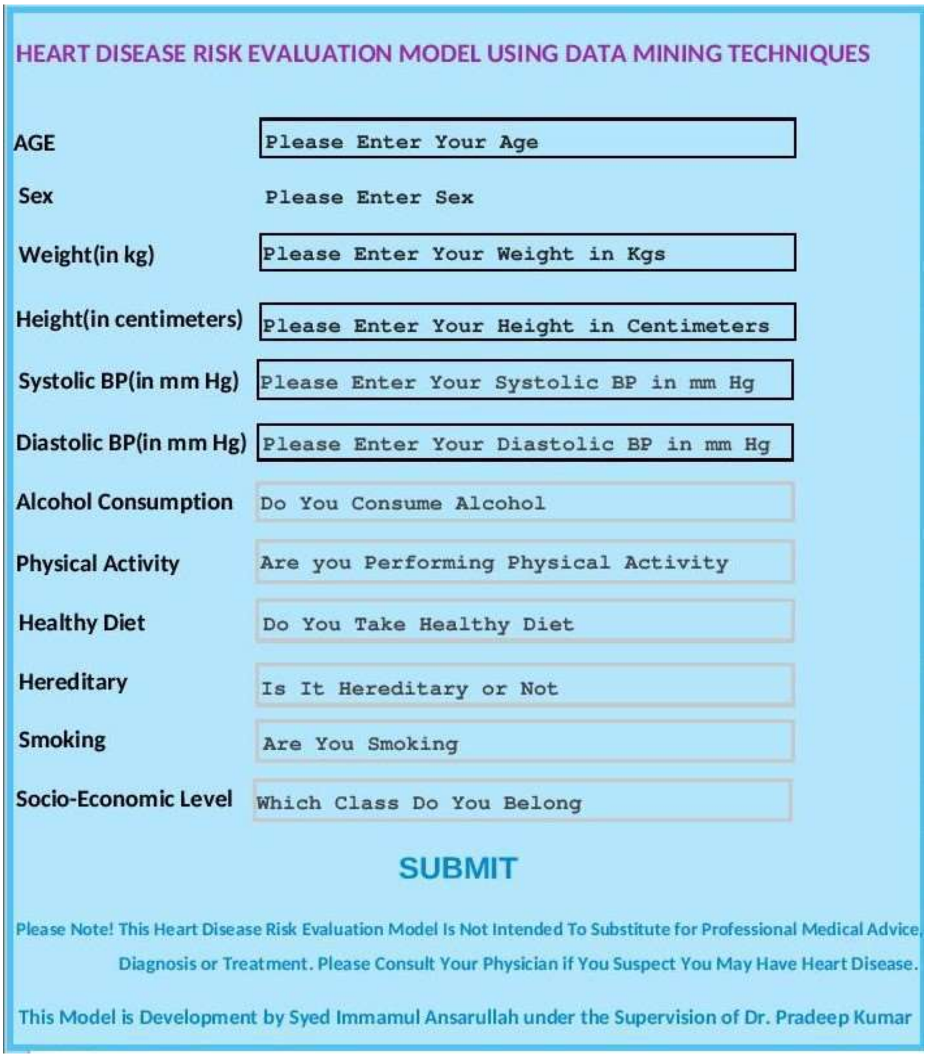
ذریعے صارف کے سامنے پیش کریں۔ User انٹرفیس "communication" اسکرینوں کی اجازت دیتا ہے جہاں user input ڈیٹا میں داخل ہوتا ہے اور expert سسٹم ہارٹ ڈیزیز رسک ڈگری return کرتا ہے جیسا کہ inference engine کے حساب سے ہوتا ہے۔

5.7 ہارٹ ڈیزیز رسک اوالویشن ماڈل (HDREM)

"In Machine Learning All Models Are Wrong But Some Are Useful" [94]

امراض قلب کے خطرے سے متعلق تشخیصی ماڈل تیار کیا گیا ہے جس کو عوامی سطح پر اسکیننگ کے لئے لاگو کیا جا سکتا ہے تاکہ وہ خطرے کی سنگین بیماری کی پیش گوئی اور تشخیص کر سکیں اور فوری مداخلت کی سہولت کے لئے معلومات فراہم کر سکیں۔ باب 4 اور باب 5 میں ڈیولپڈ ماڈلز نے دل کے عارضے میں مبتلا افراد کی پیش گوئی کرنے میں مختلف خطرات کے اوصاف کا استعمال کیا ہیں۔ نتائج سے پتہ چلتا ہے کہ عمر، سسٹولک بی پی، ڈیاسٹولک بی پی، بی ایم آئی، صحت مند غذا، موروثی اور جسمانی سرگرمی کا مجموعہ بہترین نتائج فراہم کرتا ہے۔ یہ نتائج کافی حد تک بہترین لگتے ہیں اور یہ show کرتے ہے کہ مشین لرننگ الگورتھمس جیسے کہ ڈسیشن ٹری، رینڈم فاریسٹ، کے نیریسٹ نیبر، سپورٹ ویکٹر مشین اور نیوی بائیس کا استعمال ہارٹ ڈیزیز کی تشخیص کے لئے اسکریننگ ٹیسٹ بنانے کے لئے کیا جا سکتا ہے۔ ہارٹ ڈیزیز چارٹ بنانے کے لئے قواعد نکالے جاتے ہیں کیونکہ صحت کی دیکھ بھال کے ماہرین امراض قلب کے مریضوں کے خطرہ کی ڈگری کی تشخیص کرنے کے لئے کمیونٹی اسکریننگ ٹیسٹ کی حمایت کرتے ہیں۔

HDREM ڈویلوپمنٹ منصوبہ دو بڑے مراحل پر مشتمل ہے۔ پہلے مرحلے میں اوصاف کو لوڈ کرنا اور ڈیولپڈ ماڈلز کو کشمیر ہارٹ ڈیزیز ڈیٹا سیٹ کی غیر ناگوار خصوصیات میں لاگو کرنا اور پھر تشخیصی قواعد نکال کر ذخیرہ کرنا شامل ہیں۔ دوسرے مرحلے میں، user اپنے ڈیٹا میں داخل ہوتا ہے۔ ان خصوصیات کو ذخیرہ تشخیصی قوانین کے ذریعہ استعمال کنندہ کو امراض قلب کے خطرے کی ڈگری کا حساب لگانے کے لئے استعمال کیا جاتا ہے جو user کو ظاہر ہوتا ہے۔ ایچ ڈی آر ای ایم کو Python جو پیٹر نوٹ بک کا استعمال کرتے ہوئے develop کیا گیا ہے۔ 5.15 figure ایچ ڈی آر ای ایم کی ابتدائی اسکرین کو ظاہر کرتی ہے جہاں صارف اپنا ڈیٹا داخل کرتا ہے اور پھر ہارٹ ڈیزیز رسک ڈگری کا حساب ظاہر ہو جاتا ہے۔



HEART DISEASE RISK EVALUATION MODEL USING DATA MINING TECHNIQUES

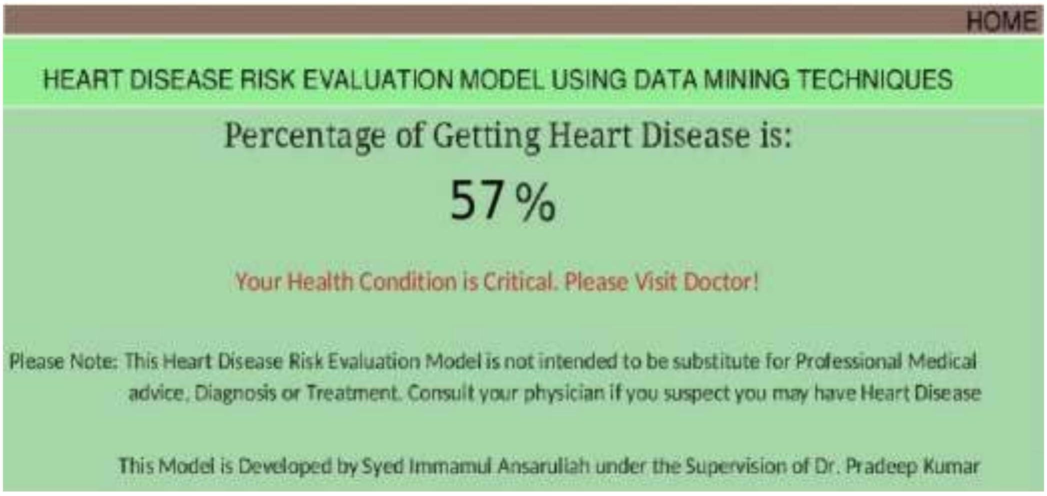
AGE	Please Enter Your Age
Sex	Please Enter Sex
Weight(in kg)	Please Enter Your Weight in Kgs
Height(in centimeters)	Please Enter Your Height in Centimeters
Systolic BP(in mm Hg)	Please Enter Your Systolic BP in mm Hg
Diastolic BP(in mm Hg)	Please Enter Your Diastolic BP in mm Hg
Alcohol Consumption	Do You Consume Alcohol
Physical Activity	Are you Performing Physical Activity
Healthy Diet	Do You Take Healthy Diet
Hereditary	Is It Hereditary or Not
Smoking	Are You Smoking
Socio-Economic Level	Which Class Do You Belong

Please Note! This Heart Disease Risk Evaluation Model Is Not Intended To Substitute for Professional Medical Advice, Diagnosis or Treatment. Please Consult Your Physician if You Suspect You May Have Heart Disease.

This Model is Development by Syed Immamul Ansarullah under the Supervision of Dr. Pradeep Kumar

Figure 5.15 The Heart Disease Risk Evaluation Model Interface

ترسیم 5.16 صارف کی طرف سے ڈیٹا کا نتیجہ ہے جو شروعاتی اسکرین کے مختلف وابستہ قدروں میں داخل ہوتا ہے۔ یہاں HDREM درج کردہ ڈیٹا کے لئے امراض قلب کے اعلی خطرہ کا حساب لگاتا ہے۔



HEART DISEASE RISK EVALUATION MODEL USING DATA MINING TECHNIQUES

Percentage of Getting Heart Disease is:

57 %

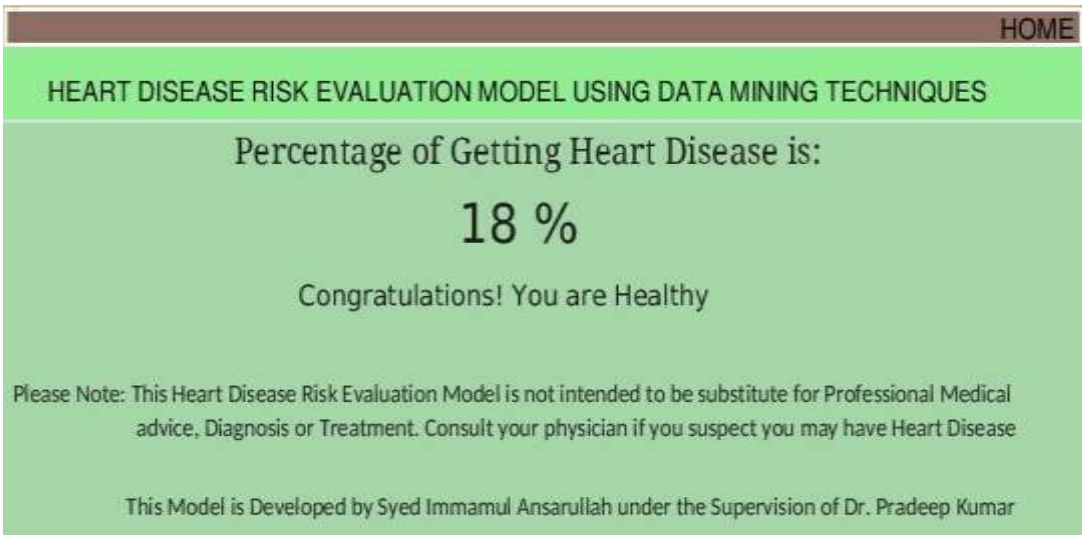
Your Health Condition is Critical. Please Visit Doctor!

Please Note: This Heart Disease Risk Evaluation Model is not intended to be substitute for Professional Medical advice, Diagnosis or Treatment. Consult your physician if you suspect you may have Heart Disease

This Model is Developed by Syed Immamul Ansarullah under the Supervision of Dr. Pradeep Kumar

Figure 5.16 High-Risk Heart Disease Evaluation Example

Figure 5.17 HDREM ماڈل کی دوسری مثال ہے، اس معاملے میں امراض قلب کا ایک کم خطرہ اسٹارٹ اسکرین میں داخل کردہ ڈیٹا کی بنیاد پر حساب کیا جاتا ہے۔



HEART DISEASE RISK EVALUATION MODEL USING DATA MINING TECHNIQUES

Percentage of Getting Heart Disease is:

18 %

Congratulations! You are Healthy

Please Note: This Heart Disease Risk Evaluation Model is not intended to be substitute for Professional Medical advice, Diagnosis or Treatment. Consult your physician if you suspect you may have Heart Disease

This Model is Developed by Syed Immamul Ansarullah under the Supervision of Dr. Pradeep Kumar

Figure 5.17 Low-Risk Heart Disease Evaluation Example

ان مثالوں سے ظاہر ہوتا ہے کہ HDREM عوامی سطح کی اسکریننگ ٹیسٹ کے طور پر کام کر سکتا ہے۔ صارف انٹرفیس کی سادگی healthcare practitioners کو انتہائی کم لاگت والی غیر ناگوار صفات کا استعمال کرتے ہوئے امراض قلب کے زیادہ خطرہ والے مریضوں کی شناخت کرنے کی اجازت دیتی ہے۔ HDREM کو موبائل کے ساتھ ساتھ desktop applications پر بھی لاگو کیا گیا ہے۔

5.8 باب کا خلاصہ

اس باب میں ہم نے hyperparameter optimization تکنیکس اور ان کی مختلف اقسام کو متعارف کرایا ہیں۔ ہم نے ابتدائی پیش گوئی میں صحت کی دیکھ بھال کرنے والے پیشہ ور افراد کی مدد کے لئے ڈیولپڈ ماڈلز کو بہتر بنا یا ہیں جس کی وجہ سے خطرناک بیماری اور پیچیدگیوں میں اضافہ کم ہوا ہیں۔ امراض قلب کے خطرے سے متعلق تشخیصی ماڈل جوپیٹر نوٹ بک ویب ایپلی کیشن پر مبنی ہے اور اس کی کارکردگی کو متنازع پیمائش کی جانچ کرنے کے لئے مختلف اقدامات جیسے TPR، TNR، Precision,Accuracy,Misclassification rate and AUROC تکنیکس کے ذریعہ مرتب کیا گیا ہے۔ تجرباتی نتائج میں ظاہر ہوتا ہے کہ ڈیولپڈ رینڈم فارسٹ ہارٹ ڈیزیز رسک اوالویشن ماڈل دوسرے ڈیولپڈ ماڈلز سے بہتر نتائج فراہم کرتا ہیں۔ اس باب میں ہارٹ ڈیزیز رسک اوالویشن ماڈل کے اجزاء اور بیماری کی ابتدائی پیش گوئی کے لئے خطرے کے قواعد کا نکالا ہوا سیٹ بھی بیان کیا گیا ہے۔

باب 6

نتیجہ اور مستقبل کے کام کا منصوبہ

اس باب میں تحقیقی نتائج کا خاکہ پیش کیا گیا ہے، تحقیقات کی حدود پر تبادلہ خیال کیا گیا ہے، اور آئندہ ہونے والے تحقیقی پہلوؤں کی وضاحت کی گئی ہے۔ متنوع خصوصیات سے امراض قلب کی پیش گوئی کرنا ایک کثیر جہتی نقطہ نظر ہے جس کے بعد اکثر غیر پیش کردہ اثرات مرتب ہوتے ہیں۔ بڑھتی ہوئی صحت کی دیکھ بھال کے اخراجات، بار بار اسپتال میں داخل ہونے اور قبل از وقت اموات نے امراض قلب کو دنیا بھر میں ایک وبا میں تبدیل کردیا ہے جو DALY (Disability Adjusted Life Years) اور YLL (Years of Life Lost) کے چارٹ میں سب سے اوپر ہے، WHO کے ذریعہ بتایا گیا ہے کہ دنیا میں امراض قلب کا سب سے زیادہ بوجھ ہے۔ اگرچہ امراض قلب سب سے زیادہ پھیلی دائمی حالت میں سے ایک ہے جس کی وجہ سے دنیا بھر میں اموات کی زبردست شرح ہوتی ہے، لیکن اسے ایک قابل علاج اور قابل کنٹرول بیماری کے طور پر تسلیم کیا جاتا ہے اگرچہ نئی تشخیصی ایجادات اب غور و فکر کے معیار میں تبدیل ہو چکی ہیں لیکن پھر بھی یہ طرز عمل حد سے زیادہ اور عملی طور پر پیچیدہ ہیں جو دیہی علاقوں میں، بنیادی صحت کی دیکھ بھال کے انتظامات اور عوامی سطح پر اسکریننگ کی تشخیصی پر ان کے استعمال کو محدود کرتے ہیں۔ لہذا ہم امراض قلب سے متاثرہ افراد کی ابتدائی شناخت اور اس کی آسانی کے لئے ایک رسک اوالویشن ماڈل تیار کیا ہیں۔ اس مقصد کو حاصل کرنے کے لئے، تحقیق ایک سوال پیدا کرتی ہے۔

Can data mining support medical specialists in the early identification and examination of heart disease at a public setting?

محققین نے مختلف ڈیٹا مائننگ ٹاسکس، مختلف ڈیٹا سیٹس، مختلف مشین لرننگ الگورتھم، اور بہت سارے techniques کا استعمال کرتے ہوئے امراض قلب کی نشاندہی کرنے میں فیصلہ کن شراکت کی ہے تاہم ہر ڈیزائن میں مسائل ہیں۔ اس تحقیق کا بنیادی مقصد مندرجہ بالا سوال کا جواب دینا ہے۔ غیر معمولی رسک تشخیص اور انتہائی پیش گوئی کی صلاحیت کے ساتھ امراض قلب کے مریضوں کی نشاندہی کے لئے میڈیکل پریکٹیشنرز کو ایک عوامی سطح کا اسکریننگ ماڈل فراہم کرنا۔

امراض قلب کے خطرے کا جلد پتہ لگانے کے مقصد کو حاصل کرنے کے لئے یہ تحقیق درج ذیل اہم سوالات پر مرکوز ہے:

i. کیاامراض قلب کے مریضوں کی پیش گوئی میں اہم خصوصیات کا تعین کیا جاسکتا ہے؟

ii. باب 3 وضاحت کرتا ہے کہ امراض قلب کے خطرے کی تشخیص کے لئے نمایاں اوصاف قابل شناخت ہیں۔

iii. امراض قلب کے مریضوں کی بروقت تشخیص کے لئے کیا نان انویسو رسک فیچرز پر ڈیٹا مائننگ techniques کو استعمال کیا جاسکتا ہے؟

iv. باب 4 اس پر تبادلہ خیال کرتا ہے کہ ہم ڈیٹا مائننگ techniques کو مؤثر طریقے سے پیش گوئی کی درستگی کے ساتھ امراض قلب کے ابتدائی مرحلے میں تشخیص کرنے کے لئے غیر ناگوار صفات پر مؤثر طریقے سے استعمال کر سکتے ہیں۔

v. کیا دل کے عارضے میں مبتلا افراد کی جلد پیش گوئی کے لئے نان انویسو رسک فیچرز کی کارکردگی بڑھانے کے لئے ہائپر پیرامیٹر کی اصلاح کے طریقوں کو استعمال کیا جاسکتا ہے؟

vi. باب 5 اس پر گفتگو کرتا ہے کہ ہم امراض قلب کے خطرے کی جانچ پڑتال کے ماڈل کی درستگی کو بڑھانے کے لئے دل کی بیماریوں کی نمایاں خصوصیات میں اصلاح کے طریقوں کو کامیابی کے ساتھ استعمال کر سکتے ہیں۔

vii. کیا نان انویسو رسک فیچرز کا استعمال کرتے ہوئے امراض قلب کے خطرے سے متعلق تشخیصی ماڈل تیار کیا جاسکتا ہے؟

viii. باب 6 بیان کرتا ہے کہ غیر حملہ آور اعداد و شمار کے اوصاف کا استعمال کرتے ہوئے قابل اعتماد امراض قلب کے خطرے سے متعلق تشخیصی ماڈل جاسکتا ہے۔ تیار کیا

اس تحقیقی کام میں درپیش سوالوں کے جوابات کی تائید اور تصدیق کرنے والے ابواب 3 سے 6 کے نتائج بھی اس بنیادی سوال کی تائید کرتے ہیں کہ عوامی اسکریننگ کے ماحول میں ڈیٹا مائننگ میڈیکل پریکٹیشنرز کی مدد کر سکتی ہے۔

6.1 تحقیقی کام کا خلاصہ

ہارٹ ڈیزیز رسک اوالویشن کے لئے جن آلات کا دور حاضر میں رواج ہے وہ میڈیکل اور لیبارٹری ٹیسٹس پر مبنی ڈیٹا کو استعمال میں لاتے ہیں اور یہ طریقے پر خطر ماحول میں مرض کی صحیح تشخیص کرنے کی صلاحیت نہیں رکھتے۔ ایسے میں ایک ایسا طریقہ دریافت کرنے کی ضرورت ہے جو جان لیوا صورتحال میں بھی مریض کی صحیح تشخیص میں مددگار ثابت ہو تو ایسے میں مریض ہسپتال پہنچنے سے قبل ہی خطرات سے آشنا ہو کر صحیح اقدام اٹھا

سکتا ہے۔ دیہی علاقوں میں ایسی دریافت سے مہنگے علاج و معالجہ سے بچا جا سکتا ہے۔ یہ طریقے بنیادی ہفظانِ صحت کیلئے مفید ہو سکتے ہیں اور مریض کی قبل از وقت تشخیص کیلئے معاون ہو سکتے ہے۔ اس تحقیقی کام میں ہم نے ایک ایسا ہی تشخیصی ماڈل دریافت کیا ہے جو بالفعل نان انویسیو طرز پر کام کرتا ہے۔ اس کے نفاذ کے لئے ہم نے جو پیٹر نوٹ بک ویب ایپلیکیشن کا استعمال کیا ہے۔ ہم نے کشمیر کے ان افراد کا معائنہ کیا ہے جو دل کے امراض میں مبتلا ہیں۔ مریضوں سے حاصل شدہ اعداد و شمار پر ہم نے ,رینڈم فارسٹ، نیوی بائیس، کے نیریسٹ نیبر، سپورٹ ویکٹر مشین اور ڈیسیشن ٹری الگورتھم کا نفاذ عمل میں لایا ہے تاکہ بیماری کی علامات کا تعین کیا جا سکے۔ تمام کلاسیفائرس کیلئے ہم نے sensitivity, specificity, accuracy, precision, error rate and AUROC کو چیک کیا ہے تاکہ تجویز کردہ ماڈل کی پوری جانچ کی جا سکے۔ یہ زور دینے کے قابل ہے کہ اکثر معاملات میں صحیح نتائج حاصل کرنے کے لئے ہائپر پیرامیٹر ٹیوننگ کی ضرورت ہوتی ہے۔ تجربات کی بنیاد پر ہم نے دیکھا کہ رینڈم فاریسٹ طریقے پر مبنی ماڈل دوسرے طریقوں ,sensitivity, specificity, accuracy, precision, error rate and AUROC کے حوالے سے فوقیت رکھتا ہے۔ ہم نے جو نتائج حاصل کئے ہیں ان کا مقابلہ ہم نے دوسرے تقریباً تمام طریقوں سے کیا ہے اور یہ ثابت کیا ہے کہ ہمارا ماڈل ان سب سے بہتر ہے۔ اس تحقیقی کام کا مقصد یہ ہے کہ آیا ڈیٹا مائننگ کی تکنیک نان انویسیو طور پر دلی امراض کی تشخیص کے لئے کارآمد ثابت ہو سکتی ہے کہ نہیں اور اس تکنیک کو رفاع عامہ کے لئے کیسے مفید بنایا جا سکتا ہے تاکہ ایک رسک اویلویشن ماڈل تیار ہو سکے۔ ذیل میں ہم نے اس تحقیقی کام کی دل کے امراض کی تشخیص کے حوالے سے انجام دی گئی خدمات کا تفصیل سے جائزہ لیا ہے۔

6.1.1 ہارٹ ڈیزیز رسک تشخیص میں اہم خصوصیات

باب 3 میں امراض قلب کی ابتدائی تشخیص کے لیے اہم اوامر کی نشاندہی کی گئی ہے۔ اس باب میں ان 5 علامات کو بیان کیا گیا ہے جو امراض قلب کا موجب بنتے ہیں۔ ان ۵ علامات کی بنیاد پر امراض قلب کی تشخیص کی گئی ہے۔ ہر علامت کو ایک وزن تفویض کیا گیا ہے اور اس وزن کی بنیاد پر خطرے کو ماپا گیا ہے۔ آخر میں تمام علامات کے وزن کو حاصل کر کے ایک ماڈل کے اعتبار سے تشخیصی عمل تشکیل دیا گیا ہے۔ جس علامت کو مقدار کے اعتبار سے زیادہ وزن دیا گیا ہے وہ علامت دل کو امراض کی تشخیص کے لئے اتنی ہی اہم گردانی گئی ہے اور اس ماپ سے قبل از وقت مرض کی

تشخیصی میں مدد ملی ہے۔ تفویض کردہ اعدادی وزن کو پھر طبّی ماہرین نے منظوری دی اور تصدیق کر دی کی ڈیٹا مائننگ کا یہ تشخیصی طریقہ کار دل کے امراض کی تشخیصی کے لئے استعمال میں لایا جا سکتا ہے۔

6.1.2 ہارٹ ڈیزیز رسک اوالویشن میں نان انویسو رسک فیچرز کی اہمیت

باب 4 اور 5 میں دل کے امراض کے خطرات کی تشخیصی علامات کی اہمیت کو تفصیل سے بیان کیا گیا ہے۔ یہ نان انویسیو علامات نہایت ہی اہمیت کی حامل ہیں کیونکہ ان کی نشاندہی آسان ہے اور یہ کم لاگت پر حاصل کی جا سکتی ہیں۔ عوامی سطح پر اسکریننگ ٹیسٹ کر کے ان علامات کی شناخت کی جا سکتی ہے اور خطرات کا تعین کیا جا سکتا ہے۔ باب ۵ میں علامات کے اس مجموعے کو تلاش کرنے کی کوشش کی گئی ہے جو دل کے امراض کی تشخیص میں بہتر طور پر رہنمائی کر سکے۔ ٹیبل 5.7 میں ان نتائج کو دکھایا گیا ہے جو علامات کے مختلف مجموعوں پر ڈیٹا مائننگ تکنیکوں کے استعمال کے بعد حاصل ہوئے ہیں۔ علامات کے مختلف مجموعوں میں جو عناصر استعمال ہوئے ہیں ان میں عمر، جنس، اسٹالک بی پی، BMI اور ہیریڈیٹی شامل ہیں۔ ان عناصر کی جانچ کے بعد جو نتائج حاصل ہوئے ہیں وہ دلچسپ ہونے کے علاوہ تشخیصی عمل کے لئے بہت اہم ثابت ہو سکتے ہیں۔

6.1.3 ہارٹ ڈیزیز رسک اوالویشن ماڈل(HDREM)ڈیولپمنٹ

اگرچہ دل کے امراض کی تشخیصی مختلف ٹیسٹس جیسے کہ ای۔ سی۔ جی، سٹرس ٹیسٹ، Cardiac angiogram وغیرہ کے مدد سے کیا جاتا ہے لیکن یہ ٹیسٹس مہنگے ہوتے ہیں اور ان کو عمل میں لانا بہت دشوار ہوتا ہے۔ خاص کر دیہی علاقوں میں یہ ٹیسٹ کرنا ممکن نہیں ہوتے۔ ان محدودیات کو نظر میں رکھ کر ہم نے ایک ایسا رسک ماڈل تیار کرنے کی کوشش کی ہے جو قبل از وقت مرض کا تعین کرنے میں مددگار ثابت ہو۔ اس تحقیقی کام میں ہم نے فیچر سلیکشن کی ۵ تکنیکس استعمال کی ہیں۔ اس کے لئے ہم نے کشمیر کے ان افراد سے ڈیٹا حاصل کئے ہیں جو دلی امراض میں مبتلا ہیں۔ ان ڈیٹا کا اوسط نکال کے ہم نے optimal علامات استعمال میں لایے ہے۔ ہم نے ۵ مستند classification algorithms جیسے کہ رینڈم فارسٹ، نیوی بائیس، کے نیریسٹ نیبر، سپورٹ ویکٹر مشین اور ڈسیشن ٹری پر تجربات کئے ہیں۔ اس عمل کے لئے ہم نے نان انویسیو علامات کا استعمال کیا ہے جو کم لاگت پر حاصل ہوتی ہیں اور تشخیصی اہمیت رکھتی ہیں۔ دل کے امرض کی شناخت کے لئے ہم نے ڈسیشن ٹری اور رینڈم فارسٹ کے عوامل کو استعمال کیا ہے۔

6.2 تحقیق کی محدودیات

اس تحقیق کی چند محدودیات ذیل میں بیان کی گی ہیں:

i. تحقیق میں ہارٹ ڈیزیز رسک اوالویشن کے لئے ڈیٹا مائننگ کی بنیادی تکنیکس جیسے کہ رینڈم فارسٹ، نیوی بائیس، کے نیریسٹ نیبر، سپورٹ ویکٹر مشین اور ڈیسیشن ٹری استعمال کی گئی ہیں۔

ii. تحقیق میں صرف چند ہی نان انویسیو علامات کو تشخیصی عمل کے لئے آزمایا گیا ہے اور دیگر فیچرز جیسے کہ Socio-economic level, depression level, low education status اور ethnicity وغیرہ کو استعمال نہیں کیا گیا ہے

iii. طب کے صرف چند ہی شعبوں سے ماڈل کا جائزہ کروایا گیا اور ان کی تصدیق حاصل کی گئی نیز کمیونٹی سطح پر صرف دوا سازوں سے مشورہ لیا گیا اور باقی شعبوں کو نظر انداز کیا گیا ہے۔

iv. ڈیٹا سیٹ کشمیر کے ان افراد پر مبنی ہے جو دل کے امراض میں ملوث ہیں لیکن ان کی تعداد بہت کم ہے۔

v. جو HDREM کا ماڈل تیار کیا گیا ہے اس میں کشمیر کی آبادی کو ہی استعمال میں لایا گیا ہے اور دوسری ریس کو خارج کر دیا گیا ہے۔

vi. امراض قلب کی مختلف اقسام کی تشخیص نہیں کی گئی ہے جیسے کہ congenital heart disease, arrhythmia اور cardiac arrest وغیرہ۔

6.3۔ مستقبل کے کام کا منصوبہ

ماڈل کو بہتر بنانے کے لئے مستقبل میں اس تحقیق پر مزید کام کیا جا سکتا ہے ذیل میں اس حوالے سے کچھ تجاویز بیان کی گئی ہیں:

i. ماڈل کی کارکردگی کو مزید بہتر بنانے کے لئے ڈیٹا مائننگ کی دوسری اہم اور نئی تکنیکس جیسے کہ نیورل نیٹورکس، Genetic Algorithm, Ensembling تکنیکس وغیرہ کو استعمال میں لایا جا سکتا ہے اور گزشتہ نتائج کا موازنہ کیا جا سکتا ہے جو نان انویسیو علامات اس ماڈل میں استعمال کی گئی ہیں، مستقبل میں ان کی جگہ جراحتی علامات کو استعمال کر کے خطرات کا تعین کیا جا سکتا ہے۔

ii. دل کے امراض کی تشخیص میں علامات کے وزن کو تبدیل کر کے تمباکو نوشی کا مختلف عمر کے لوگوں پر دلی امراض کے حوالے سے اثر کو دیکھا جا سکتا ہے۔

iii. ماڈل کو مزید بہتر بنانے کے لئے ریئل ٹائم ڈیٹا کا استعمال کر کے مزید علامات کو شامل کیا جا سکتا ہے۔ نیز مریضوں کی تعداد میں بھی اضافہ کیا جا سکتا ہے۔

iv. HDREM ماڈل کی کارکردگی کو موثر بنانے کے لئے دواسازوں کے علاوہ طب کے دوسرے شعبوں سے بھی مشورہ لیا جا سکتا ہے۔

v. مستقبل میں مزید اور ڈیٹا فارمیٹس کو استعمال میں لایا جا سکتا ہے جیسے امیج اور کلینکل رکارڈس۔

vi. مستقبل میں ڈیٹا مائننگ techniques کا استعمال کرتے ہوئے one-size-fits-all رسک اوالویشن ماڈل تیار کریں گے جو امراض قلب کے علاج معالجے کو بھی تجویز کر سکے گے۔