ڈیٹا مائننگ تکنیک کے ذریعہ انتخابی نتائج کی پیشن گوئی

مقالہ

ڈگری برائے

ڈاکٹر آف فلاسفی

مقالہ نگار

عبدالمنان کولی

Enrolment Number (A165081)

زیر نگرانی

ڈاکٹر مقیم احمد

اسسٹنٹ پروفیسر

شعبہ کمپیوٹر سائنس وانفارمیشن ٹیکنالوجی

اسکول آف ٹکنالوجی

ستمبر۔2020

شعبہ کمپیوٹر سائنس وانفارمیشن ٹیکنالوجی

مولانا آزاد نیشنل اردو یونیورسٹی (سنٹرل یونیورسٹی)، گچی بولی، حیدرآباد (تلنگانہ)، انڈیا

मौलाना आज़ाद नेशनल उर्दू यूनिवर्सिटी مولانا آزاد نیشنل اُردو 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

On the basis of declaration submitted by Abdul Manan Koli, student of Ph.D, I hereby certify that the thesis titled **"The Prediction of Election Outcomes Through Data Mining Techniques"** which is submitted to the Department of Computer Science & Information Technology, School of Technology, Maulana Azad National Urdu University (A Central University), Gachibowli, Hyderabad, India in partial fulfillment of the requirement for the award of the degree of Doctor of Philosophy, is an original contribution with existing knowledge and faithful record of research carried by him under my guidance and supervision.

To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this university or elsewhere.

Dr. Muqeem Ahmed
(Research Supervisor)
Department of CS&IT
MANUU Hyderabad

Date:
Place: Department of CS&IT
School of Technology

Dr. Pradeep Kumar
Head Department of CS&IT
MANUU Hyderabad

DECLARATION

I, **Abdul Manan Koli**, solemnly declare that the thesis entitled "**The Prediction of Election Outcomes Through Data Mining Techniques"** is my original work. The study has been conducted under the guidance of **Dr. Muqeem Ahmed** with the Department of Computer Science & Information Technology, School of Technology, **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: ___/___/2020

Abdul Manan Koli
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 رب العزت** کا شکر ادا کرتا ہوں جو مجھے علمی ہمت اور تمام نعمتیں مہیا کرنے کے لئے علم کا حتمی ذریعہ ہے، جس کی برکت کے بغیر یہ کام ایک انجام تکمیل تک نہیں پہنچتا۔

ہر وقت میری مدد کے لیے تیار ،انتہائی مستعد اور مشفق نگراں ڈاکٹر مقیم احمد صاحب ، اسسٹنٹ پروفیسر ، اسکول آف سائنس ڈیپارٹمنٹ آف کمپیوٹر سائنس اینڈ انفارمیشن ٹکنالوجی کا بے حد شکر گزار ہوں کہ ہوں جنہوں نے مجھے تحقیق میں اپنی سمت کا تعین کرنے کی مکمل آزادی اور اجازت دی۔ اگرچہ یہ مشکل اور کسی حد تک پیچیدہ ثابت ہوا، لیکن اس کے ساتھ ہی میں ان کے دانشمندانہ طریقہ اور حکمت و دانائی کی بہت قدر کرتا ہوں۔ وہ ہمیشہ تنقیدی لیکن تعمیری تبصرے کے لئے تیار رہتے۔

میں پروفیسر عبدالوحید ، ڈین ، اسکول آف ٹکنالوجی ، ڈاکٹر پردیپ کمار ہیڈ ، شعبہ کمپیوٹر سائنس وانفارمیشن ٹکنالوجی، جناب جمیل احمد، اسسٹنٹ پروفیسر، شعبہ کمپیوٹر سائنس و انفارمیشن ٹکنالوجی کا بھی شکریہ ادا کرنا چاہتا ہوں کہ مجھے

اس تحقیق کے دوران جب بھی کسی تعاون کی ضرورت پیش آئی تو ان تمام لوگوں نے تعمیری تنقید اور قیمتی تجاویز کے ذریعہ میرا بھرپور تعاون کیا۔ میں اس تحقیق کے دوران تکنیکی رہنمائی اور قیمتی آراء کے لئے سوفٹ ویئر انڈسٹری سے تعلق رکھنے والے پروفیشنل جناب محبوب باشا کا بھی دلی شکریہ ادا کرنا چاہتا ہوں۔

میں تمام فیکلٹی ممبران اور شعبہ سی ایس اینڈ آئی ٹی، اسکول آف ٹکنالوجی ،مولانا آزاد نیشنل اردو یونیورسٹی کے دیگر فیکلٹی ممبران اور دوسرے معاون عملہ کا ان کے مستقل تعاون کے لئے انتہائی شکریہ ادا کرتا ہوں۔ مجھے اپنے تحقیقی ساتھیوں بطور خاص طور، سید امام الانصار، حمزہ مطہر، عطاءاللہ نیازی، عرفاز احمد اور جاوید اقبال کا بھی شکریہ ادا کرنا ضروری ہے کہ انہوں صحتمند مباحثہ اور مفید خیالات کے ذریعہ میرا ہر قدم پر تعاون کیا۔

میں اپنے مطالعے کے دوران اخلاقی مدد فراہم کرنے پر شعبہ CS & IT کے دیگر تحقیقی اسکالرز اور دیگر افراد کی خدمت میں بھی اظہار تشکر پیش کرنا چاہتا ہوں۔

آخر میں، مجھے اپنے پیارے والدین، بیوی اور کنبہ کے دوسرے ممبروں کی محبت اور قربانی کا اعتراف کرنے پر فخر محسوس ہوتا ہے جنہوں نے مجھے اپنے مطالعے کے

پورے عرصے میں لامحدود پیار، محبت اور تعاون فراہم کیا۔ 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۔ میری ذخیرہ الفاظ اس پیار، نگہداشت، پریرتا اور شہادت کو تسلیم کرنے میں ناکام رہتے ہیں جو میں نے اپنی زندگی کے ہر مرحلے میں خصوصا تحقیقی دوران اپنے پورے کنبہ کے ممبروں سے حاصل کیا۔

عبدالمنان کولی

خلاصہ

انتخابی نتائج کی پیش گوئی تاریخی دور سے ہی اہمیت کے حامل ریسرچ کا ایک دلچسپ موضوع بنا ہوا ہے اور آج بھی موجودہ دور کا ایک دل چسپ موضوع ہے۔ زبردست بہتری اور مشکل پیچیدگیاں کے باوجود، انتخاب کی پیش گوئی محققین اور سیاسی پیش گوئی کرنے والی تنظیم کے لئے ایک متاثر کن کام ہے، ووٹنگ کے وقت سے پہلے منظر نامے تبدیل کرنے کے ساتھ۔ انتخابی نتائج کی پیش گوئی کرنے کے بے شمار طریقے ہیں، جیسے سوشل میڈیا (ٹویٹر، فیس بک ڈیٹا) کے ذریعے انتخابی نتائج کی پیش گوئیاں کرنا۔ تاہم، ایسی سوشل میڈیا سائٹوں کے ساتھ اصل چیلنج fake Ids کی تعداد کی موجودگی ہے جسے نظر انداز نہیں کیا جا سکتا۔ ادب اور عملی تجربات میں، بہت سارے محققین نے بے ترتیب نمونے لینے کا استعمال کیا ہے اور انتخابات کے نتائج کی پیش گوئی کی ہے، لیکن اس طرح کے نمونے لینے کا اصل مسئلہ یہ ہے کہ وہ کم سروے کی وجہ سے متعصبانہ نتائج پیش کرتے ہیں اور اس سے ووٹرز کے خیالات کی بنیاد پر نتائج ظاہر ہوتے ہیں۔ جنہوں نے سروے میں حصہ لیا ہے اور ووٹرز کی ایک بڑی تعداد کی رائے کو خارج کیا ہے۔ اگرچہ computational ٹکنالوجی میں جدید ترین پیشرفت اب پیش گوئی کا معیار بن چکی ہے لیکن یہ ٹیکنالوجیز نفیس نہیں ہیں اور وقت طلب ہیں۔

ادبیات اور تکنیکی رپورٹوں کا جائزہ لینے کے بعد، یہ پتہ چلا ہے کہ گذشتہ انتخابی پیش گوئی کے ماڈلز میں بہت ساری کوتاہیاں اور حدود ہیں۔ موجودہ انتخابی پیش گوئی کے ماڈل صرف ان علاقوں میں قابل اطلاق ہیں جن میں جدید ترین انفراسٹرکچر موجود ہے لیکن جموں و کشمیر (جے اینڈ کے) کی صورت میں، اس قسم کے ماڈل لاگو نہیں ہوتے ہیں کیونکہ خطے میں انٹرنیٹ کی ناکہ بندی کی وجہ سے عوام سیاسی جماعتوں کے بارے میں اپنے جذبات کا اظہار نہیں کر سکتے ہے۔ جموں و کشمیر (جے اینڈ کے) میں انتخابی نتائج کی پیش گوئی کلے، کسی کو انتخابی پیش گوئی کے اہم پیرامیٹرز کی بنیاد پر ماڈل بنانا ہوگا۔

اس مقالہ میں، محقق نے انتہائی اہم پیرامیٹرز کو مدنظر رکھتے ہوئے ایک ماڈل تیار کیا جو ہیں centralgovernment influence, religion followers, party wave, party abbreviations, sensitive, vote bank, hereditary factor, incumbent party اور caste factor. ہے۔ ان پیرامیٹرز کی شناخت سیاسی ڈومین کے ماہرین نے کی ہے اور انتخابی نتائج کی پیش گوئی میں ان کی reliability کی تفتیش کی جاتی ہے مختلف خصوصیت کے انتخاب کی تکنیک جیسے, filter method, wrapper method, embedded method اور آخر کار ان کا اوسط وسیلہ کا حساب لیا جاتا ہے۔ ان سب تکنیکوں کو جیسے Decision Tree, - KNN, Random Forest, Support Vector Machine, کو investigate کیا ہے اور آخر میں ان سب کو ایک ماڈل میں جوڑا تاکہ انتخابی پیش

گوئیاں جانچ کی جائیں۔ مزید یہ کہ کم سے کم غلطی کی شرح کے ساتھ انتخابی پیش گوئیاں زیادہ درست طریقے سے پیش گوئی کرنے کے لئے، ہائپر پرمیٹر کی اصلاح کی جاتی ہے۔

یہ ماڈل جپٹر نوٹ بک ویب ایپلی کیشن کا استعمال کرتے ہوئے تیار کیا گیا ہے اور اس کی کارکردگی کو نہ صرف سیاسی ڈومین کے ماہرین ہی جانچتے ہیں لیکن اور اقدامات کے ذریعے بھی جیسے sensitivity, specificity, precision, misclassification rate, accuracy, AUROC, اور cross-validation اعداد و شمار کی جانچ جیسے (T-paired test) لگایا ہے۔ کارکردگی میں اضافہ اور غلطی کی شرح کو کم سے کم کرنے کے لئے، انتخابی پیش گوئی کا ماڈل بہتر بنایا گیا ہے۔ تجرباتی نتائج سے پتہ چلتا ہے کہ انتخابات کی پیشن گوئی کے ماڈل ہائپر پرامیٹر کی ترتیبات پر انتخابی پیش گوئی کے دوسرے ماڈلز کو ان readings پر پیچھے چھوڑ دیتے ہیں جسمے sensitivity 86%, specificity 92%, accuracy 89%, precision 91%, 10.1 % miss-classification rate اور - AUROC score 90 % انتخاب پیش گوئی کی خصوصیت کے مرکب سبسیٹ [Central govt. Influence, Religion Followers, Party Wave, Party Abbreviations, and Sensitive Areas] نے ensemble ماڈل کا استعمال کرتے ہوئے سب سے زیادہ اسکور ظاہر کیے۔

اس ماڈل سے سیاسی پیش گوئی کرنے والوں اور عام لوگوں کو حقیقی نتائج کے اعلان سے قبل سیاسی جماعتوں یا آزاد امیدواروں کے انتخابی حلقے کے مطابق جیتنے یا ہارنے کے امکان کے بارے میں نظریہ حاصل کرنے میں مدد ملے گی۔ یہ ماڈل ان انتخابی حلقوں/علاقوں کے لئے لاگو ہے جہاں لوگوں کو باقاعدہ پیش گوئی کے لئے سوشل میڈیا اور ایگزٹ پول ٹکنالوجیوں کی باقاعدگی سے رسائی حاصل نہیں ہے۔ ماڈل کے ذریعہ تیار کردہ انتخابی پیش گوئی کے قواعد کو مختلف ڈومین ماہرین نے جانچا اور اس کی توثیق کی ہے۔ نکالی جانے والی انتخابی پیش گوئی کے قاعدے تجویز کردہ ہیں لیکن حتمی نہیں ہیں کیونکہ یہ جموں و کشمیر (جے اینڈ کے) پر مبنی ہیں۔

فہرست مشمولات

فہرست مخففات

ABS	Australian Bureau of Statistics
AUROC	Area Under the Receiver Operating Curve
BJP	Bharatiya Janata Party
CRISP_DM	Cross-Industry Standard Process for Data Mining
DMP	Default Model Parameter
DMX	Data Mining Extension
DSS	Decision Support System
DTC	Decision Tree Classifiers
EDA	Exploratory Data Analysis
EM	Ensemble Model
EPM	Election Prediction Model
FNR	False Negative Rate
FPR	False Positive Rate
GBC	Gradient Boosting Classifier
GNI	Gross National Income
HPT	Hyper-Parameter Tuning
HV	Hard Voting
IEPS	Intelligent Election Prediction System
INC	Indian National Congress
JKN	Jammu and Kashmir National Conference
JKPDP	Jammu and Kashmir People Democratic Party
KDD	Knowledge Discovery from Data
KNN	K Nearest Neighbor
MLA	Member of Legislative Assemble

MLC	Member of Legislative Councils
NDA	National Democratic Alliance
PCA	Principal Component Analysis
RFC	Random Forest Classifiers
RFE	Recursive Feature Elimination
SAS	Statistical Analysis Software
SEMMA	Sample Explore Modify Model Assess
SVM	Support Vector Machine
SV	Soft Voting
TNR	True Negative Rate
TPR	True Positive Rate
WEKA	Waikato Environment for Knowledge Analysis
XGB	Extreme Gradient Boosting

فہرست جدول

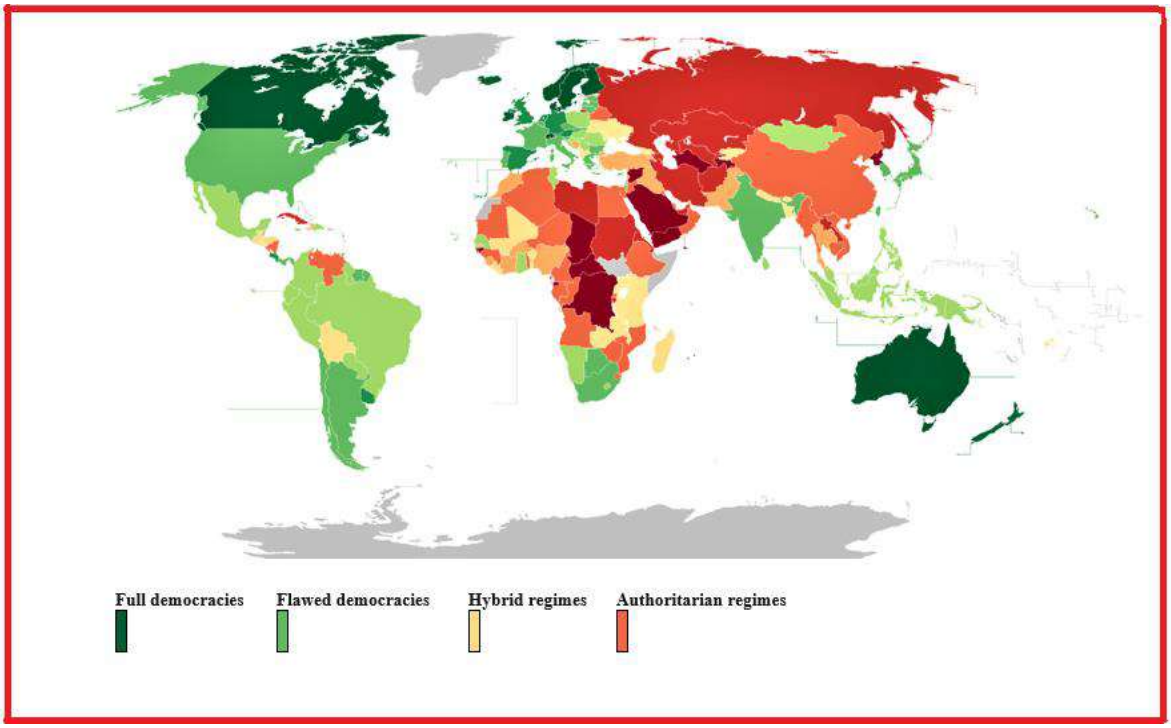
فہرست ترسیم

باب 1

1. تعارف

1.1 پس منظر

انتخابی پیشن گوئی کا میدان نہ صرف سیاسی پیش گوئین کے لئے بلکہ عام لوگوں کے لئے بھی دلکشی کا باعث بنا ہوا ہے، اور پوری دنیا کے لاکھوں لوگوں کے لئے یہ ایک اہم مقام ہے۔ اس کے نتیجے میں متعدد سیاسی سائنس دانوں نے انتخابی پیشن گوئی کے نظام کے میدان میں داخل ہونے کے ارادے سے حتمی نتائج کا اعلان ہونے سے قبل درست پیشن گوئیاں کرنے کا ارادہ کیا ہے۔ انتخابات ایک انتخاب کا عمل ہے جس میں عام عوام اپنے ووٹ کا استعمال کرکے اور ایک مخصوص وقت کے لئے کسی سیاسی رہنما یا نمائندے کا انتخاب کرتے ہیں۔ ایسے منتخب نمائندے ایک مخصوص وقت میں لوگوں کی فلاح و بہبود کے لئے فیصلہ لیتے ہیں۔ دنیا کے کل 195 ممالک میں سے 123 ممالک نے جمہوری حکومت کی ساخت پر انحصار کیا، جس کا مطلب ہے کہ موجودہ دنیا کی آبادی بنیادی طور پر جمہوری قسم کی حکومتوں کے زیر انتظام ہے[1] جیسا کہ ترسیم 1 میں دکھایا گیا ہے۔



ترسیم 1: دنیا کا جمہوریت انڈیکس نقشہ۔

مختلف ممالک انتخابی عمل کے وقوع پزیر ہونے کے لیے مقررہ اصولوں کے ساتھ اپنا تحریری دستور پیش کرتے ہیں۔ مختلف میقات کے ساتھ کچھ ممالک میں بالواسطہ انتخاب ہوتا ہے (مثلا ریاست ہائے متحدہ امریکہ) اور کچھ ممالک میں براہ رائست قسم کے انتخابات ہوتے ہیں (جیسے یورپی پارلیمنٹ)۔ لہذا یہ بات واضح ہے کہ بیشتر ممالک کو آئینی اختیار کے ذریعہ کنٹرول کیا جاتا ہے جسے بنیادی طور پر نمائندے کے اقتدار پر شہریوں کی سیاسی شمولیت کی رکاوٹوں کی ایک خود مختار تنظیم کے طور پر دکھایا گیا ہے۔

1.2 محرک

اس کام میں، محقق نے جموں و کشمیر (جے اینڈ کے) کے انتخابی نتائج کی پیش گوئی کی ہے کیونکہ انٹرنیٹ کی مسلسل ناکہ بندی کی وجہ سے سوشل میڈیا تجزیہ متروک ہے۔ ہماری بہترین معلومات کے مطابق یہ تحقیقی کام نادر اور نیا ہے جو آج تک نہیں ہوا۔ جموں و کشمیر کا انتخابی عمل بقیہ ہندوستان سے مختلف ہے کیونکہ جموں و کشمیر میں انتخابات پانچ سال کے بجائے چھ سال بعد ہوتے ہیں۔ جموں و کشمیر کے پہلے اسمبلی انتخابات 1957 میں ہوئے تھے اور اب تک جاری ہیں [2]۔ جموں و کشمیر کا علاقہ 1957 کے بعد سے بڑے پیمانے پر سماجی بدامنی کا شکار ہے اور اسی وجہ سے یہاں اکثر انٹرنیٹ ناکہ بندی ہوتی رہتی ہے۔

ان وجوہات کی وجہ سے بہت کم لوگ انٹرنیٹ تک رسائی حاصل کر رہے ہیں لہذا وہ سیاسی احوال کے بارے میں اپنے خیالات و جذبات کا اظہار کرنے سے قاصر ہیں۔ اس کو ذہن میں رکھتے ہوئے ہم نے جموں و کشمیر کے لئے اہم پیمانے کی بنیاد پر انتخابی پیشن گوئی کا ماڈل تیار کیا ہے۔ پیمانے کا تعین جموں و کشمیر

کے مختلف علاقوں سے ماہرین علم وادب کے ذریعہ پیش کردہ تجزیوں کا بغور مطالعہ کرنے کے بعد کیا گیا ہے۔

جموں وکشمیر میں دو ڈویژن ہیں، پہلا جموں ڈویژن اور دوسرا کشمیر ڈویژن۔ جموں وکشمیر میں کل 87 انتخابی نشستیں ہیں جن میں سے 50 نشستیں صوبہ کشمیر میں ہیں جبکہ 37 نشستیں جموں کے زیر انتظام ہیں [3]۔ کل 22 اضلاع ہیں جن میں سے 11 اضلاع جموں ڈویژن میں اور 11 اضلاع کشمیر میں ہیں۔ جموں و کشمیر کے جغرافیائی اور سماجی ثقافت کے نسلی تنوع کی وجہ سے، یہ معاشرے کی مبہم تصویر کو پیش کرتے ہوئے اعلی سطح پر کام کرتا ہے۔ یہاں شروع کرنے کے لئے جموں وکشمیر میں مذہب، خطہ اور ذات پات کا تنوع ہے۔ جموں وکشمیر میں تین بڑے مذاہب یعنی اسلام، بدھ مت اور ہندومت وغیرہ کے پیروکار ہیں۔ جموں وکشمیر میں مسلمانوں کی اکثریت ہے جس کے بعد ہندو مذہب، سکھ مذہب اور بدھ مذہب وغیرہ۔ مذہبی تنوع کے ساتھ ساتھ، جموں وکشمیر زبان، ثقافت، ذات اور قبیلہ پر مبنی تنوع کا بھی گواہ ہے۔ جموں وکشمیر کے دونوں خطے نہ صرف ثقافتی اور مذہبی اعتبار سے مختلف ہیں بلکہ جغرافیائی خطوں کے لحاظ سے بھی متنوع ہیں۔ جب موسموں کے مطابق ریاست کا دارالحکومت تبدیل ہوتا ہے تو، سری نگر

(کشمیر ڈویژن میں) موسم گرما کا دارالحکومت بن جاتا ہے اور جموں موسم سرما میں دارالحکومت رہتا ہے۔ ہندوستان کو آزادی ملنے کے بعد جموں و کشمیر بڑی سطح پر بدامنی کا شکار ہوا، جس سے جموں و کشمیر سیاسی، تعلیمی اور معاشی طور پر بہت حد تک متاثر ہوا[4]۔ جموں و کشمیر 1957 سے سیاسی سطح پر بڑی کشمکش کا شکار ہے اور یہ عمل اب بھی جاری ہے، خواہ وہ صدر راج کا نفاذ ہو یا اتحادی حکومت کا گرنا وغیرہ۔ جموں و کشمیر کے پہلے اسمبلی انتخابات سے لے کر 1996 تک، نیشنل کانفرنس (جے کے این) اہم سیاسی جماعت رہی کیونکہ جے کے این نے زیادہ تر انتخابات میں کامیابی حاصل کی، تاہم جموں و کشمیر پیپلز ڈیموکریٹک پارٹی (جے کے پی ڈی پی) کی تشکیل کے بعد 2002 میں اور بھارتیہ جنتا پارٹی (بی جے پی) 2008 میں، اس کے بعد حالات بدل گئے، سیٹ شیئر مختلف سیاسی جماعتوں کے مابین تقسیم ہو گیا، جیسا کہ جدول 1.2 میں بیان کیا گیا ہے[3]۔

جدول 1.2: جموں و کشمیر میں حکومتوں کا قیام سال بہ سال

Year	Election	Government formed
1957	First Assembly	JKN
1962	Second Assembly	JKN
1967	Third Assembly	INC
1972	Fourth Assembly	INC and JKN

1977	Fifth Assembly	JKN
1983	Sixth Assembly	JKN
1987	Seventh Assembly	JKN
1990-1996	No Election	President Rule
1996	Eighth Assembly	JKN
2002	Ninth Assembly	JKPDP and INC
2008	Tenth Assembly	JKN and INC
2014	Eleventh Assembly	JKPDP and BJP

1.3 ڈیٹا مائننگ اور الیکشن پیشن گوئی میں اس کا استعمال

ڈیٹا مائننگ کو ڈیٹا ویئر ہاؤس میں محفوظ خام ڈیٹا سے مفید معلومات نکالنے کے عمل کے طور پر بیان کیا جا سکتا ہے۔ گارٹنر گروپ کے مطابق، ''ڈیٹا مائننگ کا بنیادی کام مختلف ذخیروں میں ذخیرہ شدہ ڈیٹا کی بڑی مقدار کو جانچ کر پیٹرن کی شناخت، شماریاتی اور ریاضی کی تکنیک جیسی مختلف تکنیکوں کا استعمال کرتے ہوئے بامعنی نئے ارتباط، نمونوں اور رجحانات کو دریافت کرنا ہے''۔ تاہم ، کیبینا گروپ نے ڈیٹا مائننگ کو ''بین مضامین کے میدان کے طور پر بیان کیا ہے جو مشین لرننگ، پیٹرن کی پہچان، شماریات، ڈیٹا بیس اور تصور کی تکنیک کو ایک ساتھ لاتا ہے تاکہ ایک بہت بڑے ڈیٹا بیس سے معلومات نکالنے کے معاملے کو حل کیا جا سکے''[5]۔ لہذا، ڈیٹا مائننگ کی

مذکورہ بالا تعریفوں سے ہم یہ نتیجہ اخذ کر سکتے ہیں کہ ڈیٹا مائننگ بنیادی طور پر میٹا ڈیٹا کے ذخیروں سے معلومات نکالنے کا عمل ہے اور ان کی بامقصد معلومات میں درجہ بندی یا تلخیص کرنا۔ ڈیٹا مائننگ کا اطلاق کسی بھی طرح کے ڈیٹا پر ہو سکتا ہے چاہے وہ ڈیٹا ویئر ہاؤس، ٹرانزیکشنل ڈیٹا بیس، ریلیشنل ڈیٹا بیس ، ٹائم سیریز ڈیٹا بیس اور ورلڈ وائڈ ویب وغیرہ ہو۔

محقق اس تحقیقی کام میں ڈیٹا مائننگ کا استعمال کر رہے ہیں کیونکہ ڈیٹا مائننگ موجودہ دور کا ایک انتہائی استعمال شدہ محرک موضوع ہے، کیوں کہ اس میں مختلف شعبہ حیات میں بہت سے نمایاں کارنامے ہیں ، چاہے وہ تجزیاتی ہو یا صحت ، تعلیم اور سیاسی پیشن گوئی سے متعلق ہو۔ اس سے مختلف ایپلیکیشن میں ڈیٹا مائننگ کے زیادہ کامیاب اور وسیع پیمانے پر استعمال کو ترجیح دی گئی ہے، جیسے : ویب مائننگ، کاروباری ذہانت ، لوڈ کی پیشن گوئی ، تشخیص ، مارکیٹنگ اور فروخت ، تیل کی تطہیر اور اسکریننگ امیجز وغیرہ۔ ڈیٹا مائننگ انتخابی پیشن گوئی کا ایک انتہائی اہم اور ابھرتا ہوا ذریعہ ہے اس کی اہمیت اس لئے زیادہ ہے کہ یہ انتخابی نتائج کے حتمی اعلان سے قبل انتخابی نتائج کی درست اور ابتدائی

پیشن گوئی فراہم کرتا ہے۔ ڈیٹا کا ایک بڑا حصہ موبائل فون، سینسر نیٹ ورک اور سوشل میڈیا وغیرہ کے ذریعہ خطرناک شرح پر تیار کیا جاتا ہے۔ یہ ڈیٹا سیٹ نہ صرف سائز میں وسیع ہیں بلکہ یہ فطرت میں بھی پیچیدہ ہیں جیسے ساخت، نیم ساخت اور غیر ساخت [6]۔ اس طرح کے بڑے اور میٹا ڈیٹا سے RDBMS جیسے روایتی ڈیٹا بیس ذرائع کو سخت خطرہ لاحق ہے کیونکہ یہ ذرائع اتنی بڑی تعداد میں ڈیٹا کو سنبھالنے سے قاصر ہیں۔ اس طرح کے چیلنجوں پر قابو پانے کے لئے ڈیٹا مائننگ جیسی تکنیک سامنے آئی، جو اس طرح کے میٹا ڈیٹا کو آسانی سے سنبھال سکتی ہے اسے ذخیرہ کر سکتی ہے اور ان پر عملدرآمد کر سکتی ہے اور اتنے وسیع ڈیٹا سے معلوماتی تجزیات حاصل کر سکتی ہے [7]-[9]۔ آج ہر دن تقریباً 2.5 ملین ڈیٹا تیار کیا جا رہا ہے اور ہر سیکنڈ کے ساتھ مسلسل اعداد و شمار بڑھ رہے ہیں [10]-[12]۔ مختلف محققین نے اپنے تحقیقی کام میں ڈیٹا مائننگ کی تکنیک کا اطلاق کیا اور انتخابی نتائج کی پیشن گوئی کی [13]، [14] مسٹر ڈونالڈ ٹرمپ جیسے کچھ سیاست دانوں نے بھی 2016 میں اپنی انتخابی مہم کے دوران کیمبرج اینالیٹیکل آرگنائزیشن کے لئے ڈیٹا سائنس دانوں کی خدمات حاصل کی تھیں جو فتح میں معاون ثابت ہوئی [15]۔

ڈیٹا مائننگ میں سیاسی حلقہ اثر بطور خاص انتخابی پیشن گوئی کے ڈیٹا بیس میں پوشیدہ نمونوں کی

جانچ کرنے کی حیرت انگیز صلاحیت موجود ہے۔ انتخابی تجزیہ اور پیشن گوئیوں کے دوران سب سے بڑا

چیلنج یہ سامنے آتا ہے کہ ایک ایسی جدید ٹکنالوجی تشکیل دی جائے جو مخصوص اقدامات کے ذریعہ قابل اعتماد

مفروضے کو پیش کر سکے، جو سیاسی پیشن گوئی کی تحقیق پر اعتماد کر سکے اور سیاسی پیشن گوئی کے ماحول میں

اس کا اطلاق کر سکے۔ ٹریننگ ڈیٹا سیٹس سے نظریاتی مضمرات نکالنا یا دریافت کرنا ایک چیلنجنگ

عمل ہے ، اور یہاں تک کہ معروف با اثر ماہرین بھی اس طرح کے جمع کردہ اعداد و شمار سے

مغلوب ہوگئے۔ اس لئے ڈیٹا مائننگ اور مشین لرننگ ٹکنالوجی اس میدان میں آئی کیونکہ

یہاں اس کے لئے صرف ایک ٹریننگ مشین لرننگ الگوریتھم کی ضرورت ہے پھر وہ خود سے ضروری

پروسیسنگ کریں گے،اور یہ سیاسی تجزیہ کاروں کے لیے درست فیصلہ سازی میں معاون ثابت ہو سکتا ہے۔

ڈیٹا مائننگ سیاسی ماحول کے منظر نامے کو بدل رہی ہے۔ اب زیادہ تر جیتنے والی سیاسی جماعتیں روایتی

طریقوں سے ہٹ کر ڈیٹا مائننگ اور مشین لرننگ کی طرف بڑھ رہی ہیں، کیونکہ روایتی طریقے میں

وقت زیادہ لگتا ہے اور موثر کم ہوتا ہے[16] ۔ لیکن سیاسی پیشن گوئی کے میدان میں ڈیٹا مائننگ کے

داخل ہونے کے بعد چیزیں تبدیل ہونا شروع ہو گئیں، کیونکہ یہ صرف ایک پیغام نشر کر کے لاکھوں ووٹروں کو اپنی طرف متوجہ کرتی ہے، مذہب، ذات، علاقہ ، جنس اور مقام وغیرہ سے قطع نظر۔ مشین لرننگ تکنیک کے ساتھ ڈیٹا مائننگ کا ستعمال کر کے سیاسی جماعتیں مختلف حلقوں میں رائے دہندگان کے نظریات یا مطالبات جاننے کے قابل ہوئیں اور پھر انہوں نے مختلف مقامات پر انہی چیزوں کے بارے میں بات کرنا شروع کیا جو ان کے لئے اہم تھیں[16]۔ رائے دہندگان تک اسمارٹ یا آسان اپروچ کا مطلب ہے میسیج کرنا، جو لوگوں کے لئے سب سے زیادہ موزوں ہے، یہ کسی بھی شکل میں ہو سکتا ہے جیسے سوشل میڈیا یا بینر کی شکل میں۔ ڈیٹا مائننگ کے تجزیات سیاسی جماعتوں کو ذاتی سطح پر اپنے رائے دہندگان کے خیالات جاننے کے قابل بناتے ہیں۔ ڈیٹا مائننگ اور ہائبرڈ ٹول کی جدید ٹکنالوجی جیسے اینا کونڈا وغیرہ، سیاسی جماعتوں کو ضرورتوں اور ووٹروں سے رابطہ کرنے کے مختلف طریقوں کو سمجھنے میں مدد دیتی ہے۔

روایتی طور پر، سیاسی جماعتیں پوری مہم کے دوران عموماً آبادی کے اس حصے پر انحصار کرتی ہیں جو پورے دل سے انہیں ووٹ دیتی ہیں اور آبادی کے بقیہ حصے کو نظر انداز کرتی ہیں۔ لیکن اب ، ڈیٹا مائننگ اور مشین

لرننگ کی تکنیک کو بروئے کار لا کر، ووٹرز کے ایک بڑے حصے کو نشانہ بنایا جا سکتا ہے جو واقعتا سیاست کے بارے میں زیادہ نہیں جانتے ہیں۔ ایسے ووٹرز کو floating ووٹر کہا جاتا ہے اور ایسے floating ووٹروں کو اسمارٹ ٹارگٹ کرنے میں ڈیٹا مائننگ بہت مفید ہے۔ان ووٹرز کو اپنی دلچسپی کی بنیاد پر مختلف گروپوں میں تقسیم کرکے اس کے مطابق نشانہ بنایا جاتا ہے۔ لہذا ، ایسے floatingووٹروں کو سیاسی جماعتوں کی طرف موڑ کر ان کے ووٹ حال کرنے کے امکانات بڑھ سکتے ہیں جو بصورت دیگر حاصل کرنا ناممکن ہے۔

اس تحقیق میں، ہم نے پچھلے انتخابات کے اعداد و شمار تک رسائی کے لیے ڈیٹا مائننگ کی تکنیک کا استعمال کیا ہے اور انتخابی نتائج کی پیشن گوئی کی ہے۔ تو یہ واضح ہے کہ ڈیٹا مائننگ ٹیکنالوجی، ذخیرہ شدہ ڈیٹا یا روزانہ تیار کردہ ڈیٹا پر ناگزیر اثر ڈالتی ہے۔ مشین لرننگ کے ساتھ ڈیٹا مائننگ مکمل طور پر ڈیٹا اکٹھا کرنے اور کچھ معلوماتی فیصلہ کرنے کے بارے میں ہے جو سیاستدانوں کو اپنے حلقے کے لوگوں کے خیالات اور جذبات جاننے اور اس کے مطابق کام کرنے میں مدد دیتی ہے۔ ٹکنالوجی میں ترقی سے عوام اور سیاستداں دونوں کو فائدہ ہو سکتا ہے کیونکہ لوگ اپنے خیالات کا اظہار کر سکتے ہیں اور سیاستداں

آسانی سے ان درپیش مشکلات کو جان سکتے ہیں اور ان کو حل کر سکتے ہیں۔ یہ لوگوں کو سیاستدانوں کے پچھلے ریکارڈ دیکھ کر اپنے لئے صحیح سیاستدانوں کا انتخاب کرنے کا موقع فراہم کرتا ہے۔ اس طرح، ڈیٹا مائننگ کی تکنیک انتخابات کو سیاسی مہموں سے بالاتر ہوکر پوری قوم کے لئے حقیقی تبدیلی اور جیت کے حالات بنا سکتی ہے۔

1.4 تحقیقی کام کی تفصیل

اس کام میں، انتخابات کی پیشن گوئی کا ماڈل تیار کرنے کے لئے ڈیٹا مائننگ کی تکنیک استعمال کرتے ہیں ۔ ڈیٹا پری پروسیسنگ، ڈیٹا مینپولیشن اور کلاسیفیکیشن کے لیے لائبریری پیکج میں تغیر پذیری کی وجہ سے الیکشن پیشن گوئی کا ماڈل اینا کونڈا ٹول میں بنایا گیا ہے۔ اس تحقیقی کام میں فلٹر میتھڈ، ریپر میتھڈ اور ایمبیڈڈ میتھڈ جیسے تین مختلف فیچر سلیکشن تکنیک کو انتخابی پیرامیٹرز کے اوسط وزن کے ساتھ ساتھ انتخابات کی پیشن گوئی کے لئے انتہائی اہم پیرامیٹرز کا انتخاب کرنے کے لئے لاگو کیا جاتا ہے۔ انتخابی پیشن گوئی کے پیمانے کے اہم سیٹ کو منتخب کرنے کے بعد، ہم نے سیاسی پیشن گوئی کرنے والے ماڈل کی تیاری کے لئے ڈیٹا مائننگ کی الگورتھم کلاسیفیکیشن جیسے ڈیسیزن ٹری، رینڈم فاریسٹ، کے نیرسٹ نائیبر اور

سپورٹ ویکٹر مشین کا استعمال کیا۔ الخوارزمی کے مذکورہ بالااصول کے مطابق انتخابی ڈیٹاسیٹ کی مائننگ کے بعد ، ہم نے دیکھا کہ دوسرے موجودہ انتخابی پیشن گوئی ماڈلز کے مقابلہ میں رینڈم فاریسٹ ماڈل نے بہترین نتائج پیش کیے۔ ترقی یافتہ انتخاب کی پیشن گوئی کرنے والے ماڈلز کی پیشن گوئی کی کارکردگی کو بہتر بنانے اور ضرورت سے زیادہ دشواریوں پر قابو پانے کے لئے ، ہم نے ہائپر پیرامیٹر آپٹیمائزیشن اور استعمال شدہ مشین لرننگ ماڈلز کی درستگی کی جانچ کی۔ ہم ہائپر پیرامیٹر آپٹمائزیشن کے انتخابی پیشن گوئی کے ماڈلز کے نقلی نتائج کا جائزہ لیتے ہیں کہ یہ دیکھا گیا ہے کہ رینڈم فارسٹ ماڈل نے انتخابی پیشن گوئی کے ماڈلز میں موجود دوسروں کو مات دیدی۔

انتخابی پیشن گوئی کرنے والے ماڈلز کی استعداد کار کو بہتر بنانے کے لئے ، ہم نے اعلی درستگی کے ضمن میں ماڈل کی کارکردگی کو بڑھانے اور ماڈل کی غلط درجہ بندی کی شرح کو کم کرنے کے لئے سافٹ ووٹنگ انسیمبلنگ تکنیک کا استعمال کیا۔ سافٹ ووٹنگ انسیمبلنگ تکنیک کا اطلاق اس لیے کیا جاتا ہے کیونکہ یہ استعمال کیے جانے والے تمام ماڈلز کے امکانات کے ذرائع فراہم کرتا ہے۔ آخر میں ، ترقی یافتہ انتخابی پیشن گوئی ماڈل کی اہمیت کی جانچ کرنے کے لئے ہم نے پی ویلیو کا استعمال کرتے ہوئے اسٹیٹسٹیکل ٹی-پیئرڈ

ٹیسٹ کا اطلاق کیا۔ انتخابی پیشن گوئی ماڈل کے نتائج سے یہ بات سامنے آتی ہے کہ مجوزہ انتخابی پیشن گوئی ماڈل اور تمام ماڈل کے مابین اہم فرق ہے۔ ماڈل کے نتائج سے پتہ چلتا ہے کہ ایک مجموعی ماڈل کی درست پیشن گوئی نے دوسرے مجوزہ ماڈلز کو مات دیدی۔ آخر میں، انتخابی پیشن گوئی ماڈل کے نتائج کی خصوصا جموں و کشمیر اور عام طور پر ہندوستان کے بڑے ماہرین سے توثیق کرائی جاتی ہے۔

1.5 انتخابات کی پیشن گوئی کے طریقے

حالیہ برسوں میں پیشن گوئی کے بازاروں نے مستقبل کے رجحانات کی پیشن گوئی کرنے کے ایک آلے کے طور پر کافی پذیرائی حاصل کی ہے اور سیاسی پیشن گوئی ان پیشن گوئی کے ماحول کا مرکزی موضوع بنی ہوئی ہے [17]۔ سن 1936 میں سائنسی پولنگ کی ایجاد سے پہلے سیاسی پیشن گوئی کرنے والوں کے لیے رائے عامہ کا اندازہ لگانا اور انتخابی نتائج کی پیشن گوئی کرنا کافی مشکل تھا [18]۔ 1936 کے بعد سے، سروے یا براہ راست مواصلات کا استعمال کرتے ہوئے ترغیبی رائے عامہ کے جائزے کا آنا سیاسی پیشن گوئی کے لئے بنیادی کردار رہا ہے [17]۔ ابھی حال ہی میں، 1988 میں لووا الیکٹرانک مارکیٹس کے ساتھ پیشن گوئی کی منڈیوں نے عروج حاصل کرنا شروع کیا۔ لیکن اعداد و شمار اور کمپیوٹر تکنیک بطور خاص

ڈیٹا مائننگ اور مشین لرننگ کے متعارف ہونے کے ساتھ ہی سیاسی پیشن گوئی میں انتخابی پیشن گوئی کا پورا منظر نامہ زیادہ دلچسپ ہو گیا، کیونکہ بڑھتے ہوئے انتخابی اعداد و شمار کو سنبھالنا کافی آسان ہو گیا [19]۔ اس نے سیاسی پیشن گوئی کو ایک بڑا کاروبار بنا دیا، خواہ وہ نیوز چینلز جیسے ایگزٹ پول ہو یا سیاسی پیشن گوئی کرنے والی تنظیم مثال کے طور پر، فائیو تھرٹی ایٹ [20] وغیرہ، یا پوری دنیا میں بہت سارے محققین ۔ محققین انتخابی پیشن گوئی کے لئے بے شمار تراکیب اور طریقے استعمال کر رہے ہیں ان میں سے کچھ معاشیات کے پیمانے کا استعمال کرتے ہیں [21] - [24] جبکہ دوسرے ڈیٹا مائننگ تکنیک کے ذریعہ انتخابی پیشن گوئی کے لیے سوشل میڈیا [25]-[27] جیسے بنیادی پیمائش کو شامل کرتے ہیں جیسے مشین لرننگ، بگ ڈیٹا اینالسٹک اور ڈیپ لرننگ وغیرہ۔ جب انتخابی پیشن گوئی کا دور مزید عام ہو گیا، جب براک اوبامہ نے اپنے انتخابی عمل میں سوشل میڈیا خصوصا ٹویٹر کو منصوبہ بند انداز میں استعمال کیا [28]۔ پھر ان کے نقش قدم پر چلتے ہوئے، پوری دنیا کے متعدد سیاست دانوں اور محققین نے بالترتیب اپنے انتخابی عمل اور تحقیقی کام میں ان تراکیب کو استعمال کیا۔

سوشل میڈیا کے ساتھ بنیادی پریشانی یہ ہے کہ ایک تو، ان کے پاس جعلی آئی ڈی ہیں [29] اور دوسرا، وہ

صرف ان لوگوں کے خیالات یا فکر فراہم کر رہے ہیں جو ان تک رسائی حاصل کر رہے ہیں اور ان لوگوں کے جذبات کو خارج کر رہے ہیں جو ان تک رسائی حاصل نہیں کر رہے ہیں [30]۔ لہذا، اس طرح کے معاملات پر قابو پانے کے لئے، انتخابی نتائج کی پیشن گوئی کچھ پیمانوں پر مبنی ہونی چاہیے جیسے ذوالغدر اور ان کی ٹیم [31] نے امریکہ کے انتخابی نتائج کی پیشن گوئی بعض اوصاف کی بنا پر کی ہے، اسی طرح دوسرے محققین جیسے سنگھ اور ان کی ٹیم [32] نے پیمانوں کی بنیاد پر ریاست پنجاب (ہندوستان) کے انتخابی نتائج کی پیشن گوئی کی۔ اسی طرح، متعدد محققین جیسے ہمل گروپ [33]، ارولپالم گروپ [34]، سنگھ اور ان کی ٹیم [35]، عالم گروپ [36]، گل [37]، ساونت گروپ [38]، جگدیو گروپ [14] اور بہت سے دوسرے لوگوں نے کمپیوٹر تکنیک کے مدد سے پیمانوں کی بنیاد پر مختلف علاقوں کے انتخابی نتائج کی پیشن گوئی کی ہے۔ ذوا لغدر گروپ سے لے کر جگدیو گروپ تک کے تحقیقی کام کی بنیاد پر ہم جموں و کشمیر (ہندوستان) کے انتخابی نتائج کی پیشن گوئی کرتے ہیں۔ پیرامیٹرک نقطہ نظر کا استعمال کرتے ہوئے انتخابی نتائج کی پیش گوئی کرنے کا مرکزی موضوع یہ ہے کہ جموں و کشمیر میں اکثر انٹرنیٹ بند ہوتے ہیں [39]-[41] لہذا ایسے حالات میں، لوگ سوشل میڈیا کا استعمال کرنے اور سیاسی جماعتوں کے متعلق اپنے جذبات کا اظہار

کرنے سے قاصر ہیں۔ اس تحقیقی کام میں، ہم پیمانوں کی بنیاد پر جموں و کشمیر کے انتخابی نتائج کی پیشن گوئی کر رہے ہیں، اور ان پیمانوں کا انتخاب ماہرین سے مشورہ کرنے اور رائے شماری کی پیشن گوئی کرنے والے متعدد افراد کے ساتھ کراس چیک کرنے کے بعد کیا گیا ہے۔

1.6 مقاصد

اس کام کا مقصد مندرجہ ذیل ہے۔

i. ڈیٹا مائننگ اور مشین لرننگ کی تکنیکوں کا استعمال کرتے ہوئے انتخابی پیشن گوئی کے بارے میں منظم ادبی جائزہ لینا۔

ii. ان اہم پیمانوں کی نشاندہی کرنا جو انتخابی نتائج کا درست اندازہ لگا سکتے ہیں۔

iii. الخوارزمی کے مناسب ڈیٹا مائننگ کا استعمال کرتے ہوئے انتخابی پیشن گوئی کے ماڈل تیار کرنا۔

iv. مختلف میٹرکس کا استعمال کرتے ہوئے ترقی یافتہ انتخابی پیشن گوئی ماڈل کی کارکردگی کا تجزیہ کرنا۔

v. ریئل ورلڈ ڈیٹا سیٹس، ماہرین کی آرا اور موجودہ انتخابی پیشن گوئی کے ماڈل کا استعمال کرتے ہوئے ماڈل کی توثیق کرنا۔

1.7 مسئلہ کا بیان

مذکورہ تحقیقی مقاصد کی تکمیل کے لئے ، یہ تحقیقی کام نظر ثانی شدہ کام سے پہلے ہے اور سپورٹ ویکٹر مشین، رینڈم فاریسٹ، ڈیسیزن ٹری اور K- نیرسٹ نائیبر جیسے اعلی درجے کو ایک درجہ بندی میں جوڑتا ہے اور موثر درجہ بندی الگورتھم ساتھ انتہائی موزوں پیرامیٹرز کو شامل کر کے نظر ثانی شدہ ادب میں حدود اور خامیوں کو دور کرتا ہے۔

موجودہ انتخابی پیشن گوئی کے نظام ناکافی پیمانوں پر تیار کیے گئے ہیں جس کے نتیجے میں پیشن گوئیاں غلط ثابت ہوتی ہیں اور جن آلات کا استعمال کیا جاتا ہے وہ بڑے ڈیٹا سیٹس کو سنبھالنے کی سکت نہیں رکھتے ہیں کیونکہ وہ جائزوں کے بے ترتیب نمونوں پر مبنی ہیں۔ موجودہ انتخابی پیشن گوئی کے ماڈلز کی دوسری تحدید یہ ہے کہ وہ متعصبانہ نتائج پیش کرتے ہیں کیونکہ اس میں آبادی کا صرف وہ حصہ شامل ہے جو ایکزٹ پول میں حصہ لے کر اپنے جذبات کا اظہار کر رہا ہے اور آبادی کا بقیہ حصہ خارج کر دیتے ہیں جو ان طریقوں میں شامل نہیں ہیں۔

محقق ڈیٹا مائننگ اور مشین لرننگ کی درجہ بندی کی تکنیکوں کا استعمال کر کے انتخابی پیشن گوئی کا ماڈل

بناتا ہے۔ پیشن گوئی ماڈل مختلف تکنیک جیسے سپورٹ ویکٹر مشین، ڈیسیزن ٹری، رینڈم فارسٹ اور کے نیرسٹ نائیبر کے اہم پیمانوں پر تیار کیا گیا ہے۔ ماڈل کی استعداد کار کو بہتر بنانے کے لئے ہم نے انسیمبلنگ تکنیک کا اطلاق کیا ہے اور ماڈل کی توثیق کے لئے شماریاتی ٹی-پیئرڈ ٹیسٹ لاگو کیا گیا ہے۔ ڈیٹا کے عمل درآمد اور اعداد و شمار کے تجربے کے لئے پیشن گوئی ماڈل اینا کونڈا ٹول کے ساتھ جوپیٹر نوٹ بک ویب ایپلیکیشن میں تیار کیا گیا ہے، تاکہ حلقہ کی سطح پر انتخابی نتائج کی پیشن گوئی کی جا سکے۔

1.8 مقالے کا خاکہ

یہ مقالہ مندرجہ ذیل ابواب پر مشتمل ہے۔

باب 1 ایک تعارفی باب ہے جو تحقیقی کام کے پس منظر پر گفتگو کرتا ہے۔ اس کا آغاز انتخابی عمل، جمہوریت، جموں و کشمیر کے انتخابی عمل، انتخابی پیشن گوئی کی تکنیک کے جائزہ کے ساتھ ہوتا ہے۔ اس باب میں انتخابی پیشن گوئیوں میں ڈیٹا مائننگ اور اس کا نفاذ، مسائل کا بیان اور تحقیق کے مقاصد پر بھی تبادلہ خیال کیا گیا ہے۔

باب 2 ڈیٹا مائننگ اور مشین لرننگ کی تکنیکوں کا استعمال کرتے ہوئے انتخابی نتائج کی پیشن گوئی کے بارے میں ادب کا ایک تفصیلی جائزہ پیش کرتا ہے۔ اس باب میں ہم پچھلے ادب کی حدود کو بھی بیان کیا ہے اور اپنے تحقیقی کام میں ان پر قابو پانے کی کوشش کی ہے۔

باب 3 انتخابی پیشن گوئی ماڈلز کے لیے ڈیٹا مائننگ کے آلے اور تکنیک کی وضاحت کرتا ہے۔ اس باب میں ہم نے فلٹر میتھڈ، ریپر میتھڈ اور ایمبیڈڈ میتھڈ جیسے فیچر سلیکشن تکنیک کی تجویز پیش کی۔ ہم نے مختلف ماڈل ایویلوایشن میٹرکس کے ساتھ الخوارزمی کی درجہ بندی جیسے سپورٹ ویکٹر مشین ،ڈیسیزن ٹری، رینڈم فارسٹ اور کے نیرسٹ نائیبر وغیرہ کی تجویز پیش کی ہے۔

باب 4 نالج ڈسکوری ڈیٹا (کے ڈی ڈی) ڈیٹا مائننگ کے طریقہ کار کی وضاحت کرتا ہے جو مقالہ کے مقاصد کو ظاہر کرتا ہے۔ ہم نے تحقیقی ڈیزائن کی بھی تجویز پیش کی جو انتخابی نتائج کی درست پیشن گوئی کے لئے انتخابی پیشن گوئی کے ماڈل کی تشکیل کے لئے ہر اقدام کی باقاعدہ وضاحت کرتی ہے۔ اس باب میں، ہم ڈیٹا کوالٹی کو بہتر بنانے کے لئے تحقیقاتی ڈیٹا تجزیہ، اور قبل از پروسیسنگ تکنیک کی بھی وضاحت کرتے ہیں، جو بعد میں کان کنی کے عمل کی درستگی اور کارکردگی کو بہتر بناتا ہے۔

باب 5 ڈسیزن ٹری، کے نیئرسٹ نائبر، سپورٹ ویکٹر مشین اور رینڈم فارسٹ جیسے ترقی یافتہ انتخابی پیشن گوئی ماڈل کے نتائج کی وضاحت کرتا ہے۔ اس باب میں ہم ہر ترقی یافتہ پیشن گوئی کے ماڈل کی specificity،sensitivity، F1 اسکور، AUROC اسکور، درستگی اور غلط درجہ بندی کی شرح جیسے تشخیص کی پیمائشیں بھی انجام دیتے ہیں۔

باب 6 ترقی یافتہ پیشن گوئی کے ماڈلز پر ہائپر پیرامیٹر آپٹمائزیشن کی تکنیک اور اس کی ایپلیکیشن کی وضاحت کرتا ہے۔ اس باب میں ہم نے بہترین انتخابی نتائج حاصل کرنے کے لئے ترقی یافتہ ماڈلز کے سافٹ ووٹنگ انسیمبلنگ کا استعمال کیا ہے۔ آخر میں، ہم ماڈل کی توثیق کرنے اور اس کی اہمیت کی جانچ کرنے کے لئے شماریاتی ٹی-پیئرڈ ٹیسٹ کا استعمال کرتے ہیں۔

باب 7 میں انتخابی پیشن گوئی کے جدید ماڈل کے خلاصہ، حدود اور مستقبل کے کام کی وضاحت کی گئی ہے۔

باب 2

2. ادب کا جائزہ

اس باب میں انتخابی پیشن گوئی کے ماڈل کی ترقی کے لئے مختلف محققین کی جانب سے ڈیٹا مائننگ کی مختلف تکنیکوں کا استعمال کرتے ہوئے کی گئی اہم خدمات کا خاکہ پیش کیا گیا ہے۔اس باب میں ابتدائی تشخیص اور شناخت کے اوصاف کی اہمیت پر بھی روشنی ڈالی گئی ہے جس کی بنیاد پر انتخابی پیشن گوئی کی جاسکتی ہے۔آخر میں ،مروجہ ادب میں پائے جانے والے تحقیقی خلیوں پر تبادلہ خیال کیا گیا۔

2.1 مختلف ڈیٹا مائننگ ٹاسکس اور تکنیکوں کے استعمال سے الیکشن کی پیشن گوئی

الیکشن کی پیشن گوئی ایک پیچیدہ عمل ہے جو غلط مفروضوں سے پاک نہیں ہے۔انتخابی نتائج کی پیشن گوئی کے لیے، مختلف محققین نے ڈیٹا مائننگ کی متعدد تکنیکوں کو راپول ڈیٹا لاگو کیا اور انتخابی نتائج کی پیشن گوئی کی۔پیش کی گئی تحقیق ایک تنقیدی ادب کے جائزے پر مبنی ہے

تاکہ سیاسی پیشن گوئی کے مقاصد کے لئے ڈیٹا مائننگ کی تکنیک کے ساتھ لازمی پیرامیٹرز کی صلاحیت اور اس کی اہمیت کا صحیح تاثر لگایاجاسکے۔

2.1.1 پیرامیٹرک نقطہ نظر کااستعمال کرتے ہوئے انتخابی پیشن گوئی

پیرامیٹرک نقطہ نظر کواستعمال کرنے کابنیادی مقصد انتخابی پیشن گوئی کے لئے انتہائی اہم پیرامیٹرز کا پتہ لگانا ہے۔ مختلف ممالک اور معاشروں میں سیاسی پیشن گوئی کے متعدد پیرامیٹرز موجود ہیں۔ مندرجہ ذیل محققین نے ڈیٹا مائننگ تکنیک کااستعمال کرتے ہوئے انتخابی پیشن گوئی کے لئے مخصوص پیرامیٹرک نقطہ نظر کااطلاق کیاہے۔

Erikson and Wlezien (2008) [42] امریکی صدارتی انتخابات کی پیش گوئی کے لئے ایک معاشی پیشن گوئی کے ٹول کو تیار کیا۔ تجزیہ کے لئے منتخب کردہ پیرامیٹر معاشی ، دوسرا غیر معاشی تھا جس میں صدارتی منظوری یا trial heats polls اور تیسرے نمبر پر پچھلے انتخابات کے اعداد وشمار شامل ہیں۔رووٹ مین اسکوئرارر کے ساتھ درجہ بندی کا اطلاق کرنے کے بعد، محقق نے انکشاف کیا ہے کہ اگر اہم اقتصادی اشارے (ایل ای

آئی) میں سہ ماہی(0.1%) کا اضافہ ہوا ہے تو آئندہ انتخابات میں برسر اقتدار پارٹی اچھا مظاہرہ کرسکتی ہے۔لہذا اس تحقیق سے یہ بات نوٹ کی جاسکتی ہے کہ رائے دہندگان کی قسمت کا فیصلہ کرنے میں معیشت اہم کردار ادا کرتی ہے ۔

Pavia et al. (2008) [43] جغرافیائی عنصر کے ساتھ معیشت ، ترقی وغیرہ جیسے اہم پیرامیٹرز پر مبنی انتخابی نتائج کی پیش گوئی کرنا۔عام جغرافیائی اور عام کوکریگ جیسی دو جیو شماریاتی تکنیک کا استعمال پولنگ اسٹیشنوں کے مقامی مقامات پر مبنی انتخابی نتائج کی پیش گوئی کے لئے کیا گیا تھا۔ محقق کا تجزیہ کرنے کے بعد یہ ثابت ہوا کہ دونوں مقامی (کرینگنگ) اور مکانی (کوکریگنگ) کی پیشن گوئی انتخابات کے نتائج کو مضبوطی سے بہتر بناتی ہے۔مزید یہ کہ ، پیش کش میں یہ بھی بتایا گیا کہ جب مقامی پولنگ اسٹیشنوں کی تعداد زیادہ ہوتی ہے اور ان کی جگہیں مختلف ہوتی ہیں تو مقامی پیشن گوئی وقتی پیشن گوئی کے مقابلے میں بہتر نتائج حاصل کرتی ہے۔

Norpotha and Gschwend (2010) [44] اس تحقیقی کام کا مقصد جرمنی کے انتخابات کے تین پیرامیٹرز پر مبنی نتائج کی پیشن گوئی کرنا ہےجسے the fame factor, the long-term

partisan balanceاور the cost of ruling۔ تجرباتی نتائج سے ظاہر ہوا ہے کہ مذکورہ انتخابی عوامل انتخابی پیش گوئی میں بھاری اور بھرپور کردار ادا کرتے ہیں جبکہ انتخابی پیش گوئی کے زیادہ سے زیادہ نتائج(1.3٪) سے کم کی غلطی کی شرح کے ساتھ ہوتے ہیں۔

Lewis-Beck and Nadeau (2011) [45]اس تحقیقی کام کا بنیادی مقصد امریکی صدارتی انتخابات کی پیشن گوئی میں معاشی ماڈل تیار کرنا ہے۔معاشیات کی تین اہم جہتوں کا استعمال کرتے ہوئے جیسے توازن ، مقام ، اور وراثت ذیلی جہتوں کے ساتھ ، عمر ، جنس ، نسل ، تعلیم ، آمدنی ، طبقہ ، وراثت وغیرہ پر اس سروے کی بنیاد رکھی گئی تھی۔ محقق نے کئے گئے سروے میں لاجسٹک ریگریشن مساوات کا اطلاق کیا تھا اور اس بات پر زور دیا تھا کہ معاشی جہت حکومت سازی اور امیدواروں کے انتخاب میں اہم کردار ادا کرتی ہے۔مجوزہ معاشی ماڈل میں یہ دکھایا گیا ہے کہ براک اوباما کے 2008 کے انتخابات میں کامیابی کے زیادہ امکانات ہیں ،جیسا کہ اقتصادی پیرامیٹر کی رائے تھی۔

Emre Toros (2011)[46] محقق نے ترکی میں تین نظریاتی مقدمات کی بنیاد پر انتخابات

کی پیشن گوئی کرنے کے لئے ایک ماڈل کی تجویز پیش کی۔ محقق نے پایا کہ ووٹ شیئر میں تبدیلی تین بنیادی اسباب پر منحصر ہے جیسے ''معاشی حالات ، مقامی انتخابات کی کامیابی اور سیاسی ڈھانچہ'' محقق نے مجوزہ ماڈل کے نقائص کی شرح نکالنے اور اس کے معیار کی جانچ کرنے کے لیے Lewis Beck کے ٹولز کا استعمال کیا۔ جانچ پڑتال کے بعد انہوں نے پایا کہ انتخابی نتائج کی پیشن گوئی کرنے کے لیے یہ اسباب فیصلہ کن ہیں۔

S Singh (2012) [47] نو متفرق انتخابی پیرامیٹرز کی بنیاد پر ایک مبہم منطق پر مبنی انتخاب کی پیش گوئی کا ماڈل تیار کیا۔ محققین نے ترقی یافتہ ماڈل کے آؤٹ پٹ کا تعین کرنے کے لئے مصدنی تکنیک کا استعمال کیا۔ اس طریقہ کار کو نافذ کرنے کے بعد یہ نتیجہ اخذ کیا گیا کہ یہ تکنیک انتخابی پیشن گوئی میں اہم کردار ادا کرتی ہے۔

Kodinariya (2012) [48] محققین نے کچھ اہم انتخابی پیشن گوئی کے پیمانوں جیسے امیدوار، وقت، ووٹر کا تعلیمی معیار، مزہب عمر اور شیشن وغیرہ پر منحصر انتخابی اعداد و شمار کا ڈیٹا ویئر ہاؤس تیار کیا ۔ ویژولائزیشن کے لیے Microsoft SQL server 2000 کا استعمال کرتے ہوئے مجوزہ

نقطہ نظر کا استعمال پکسل پر مبنی تکنیک کے ساتھ ہوتا ہے۔ ڈیٹاسیٹ کے مناسب تصور کے بعد محقق نے دعوی کیا کہ اس تکنیک کو عام لوگوں میں شعور پیدا کرنے کے لئے استعمال کیا جاسکتا ہے۔

Singh et al. (2013) [32] fuzzy cognitive map کا استعمال کرکے ایک پیش گوئی کرنے والا ماڈل تیار کیا۔ جو دس مختلف متغیرات پر مبنی امیدوار کے جیتنے کے امکانات کی پیشن گوئی کے لئے استعمال ہوتا ہے۔مناسب تجزیہ کرنے کے بعد ، یہ تسلیم کیا گیا کہ مجوزہ ماڈل کو نتیجہ کی پیشن گوئی کے لئے اصل نتائج کے اعلان سے پہلے ہی استعمال کیا جاسکتا ہے ، لیکن بہتر پیشن گوئی کرنے کے لئےمزید پیرامیٹر کو شامل کرنے کی ضرورت ہے ۔

Dahlberg Stefan et al. (2013) [49] تجزیہ کیا کہ پارٹی اور سیاسی نظام جیسے پیرامیٹرز انتہائی باہم وابستہ ہیں اور ان کا براہ راست اثر ووٹرز پر پڑتا ہے۔ان کے تجزیے کے لئے، محقق نے 86811 جواب دہندگان کے 32 مختلف ممالک کا سروے کیا۔ نتائج سے پتہ چلتا ہے کہ سسٹم سے

متعلق متغیرات کا ووٹروں کے پر سب سے چھوٹا اثر پڑا جبکہ پارٹی اور پھر انفرادیت سے متعلق متغیرات کا سب سے زیادہ اثر ووٹرز پر پڑا۔

Hummel and Rothschild (2014) [33] مختلف پیرامیٹر کی بنیاد پر ایک ماڈل تیار کیا۔ یہ ماڈل linear regression کا استعمال کرتے ہوئے تیار کیا گیا ہے، محقق نے یہ نتیجہ اخذ کیا کہ یہ ماڈل کسی اعلیٰ سیاستداں کی پیشن گوئی کے لئے بہتر ہوسکتا ہے جس میں کامیابی کے امکانات صدر کے لئے 90 فیصد، سینیٹر کے لئے 82 فیصد اور گورنر کے لیے 79 فیصد بالترتیب کم غلطی کی شرح کے ساتھ۔

Yu Wang et al. (2016) [50] مخصوص پیرامیٹر کی بنیاد پر انتخابی پیشن گوئی کے لئے ایک ماڈل تیار کیا۔ یہ ماڈل آنلائن حاصل کردہ ٹوئیٹر ڈیٹا پر تربیت یافتہ ہیں۔ انھوں نے اپنے تحقیقی کام کے لیے چار پیرامیٹرز یعنی معاشرتی سرمائے ، جنس ، عمر ، اور نسل لئے ،اور درجہ بندی اور تجزیہ کے لئے convolutional neural network اور Face++ API آلہ کا استعمال کیا۔

ٹویٹر کے اعداد و شمار کے سیٹ کی مائننگ کے بعد انہوں نے یہ نتیجہ اخذ کیا کہ خواتین صارف مرد صارف کے مقابلے unfollow کرنے میں زیادہ مائل ہیں۔

Mohammad Zolghadr et al. (2017) [51] ایس وی ایم ، اے این این اور لینیئر ریگریشن کا استعمال کرتے ہوئے انہوں نے ایک انتخابی پیشن گوئی ماڈل تیار کیا۔ ماڈل کے کارکردگی کی توثیق مختلف پیرامیٹر کے ذریعے کی گئی مثلا پیشن گوئی کی معمولی کمی (MAPE) ، اور روٹ مین اسکوائر ارر (RMSE)۔ لرننگ الگورتھم کی مناسب جانچ پڑتال کے بعد یہ انکشاف ہوا ہے کہ اے این این کے مقابلے میں ایس وی ایم کے پیشن گوئی کے بہتر نتائج ہیں۔ یہ پوری پیشن گوئی کچھ اہم آزاد متغیرات جیسے جی ڈی پی ، بے روزگاری کی شرح ، ذاتی آمدنی ، وغیرہ۔ تجرباتی نتائج سے معلوم ہوا کہ ڈیٹاسیٹ کی بڑی مقدار پر مختلف الگورتھم لگا کر ماڈل کی درستگی میں مزید بہتری لائی جا سکتی ہے۔

2.1.2 سوشل میڈیا کا استعمال کرتے ہوئے انتخابی پیشن گوئی

سوشل میڈیا ایک ایسا میڈیم سمجھا جاتا ہے جہاں ہر شخص اپنے جذبات کا اظہار کرتا ہے۔ فیس بک اور ٹویٹر جیسے سوشل میڈیا پلیٹ فارم کے ذریعے لائے گئے ڈیٹا کو انتخابی نتائج کی پیش گوئی کے لئے

استعمال کیا جا سکتا ہے۔۔ مندرجہ ذیل محققین نے سوشل میڈیا ڈیٹا کے ساتھ ڈیٹا مائننگ کی تکنیک کا استعمال کرتے ہوئے انتخابی نتائج کی پیشن گوئی کرنے میں اہم کردار ادا کیا ہے۔

Conover et al. (2011) [52] کلاسیفیکیشن الگورتھم کا استعمال کرتے ہوئے ایک انتخابی پیشن گوئی ماڈل تیار کیا۔ استعمال کیا گیا یہ انتخابی پیشن گوئی ماڈل ٹویٹر استعمال کرنے والوں کے میٹا ڈیٹا پر تربیت یافتہ ہے، جس میں مختلف صارف کے ذریعہ تیار کردہ وسیع ڈیٹا کی چھپی ہوئی معلومات کی نشاندہی کرنے کے لئے محققین نے Semantic تجزیہ کا اطلاق کیا ہے۔ محققین نے درجہ بندی اور مواصلات کے نیٹ ورک کے لیے کلسٹرنگ الگورتھم کا استعمال کیا ہے۔ یہ طریقہ انتخابی تجزیہ کی پیشن گوئی کے لئے 91 فیصد درستگی کی شرح کے ساتھ استعمال کیا جا سکتا ہے ۔

Lei Shi et al. (2012) [53] ٹویٹر ڈیٹا کا استعمال کرتے ہوئے انتخابات کی پیشن گوئی کرنے کے لئے ایک پیشن گوئی ماڈل تجویز کیا ہے۔ ٹویٹس کی شکل میں ستمبر سے فروری 2012 تک ٹویٹر کا ڈیٹا اکٹھا کیا گیا تھا۔ پیشن گوئی کی پریشانیوں کی تحقیقات کے لیے Lasso (Least

(Absolute Shrinkage and Selection Operator)ریگریشن ایلگورتھم کا استعمال کیا گیا تین ریاستوں سے حاصل کی جانے والی پیشن گوئی کے نتائج کا موازنہ ریئل کلیئر پولٹکس ویب سائٹ سے کیا گیا ہے۔ تین ریاستوں کے نتائج مرتب کرنے کے بعد یہ انکشاف ہوا کہ امریکی صدارتی انتخابات کی پیشن گوئی کرنا ممکن ہے۔

Mahmood Tariq et al. (2013) [54] ٹویٹر کا ڈیٹا اکٹھا کیا اور انتخابی پیشن گوئی کے ابتدائی نتائج کے لئے نیوی بائز، سپورٹ ویکٹر مشین اور ڈسیزن ٹری الگورتھم کا اطلاق کیا۔ انتخاب کی پیشن گوئی کا ماڈل درجہ بندی الگورتھم کے ساتھ ریپڈ مائنر ٹول پر تیار کیا گیا تھا۔ نتیجہ یہ ظاہر کرتا ہے کہ ڈسیزن ٹری کی درجہ بندی کرنے والے عملے، دوسرے پیشن گوئی کرنے والے ماڈلز کی حکمت عملی سے بہتر کارکردگی کا مظاہرہ کرتے ہیں۔

Spyros Polykalas et al. (2013) [55] جرمنی کے انتخابی نتائج کو دو مشہور سیاسی جماعتوں میں تجزیہ کرنے کے لئے گوگل کے رجحانات کا استعمال کیا۔ گوگل کے رجحانات سے جمع کردہ ڈیٹا پر ڈیٹا مائننگ الگورتھم کا اطلاق کرنے کے بعد، محقق نے دعویٰ کیا کہ اس طریقہ

سے انتخابی نتائج کی پیشن گوئی بالکل صحیح طور پر 2013 اور 2009 کے انتخابات کے لئے کی جاسکتی ہے لیکن 2005 میں انتخابی کارکردگی مطلوبہ نمبر تک نہیں پہنچ سکی تھی۔اس کی وجہ یہ ہے کہ 2009 اور 2013 کے مقابلے میں 2005 میں لوگ انٹرنیٹ کااستعمال کم کرتے تھے۔

Fernanda et al. (2013) [56] چار سوشل میڈیا پلیٹ فارم اور سات ریپبلکن امیدواروں کااستعمال کرتے ہوئے 2012ء کے امریکی صدارتی انتخابات سے پہلے کے سوشل میڈیا کے امکانات کا تجزیہ کیا ہے۔اعداد و شمار کو ان کے API اور ٹیکنورتی کااستعمال کرتے ہوئے جمع کیا گیا تھا، پھر جمع کردہ اعداد و شمار کا حجم، توجہ اور مقبولیت کی پیمائش کے بارے میں سوشل میڈیا کی صلاحیت کو حاصل کرنے کے لئے گہرائی سے تجزیہ کیا گیا تھا۔۔حتمی نتائج کا موازنہ Gallup pollکے نتائج سے کیا گیا جس سے ظاہر ہوتا ہے کہ اس مجوزہ نقطہ نظر سے 2012 کے ابتدائی نتائج کے ساتھ ساتھ اس انتخابی عمل کے بارے میں رائے عامہ کے جائزوں سے اچھااثر پڑا۔

Song et al. (2014) [57] ٹویٹر کے ڈیٹا پر ڈیٹا مائننگ کی مختلف تکنیک کا استعمال کرتے ہوئے ایک انتخابی پیشن گوئی ماڈل تیار کیا جسے multinomial topic modeling, network analysis, and co-occurrence retrieval techniques کا استعمال کرتے ہیں۔ تجزیہ کرنے کے بعد، یہ انکشاف ہوا کہ اس تکنیک کا استعمال ٹویٹر میں پیدا ہونے والے متحرک معاشرتی رجحانات اور مواد پر مبنی نیٹ ورک میں ہوسکتا ہے ۔

Ceron Andrea et al. (2014) [58] 2012 میں ہونے والے امریکی صدارتی اور اٹلی پرائمری انتخابات یعنی دو ممالک کے لئے ایک پیش گوئی ماڈل تیار کیا ہے ۔انہوں نے ٹویٹر ڈیٹا کا استعمال کیا اور Hopkins اور king کا بہت ہی کم غلطی کے ساتھ استعمال کرتے ہوئے نگرانی والے حاس تجزیے کیے۔پیشن گوئی شدہ نتائج کا موازنہ امریکی اور اٹلی دونوں ممالک کے لئے روایتی انتخابات کے سروے کے ساتھ کیا گیا ہے۔ مزید برآں ، وہاں پیشن گوئی کے نتیجہ نے دونوں ممالک کے لئے کہیں بہتر درستگی کا مظاہرہ کیا۔سروے کے دوسرے روایتی سروے

کے طریقوں سے جب موازنہ کیا گیا تو امریکہ کے لئے 0.02 فیصد اور اٹلی کے لئے 1.96 فیصد بالترتیب کم معنوی غلطی ہوئی۔

Anjaria Malhar et al. (2014) [59] ٹویٹر ڈیٹا پر مشتمل سوپر وائزڈ مشین لرننگ تکنیک کا استعمال کرتے ہوئے انتخابات کی پیشن گوئی کی۔انہوں supervised مشین لرننگ تکنیک جیسے سپورٹ ویکٹر مشینیں، Naive Bayes , maximum entropy اور artificial neural networkاستعمال کیے۔ صحیح جانچ پڑتال کے بعد انہوں نے پایا کہ ایس وی ایم ا انتخابات کے نتائج کی پیشن گوئی کرنے میں 88 فیصد کی درستگی کے ساتھ دوسرے تمام درجہ بندیوں کو پیچھے چھوڑ دیتا ہے-

Wani and Alone (2014) [60] سوشل میڈیا ڈیٹا پر مبنی ایک پیشن گوئی ماڈل تیار کیا جو ہماری روزمرہ کی زندگی میں پیدا ہوتا ہے۔ سیاسی جماعتوں کے بارے میں صارف کی رائے حاصل کرنے کے لئے ٹویٹر ڈیٹا کو سروے کے مقصد کے لئے استعمال کیا گیا تھا۔انتخابی پیش گوئی کا ماڈل KNNالگورتھم کا استعمال کرکے بنایا گیا تھا۔

Nam Yoonjae et al. (2015) [61] نیٹ ورک تجزیہ اور لینیئر ریگریشن تکنیک کا استعمال کرتے ہوئے ایک انتخابی پیشن گوئی ماڈل تیار کیا گیا۔ سوشل میڈیا سے ڈیٹا سیٹ کی صحیح طریقے سے مائننگ کے بعد ، محقق نے پایا کہ ٹویٹر نے سب سے زیادہ متعصبانہ نتائج برآمد کیے جبکہ آنلائن خبریں کم متعصبانہ رہیں۔ سوشل نیٹ ورک تجزیہ کے لئے دو اہم سیاسی جماعتوں کا انتخاب کیا گیا جس میں پارک جیون ہی (Park Geun -hye) نامی امیدوار نے چاروں سوشل میڈیا ماحول میں مرکزی شخصیات کو یاد دلاتے ہیں اور اس کے حریف امیدواروں نے غیر معمولی حد تک بہتر کارکردگی کا مظاہرہ کیا اور اس طرح انتخابات جیتنے کا زیادہ امکان ہے۔

Kagan et al. (2015) [62] ٹویٹر ڈیٹا پر مبنی انتخابی پیش گوئی کا ماڈل تیار کیا۔ ، انہوں نے diffusion estimation model اور sentiment analysis algorithm کا استعمال کیا۔ ہندوستانی الیکشن ٹویٹ ڈیٹا بیس (IET-Db) کو ماڈل کی تربیت اور انتخابی نتائج کی پیش گوئی کے لئے استعمال کیا گیا تھا۔ انتخابی پیش گوئی کا یہ ماڈل دوسرے ماڈل سے بہتر پایا گیا تجرباتی نتائج کی بنیاد پر۔

Ullah and Irfan (2015) [63] اسلام آباد میں منعقدہ پاکستان کے 2015بلدیاتی انتخابات کی پیشن گوئی ٹویٹر کے اعداد و شمار کا استعمال کیا ہے ۔ پاکستان کی دو اہم سیاسی جماعتوں کے لیے ٹویٹر کے اعداد و شمار کو ٹویٹر API سے منفی ، مثبت اور غیر جانبدار ٹویٹس کی شکل میں جمع کیا گیا تھا۔ 2588 ٹویٹس کا تجزیہ کرنے کے بعد محقق نے پیشن گوئی کی ہے کہ مسلم لیگ نواز 24.11 اور پی ٹی آئی 25.89 نشستیں جیت سکتی ہے ، لیکن اصل نتیجہ میں مسلم لیگ نواز نے 21 سیٹیں اور پی ٹی آئی نے 17 نشستوں پر کامیابی حاصل کی۔ اس ماڈل کی کل درستگی0.330 فیصد ہے۔اس تحقیقی کام میں بنیادی مسئلہ یہ ہے کہ اس نے انتخابی پیشن گوئی کے لئے مقرر کردہ ٹویٹر ڈیٹا کا استعمال کیا جو اسلام آباد کی صرف 10فیصد آبادی ہی استعمال کرتی ہے ، لہذا اس ماڈل کا جانبدارانہ نتیجہ برآمد ہوسکتا ہے کیونکہ اس نے پیشن گوئی کے لئے90 فیصد آبادی کو نہیں دیکھا۔

Conway A. Bethany et al. (2015) [64] ٹویٹر، فیس بک اور نیوز آرٹیکلز جیسے مختلف سوشل سائٹوں سے حاصل کردہ ڈیٹا پر مبنی انتخابی پیش گوئی کا ماڈل تیار کیا جس نے ٹویٹس اور مستند شائع

مضامین کو جمع کرنے کے لئے ADA کان کن اور لفظی ریاستی تکنیک کا استعمال کیا۔ نتائج سے پتہ چلتا ہے کہ ان اعداد و شمار کے مابین غیر بے ترتیب تعلقات موجود ہیں۔ مصنفین نے بتایا کہ اگرچہ ٹویٹر مہم چلانے کا ایک لازمی طریقہ ہے لیکن روایتی تکنیکیں بھی اس میں نمایاں کردار ادا کرتا ہے۔

Tsakalidis Adam et al. (2015) [65] ٹویٹر کی بنیاد پر انتخابی نتائج کی پیشن گوئی کرنے کے لئے ایک ماڈل تجویز کیا۔انہوں نے ٹویٹر API سے ٹویٹس کی شکل میں ڈیٹا اکٹھا کیا اور لوگوں کے جذبات لانے کے لئے لیکسیکون پر مبنی نقطہ نظر کو انجام دیا۔ تجزیہ کرنے اور موازنہ کرنے کے لئے تین الگورتھم کا انتخاب کیا گیا وہ یہ ہیں linear regression ، Gaussian process، اور sequential minimal optimization for regression ، ویکا ٹول کے ذریعہ ، اور غلطی کی نشاندہی کرنے کے لئے مین ایبسلوٹ ارر۔ڈیٹا سیٹ کی مناسب مائننگ کے بعد ، محقق انکشاف کرتے ہیں کہ گاوسی عمل نے سب سے کم ایم اے ای (1.31) حاصل کیا ، اس کے بعد ترتیب وار مینیمل اپٹمائزیشن (1.35) اور لینیئر ریگریشن۔ایسا لگتا ہے کہ گاوسین نے اس مجوزہ کام یعنی ٹویٹر پر مبنی پیشن گوئی میں دوسروں کے مقابلے میں بہتر کارکردگی کا

مظاہرہ کیا۔اس مجوزہ ماڈل کے ساتھ بنیادی پریشانی یہ ہے کہ انہوں نے صرف ٹویٹر ڈیٹا کو استعمال کرکے،انتخابات کی پیشن گوئی کے اصول کو جانے بغیر پیشن گوئیاں کی تھیں جوالیکشن کی حقیقی سچائی کی یقین دہانی نہیں کرسکتا،چونکہ ٹویٹر ڈیٹا بہت بڑی آبادی کے حاس تجزیہ کے لئے بہت چھوٹا نمونہ ہے۔

Singhal Kartik et al. (2015) [66] 2014میں دہلی(ہندوستان) کے لئے انتخابی پیش گوئی کا ایک ماڈل تیار کیا گیا جس میں semantic and context-aware rule سے واقف اصولوں کا استعمال کیا گیا تھا۔مزید،لیکن اور rule based جذبات کے تجزیے کے لئے ہائبرڈ نقطہ نظر استعمال کیا گیا تھا۔اس لفظ کے جذبات کی گنتی کے لیے اور اضافی لفظ کو ہٹانے کے لئے جو جذبات سے متعلق نہیں ہے Stanford parser toolاستعمال کیا گیا تھا۔ مناسب انکشافات کے بعد یہ مشاہدہ کیا گیا کہ یہ طریقہ انتخابی نتائج کی پیشن گوئی کے لئے استعمال کیا جاسکتا ہے ۔تاہم،یہ طریقہ شہری علاقوں کے لئے موزوں ہے کیونکہ ٹویٹر استعمال کرنے والوں کی تعداد دیہی علاقوں کے مقابلے میں زیادہ ہے۔

Jadhav and Deshmukh (2016) [67] بہار سال2015کے ریاستی انتخابات کی پیشن گوئی کرنے کے لئے ایک طریقہ کار تیار کیا -وہ اپنے تجزیہ کار کے لئے ٹویٹس کا انتخاب کرتے ہیں جو ٹویٹر API سے جمع کیے گئے تھے۔مزید انہوں نے ٹویٹس کو مثبت، منفی اور غیر جانبدار درجہ بندی کی۔ Naive Bayesاور R Tools کو حاس تجزیہ اور درجہ بندی کے لئے غلطی کی نشاندہی کرنے کے لئے استعمال کیا گیا تھا۔مناسب منائنگ کے بعد انھوں نے انکشاف کیا ہے کہ عظیم اتحاد کے مقابلے میں این ڈی اے کے زیادہ مثبت ٹویٹس ہیں۔لیکن عظیم اتحاد نے دراصل حکومت کی تشکیل کی تھی۔ بہار میں اس کی وجہ یہ ہے کہ یہ سروے زیادہ تر شہری بہار میں کیا جاتا ہے اور شہری بہار میں لوگ بی جے پی کو دوسری مخالف جماعتوں کے مقابلے میں زیادہ پسند کرتے ہیں۔

Ismail (2016) [68]امریکہ کے صدارتی امیدواروں کے انتخابات کی پیش گوئی کرنے کے لئے ایک ماڈل تیار کیا جو سال2016میں ہوئے تھے۔ وہ اپنے تجزیہ کام کے لئے دو اہم امیدواروں کا انتخاب کرتے ہیں یعنی ڈونلڈ ٹرمپ اور ہلیری کلنٹن۔تجزیہ کے لئے ڈیٹا ٹویٹر API سے

جمع کیا گیا تھا اور10000ٹویٹس جمع کرکے دونوں امیدواروں کا تجزیہ کیا گیا تھا۔ ماڈل کی پروسیسنگ Polarity Lexicon mode اور R Tool کا استعمال کرکے کی گئی۔ صارف کے جذبات کا صحیح تجزیہ کرنے کے بعد، محقق نے انکشاف کیا ہے کہ کلنٹن کے مقابلہ میں ٹرمپ کے انتخاب جیتنے کا زیادہ امکان ہے-

Oliveira et al. (2016) [69] ٹویٹر ڈیٹا سیٹس سے اہم معلومات نکالنے کلے Natural Languageپروسیسنگ کی تکنیک کے ساتھ ساتھ ریپڈ مائنر ٹول کااستعمال کرکے انتخاب کی پیش گوئی کاایک طریقہ تجویز کیا۔ مجوزہ انتخابی پیش گوئی ماڈل ٹویٹر اور روایتی اعداد وشمار پر تربیت یافتہ ہے۔ تجرباتی نتائج نے انکشاف کیا کہ ترقی یافتہ انتخابی پیش گوئی ماڈل کم غلطی کی شرح کے ساتھ انتخابی نتائج کی پیش گوئی کرنے میں موثر درستگی رکھتا ہے۔

[70] Wang and Lei (2016) اس تحقیقی کام کا بنیادی موضوع ہائبرڈ ماڈل بنانا ہے جس میں اشارے یعنی peer-to-peer ratings, sentiment scores, and candidate mentioning volumesکااستعمال ہوتا ہے۔ریگریشن تجزیہ کے ساتھ ماڈل کی درستگی ٹائم

سیریز، تجزیہ کی تشخیص اور پیشن گوئی کے لیے بالترتیب کی گی ہے۔ تجرباتی نتائج سے ظاہر ہوا ہے کہ ترقی یافتہ ماڈل بہترین کارکردگی کے ساتھ موجودہ انتخابی پیش گوئی کے ماڈل کے مقابلے میں بھی بہتر کارکردگی کا مظاہرہ کرتا ہے۔

Sharma and Moh (2016) [71] ہندی زبان کی شکلوں یعنی ٹویٹس میں جمع کردہ ٹویٹر ڈیٹا کا استعمال کرتے ہوئے انتخابی نتائج کی پیشن گوئی کرنے میں حاس تجزیے کو استعمال کرنے کی اہمیت دی۔انہوں نے اپنے کام میں لغت پر مبنی ، نیوی بایس ، اور سپورٹ ویکٹر مشین الگورتھم کو H-SWN(Hindi-sentwordnet) کے ساتھ استعمال کیا۔ ہندی ٹویٹس کے ڈیٹاسیٹ کی مائننگ کے بعد وہ انکشاف کرتے ہیں کہ انتخابی نتائج کی پیشن گوئی کرنے میں اس کا استعمال کیا جاسکتا ہے۔

Maurice Vergeer (2017) [72] سال 2010 سے 2012 تک پانچ متنوع ممالک کے ٹویٹر ڈیٹا کا استعمال کرتے ہوئے انتخابی پیش گوئی کا ماڈل تیار کیا R mass پر پیکیج کے ساتھ منفی دوئم اور رجعت

تراکیب کو استعمال کر کے۔ انتخابی پیش گوئی ماڈل سے حاصل تجرباتی نتائج زیادہ سے زیادہ ہیں تاہم اضافی بہتری کی ضرورت ہے۔

Mellon and Prosser (2017) [73] الگورتھم linear regression کا استعمال کر کے ٹویٹر اور فیس بک کے ڈیٹا کی بنیاد پر انتخابی پیش گوئی کا ماڈل تیار کیا۔ اعداد و شمار متفاوت اعداد و شمار کے گروپوں جیسے ڈیمو گرافکس، سیاسی رویوں وغیرہ سے حاصل کیا گیا ہے۔ اعداد و شمار کے سیٹ کی مناسب جانچ پڑتال کے بعد، یہ دیکھا گیا ہے کہ فیس بک اور ٹویٹر کے صارفین مختلف سیاسی نظاموں کے بارے میں الگ الگ رائے اور رائے رکھتے ہیں۔ جس میں عمر، تعلیم، صنف، ووٹ کا انتخاب اور ٹرن آؤٹ شامل ہیں اس انتخابی پیش گوئی ماڈل کے تجرباتی نتائج سے پتہ چلتا ہے کہ ماڈل مثالی ہے تاہم زیادہ سے زیادہ درستگی کی ضرورت ہے

Ranjan and Reza (2017) [74] گجرات ریاست (ہندوستان) کے انتخابی نتائج کی پیش گوئی کی ہے کہ ٹویٹر کے اعداد و شمار کو جذبات کے تجزیے کے لئے deep learning کے طریقہ کار اور decision tree کو استعمال کر کے پروسیسنگ کے مقصد کے لئے ورڈ 2 ویک ماڈل کے ساتھ استعمال

کیا ہے۔کانگریس اور بی جے پی کے نام سے دو اہم سیاسی جماعتوں کا انتخاب کیا گیا اور python library میں این ایل ٹی کے(نیچرل لینگویج ٹول کٹ)کے ساتھ Tweepy کا استعمال بالترتیب ٹویٹس کو جمع کرنے اور قبل از پروسیسنگ کے لئے کیا گیا۔ٹویٹر ڈیٹا کی کان کنی کے بعد ،یہ مشاہدہ کیا گیا ہے کہ اس طریقے کو سیاسی پیش گوئی کے لئے استعمال کیا جا سکتا ہے

Jain and Kumar (2017) [75] مصنفین نے 2015 میں ہونے والے دہلی کے ریاستی انتخابات کے نتائج کی پیشن گوئی کرنے کے لئے ایک ماڈل کی تجویز پیش کی۔ اس ماڈل کو سپورٹ ویکٹر مشین، نیوی بایس کلاسیفائر decision tree، اور رینڈم فاریسٹ الگورتھم استعمال کر کے تیار کیا گیا تھا۔ دوسروں کے مقابلے میں سپورٹ ویکٹر مشین اچھی کارکردگی کا مظاہرہ کیا گیا تھا جس میں 79.4٪ درستگی کے ساتھ درجہ بندی کرنے کی تجویز پیش کی گئی تھی۔ ترقی یافتہ ماڈل نے عام آدمی پارٹی (اے اے ایم) کے لئے ایک بھاری اکثریت کی فتح کی توقع کی تھی جس سے پتہ چلتا ہے کہ اس ماڈل کو پیش گوئی کے کاموں کے لئے استعمال کیا جا سکتا ہے۔

Navya et al. (2017) [76]ٹویٹر ڈیٹا کا استعمال کرتے ہوئے ہندوستانی انتخابات کی

پیشن گوئی کرنے کے لئے ایک ماڈل تجویز کیا۔ محقق نے جاوا کی زبان کے ساتھ تین درجہ بندی کا استعمال کیا یعنی ARIMA ، CVAR اورImproved CVAR ۔ تین درجہ بندی کے ساتھ ٹویٹر ڈیٹاسیٹ کی مناسب مائننگ کے بعد محقق نے پایا کہ Improved classifier میں ہفتہ وار ٹویٹر کے رجحان کی پیشن گوئی کرنے کی صلاحیت ہوتی ہے ۔

Goyal Shubham (2017) [77] انتخابی پیش گوئی کا ماڈل تیار کیا جس نے K-NNاور نیوی بائیس الگورتھم کا استعمال کیا۔ مجوزہ ماڈل کو ٹویٹر ڈیٹا اور مختلف مصنفین کے تبصروں پر تربیت دی گئی تھی۔ اعداد و شمار کے تجزیہ کے بعد پتہ چلا کہ سوشل میڈیا ڈیٹا کو لوگوں کے جذبات کی پیش گوئی کے لئے استعمال کیا جا سکتا ہے۔ ماڈل انتخابی نتائج کی پیش گوئی کرنے کے لئے کافی اچھا ہے تاہم اس میں مزید اضافہ لازمی ہو سکتا ہے۔

Safiullah et al. (2017) [78] ٹویٹر ڈیٹا کے ڈیٹا تجزیہ کلےregression analysis کی تکنیک کا استعمال کیا۔ تیار شدہ ماڈل کی کارکردگی کو روٹ میین اسکوائر ڈ ایرر کا استعمال کر کے چیک کیا جاتا ہے۔ ڈیٹاسیٹ کی صحیح جانچ پڑتال کے بعد یہ بات سامنے آئی کہ سوشل میڈیا

خصوصا ٹویٹر کو انتخابی پیشن گوئی میں صحت مند اشارے کے طور پر استعمال کیا جاسکتا ہے۔

Gaurav et al. (2017) [79] آن لائن ٹویٹر ڈیٹا پر کے text and mining methods کئی کے طریقوں کا استعمال کر کے انتخابی پیش گوئی کا ماڈل تیار کیا۔ انتخابات کے نتائج کی پیشن گوئی کے لیے natural language پروسیسنگ کے طریقوں اور ٹویٹر سرچ API کا استعمال کرتے ہوئے ان ٹویٹس کے تجزیے آر اسٹوڈیو میں کیے گئے تھے۔

Suarez Hernandez A et al. (2017) [80] موڈ تجزیہ کے طریقہ کار کی بنیاد پر انتخابی پیش گوئی کا ماڈل تیار کیا ۔ اس مجوزہ طریقہ کار کا اطلاق ٹویٹر پر ظاہر کردہ معاشرتی جذبات کی پیش گوئی کے لئے کیا جاتا ہے۔ صارف کے ٹویٹس کو مثبت اور منفی لیبلز میں درجہ بندی کرنے کے لئے naïve bayes الگورتھم کا استعمال کیا گیا تھا۔ مجوزہ ماڈل اسٹریٹجک کے تجرباتی نتائج نے یہ ثابت کیا کہ اسے مختلف سیاسی امور خصوصا پولنگ کے اوقات میں صارفین کے آن لائن طرز عمل کو دیکھنے کے لئے ایک بہترین ماڈل کے طور پر استعمال کیا جاسکتا ہے۔

Singh and Sawhney (2017) [81] مختلف ممالک سے موصولہ ٹویٹس کی شکل میں ٹویٹر ڈیٹا کا استعمال کرتے ہوئے ایک سیاسی پیش گوئی کرنے والا ماڈل بنایا۔ مناسب مائننگ کے بعد انہوں نے انکشاف کیا کہ جن ممالک میں انٹرنیٹ صارف 80 فیصد سے زیادہ ہیں وہ اس تجزیے کے لئے موزوں ہیں۔دوسری طرف وہ ممالک جن کے پاس انٹرنیٹ استعمال کنندہ 80 فیصد سے کم ہیں وہ اس پیشن گوئی کے لیے موزوں نہیں ہیں۔

Narwal Neetu et al. (2018) [82] مصنفین ٹویٹر کے ذریعے 2018 میں ہونے والے دہلی ایم سی ڈی(MCD) انتخابات کی پیشن گوئی کرنے کی کوشش کرتے ہیں by using آر ٹول اور K-means clustering algorithm۔تجزیہ کے لئے تین اہم سیاسی جماعتوں کا انتخاب کیا گیا تھا ، عآپ ، بی جے پی اور کانگریس۔ مناسب تجزیہ کے بعد محقق نے انکشاف کیا ہے کہ بی جے پی کو دیگر مخالف جماعتوں کے مقابلے میں زیادہ مثبت تبصرے اور ٹویٹس ملے ہیں ، لہذا بی جے پی کے انتخابات جیتنے کا زیادہ امکان ہے۔

Hasan et al. (2018)[83]درجہ بندی الگورتھم یعنی سپورٹ ویکٹر مشین اور نیو یو بایس کے

ساتھ ویکاٹول کااستعمال کرتے ہوئے انتخابی پیش گوئی کاماڈل تیار کیا۔ محقق نے نگرانی کے ساتھ three sentimental lexicons viz (W-WSD, SentiWordNet, TextBlob) کا استعمال کیا۔ محقق نے تجزیہ کیا کہ تینوں جذباتی تجزیہ کاروں میں سے،TextBlob میں دیگر جذباتی تجزیہ کار کے مقابلے میں زیادہ درستگی کی شرح ہے۔

Imane El Alaoui et al. (2018) [84] ٹویٹر ڈیٹا کے ساتھ big data techniques کا استعمال کرکے انتخابی پیش گوئی کاماڈل بنایا۔ ٹویٹر ڈیٹا سیٹ تجزیہ کے لئے استعمال کیا گیا تھا جو apache Kafka کا استعمال کرتے ہوئے اکٹھا کیا گیا تھا ، ایچ ڈی ایف سی میں محفوظ تھا اور procssing spark کے لئے ملازم تھا۔ٹویٹس کا sentimental analysis دو امیدواروں کے لئے کیا گیا تھا اور اس کا نتیجہ سے ظاہر ہوتا ہے کہ مجوزہ نظام مذکورہ بالا ماڈل کے مقابلے میں زیادہ درستگی کا حامل ہے۔

Mazumder Pritom et al. (2018)[85] انتخابات کی پیش گوئی کلے ایک انتخابی پیش گوئی کاماڈل پیش کیاے۔اپنے تجزیہ کار کے لئے جو Harvard Dataverse, Twarc library اور ٹویٹر

API سے جمع کیا گیا ماڈل ٹویٹر ڈیٹا، ٹورک لائبریری اور ٹویٹر API پر تیار کیا گیا ہے۔ مجوزہ انتخاب کی پیشن گوئی کا ماڈل اے این ایف آئی ایس (adaptive neuro fuzzy انفینس سسٹم) کی بنیاد پر تیار کیا گیا ہے۔ RMSE (root means square error) کا حساب لگانے کے بعد وہ 16 فیصد غلطیاں لے کر آئے تھے جو 84 فیصد درستگی کی طرف جاتا ہے۔

Thampi et al. (2018) [86] جذباتی تجزیوں کا استعمال کرتے ہوئے 2014 کے ہندوستانی انتخابی نتائج کی پیش گوئی کرنے کے لئے ایک ماڈل تجویز کیا۔ وہ اپنے تجزیہ کار کاموں کے لئے ٹویٹر ڈیٹا کے ساتھ کانگریس اور بی جے پی کے نام سے سیاسی طور پر دو پارٹیوں کا انتخاب کرتے ہیں۔اس کام میں تین اہم کلاسفائیر استعمال کیے گئے تھے وہ تھے، نیوی بائز، سپورٹ ویکٹر مشین اور maximum entropy classifier ۔ ٹویٹس کی مثبت ، منفی اور غیر جانبدارانہ درجہ بندی کرنے کے بعد ، محقق نے بی جے پی کے لئے تقریبا-65 60 فیصد کی درستگی کے ساتھ فتح کی پیشن گوئی کی ہے۔

2.1.3 گزشتہ انتخابات کے اعداد و شمار کا استعمال کرتے ہوئے انتخابی پیشن گوئی

اس حصے میں محققین ایسے تحقیقی کام پر تبادلہ خیال کریں گے جو ماضی کے انتخابی ڈیٹا سیٹ پر مشتمل ہے اور اس کی بنیاد پر مستقبل کے انتخابی نتائج کی پیشن گوئی کرتے ہیں [87]۔ مندرجہ ذیل محققین نے متعدد ممالک کے ماضی کے انتخابات کا ڈیٹا سیٹ مختلف وقت کے ساتھ استعمال کیا اور آئندہ انتخابات کے نتائج کی پیشن گوئی کی ہے۔

جی ایس گل (2008) [88] ہندوستان کے لوک سبھا انتخابات کی پیشن گوئی کرنے کے لئے ایک ماڈل کی تجویز پیش کی۔ نیورل نیٹ ورک کے مرحلے کو استعمال کرتے ہوئے بنایا گیا تھا جس میں عام عوام کے سروے اور نو انتخابات کے اعداد و شمار شامل تھے۔اس مقصد کے لئے صرف ایک ان پٹ اور آؤٹ پٹ پرتوں پر مشتمل ایک دو پرت والا نیٹ ورک استعمال کیا گیا تھا۔ نیورل نیٹ ورک کے استعمال کے بعد ،محقق نے دعویٰ کیا کہ یہ ماڈل غلط پیشن گوئی کے ساتھ چھ درست پیشن گوئی کر سکتا ہے۔ لہذا اس کو مزید بہتری کے ساتھ پیشن گوئی کرنے والے ٹولز کے طور پر استعمال کیا جا سکتا ہے۔

Wiji Arulampalam et al. (2008) [34] الگورتھمlinear regression کو شامل کرکے نظریاتی انتخابات کی پیش گوئی کا ماڈل تیار کیا۔ اس ماڈل کو 1975 سے 1996 کے دوران مختلف ریاستوں کے ڈیٹا سیٹ پر تربیت دی جاتی ہے۔ تجرباتی نتائج سے پتہ چلتا ہے کہ وہ ریاستیں جن کا مرکز حکومت سے مضبوط رشتہ ہے، انہیں زیادہ سے زیادہ گرانٹ ملتا ہے جبکہ دوسری ریاستیں جو غیر منسلک یا جزوی طور پر یونین حکومت کے ساتھ منسلک ہیں، کم گرانٹ وصول کرتی ہیں۔

Campbell and Lewis-Beck (2008) [89] **سال 1979 سے 2008 کے درمیان منظم ادب کے جائزے پر مبنی انتخابی پیش گوئی کا ماڈل تیار کیا۔** اس ماڈل کو تجرباتی نتائج کی بنیاد پر انتخاب کی پیش گوئی اور تجزیہ کے لئے استعمال کیا گیا تھا۔ تاہم، ماڈل کے ڈیزائن کردہ مرحلے میں اعداد و شمار، پیرامیٹرز اور ٹیکنالوجیز کی اصلاحات کے ساتھ مزید بہتری کی ضرورت ہے۔

Murray et al. (2009) [90] اس تحقیقی کام کا بنیادی مقصد امریکی صدارتی انتخابات کے لئے دو مختلف طرح کے ووٹر ماڈل کی تجویز پیش کرنا ہے۔ دو سروے کیا گیا تھا جو متغیرات، ووٹ کا ارادہ اور سابقہ صدارتی ووٹنگ 1964 ، 1976 ، 1980 اور 1988 کے انتخابات

پر مشتمل تھا۔ Iterative Expert Data Mining likely voter 2 (IEDM LV2) technique کو ڈسیزن ٹری کے ساتھ استعمال کرنے کے بعد تفتیش کار نے زور دے کر کہا کہ یہ ماڈل دوسرے ماڈل کے مقابلے میں انتخابی نتائج کی پیشن گوئی 78 فیصد کی درستگی اور کم لاگت کے ساتھ کرتا ہے-

Munzert et al. (2017) [91] سروے کے پول اور پچھلے انتخابات کے نتائج کی بنیاد پر 2013 کے جرمن انتخابات کی پیشن گوئی کے لئے ٹائم سیریز کا طریقہ تجویز کیا تھا۔ مناسب طریقے سے مائننگ کے بعد مصنفین نے دعویٰ کیا کہ یہ ماڈل 299 انتخابی حلقوں کی صحیح پیشن گوئی کرتا ہے جس میں 2.5 فیصد معمولی نقص ہے–

Dassonneville et al. (2017) [92] نیدرلینڈ کے لئے الگورتھم least square regression کا استعمال کرتے ہوئے 2017 میں انتخابی پیشن گوئی کا ماڈل تیار کیا۔ پیشن گوئی گزشتہ 18 انتخابات پر مبنی تھی جس میں جی ڈی پی ، بے روزگاری ، سابقہ کامیابی اور اپنے عہدے میں طویل

مدت کے تین پیرامیٹرز تھے۔ تجرباتی نتائج سے پتہ چلتا ہے کہ اس ماڈل کو minimum least square error کے ساتھ سیاسی پیش گوئی کے لئے استعمال کیا جا سکتا ہے۔

2.1.4 ہائبرڈ نقطہ نظر کا استعمال کرتے ہوئے انتخابی پیشن گوئی

سنگل سسٹم کے ڈیزائن کے اندر متعدد طریق کار کا امتزاج ہائبرڈ سسٹم کا نتیجہ ہے۔ ہائبرڈ سسٹم تمام طریق کار سے سب سے بہتر طریق کار اخذ کرتے ہیں اور انتخابی پیشن گوئی کے لئے ایک بہترین حل فراہم کرتے ہیں۔ مندرجہ ذیل محققین نے انتخابی نتائج کی پیشن گوئی کے لئے ڈیٹا مائننگ کی متعدد تکنیک استعمال کی ہیں۔

Draper and Riesenfeld (2008) [93] جدید ترین صارف انٹرفیس کے ساتھ انتخابی پیش گوئی کا ماڈل تیار کیا جس سے صارفین کو میٹا ڈیٹا پر استفسارات پیدا کرنے اور کم سے کم وقت کی ضرورتکے نتیجے میں نتائج کو دیکھنے کی اجازت مل جاتی ہے۔ وہ ٹیبلولر ڈیموگرافک ڈیٹا سے استفسار کرنے اور تجزیہ کرنے کے لئے ایک نیا انٹرایکٹو ویژولائیزیشن تجویز کر سکتے ہیں نئے

صارف کے ساتھ ساتھ تربیت یافتہ پیشہ ور صارف بھی اسی طرح کے اندازاور اوقات میں وسیع ڈیٹا کا پتہ لگاسکتے ہیں۔مناسب جانچ پڑتال کے بعد یہ انکشاف ہوا کہ رائے عامہ کے اعداد و شمار کے تجزیہ میں یہ تکنیک موثر کردار ادا کرتی ہے ۔

Rigdon et al. (2009) [94] امریکی صدارتی انتخابات2004کی پیشن گوئی کرنے کے لئے ایک طریقہ کار تجویز کیا تھا جس میں بایسینalgorithm کا استعمال ہوتا ہے۔ اس نے اعتراف کیا کہ اعلی پیش گوئی کرنے کی اہلیت کے ساتھ انتخابی انتخاب کے زیادہ سے زیادہ نتائج حاصل کرنے کے لئے ڈیٹاسیٹ کو حتمی دیکھ بھال کے ساتھ پہلے سے عملدرآمد کرنا چاہئے کیونکہ noisyڈیٹاسیٹ نے مجوزہ انتخاب کی پیش گوئی ماڈل کی کارکردگی کو خراب کردیتاے-

Vreese (2009)[95]2004کے انتخابات کے لئے آٹھ یورپی ممالک کی سیاسی انتخابی مہموں کے وسیع دائرہ کار کا تجزیہ کیا۔ تجزیہ کے بعد یہ مشاہدہ کیا گیا کہ انتخابی مہم کے سلسلے میں مختلف ممالک مختلف ہوتی ہیں-لوگ یوروپ میں پہلے آرڈر والے انتخابات کے مقابلے میں دوسرے آرڈر

کے انتخابات میں کم سرگرم ہیں۔

Armstrong et al. (2010) [96] امریکی صدارتی انتخابات کے نتائج کی پیشن گوئی کرنے میں چہرے کے تاثرات کو استعمال کرنا ہے۔اپنے تجزیے کے لئے انہوں نے ڈیموکریٹک پارٹی کی نامزدگی سے 11 اور ریپبلکن پارٹی کی نامزدگی سے بالترتیب 13 امیدواروں کی چہرے کی اہلیت کی درجہ بندی کا انتخاب کیا۔پیشن گوئی کرنے کے لئے محقق نے امیدواروں کی تصاویر استعمال کیں اور فیصلے طلب کیے ۔ درجہ بندی کے بعد ،محقق نے دعویٰ کیا کہ اوباما کو زیادہ rating درجہ حاصل ہے اورانتخابات میں کامیابی کی امیدان کی بہت زیادہ ہے۔

Lewis-Beck and Stegmaier (2011) [97] کام کا بنیادی مقصد سروے کے ذریعے جمع لوگوں کی رائے پر مبنی برطانوی انتخابات کی پیشن گوئی کرنا ہے۔محقق نے اس کام کے لیے دو طریقہ کار اپنائے۔ (1) شماریاتی ماڈل اور (2) رائے شماری۔ حاصل سروے کی بنیاد پر 1992 سے 2005 تک کے چار انتخابات کا تجربہ کیا گیا ، جس میں بتایا گیا کہ مجوزہ طریقہ انتخابی نتائج کی درست پیشن گوئی کرتا ہے۔آخر میں،انہوں نے ٹیلیفون سروے کے

ساتھ انٹرنیٹ سے آن لائن سروے شامل کرکے آئندہ 2010 کے انتخابات کی پیشن گوئیاں کیں ، جس میں انہوں نے پیشن گوئی کی تھی کہ قدامت پسند پارٹی انتخابات جیت سکتی ہے اور در حقیقت قدامت پسند پارٹی انتخابات جیت گئی۔اس طرح یہ طریقہ انتخابات کے نتائج کی پیشن گوئی کے لئے استعمال کیا جاسکتا ہے۔

Ford et al. (2015) [98] رطانوی رائے شماری میں رائے دہندگان کی ترجیحات سے پارلیمانی انتخابات کے نتائج کی پیشن گوئی کرنے کے لئے مصنفین نے ایک ایسا طریقہ تیار کرنے کی کوشش کی جو تین مراحل پر مشتمل ہے۔پہلے مرحلے میں مختلف پولنگ ایجنسیوں سے ڈیٹا اکٹھا کرنا تھا اور پھر سیاسی جماعتوں کے لئے ووٹر کے ارادے کی پیشن گوئی کرنا تھا اور پھر نتیجہ کا تخمینہ لگانا تھا۔ مجموعی طور پر یہ طریقہ قومی، مقامی اور علاقائی اعداد و شمار کو جوڑتا ہے اور انتخابی نتائج کی پیشن گوئی کرتا ہے۔

Khatua et al. (2015) [99]ہندوستان کے انتخابی نتائج کی پیش گوئی کے لئے انتخابی پیش گوئی کا

ماڈل تیار کیا۔ حساس تجزیہ کے لئے لیکسیکن کا طریقہ کار اور غلطی کی نشاندہی کے لئے بہت کم غلطی کے ساتھ ووٹ سوئنگ کے لئے OLS ریگریشن ماڈل کا استعمال کیا گیا تھا۔ آخر میں، یہ مشاہدہ کیا گیا ہے کہ یہ طریقہ سیاسی پیشن گوئی کے لئے استعمال کیا جا سکتا ہے۔

Rallings et al. (2015) [100] انتخابی پیشن گوئی کا نمونہ پیش کیا جس میں انہوں نے برطانیہ کے پارلیمانی انتخابات کے قومی سطح کے نتائج کی پیشن گوئی کے لئے مقامی انتخابی نتائج کا استعمال کیا۔ اس پیشن گوئی کا بنیادی مقصد انتخابی حلقوں کی نشاندہی کرنا ہے، جہاں مقامی سطح کا پیٹرن دی گئی پارٹی یعنی لبرل ڈیموکریٹ، لیبر اور کنزرویٹو پارٹی کی حمایت کرتا ہے یا متوقع حد میں گرتا ہے یا نہیں۔ مناسب تجزیہ کے بعد محقق نے پیشن گوئی کی کہ کنزرویٹو 275 ،لی 280، لبرل ڈیموکریٹس 22 اور ایس این پی 48 نشستیں جیت سکتے ہیں۔ لہذا اس ماڈل کو سیاسی پیشن گوئی کے لئے استعمال کیا جا سکتا ہے۔

You et al. (2015) [101] انتخابی نتائج کی پیش گوئی کے لئے ایک رائے شماری کی پیش گوئی کرنے والا

ماڈل بنایا۔ پیشن گوئی کے لئے ایکcompitative vector auto regression (سی وی اے آر)کی

تجویز پیش کی۔وہ تجزیہ کے کام کے لئے فلکر ڈیٹا کا انتخاب کرتے ہیں - انہوں نے اپنے نتائج

کا موازنہ مختلف پیشن گوئی کرنے والے ماڈلز کے ساتھ کیا جیسے VAR ، AR اور یہ پتہ چلا کہ

سی وی اے آر ماڈل امریکی انتخابی پیشن گوئوں کے دوسرے ماڈل کے مقابلے میں بہتر ہے۔

Jagdev et al. (2016) [9]ڈیٹاکے تجزیہ کے لئے Hadoop and Map reduce جیسی

بڑی ڈیٹا ٹکنالوجی استعمال کی گئیں۔ہڈوپ ٹول کا استعمال وسیع الجزائید اعداد و شمار

کے حصول اور ذخیرہ کرنے کے لئے کیا گیا تھا اور Map reduce اور پروسیسنگ کے لیے الگورتھم

کو کم کرتا ہے ، اور آر ڈی بی ایم ایس جیسے روایتی ڈیٹا بیس ٹول کے مقابلے میں کم لاگت کی

کھپت کے ساتھ انتخابات سے متعلق معلومات کو کم وقت میں نکالا گیا تھا۔اعداد و

شمار کی مناسب مائننگ کے بعد یہ تسلیم کیا گیا کہ مجوزہ طریقوں کو سیاسی پیشن گوئی کے

لئے استعمال کیا جاسکتا ہے۔

Xie et al. (2016) [102]تائیوان میں ہونے والے صدارتی انتخابات 2016 کی پیشن گوئی ہیں عوام کی رائے کا آن لائن اور آف لائن ڈیٹا اکٹھا کیا۔ محققین نے سگنل پروسیسنگ کے لئے Kalman Filter Techniques اور Real time Burst detection کے لئے Moving Average Model کا استعمال کیا، وسیع ڈیٹا میں غلطی کی شرح 3 فیصد سے بھی کم کے ساتھ تائیوان صدارتی انتخاب کی پیشن گوئی کرنے کے لیے۔ حاصل شدہ نتیجہ امید افزا لگتا ہے اور اسے انتخابی پیش گوئوں کے لئے استعمال کیا جا سکتا ہے۔

Xu and Liu (2017) [103] نے ہانگ کانگ کے legislative council کے انتخابات 2016 کی پیشن گوئی کرنے کے لئے ایک ماڈل data mining and machine learning techniques پر تجویز کیا۔ محقق نے اپریوری الگورتھم کے ساتھ Delegating and weighting کے عمل کو استعمال کرکے ماڈل کے درستگی کے نتائج کو بہتر بنایا۔ تجزیہ کے لئے اعداد و شمار عام رائے عامہ اور سروے کی شکل میں جمع کیا گیا تھا۔ محقق نے یہ نتیجہ اخذ کیا کہ یہ ماڈل 82.5 فیصد درستگی کے ساتھ انتخابی نتائج کی درست پیشن گوئی کر سکتا ہے۔

Sathiaraj et al. (2017) [104]لوزیانا کے تین انتخابات کے انتخابی نتائج کو زیادہ درست طریقے سے پیشن گوئی کرنے کے لئےcampaign specific voter share algorithms (CVS) کا استعمال کرتے ہوئے ایک ہائبرڈ مشین لرننگ اپروچ تیار کیا گیا تھا۔ مجوزہ انتخابی پیش گوئی کے ماڈل نے زیادہ سے زیادہ درستگی اور کم سے کم غلطی کی شرح کے ساتھ نتیجہ تیار کیا۔ اس پیشن گوئی کرنے والے طریقہ کار کی درستگی کے نتائج کا موازنہ اصل نتائج سے کیا گیا جہاں یہ مشاہدہ کیا گیا کہ اس طریقہ کار سے انتخابات کے نتائج کی پیشن گوئی کی جاسکتی ہے۔

Cristina et al. (2018) [105] برازیل کے 2016 municipality انتخابات کی پیشن گوئی کرنے کی کوشش کرتے ہیں۔ انہوں نے چار مختلف اخبارات کی پسند ، ناپسند اور تبصرے کی شکل میں آن لائن خبروں کے جذبات تجزیہ کیے۔وہ Cross validationاور بالترتیب Attribute selectionکے لئے weka tool اور M5 methods کے ساتھ ملٹی ویریٹ لینئر ریگریشن تکنیک استعمال کرتے ہیں۔پیمائش کی غلطی کو ناپنے کے لیے Mean absolute errorکا استعمال کیا گیا۔ مناسب مائننگ کے بعد مصنفین کا ماننا ہے کہ انتہائی مثبت

تبصرے اورلائک حاصل کرنے والے امیدواروں کے انتخابات جیتنے کا زیادہ امکان ہے اور لہذا یہ طریقہ انتخابی نتائج کی پیشن گوئی کے لئے استعمال کیا جاسکتا ہے ۔

Sharar and Abd-el-Barr (2018) [106] انتخابی نتائج کی پیش گوئی کے لئے آن لائن سروے کے اعداد و شمار کا استعمال کیا۔ آن لائن سوالنامہ مارچ سے اپریل 2016 کے درمیان تقسیم کیا گیا تھا اور کل 637 نمونے حاصل کیے گئے تھے۔ سروے کے حاصل کردہ نمونے کا تجزیہ شماریاتی ٹول(ایس پی ایس ایس)کے ذریعہ کیا گیا۔ تجرباتی نتائج سے پتہ چلتا ہے کہ سوشل میڈیا ڈیٹا خاص طور پر ٹویٹر ڈیٹا انتخابی نتائج کے لئے ترقی یافتہ ممالک میں اہم کردار ادا کرتا ہے

Pranay Patel (2018) [107] انتخابی تجزیہ اور پیشن گوئوں کے لئے سائیکومیٹرک بگ ڈیٹا اینالٹیکس (تیسری پارٹی کیمبرج اینالیٹیکا) کے ذریعہ اہم کردار ادا کرنے کی وضاحت کی۔۔ کیمبرج کا یہ تجزیہ نگار فیس بک ، الیکشن کمیشن ، automotive store وغیرہ سے عام لوگوں کا ریکارڈ اکٹھا کرتا ہے اور وہ ڈونلڈ ٹرمپ کی انتخابی مہم کی ٹیم کی تشہیر کرتے ہیں تا کہ وہ اپنی نفسیاتی خواہش

کے مطابق ووٹر کو نشانہ بنائیں۔یہ تکنیک ٹرمپ کے لیے اچھی کارکردگی کا مظاہرہ کرتی ہے کیونکہ انہوں نے ووٹروں کی خواہش کے مطابق اپنی تقریروں اور آن لائن اور آف لائن دونوں مہموں سے عوام کو نشانہ بنایا ، جس سے ٹرمپ فتح کو موثر انداز میں لے جاتا ہے۔

2.2ریسرچ گیپس (Research Gaps)

انتخابی نتائج کی پیشن گوئی کئی طریقوں سے کی جاسکتی ہے۔تاہم، انتہائی اہم اور قابل اعتماد طریقے انتخابات کی پیشن گوئی کے اوصاف کی تشخیص پر مبنی ہیں۔مختلف محققین نے حقیقی نتائج کے اعلان سے پہلے ڈیٹامائننگ کی مختلف تکنیکوں کا استعمال کرتے ہوئےانتخابی نتائج کی پیشن گوئی کی ہے، لیکن ادب کاجائزہ لیتے ہوئےاس کو قریب سے دیکھنے سے کئی کوتاہیوں کا انکشاف ہوتا ہے جن کو ذیل میں بیان کیا گیا ہے۔

i. انتخابی پیشن گوئی کے زیادہ تر تیار شدہ ماڈلز میں عمومی صلاحیت کی کمی ہے۔

ii. زیادہ تر انتخابی پیشن گوئی کے ماڈل ٹویٹر اور فیس بک جیسے سوشل میڈیا سائٹ سے اکٹھا کیے گئے ڈیٹا پر مبنی تھےجوFake IDکی بڑی تعداد کی موجودگی کی وجہ سے غلط نتائج برآمد کرسکتے

ہیں۔

iii. انتخابات کی پیشن گوئی کرنے کے لئے ٹویٹر یا فیس بک سے اکٹھا کیا گیا ڈیٹا صرف ترقی یافتہ ممالک میں ہی کیا جاسکتا ہے کیونکہ ترقی یافتہ ممالک میں لوگوں کو انٹرنیٹ تک زیادہ رسائی حاصل ہے۔لیکن ترقی پذیر ممالک کی صورت میں یہ متعصب نتائج کی پیشن گوئی کر سکتی ہے کیونکہ آبادی کا صرف تھوڑا حصہ ہی اسے استعمال کررہا ہے۔

iv. بوٹس کے تصورات، ٹویٹر میں بوٹس کے استعمال سے، زیادہ سے زیادہ ٹویٹس پوسٹ کر سکتے ہیں جتنا وہ کسی بھی سیاسی جماعت کی حمایت کرنا چاہتے ہوں۔اور یہ بات قابل غور ہے کہ اگر سوشل میڈیا کے جذبات کی پیمائش کرنا انتخابات کی پیشن گوئیوں کے لئے ایک زیادہ قائم طریقہ بن جاتا ہے تو، دونوں طرف کے لئے یہ بہت بڑامحرک ثابت ہوگا کہ وہ بوٹس کا استعمال کرکے یہ تاثر دیں کی وہ انتخابات جیت رہے ہیں۔

v. محدود انٹرنیٹ رسائی سب سے بڑی رکاوٹ ہے اور خاص طور پر ترقی پذیر علاقوں میں اعداد وشمار کے لحاظ سے اہم نمونے جمع ہونے سے روک سکتی ہے۔گورنمنٹ سینسرشپ کی

مختلف کوششوں سے نتائج ضائع ہوسکتے ہیں اور اس کا محاسبہ کرنا مشکل ہوسکتا ہے۔جیسا کہ ہم جانتے ہیں کہ انٹرنیٹ استعمال کرنے والے زیادہ تر ممالک میں اوسط شہری کے مقابلے میں کم عمر اور دولت مند ہوتے ہیں۔ درست معلومات جمع کرنے کے لئے ان عوامل پر قابو رکھنا ضروری ہے۔

vi. موجودہ انتخابی پیشن گوئی کا نمونہ صرف ترقی یافتہ علاقوں میں ہی لاگو ہوسکتا ہے لیکن جموں وکشمیر (J&K) جیسے ترقی پذیر علاقوں میں اس کا استعمال نہیں کیا جاسکتا کیونکہ جموں وکشمیر میں اکثر بجلی کٹ جاتی ہےاور انٹرنیٹ ناکہ بندی ہوتی ہے اور کوئی بھی،سیاسی جماعتوں کے تعلق سے، اپنے جذبات کا اظہار کرنے کے لئے انٹرنیٹ کا استعمال نہیں کرسکتا ۔

مختلف ٹولزجیسے weka، Rapidminer، orange ، وغیرہ تجربات اور نقلی مقاصد کے لئے استعمال ہوتے ہیں۔ تاہم، ہر آلے میں اس کے ساتھ پیچیدگیاں ہوتی ہیں۔ پیشن گوئی کے

لئے بہت سارے میکانزم موجود ہیں لیکن سب کی اپنی حدود ہیں جیسے جی یو آئی کے لئے دستاویزات محدود ہیں، اسکیلنگ ایک مسئلہ ہے، بگ ڈیٹا کو سنبھالا نہیں جاسکتا، وغیرہ۔

vii. انتخاب کی پیشن گوئی کی تشخیص کارکردگی کے اقدامات جیسے حساسیت،صراحت،درستگی، صحت سے متعلق (precision)وغیرہ استعمال کی گئی تھی۔تاہم، محققین کے ذریعہ حسابی پیچیدگی،اسکیل ایبیلیٹی، مضبوطی(robustness)، اور فہمیت(comprehensiblility) اور انتہائی اہم شماریاتی جانچ جیسے ماڈل محققین کے ذریعہ استعمال نہیں کیے جاتے ہیں۔

viii. زیادہ تر محققین اہم خصوصیات کو حاصل کرنے کے لئے صرف ایک ہی خصوصیت کی انتخابی تکنیک کا استعمال کرتے ہیں۔تاہم، انتخاب کی پیشن گوئی کے لئے ان کی اُصل اقدار کے ساتھ نمایاں صفات اخذ کرنے کے لئے متعدد خصوصیت کی انتخاب کی تکنیک کے استعمال کی کوئی تحقیقات نہیں ہیں۔

2.3 رکاوٹوں کو دور کرنے کے لئے مجوزہ حل

مندرجہ بالا تحقیقی خلیوں کو مدنظر رکھتے ہوئے، تحقیق کے اس طرح کے خلیجوں

پر قابو پانے کے لئے ہم نے انتخابی پیشن گوئی کا نمونہ تجویز کیاجو ذیل میں بتایا گیا ہے:

i. ہم نے مرکزی حکومت کے اثر و رسوخ، ذات پات اور حاس علاقوں وغیرہ جیسے نئے پیرامیٹرز کی نشاندہی کی اور ان انتخابی پیرامیٹرز کو شامل کرکے انتخابی نتائج کی پیشن گوئی کرنے کی سطح میں اضافہ ہوا۔

ii. ہم نے چار مختلف الگورتھم کے ساتھ مل کر فیچر سلیکشن کے تین مختلف طریقوں کا اِطلاق کیا اور آخر کار ایک مضبوط انتخابات کی پیشن گوئی ماڈل کی تشکیل کے لئے سافٹ ووٹنگانسیمبلنگ کا استعمال کیا۔ آج تک ہمیں ایسا کوئی تحقیقی کام نہیں ملا جس نے ان تکنیکوں کو خصوصیت کے انتخاب اور انتخابی نتائج کی پیشن گوئوں کے لئے شامل کیا ہو۔

iii. مجوزہ ماڈل سوشل میڈیا پیرامیٹرز کے بجائے مختلف پیرامیٹرز کے مختلف سیٹوں پر بنایا گیا ہے کیونکہ جموں وکشمیر بار بار بجلی چلی جاتی ہے اور انٹرنیٹ بند ہوجاتے ہیں۔

iv. یہ تحقیقی کام اپنی نوعیت کا ایک نادرکام ہے کیونکہ ہمیں جموں وکشمیر کے علاقے کے لئے انتخابی نتائج کی پیشن گوئی سے متعلق ایسا کوئی تحقیقی کام نہیں ملا۔

2.4 خلاصہ

اس باب میں ڈیٹا مائننگ کی تکنیکوں کا استعمال کرتے ہوئے انتخابی پیشن گوئی کے طریقوں کا تفصیلی جائزہ پیش کیا گیا ہے۔زیادہ تر محققین نے مجموعی گھریلو مصنوعات، حکومتوں کا سابقہ میعاد ریکارڈ بے روزگاری کی شرح، تعلیم کی اہلیت اور صدر کی منظوری کی شرح وغیرہ جیسے کچھ پیرامیٹرز کا استعمال کرتے ہوئے انتخابی پیشن گوئی کے ماڈل تیار کیے۔تاہم ، ان میں مرکزی اثر و رسوخ ، ذات پات اور حاس علاقوں وغیرہ جیسے کچھ اہم پیرامیٹرز غائب ہیں جو ترقی پذیر ممالک میں انتخابی عمل کے دوران اہم کردار ادا کرتے ہیں۔نظر ثانی شدہ تحقیقی کام کی بنیادی حد یہ ہے کہ زیادہ تر محققین اپنی رائے شماری کی پیشن گوئی کے ماڈلز میں صرف ایک الگورتھم جیسے نیورل نیٹ ورک وغیرہ ہی استعمال کرتے ہیں۔یا دو الگورتھم کا مجموعہ جیسے سپورٹ ویکٹر مشین اور آرٹیفیشیل نیورل نیٹ ورک وغیرہ کا استعمال کرتے ہیں۔سیاسی پیشن گوئی کے لیے مختلف محققین کے ذریعہ اختیار کردہ پیرامیٹرز اور تکنیک اتنی مضبوط نہیں ہیں کہ ترقی پذیر ممالک میں درست پیشن گوئیاں کی جا سکیں۔ان مسائل پر قابو پانے کے لئے، ہم ایک

انتخابی پیشن گوئی کا ماڈل بناتے ہیں جو کچھ اہم پیرامیٹرز جیسے مرکزی حکومت کے اثر و رسوخ ، حاس علاقے اور موروثی عنصر وغیرہ پر مبنی ہےایڈوانس کمپیوٹر تکنیک جیسے سپورٹ ویکٹر مشین،ڈسیزن ٹری، رینڈم فارسٹ،اور کے نیرسٹ نائیبر الگورتھم وغیرہ کااستعمال کرتے ہیں اور آخر میں سافٹ ووٹنگ کا استعمال کرتے ہوئے ان کو اکٹھا کردیتے ہیں۔

باب 3

3 .انتخابی پیشن گوئی کے لیے تحقیقی ٹول اور تکنیک

3.1 تعارف

انتخابی ڈیٹا بیس میں زیادہ تر (ڈیٹا) خام مواد، غیر عمل شدہ، نامکمل اور پر شور ہے اور اس خام اعداد و شمار کو مفید معلومات ڈیٹا میں تبدیل کرنے کے لئے ڈیٹا مائننگ کے ٹول اور تکنیک استعمال کی جاتی ہیں۔

اس باب میں ہم خام انتخابی اعداد و شمار کو ایک مناسب شکل میں تبدیل کرتے ہیں اور انتخابی نتائج کی پیشن گوئی کے لئے ڈیٹا مائننگ الگورتھم کو کامیابی کے ساتھ استعمال کرتے ہیں۔اس تحقیق سے انتخابی نتائج کی جلد پیشن گوئی کے لئے انتخابی اوصاف کے ایک اہم ذیلی حصہ کی شناخت ہوتی ہے۔ ہم سیاسی پیشن گوئی میں عمدہ خصوصیات حاصل کرنے کے لئے فیچر سلیکشن تکنیک کا استعمال کرتے ہیں۔ اس باب میں ہم ماڈل ویلیڈیشن تکنیک پر بھی تبادلہ خیال کریں گے اور آخر میں باب کا خلاصہ اور نتائج پر یہ باب اختتام پذیر ہو گا۔

3.2-ڈیٹا مائننگ ٹولس

متعدد ڈیٹا مائننگ ٹولز ہیں جن کو ماڈل کی تیاری میں استعمال کیا جا سکتا ہے، جنہیں محققین انتخابی نتائج کی پیشن گوئی میں استعمال کرتے ہیں۔

3.2.1- ویکا (WEKA)

ویکا اسے وائیکاٹو ماحولیات کے نام سے بھی جانا جاتا ہے جو کہ ایک مشین لرننگ سافٹ ویئر ہے جسے نیوزی لینڈ کی یونیورسٹی آف وائیکٹو میں تیار کیا گیا ہے۔[108] یہ ڈیٹا (مواد) کے تجزیہ اور پیشن گوئی ماڈلنگ کے لئے بہترین اور نہایت ہی موزوں ہے، جو کہ الگورتھم اور ویژولائزیشن ٹولز (Visualization Tools) پر مشتمل ہوتا ہے جو مشین لرننگ میں معاون ثابت ہوتا ہے۔ ویکا ایک جی یو آئی (GUI) ہوتا ہے جو اس کی تمام خصوصیات تک آسان رسائی فراہم کرتا ہے جو جاوا پروگرامنگ زبان میں لکھا گیا ہے، ویکا ڈیٹا مائننگ کے بڑے کاموں ، پروسیسنگ ، ویژنلائزیشن اور ریگریشن کی حمایت کرتا ہے اور اس مفروضے پر کام کرتا ہے جو فلیٹ فائل (Flat File) کی شکل میں دستیاب ہو، ویکا ایس کیو ایل

(SQL)ڈیٹا بیس تک ڈیٹا بیس کنیکٹیویٹی کے ذریعہ رسائی مہیا کر سکتا ہے اور استفسار کے ذریعہ موصولہ ڈیٹا، نتائج کی مزید کارروائی کر سکتا ہے[109]۔

3.2.2-ریپڈ مائنر (Rapid Minor)

یہ ایک ڈیٹا مائننگ پیشن گوئی تجزیہ کا نظام ہے جو ریپڈ مائنر کے ذریعہ تیار کیا گیا ہے۔[110] یہ جاوا پروگرامنگ زبان میں لکھا گیا ہے جو ڈیپ لرننگ، ٹکسٹ مائننگ، مشین لرننگ اور پیشن گوئی تجزیہ کے لئے ایک مربوط ماحول فراہم کرتا ہے۔[111] اس آلے کو کاروباری ایپلی کیشنز، تجارتی ایپلی کیشنز، ٹریننگ، تعلیم، تحقیق، ایپلی کیشن ڈویلپمنٹ، اور مشین لرننگ جیسے وسیع الاطلاق کے لئے استعمال کیا جا سکتا ہے، ریپڈ مائنر سرور کو پبلک، پرائیویٹ، پرائیویٹ کلاؤڈ انفراسٹرکچر دونوں کے طور پر پیش کرتا ہے اور اس کی بنیاد کے طور پر ایک کلائنٹ یا سرور ماڈل ہے، ریپڈ مائنر ٹیمپلیٹ پر مبنی فریم ورک کے ساتھ آتا ہے جو کم تعداد میں غلطیوں (جس کی دستی کوڈ تحریری عمل میں عموما expected توقع کی جاتی ہے) کے ساتھ تیز تر فراہمی کے قابل ہو جاتی ہے[112]-

3.2.3-آرینج (Orange)

یہ مشین لرننگ اور ڈیٹا مائننگ کے لئے ایک بہترین سافٹ ویئر سوٹ ہے، اس سے اعداد و شمار کو دیکھنے میں مدد ملتی ہے اور یہ جزو پر مبنی سافٹ ویئر ہے[113]۔ یہ پائیتھان کمپیوٹنگ زبان میں لکھا گیا ہے، چونکہ یہ جزو پر مبنی سافٹ ویئر ہے اور آرینج کے اجزاء کو وجیٹس کہتے ہیں، یہ ویجٹ کی حد میں ڈیٹا ویژولائزیشن اور پری پروسیسنگ سے لے کر الگورتھم اور پیشن گوئی ماڈلنگ کی تشخیص و جانچ آتی ہے، آرینج میں آنے والا ڈیٹا تیزی سے مطلوبہ پیٹرن فارمیٹ میں آجاتا ہے، اور بہ آسانی ویجٹ کو حرکت دینے سے منتقل بھی ہو جاتا ہے، آرینج صارفین کو ڈیٹا کا تیزی سے موازنہ اور تجزیہ کر کے مختصر وقت میں بہتر فیصلے کرنے کی سہولت فراہم کرتا ہے[114]۔

3.2.4-میٹلیب (MATLAB)

میٹلیب (میٹرکس لیبارٹری) میٹرکس سافٹ ویئر آسان رسائی فراہم کرتا ہے [115]۔ میٹلیب تکنیکی کمپیوٹنگ کے لئے ایک اعلی کارکردگی کی زبان ہے جو حساب، ویژولائیزیشن اور پروگرامنگ ماحول کو مکمل کرتا ہے، میٹلیب ایک جدید پروگرامنگ لینگویج ماحول ہے جس میں اعداد و شمار کے جدید ڈھانچے ہیں،

اس میں بلٹ ان ایڈیٹنگ اور ڈیبگنگ ٹولز شامل ہیں، اور آبجیکٹ پر مبنی پروگرامنگ کی حمایت بھی کرتا ہے، یہ عوامل میٹلیب کو تدریس اور تحقیق کا ایک بہترین ذریعہ بناتے ہیں[116]-[117]۔مذکورہ بالا تمام ٹولز کی کچھ حدیں ہیں جو مندرجہ ذیل جدول 3.2 میں درج ہیں۔

جدول 3.2: ذکر کردہ ٹولز کی حدیں

ٹولز	حدود
ویکا	اس میں مناسب دستاویز کا فقدان ہے اور "کچن سنک سنڈروم" کا شکار ہے جہاں سسٹم کو مسلسل اپ ڈیٹ کیا جاتا ہے۔ ایکسل اسپریڈ شیٹ اور جاوا پر مبنی غیر ڈیٹا بیس سے زیادہ مربوط ہے۔
ریپڈ مائنر	ویڈیو، تصاویر اور آڈیو جیسے ملٹی میڈیا استعمال کرتے ہوئے یہ ٹول کم ماحول دوست ثابت ہوا ہے۔ ریپڈ مائنر نے پروسیسنگ میں بہت زیادہ وقت صرف چھوٹے ڈیٹا سیٹس پر بھی لیا ہے، خاص طور پر جب صارف نتائج کی بنیاد پر دستی طور پر مختلف صفات کو بہتر بنا رہا ہو۔

آرینج	مشین لرننگ الگورتھم کی محدود فہرست، مشین لرننگ مختلف لائبریریو کے مابین یکساں طور پر نہیں سنبھالی جاتی، کلاسیکی شماریات میں آر کمزور ہے، اگرچہ یہ اعداد و شمار کی بنیادی شماریاتی خصوصیات کی گ کر سکتا ہے، لیکن یہ اعداد و شمار کی جانچ کے لئے کوئی وجیٹ فراہم نہیں کرتا ہے
میٹلیب	چونکہ یہ ایک ترجمانی و تفسیری زبان ہے جس کی وجہ سے اس کی عمل آوری رفتار بڑی سست ہے اور یہ asynchronous events کو سپورٹ نہیں کرتا۔

جیسا کہ جدول 3.2 میں دکھایا گیا ہے، مذکورہ بالا ٹولز کی حدیں اور نقائص کو دور کرنے کے لئے ہم اس تحقیقی کام میں پائتھان زبان میں اینا کونڈا ٹول (پیکیج) اور جوپیٹر نوٹ بک کا استعمال کریں گے الیکشن پیشن گوئی ماڈل تیار کرنے کے لئے۔

3.2.5-اینا کونڈا (ANACONDA)

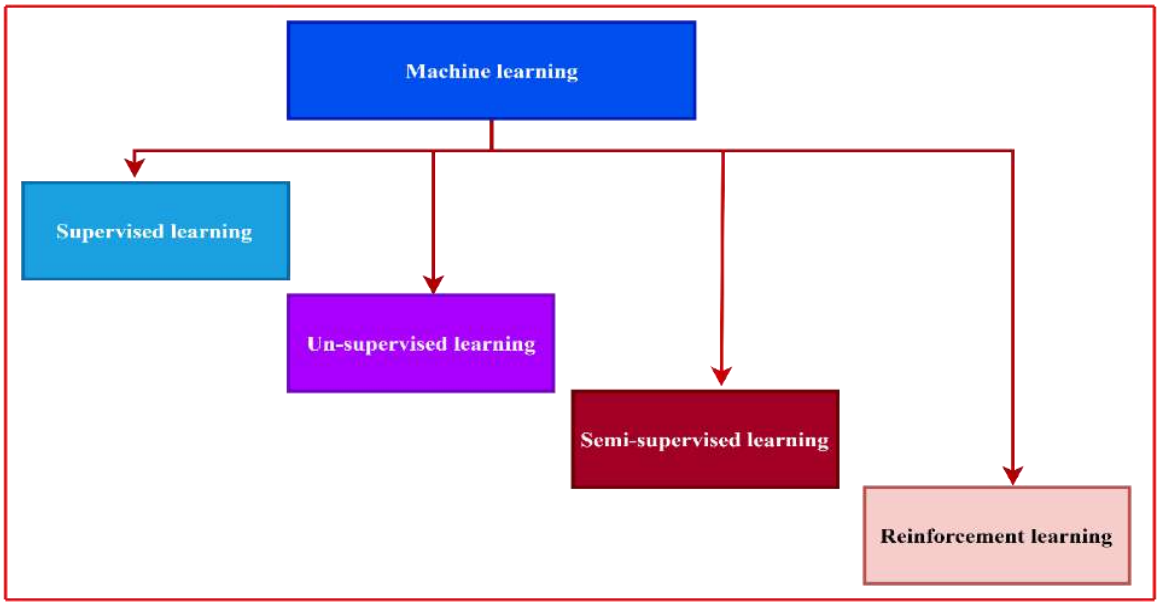
اس تحقیقی کام میں ہم اینا کونڈا کا استعمال کرتے ہیں جو ان بلٹ ڈیٹا پروسینگ اور دستکاری کی سہولیات کے ساتھ ایک اوپن سورس ٹول ہے۔اینا کونڈا بنیادی طور پر ایک ڈیٹا سائنس اسٹیک ہے جو 1000 سے زیادہ طاقتور لائبریریوں پر مشتمل ہے جس کی بنیاد پائتھان اور دیگر پروگرامنگ زبانوں پر مبنی ہے [118]۔ہم 'کونڈا' جو ایک پیکیج مینیجر ہے کا استعمال کر رہے ہیں، جو پائتھان زبان کے سیکڑوں پیکج پر مشتمل ہے اور ان پیکیج کو ڈیٹا پری پراسیسنگ، درجہ بندی اور توثیق کے لئے ہمارے تحقیقی کام میں استعمال کیا گیا ہے۔ انبیلٹ مشین لرننگ الگورتھم کی وجہ سے ، اینا کونڈا آسانی سے انتظام کرنے والا ماحولیاتی ترتیب دینے میں ہماری مدد کرتا ہے جسے ہم بس ایک کلک کے ذریعہ اپنے انتخابی پیشن گوئی کے ماڈل میں تعینات کر سکتے ہیں[119]۔

انتخابی پیشن گوئی ماڈل تیار کرنے کے لئے ہم جوپیٹر نوٹ بک ویب ایپلی کیشن کا استعمال کرتے ہیں،انتخاب کی پیشن گوئی کے ماڈل کا سامنے والا حصہ فلاسک کا استعمال کرتے ہوئے تیار کیا گیا ہے [121] ,[120]

۔تیار کردہ ماڈل ہیرو کو پر اپ لوڈ کیا گیا ہے جو کنٹینرز پر مبنی کلاؤڈ پلیٹ فارم کے طور پر بطور سروس (PaaS) ہے[122]۔

3.3 مشین لرننگ تکنیکس

مشین لرننگ ڈیٹا مائننگ کی ایک ایسی فیلڈ ہے جو کمپیوٹر کو واضح طور پر تربیت دیئے بغیر پچھلے ڈیٹا یا معلومات سے خود بخود سیکھنے کی صلاحیت فراہم کرتی ہے[123]۔مشین لرننگ کا بنیادی مقصد ایک ایسے کمپیوٹر پروگرام کی تشکیل کرنا ہے جو اعداد و شمار تک رسائی حاصل کر سکے اور خود سیکھ سکے،اس کے بعد یہ مستقبل کے نتائج کی پیشن گوئی کے لئے یہ ڈیٹا (لیبل لگایا ہوا/ غیر لیبل لگایا ہوا) کا استعمال کرتا ہے [124]۔مشین لرننگ تکنیک موجودہ دور کی سب سے زیادہ وسیع پیمانے پر استعمال کی جانے والی تکنیک ہے اور یہ اندازہ لگایا جاتا ہے کہ یہ مستقبل کی پیشن گوئی کے لئے مفید ٹول ہیں، مشین لرننگ تکنیک کی زیادہ تر چار مختلف اقسام میں درجہ بندی کی گئی ہے جیسا کہ ترسیم 3.3 میں دیے گئے اعداد و شمار میں دکھایا گیا ہے۔



ترسیم 3.3: مشین لرننگ اور اس کی اقسام [128]

Supervised Learning-1 سوپروائزڈ لرننگ: یہ ایک قسم ہے جس میں مشین لیبل والے ڈیٹا سے سیکھ سکتی ہے جس کا مطلب ہے کہ کچھ اعداد و شمار کو پہلے ہی صحیح جواب کے ساتھ ٹیگ کیا گیا ہے[125]۔ سوپر وائزڈ مشین لرننگ کی تکنیکوں کی دو قسموں میں درجہ بندی کی گئی ہے یعنی کلاسیفیکیشن اور ریگریشن۔

Unsupervised Learning- 2 ان سوپر وائزڈ لرننگ: یہ ایک قسم ہے جس میں مشینوں کو ایسی معلومات کی تربیت دی جارہی ہے جس پر نہ تو لیبل لگایا گیا ہے اور نہ ہی اس کی درجہ بندی کی گئی ہے [126]۔ یہاں مشین کا کام ڈیٹا کی کسی بھی پیشگی تربیت کے مماثلت یا تفاوت کے پیمانہ پر مبنی لیبل نہ

لگائے ہوئے ڈیٹا کو گروپس یا پیٹرن میں درجہ بندی کرنا ہے، ان سوپر وائزڈ مشین لرننگ کی تکنیک دو طرح کی ہوتی ہے یعنی کلسٹرنگ اور ایسوسی ایشن۔

3. Semi-supervised learning **سیمی سوپر وائزڈ لرننگ** اَن سوپر وائزڈ مشین لرننگ اور سوپر وائزڈ مشین لرننگ کے مابین سیمی سوپر وائزڈ مشین لرننگ ہے، اس قسم کی لرننگ میں مشینوں کو لیبل والے اور غیر لیبل والے دونوں ڈیٹا پر تربیت دی جاتی ہے، مقصد یہ ہے کہ کس طرح لیبل لگا اور غیر لیبل لگے ہوئے دونوں ڈیٹا کا انتخاب کمپیوٹر کے سیکھنے کے طرز عمل کو تبدیل کر سکتا ہے اور اس طرح کے امتزاج کے لئے الگورتھم ڈیزائن کر سکتا ہے، لہذا سیمی سوپر وائزڈ مشین لرننگ میں الگورتھم کو لیبل لگانے کے ساتھ ساتھ اور غیر لیبل لگائے ہوئے دونوں ڈیٹا پر بھی تربیت دی جاتی ہے، عام طور پر اس طرح کے امتزاج میں غیر لیبل شدہ ڈیٹا کی ایک بڑی مقدار ہوتی ہے جس میں تھوڑی مقدار میں لیبل لگا ہوا ڈیٹا ہوتا ہے، مثال کے طور پر تقریر کا تجزیہ اور انٹرنیٹ مواد کی درجہ بندی وغیرہ[127]۔

4.Reinforcement learning **ری انفورسمنٹ لرننگ**: اس کو reward-based learning کے نام

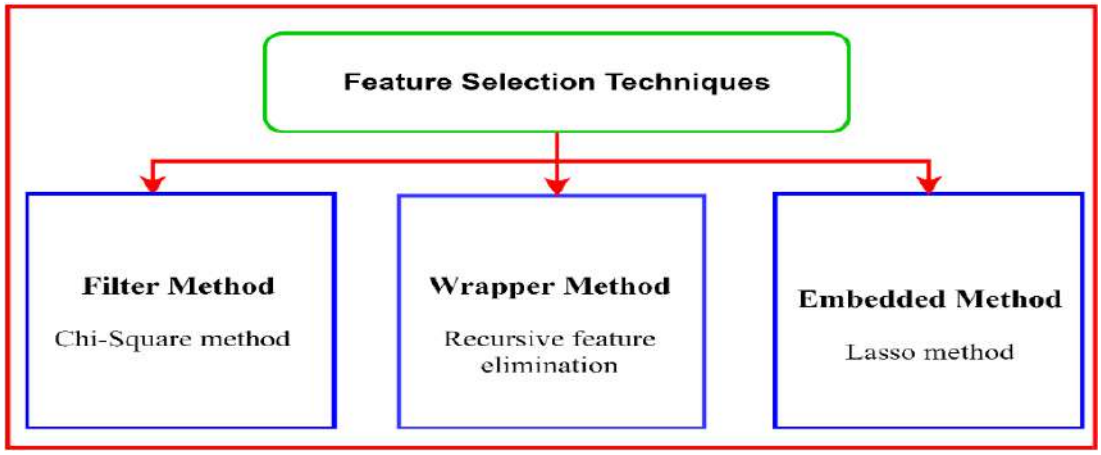
سے بھی جانا جاتا ہے، ری انفورسمنٹ لرننگ میں ایجنٹ ماحول کے ساتھ بات چیت کرتے ہوئے سیکھتے ہیں[126]۔ایجنٹ کو ہر اچھے اقدام پر انعامات اور ہر غلط اقدام پر جرمانے عائد کیے جاتے ہیں، ری انفورسمنٹ لرننگ دوسرے لرننگ طریقوں کے برعکس کام کرتا ہے کہ جنہیں یہ نہیں بتایا گیا ہے کہ کسی بھی پریشانی پر کس طرح کام کیا جائے، لیکن یہ مسئلہ پر کام کرتا ہے اور اپنا حل سامنے پیش کرتا ہے۔ [129] مثال کے طور پر، خود چلانے والی کار اور شطرنج کا کھیل وغیرہ۔اس تحقیقی کام میں ہم نے انتخابی نتائج کی جلد پیشن گوئی کے لئے سوپر وائزڈ کلاسیفکیشن مشین لرننگ کی تکنیک استعمال کی ہے۔

3.4 خصوصیت کے انتخاب کی تکنیکس Feature Selection Techniques

فیچر سلیکشن میتھڈ ان خصوصیات کو منتخب کرنے کے لئے ایک ایسی تکنیک ہے جو ماڈل کی تعمیر کے لئے انتہائی موزوں ہے،۔فیچر سلیکشن مخصوص متعلقہ تشخیصی معیار کے مطابق اصل سے متعلقہ خصوصیات کے غیر معمولی سبسیٹ کا انتخاب کرتا ہے، جو عام طور پر بہتر سیکھنے کی کارکردگی (جیسے درجہ بندی کے لئے اعلی سیکھنے کی درستگی)، کم کمپیوٹیشنل لاگت، اور بہتر ماڈل تشریح کی طرف جاتا ہے۔ فیچر سلیکشن

تکنیکوں کی نگرانی (سوپروائزڈ)، غیر نگرانی (ان سوپروائزڈ) اور نیم نگرانی (سیمی سوپروائزڈ) والے فیچر کے انتخاب میں درجہ بندی کی جا سکتی ہے۔

زیر نگرانی (سوپروائزڈ) فیچر سلیکشن میتھڈ کی مزید درجہ بندی انٹرفیلٹر میتھڈ، ریپر میتھڈ اور ایمبیڈڈ میتھڈ کے مطابق درج ذیل ترسیم 3.4 میں دکھایا گیا ہے۔

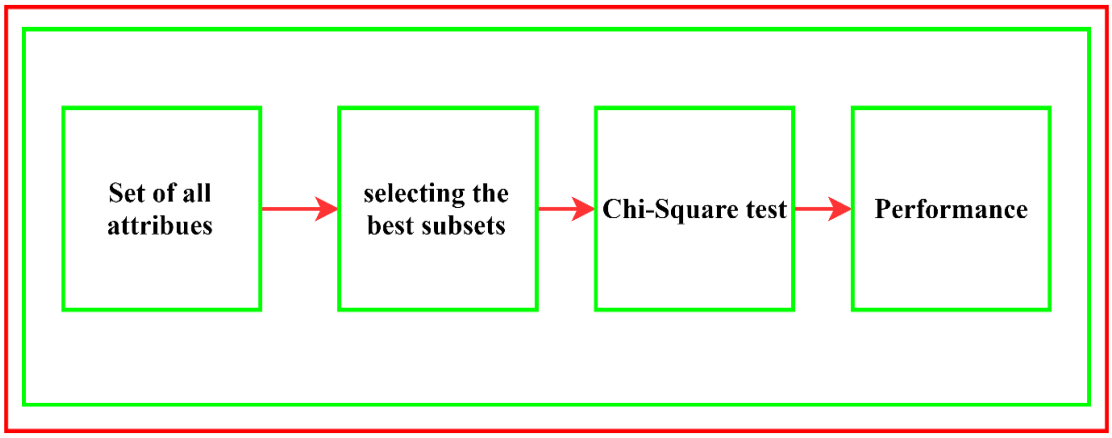


ترسیم 3.4: فیچر سلیکشن تکنیک کی درجہ بندی

یہ تکنیک غیر منسلک صفات کو کم کرتی ہیں جو الگورتھم کے running time کو کم کرتی ہیں۔ اس تحقیقی کام میں ہم نے فلٹر میتھڈ، ریپر میتھڈ اور ایمبیڈڈ میتھڈ جیسے تین سب سے زیادہ اثر انگیز فیچر سلیکشن کے طریقوں کا اطلاق کیا ہے۔

3.4.1 فلٹر میتھڈ (Filter Method)

فلٹر میتھڈ کچھ اعداد و شمار کی جانچ کی بنیاد پر ہر ایک وصف کو کچھ اسکور تفویض کرتے ہیں، فیچرس کو ڈیٹاسیٹ میں حاصل کردہ اسکور کے مطابق درجہ بندی کی جاتی ہے اور حاصل کردہ اسکور کی بنیاد پر اوصاف کو ہٹایا یا رکھا جاتا ہے[130]-



ترسیم 3.4.1 فلٹر میتھڈ برائے انتخاب پیرامیٹر سبسیٹ سلیکشن [133]

فلٹر میتھڈ دو مراحل پر مشتمل ہوتا ہے، پہلے تو یہ بعض یقینی معیار پر مبنی خصوصیات کی درجہ بندی کرتا ہے اور پھر اعلی درجہ بندی والی خصوصیات کو درجہ بندی کے ماڈل میں راغب کرنے کے لئے منتخب کرتا ہے، [131]۔ اس تحقیقی کام میں ہم ہر انتخابی خصوصیت متغیر اور ہدفی متغیر کے

درمیان Chi-square statistics کا حساب لگاتے ہیں اور متغیر اور ہدف کے مابین تعلقات کے وجود کا تجزیہ کرتے ہیں[132] ۔ ترسیم 3.3 انتخابات کی پیشن گوئی کے لئے فلٹر کے طریقہ کار کی کاردگی کو بتاتا ہے۔

3.4.2 ریپر میتھڈ (wrapper Method)

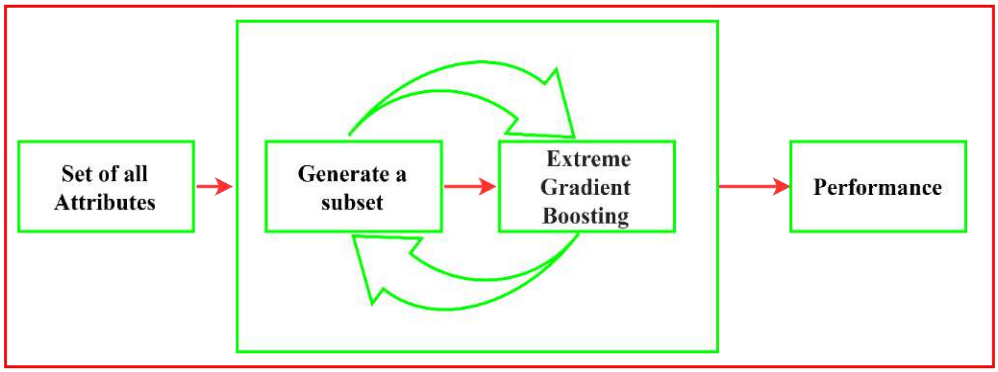
ریپر میتھڈس گریڈی سرچ الگورتھم پر مبنی ہیں جو فیچر کے تمام امتزاجوں کا جائزہ لیتے ہیں-اور اس مرکب کا انتخاب کرتے ہیں جو مشین لرننگ الگورتھم کے لئے بہترین نتیجہ پیدا کرتا ہے، ریپر میتھڈ منتخب شدہ خصوصیات کے معیار کا اندازہ کرنے کے لئے ایک مخصوص درجہ بندی کا اطلاق کرتا ہے اور منتخب کردہ مشین لرننگ کلاسیفائیر سے قطع نظر، فیچر سلیکشن کے مسئلے کو حل کرنے کا ایک آسان اور طاقتور طریقہ پیش کرتا ہے[134]۔ایک وضاحتی درجہ بندی کے پیش نظر، ایک عام ریپر طریقہ مندرجہ ذیل اقدامات انجام دے گا:

a) خصوصیات کے سبسیٹ کی تلاش۔

b) کلاسیفائیر کی کارکردگی کی بنیاد پر خصوصیات کے منتخب کردہ سبسیٹ کا جائزہ لینا۔

c) مطلوبہ معیار تک پہنچنے تک مرحلہ 1 اور مرحلہ 2 کو دہرانا۔

اس تحقیقی کام میں ہم Recursive feature elimination (انتہائی تدریجی فروغ دینے) کے طریقہ کار کا اطلاق کرتے ہیں جو کسی ماڈل میں فٹ بیٹھتا ہے اور خصوصیات کی ایک مخصوص تعداد تک پہنچنے تک کمزور ترین خصوصیت (یا خصوصیات) کو ہٹاتا رہتا ہے [135]۔ انتخابی پیرامیٹرز کی موزوں تعداد معلوم کرنے کے لئے، آر ایف ای کے ساتھ مختلف خصوصیت کے سبسٹس اسکور کرنے اور خصوصیات کا بہترین اسکورنگ مجموعہ منتخب کرنے کے لئے کراس –ویلیڈیشن کا استعمال کیا جاتا ہے [136]- [137]۔ ترسیم 3.4.2 انتخابات کی پیشن گوئی کے لئے ریپر میتھڈ کے طریقہ کار کی کارکردگی کو بتاتا ہے۔



ترسیم 3.4.2 الیکشن پیرامیٹر کے انتخاب کے لئے ریپر میتھڈ [133]

3.4.3 ایمبیڈڈ میتھڈ (Embedded Method)

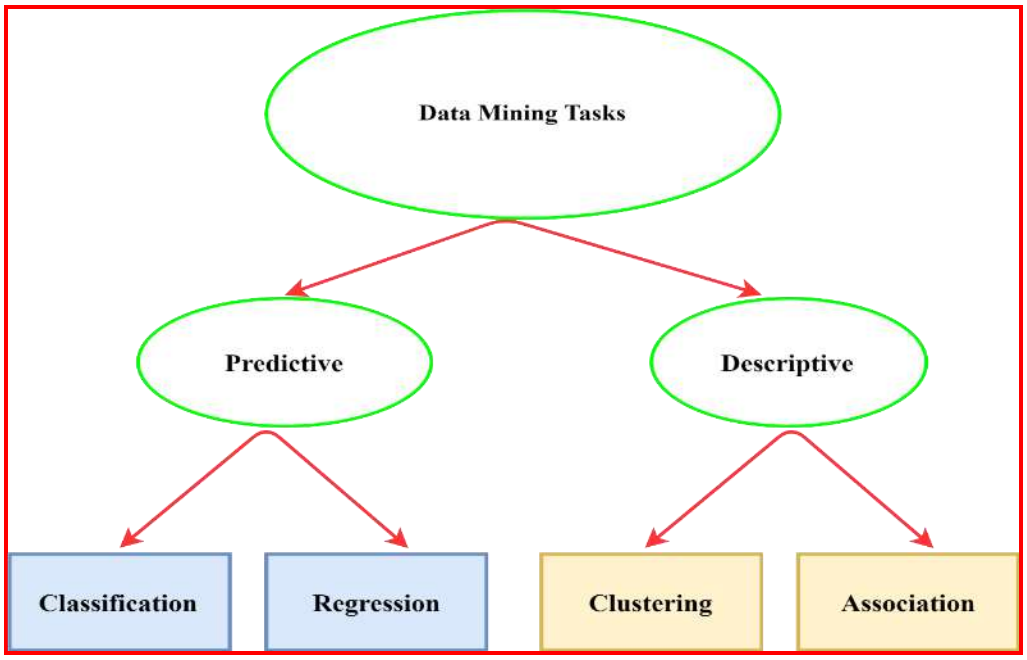
ایمبیڈڈ طریقہ (یا ہائبرڈ طریقہ) فلٹر کے طریقہ کار اور ریپر طریقہ کا ایک مجموعہ ہے جس میں خصوصیت کے انتخاب کا طریقہ سیکھنے الگورتھم کو سیکھنے میں شامل کیا جاتا ہے اور پھر اسے بہتر بنایا جاتا ہے۔ ایمبیڈڈ میتھڈ مختلف (سبسیٹ) ذیلی ذخیروں کی دوبارہ جانچ پڑتال کیلئے لیے ہوئے گنتی کو کم کرتا ہے[133]۔

اس تحقیقی کام میں ہم نے پیشن گوئی کی غلطی کو کم سے کم کرنے والی صفات میں سے سبسیٹ کو منتخب کرنے کے لئے Lasso method کو لاگو کیا[138]۔ لاسو ماڈل پیرامیٹر پر ایک رکاوٹ ڈال کر اس طرح کام کرتا ہے کہ جس کی وجہ سے کچھ متغیرات کے لئے ریگریشن کوایفیشینس صفر کی طرف سکڑ جاتے

ہیں، متغیرات سکڑنے کے عمل کے بعد صفر کے برابر ریگریشن کے مرجع کے ساتھ ماڈل سے خارج کیا جاتے ہیں، غیر صفری متغیرات ریگریشن ضوابط کے ساتھ متغیرات کو ماڈل میں شامل کیا جاتا ہے یا پھر رکھ دیا جاتا ہے[139]۔

3.5 ڈیٹا مائننگ ٹاسکس:

ڈیٹا مائننگ کا بنیادی مقصد ڈیٹا سے سیکھنا ہے، ڈیٹا مائننگ ٹاسکس ڈیٹا مائننگ کے عمل میں پائے جانے والے نمونوں کی وضاحت کرنے کے لئے استعمال کیے جاتے ہے، ڈیٹا مائننگ ٹاسکس کو عام طور پر دو بڑی اقسام میں تقسیم کیا جاتا ہے: پیش گوئی کرنے والے کام اور وضاحتی ٹاسک، جیسا کہ ذیل میں دیئے گئے ترسیم 3.5 میں دکھایا گیا ہے۔ پیش گوئی کرنے والے کاموں میں، انحصار شدہ خصوصیت کی قیمت کی کھوج کی خصوصیات کی بنیاد پر پیش گوئی کی جاتی ہے، تاہم وضاحتی کاموں میں ناول کے نمونے اخذ کیے گئے ہیں جو اعداد و شمار میں مرکزی انجمن کی وضاحت کرتے ہیں۔ وضاحتی ڈیٹا مائننگ کے کام اکثر فطرت میں تلاش کرتے ہیں اور نتائج کی توثیق اور وضاحت کے لئے اکثر پروسیسنگ کے بعد کی تکنیک کی ضرورت ہوتی ہے۔



ترسیم:3.5ڈیٹامائننگ ٹاسکس کی درجہ بندی [140]

3.5.1 پیشن گوئی کرنے والاڈیٹامائننگ ٹاسک

پیشن گوئی ماڈلنگ سے مراد وضاحتی متغیر کے طور پر ہدفی متغیر کے لئے ماڈل بنانے کا کام ہے،پیشن گوئی کرنے والی ماڈلنگ کی دو قسمیں ہیں: کلاسیفیکیشن اور ریگریشن/پیشن گوئی۔

I. **درجہ بندی:** کلاسی فیکیشن (درجہ بندی)ڈیٹامائننگ کی ایپلی کیشنز پیشن گوئی کرنے والی (سوپروائزڈ) سیکھنے والے ڈیٹامائننگ کے کام ہیں جو ہدف کی خصوصیت کی متضاد اقدار کی پیشن گوئی کرتی ہیں،اگر ہدف کی خصوصیت والی اقدار دو اقدار ہیں (جیسے ہاں اور نہیں)، تو اسے بائنری درجہ بندی کہا جاتا ہے، لیکن اگر

ہدفی وصف میں متعدد ممکنہ اقدار ہیں تو پھر اسے کثیر طبقاتی درجہ بندی(multi –class classification) کہا جاتا ہے[140] ۔

II. **ریگریشن:** ریگریشن ماڈلنگ سے مراد وضاحتی متغیر کے طور پر مستقل ہدفی متغیر کے لئے ماڈل بنانے کا کام ہے، مثال کے طور پر، اسٹاک کی قیمت کی آئندہ تین ماہ میں پیشن گوئی کرنا، کیونکہ قیمت ایک مستقل قدر والا وصف ہے دونوں کاموں کا ہدف ایک ایسا ماڈل سیکھنا ہے جو ہدفی متغیر کی پیشن گوئی کرتا اور صحیح قدروں کے مابین غلطی کو کم کرتا ہو[141] ۔

درجہ بندی اور رجعت کا مقصد ایک ایسا نظام ہونا ہے جو مطلوبہ متغیر کی پیش گوئی اور صحیح قدروں کے درمیان غلطی کو کم کردے۔

3.5.2 وضاحتی ڈیٹا مائننگ ٹاسکس

وضاحتی ماڈلنگ سے مراد ان نمونوں کو اخذ کرنے کا کام ہے جو اعداد و شمار میں بنیادی رشتوں کا خلاصہ کرتے ہیں، اس ماڈلنگ کے دو طرح کے کام ہیں: کلسٹرنگ اور ایسوسی ایشن۔

i. **کلسٹرنگ:** کلسٹرنگ ڈیٹا مائننگ ایپلی کیشنز وضاحتی (غیر سپروائزڈ) ڈیٹا مائننگ ٹاسکس ہیں جو قریبی متعلقہ مشاہدات کے گروہوں کو تلاش کرنے کی کوشش کرتے ہیں تاکہ وہ مشاہدات جو ایک دوسرے سے مشابہت رکھتے ہیں ان مشاہدات کے مقابلے میں ایک دوسرے سے ملتے جلتے ہیں جو مشاہدات دیگر مشاہدات سے متعلق ہیں[140] ۔

ii. **ایسوسی ایشن:** ایسوسی ایشن انالسس، ان نمونوں کو تلاش کرنے کے لئے استعمال کیا جاتا ہے جو ڈیٹا میں مضبوطی سے وابستہ خصوصیات کو بیان کرتے ہیں، دریافت کردہ نمونوں کو عام طور پر مضمر قواعد یا خصوصیات سبسٹس کی شکل میں پیش کیا جاتا ہے[140] ۔

3.6 ڈیٹا مائننگ کی تکنیکیں:

انتخابات کی پیشن گوئی نہ صرف اس کی بے حد پیچیدگیوں کی وجہ سے ایک مشکل چیلنج ہے بلکہ ووٹنگ کی راتوں سے پہلے ہی معاملات تبدیل ہوجاتے ہیں، انتخابی نتائج کی پیشن گوئی کرنے کے متعدد طریقے ہیں جیسے سوشل میڈیا کا استعمال کرتے ہوئے انتخابی نتائج کی پیشن گوئی کرنا یا بے ترتیب نمونے لینے کے ساتھ انتخابی پیشن گوئی کا اندازہ لگانے کے طریقے وغیرہ۔ تاہم انتخابی پیشن گوئی کے تمام موجودہ طریقوں کے

ساتھ کچھ حدود ہیں جیسے سوشل میڈیاان لوگوں کے خیالات کو خارج کرتا ہے جو سوشل سائٹس کا استعمال نہیں کررہے ہیں اور نمونے لینے کا طریقہ متعصبانہ نتائج پیش کرتا ہے کیونکہ اس میں سروے کاصرف ایک چھوٹاسا حصہ شامل ہوتاہےاس طرح کی تحقیق کی حدود پر قابو پانے کے لئے ہم نے انتخابی پیشن گوئی کے ماڈل کی تشکیل کے لئے ڈیٹامائننگ کی تکنیک کااستعمال کیا ہے جو خالصتا پیرامیٹرز پر مبنی ہے اور ان پیرامیٹرز کا انتخاب ڈومین ماہرین سے مشاورت کے بعد کیا گیا ہے۔انتخابی ڈیٹاسیٹ سے معلومات حاصل کرنے کے لئے ہم نے ڈیٹامائننگ کی مختلف تکنیکوں کااستعمال کیاہے، سیاسی پیشن گوئی میں ڈیٹامائننگ کی تکنیک کو شامل کرنے کامقصد سروے یا ایگزٹ پول کو تبدیل کرنا نہیں بلکہ ایک ایسا متبادل طریقہ فراہم کرنا ہے جہاں انسان جدوجہد کرتے ہیں،اگرچہ ڈیٹامائننگ کی متعدد بنیادی تکنیکیں دستیاب ہیں لیکن اس تحقیق کا محور درجہ بندی کے ڈیٹامائننگ کی تکنیک ہے جو سیاسی پیشن گوئی کرنے والے کو پیرامیٹرز کی شناخت کرنے اور جموں و کشمیر کے انتخابی نتائج کی پیشن گوئی کرنے میں مدد فراہم کرے گی۔

3.6.1 ڈیسیزن ٹری Decision Tree

ڈیسیزن ٹری ایک سوپر وائزڈ مشین الگورتھم ہے جسے ریگریشن اور کلاسی فیکیشن کے مسائل کو حل کرنے کے لئے استعمال کیا جا سکتا ہے، یہ بنیادی طور پر ایک درخت کی طرح ہوتا ہے جس میں ہر جڑ نوڈ ایک خاصیت پر ایک ٹیسٹ کی نشاندہی کرتی ہے، شاخ نوڈ سے مطابقت رکھتی ہے جو ٹیسٹ کے نتائج کی نمائندگی کرتی ہے، شاخ نوڈ مزید لیف نوڈ سے مطابقت رکھتا ہے جو کلاس لیبل رکھتا ہے[142]۔ڈیسیزن ٹری الگورتھم ایک گریڈی (غیر بیک ٹریکنگ) نقطہ نظر کی اتباع کرتا ہے اور اوپر سے نیچے کی تکرار شدہ، تقسیم اور فتح کے انداز میں تشکیل دیا جاتا ہے، الگورتھم ٹپلس اور ان سے وابستہ کلاس لیبل کی تربیت سیٹ کے ساتھ شروع ہوتا ہے۔ درخت کی تعمیر کی وجہ سے تربیت کا سلسلہ چھوٹے چھوٹے ذیلی حصوں میں قسیم کیا جاتا ہے، جب ڈیسیزن ٹری کی تعمیر مکمل ہو جاتی ہے تو بہت ساری شاخیں تربیتی اعداد و شمار میں آوازیں یا شور کرتی ہیں، درخت کی کٹائی اس طرح کی شاخوں سے بے ضابطگیوں کو دور کرنے کے لئے استعمال کی جاتی ہیں جس کا مقصد غیر مرئی اعداد و شمار پر درجہ بندی کی درستگی کو بہتر بنانا ہے[140]۔

مختلف قسم کے ڈیسیزن ٹری دستیاب ہیں ان میں فرق ریاضیاتی ماڈل ہے جو ڈیسیزن ٹری کے فیصلوں

کو نکالنے میں اور الگ الگ وصف کو منتخب کرنے کے لئے استعمال ہوتا ہے[143]۔انتساب کے انتخاب کے مقبول اقدامات انفارمیشن گین، گین ریشو اور جنی انڈیکس ہیں۔[144] انفارمیشن گین (معلومات حاصل کرنے کا کام) اس وقت انجام دیا جاتا ہے جب صفات فطرت میں طبقاتی ہوں جبکہ جنی انڈیکس اس وقت انجام دیا جاتا ہے جب صفات فطرت میں مستقل رہتی ہوں، انفارمیشن گین انتساب کا انتخاب ایک ایسا پیمانہ ہے جو اس بات کا اظہار کرتا ہے کہ کسی خصوصیت کی بہترین پارٹیاں مختلف طبقات یا گروہوں میں داخل ہو جاتی ہیں، انفارمیشن گین نقطہ نظر کا بنیادی مقصد یہ ہے کہ اس سے الگ ہونے والے وصف کو اس طرح سے منتخب کیا جائے کہ وہ انٹراپی کی قدر کو کم سے کم کرے، اس طرح سے معلومات کو زیادہ سے زیادہ حاصل کیا جا سکے، ہر ایک وصف کے لئے حاصل ہونے والے معلومات کا حساب مساوات 3.6.1 کے ذریعہ کیا گیا ہے۔

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

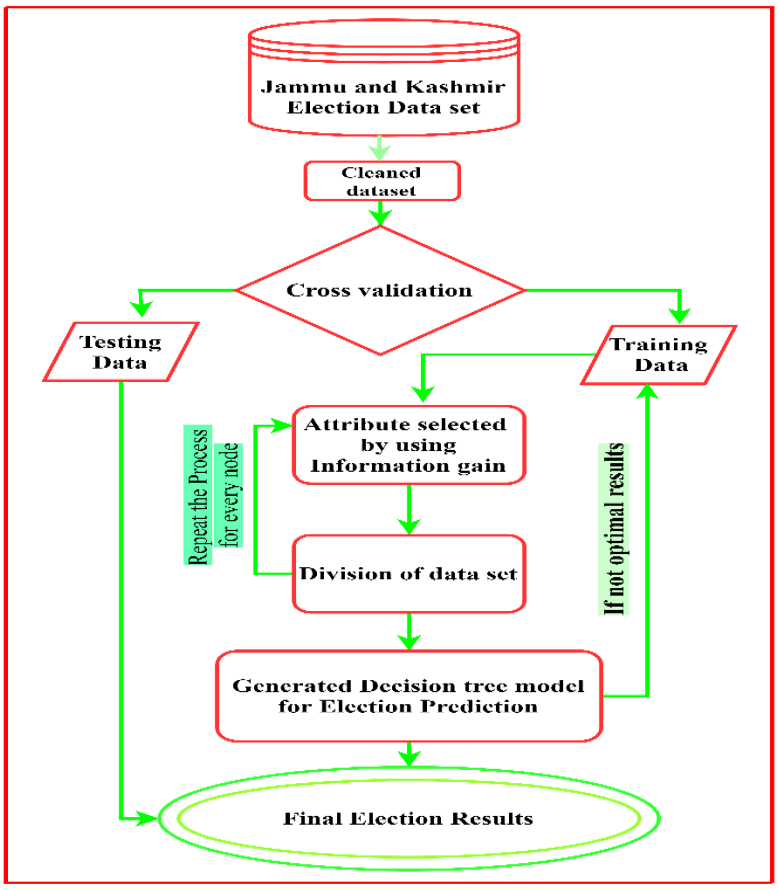
جہاں Info(D) مطلوبہ معلومات کی اوسط رقم ہوتی ہے جس سے D میں ٹیوپل کے کلاس لیبل کی نشاندہی ہوتی ہے اور مساوات 3.2 کا استعمال کر کے اس کا حساب لگایا گیا ہے، $Info_A(D)$ ایک متوقع انفارمیشن ہے جو D کی طرف سے A تقسیم کی بنیاد پر سے D ایک ٹپل کو درجہ بندی کرنے کے لئے مطلوبہ معلومات کی ضرورت ہے اور مساوات 3.3 میں اس کا حساب کیا گیا ہے۔

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

جہاں pi غیر صفر امکان ہے کہ D میں ایک صوابدیدی tuple کلاس Ci سے تعلق رکھتا ہے اور اس کا تخمینہ لگایا جاتا ہے |Ci, D| / |D|. ایک log فنکشن base2 کا استعمال کیا گیا ہے کیونکہ معلومات کو بٹس میں اَن کوڈ کیا گیا ہے۔

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

| Dj | / | D| کی اصطلاح j^{th} پارٹیشن کے وزن کا عمل کرتی ہے۔

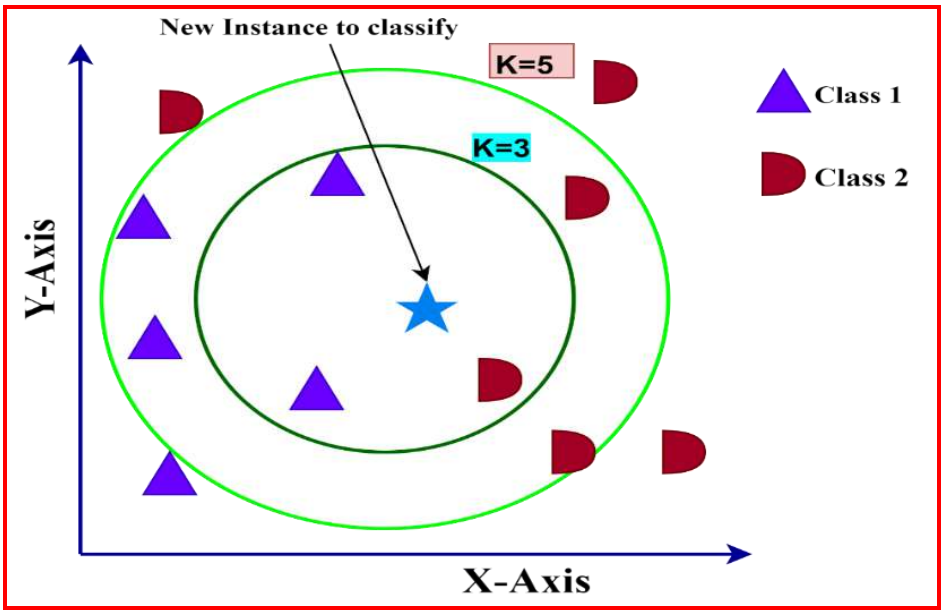


ترسیم 3.6.1: انتخابی پیشن گوئی کے لیے ڈسیزن ٹری ماڈل

ڈسیزن ٹری انتخابی پیشن گوئی ماڈل کا کام ترسیم 3.6.1 میں دکھایا گیا ہے ڈسیزن ٹری کو استعمال کرنے کا مرکزی مقصد ایک انتخابی پیشن گوئی کرنے والا ماڈل بنانا ہے جو تربیت کے اعداد و شمار سے حاصل کردہ بنیادی فیصلے کے قواعد کو سیکھ کر ہدف متغیر یا کسی طبقے کی قیمت کی پیشن گوئی کر سکتا ہے، ڈسیزن ٹری سے حاصل ہونے والے تجرباتی نتائج باب 4 میں بیان کیے گئے ہیں۔

3.6.2 کے-نیریسٹ نائیبر (K-Nearest Neighbors):

کے نیریسٹ نائیبر K-Nearest Neighbors (کے-این این) سب سے آسان الگورتھم ہے جو دستیاب تمام صورتوں کو محفوظ کرتا ہے اور قریبی اکثریت والے پڑوسیوں (مثال کے طور پر فاصلہ افعال) کی بنیاد پر نئی مثالوں کی درجہ بندی کرتا ہے[145]۔ کے این این کو شماریاتی تخمینے، پیٹرن کی پہچان میں اپنایا گیا ہے اور پہلے ہی اسے غیر پیرامیٹرک ٹول کے طور پر سمجھا جاتا تھا[146]- کے این این K-NN الگورتھم 'لیزی لرنر' کے درجہ بندی کے تحت آتا ہے کیونکہ وہ تربیت ڈیٹاسیٹ کو اسٹور کرتا ہے اور ٹیسٹ ڈیٹاسیٹ کے ساتھ فراہم ہونے تک اس کا انتظار کرتا ہے، پھر وہ دستیاب اکثریت والے نائیبر کی بنیاد پر ڈیٹاسیٹ کی درجہ بندی کے لئے مطلوبہ آپریشن انجام دیتا ہے[147] کے-این این نے فرض کیا ہے کہ بند قربت میں ایک جیسی کلاسیں موجود ہیں یعنی ایسے ادارے جو ملتے جلتے ہیں ایک ساتھ موجود ہیں، K-NN الگورتھم کے نام پر حرف تہجی 'K' کا مطلب ہے مثال کے کلاس کا تعین کرنے کے لئے نیریسٹ نائیبر کی تعداد، جیسا کہ ترسیم 3.6.2 میں دکھایا گیا ہے۔



ترسیم 3.6.2: کے نیریسٹ نائیبر (کے این این) کی درجہ بندی کی مثال [149]

K-NN کو instance-based learner بھی کہا جاتا ہے۔ لیزی لرنر یا انسٹینس بیسڈ لرنر ٹریننگ ٹپلس کے ساتھ پیش کیے جانے پر کم کارکردگی کا مظاہرہ کرتے ہیں اور درجہ بندی اور پیشن گوئی کے ساتھ کام کرتے ہوئے اچھی کارکردگی کا مظاہرہ کرتے ہیں، لہذا یہ اس کو بے حد مہنگا بنا دیتا ہے، ایگر لرنر کے برعکس کہ جب تربیت دی جاتی ہے تو اس میں درجہ بندی کرنے کے لئے ٹیسٹ ٹپل وصول کرنے سے قبل ہی درجہ بندی کا ماڈل تیار کرنے کی صلاحیت ہوتی ہے، لہذا K-NN نظر نہ آنے والے ٹپلس کی درجہ بندی کرنے کے لئے تیار اور بے چین رہتا ہے [146]۔ 'K' کی اچھی قیمت کا تعین اس تجرباتی طور پر کیا جاسکتا ہے کہ K کی قیمت 1 مقرر کی جائے اور پھر اضافہ 'K' کو مزید نائیبر کی اجازت دی

جائے، کم سے کم خرابی کی شرح دینے والی 'K' value کا انتخاب کیا گیا ہے، درجہ بندی غلطیوں کی شرح کا اندازہ لگانے کے لئے ٹیسٹ سیٹ کا استعمال کیا گیا ہے جبکہ نئے انسٹینٹ کے کلاس کی وضاحت کرتے ہوئے K-NN نے یوکلائیڈن یا مین ہٹن کے فاصلے کی پیمائش کا استعمال کرتے ہوئے فاصلہ افعال کی پیمائش کی اور پھر قیب ترین نائیبر کی بنیاد پر اس نئی مثال کی درجہ بندی کی ہے [148]۔K-NN کے بہت سے فاصلاتی پیمائشیں ہیں جن کو استعمال کیا جا سکتا ہے، جیسے (یوکلیڈین، مین ہٹن اور منکووسکی) لیکن اس تحقیق میں منکوسکی پیمانے کو انتخابی اعداد و شمار کی خصوصیات کی وجہ سے استعمال کیا گیا ہے۔ دو پوائنٹس (نقطوں) کے درمیان یکلیڈیائی ڈسٹینس وہی ہے جو ان کو جوڑنے والے راستے کی لمبائی ہے، یکلیڈین ڈسٹینس کا حساب کتاب i= (xi1, xi2 ...xip) اور j= (xj1, xj2... xjp) کے درمیان squared differences کے جو square root کے حساب سے کیا گیا ہے جیسا کہ ذیل میں دی گئی مساوات 3.4 میں دکھایا گیا ہے۔

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

یکلیڈین فاصلے کی پیمائش کو استعمال کرنے سے پہلے تمام اوصاف کی اقدار کو معمول پر لایا جاتا ہے تاکہ ابتدائی طور پر بڑی حدود والی صفات کو ابتدائی طور پر چھوٹی حدود سے بڑھ جانے والے صفات سے روک دیا جا سکے، کم سے کم معمولی قدر کو مساوات 3.5 کی طرف سے [1،0] کی حد میں عددی صفت A کی قدر V کو V میں تبدیل کرنے لے لئے استعمال کیا گیا ہے۔

$$V = \frac{V \ - \ min\, A}{max\, A \ - \ min\, A} \qquad (3.5)$$

جہاں min A اور max A خصوصیت A کی کم سے کم اور زیادہ سے زیادہ اقدار ہیں۔

اس تحقیقی کام میں K-NN تکنیک انتخابی نتائج کی پیشن گوئی کے لئے استعمال کی گئی ہے جس کے تجرباتی نتائج کو باب 4 میں زیر بحث لایا گیا ہے۔

3.6.3 سَپورٹ ویکٹر مشین (ایس وی ایم) Support Vector Machine

سَپورٹ ویکٹر مشین (ایس وی ایم) ایک سوپروائزڈ ڈیٹا مائننگ تکنیک ہے جسے کلاسی فیکیشن اور ریگریشن دونوں مقاصد کے لئے استعمال کیا جا سکتا ہے [150] سَپورٹ ویکٹر مشین کا بنیادی کام

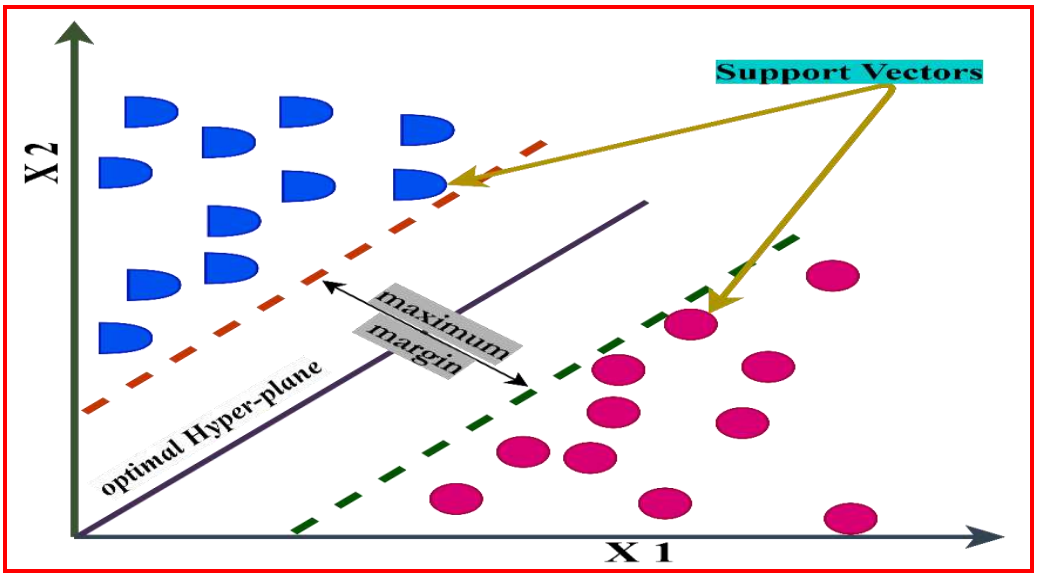
زیادہ سے زیادہ ہائپر پلین (طیارہ) تلاش کرنا ہے جو ڈیٹا سیٹ کو اپنی خصوصیات کی بنیاد پر مختلف طبقات میں تقسیم کرتا ہے۔[151] کسی ایس وی ایم (SVM) کے لئے زیادہ سے زیادہ ہائپر پلین (طیارہ) کا مطلب وہ ہے جس کا فاصلہ ہر قریب ترین کلاس کے ساتھ زیادہ سے زیادہ ہو، ایس وی ایم (SVM) کو یہ ہائپر پلین سپورٹ ویکٹرس اور مارجنز کا استعمال کرتے ہوئے ملتا ہے۔ سپورٹ ویکٹر وہ ڈیٹا پوائنٹس ہیں جو الگ کرنے والے ہائپر پلین (طیارہ) کے قریب ہوتے ہیں اور ڈیٹا سیٹ کے اہم عنصر سمجھے جاتے ہیں اور مارجن ہائپر پلیون (طیاروں) کے متوازی سلاب slab کی زیادہ سے زیادہ چوڑائی ہے جس میں داخلی ڈیٹا پوائنٹس نہیں ہیں۔ جانچ کے نمونے T کے لئے امتیازی فعل f(T) معاون ویکٹر کا ایک خطی امتزاج ہے اور جسے نیچے دیے گئے مساوات 3.6 میں دکھایا گیا ہے۔

$$f(T) = \sum_{n=1}^{\infty} \alpha_i y_i (X_i \ . \ T) \ + b \qquad 3.6$$

جہاں ویکٹر X_i معاون ویکٹرس ہیں، X_i, Y_i کے کلاس لیبل ہیں، ویکٹر T ٹیسٹ کے نمونے کی نمائندگی کرتا ہے، $X_i.T$ معاون ویکٹر X_i کے ساتھ ٹیسٹ نمونے T کا ڈاٹ پروڈکٹ ہے، α_i اور b

عددی پیرامیٹرز ہیں جو لرننگ الگورتھم کے ذریعے طے کیے جاتے ہیں۔

نیچے دیئے گئے ترسیم 3.6.3، لینیئر سپورٹ ویکٹر مشین کی وضاحت کرتی ہے، بھوری رنگ کے سرخ حلقے کلاس X1 کے ڈیٹا پوائنٹس کی نمائندگی کرتے ہیں اور نیلے رنگ (اسکائی بلیو) x2 کے ڈیٹا پوائنٹس کی نشاندہی کرتے ہیں، ایس وی ایم SVM کا مقصد ہائپر پلین (طیارہ) اور ٹریننگ سیٹ کے ساتھ کسی بھی ڈیٹا پوائنٹ کے مابین سب سے زیادہ مارجن کے ساتھ ایک ہائپر پلین (طیارہ) کا انتخاب کرنا ہے، جس سے نئے اعداد و شمار کو صحیح درجہ میں درجہ بندی کرنے کا زیادہ سے زیادہ امکان مل جاتا ہے، تاہم اگر کوئی واضح ہائپر پلین (طیارہ) موجود نہیں ہے تو یہ ضروری ہے ایس وی ایم (SVM) میں کارنیلنگ کے نام سے ایک اعلی طول و عرض والے نظارے کی طرف جائیں۔ کرنل استعمال کرنے کا مقصد یہ ہے کہ وہ ڈیٹا کو اِن پٹ کے طور پر لے سکتا ہے اور اسے مطلوبہ شکل میں مثال کے طور پر ایک جہتی اِن پٹ ڈیٹا سے لے کر دو جہتی آؤٹ پٹ ڈیٹا میں تبدیل کر سکتا ہے، خیال یہ ہے کہ اس اعداد و شمار کو اعلی اور اعلی طول و عرض میں نقش کیا جاتا رہے گا جب تک کہ ایک ہائپر پلیٹ تشکیل نہ دیا جا۔ اور اسے مختلف طبقات میں الگ کر دیا جائے [152]۔



ترسیم 3.6.3:دو طبقے کی نمائندگی کے لئے لینیئر ایس وی ایم درجہ بندی [153]

لہذا نان لینیئر علیحدگی کی صورت میں، تربیت کے اعداد و شمار کو اعلی جہتی جگہ H میں نقشہ بنایا جائے گا اور وہاں ایک زیادہ سے زیادہ ہائپر پلین (طیارہ) تعمیر کیا جائے گا، نقشہ سازی کینل فنکشن (K). کے ذریعہ انجام دی جاتی ہے جس میں H کے اندرونی مصنوعات کی وضاحت ہوتی ہے، مختلف نقشے مختلف SVM تیار کرتے ہیں جیسا کہ مساوات 3.7 میں دیے گئے امتیازی سلوک کے ذریعہ بیان کیا گیا ہے۔

$$f(T) = \sum_{i=1}^{n} \partial_i y_i K(x_i, T) + b \qquad (3.7)$$

ایس وی ایم بڑی حد تک اس کے کرنل فنکشن کے انتخاب کی خصوصیت سے نمایاں ہے مثلاً پولینومئیل کرنل(polynomial Kernel)اور گاؤشین ریڈیل(Gaussian Radial)بیس کرنل فنکشن۔تاہم ان کرنل فنکشنس کے افعال کے علاوہ، کرنل کے بھی دوسرے کام ہیں، اور y پیرامیٹرز کا تعین کرنے اور مذکورہ مساوات میں امتیازی تقریب کی تعمیر آخر کارlangrangianدوہری مقصد کی تقریب کو بالترتیب مساوات3.8اور3.9 میں بیان کرتے ہوئے زیادہ سے زیادہ ایک کواڈریٹک مسئلہ ہے۔

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

رکاوٹوں کے تحت

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

تربیت کے اعداد و شمار میں نمونے کی تعداد n ہے، تاہم مساوات3.8میں کواڈریٹک پروگرامنگ کا مسئلہ معیاری تکنیکوں کے ذریعے آسانی سے حل نہیں کیا جا سکتا کیونکہ اس میں

ایک میٹرکس(Matrix) شامل ہوتا ہے جس میں متعدد عناصر ہوتے ہیں جن کی تربیت کے نمونے کی تعداد کے مربع کے برابر ہوتی ہے، انتخابی پیشن گوئی میں ایس وی ایم تکنیک کا استعمال کیا گیا ہے کیونکہ اس کے جامع اور صحیح قواعد رائے شماری کی پیشن گوئی کرنے والوں کی مدد کے لئے انتہائی ضروری ہیں، باب 5 میں ایس وی ایم ماڈل سے حاصل کردہ انتخابی پیشن گوئی کے نتائج پر تبادلہ خیال کیا گیا ہے۔

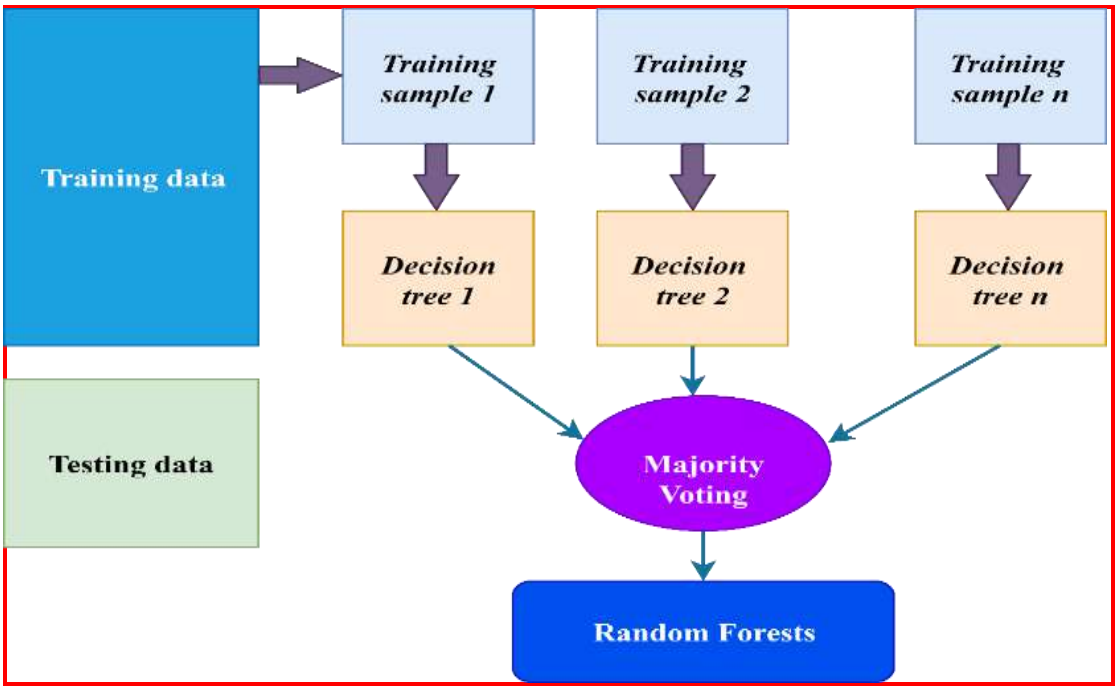
3.6.4 رینڈم فاریسٹ (Random Forest)

رینڈم فارسٹس بڑی تعداد میں ڈسیزن ٹری کا ایک جوڑا ہے جس کی درجہ بندی اور ریگریشن کے لئے اوسطاً اکثریت کے ووٹ کو جمع کر کے نئے اعداد و شمار یا طبقے کی پیشن گوئی کرنے کے لئے استعمال کیا جا سکتا ہے[154]۔ رینڈم فاریسٹ الگورتھم انفرادی ڈسیزن ٹری کے زیادہ مناسب مسئلہ پر قابو پانے اور میٹاڈیٹا سیٹ کو اعلی طول و عرض کے ساتھ سنبھالنے کے مقصد کے ساتھ بے ترتیب منتخب کردہ تربیتی سیٹ سے متعدد ڈسیزن ٹری کے ساتھ فاریسٹ تخلیق کرتا ہے، فاریسٹ کی غیر مرتب درجہ بندی میں ہر ڈسیزن ٹری کے ووٹ اور مجموعی ووٹ ٹیسٹ آبجیکٹ کی آخری کلاسوں کا فیصلہ کرتے

ہیں جبکہ ریگریشن میں انفرادی درختوں کی اسباب کی پیشن گوئی یا ریگریشن کا حساب لگایا جاتا ہے اور آخر میں اکثریت کے ووٹ ماڈل کی پیشن گوئیاں بن جاتے ہیں[155]۔

رینڈم فاریسٹ کو بیگنگ کے طریقوں سے تربیت دی جاتی ہے جو بڑے پیمانے پر غیر منسلک درختوں کا مجموعہ بناتے ہیں اور پھر ان کو متوسط بناتے ہیں[156]۔ رینڈم فاریسٹ الگورتھم میں ہر ایک درخت کو اصل اعداد و شمار کے مختلف نمونوں پر اگایا جاتا ہے اور ٹیسٹ سیٹ غلطی کا غیر جانبدارانہ تخمینہ حاصل کرنے کے لئے کراس ویلیڈیشن یا علیحدہ ٹیسٹ سیٹ کی ضرورت نہیں ہوتی ہے، کیونکہ ہر ایک کے اعادہ میں تقریبا ایک تہائی نمونے بوٹسٹریپ (Bootstrap) کی نئی ٹریننگ سیٹ سے باہر رہ گئے ہیں اور درخت کی تعمیر میں استعمال نہیں کیے گئے ہیں، خیال کیا گیا ہے۔ ترسیم 3.6.4 رینڈم فاریسٹ الگورتھم کے کام کو ظاہر کرتی ہے جس کے اقدامات کی وضاحت مندرجہ ذیل میں ہے اس طرح تعمیر شدہ درختوں میں سے ہر ایک نمونے کے لئے ٹیسٹ سیٹ کی درجہ بندی حاصل کی جاتی ہے، کسی نمونے کی آخری درجہ بندی وہ کلاس ہے جس میں جنگل کے درختوں سے بھی کئی زیادہ ووٹ ہوتے ہیں انتخاب کی پیشن گوئی اور دیگر پیشن گوئی میں رینڈم فاریسٹ کا الگورتھم استعمال کیا

جاتا ہے، باب 5 میں رینڈم فاریسٹ ماڈل سے حاصل کردہ انتخابی پیشن گوئی کے نتائج پر بحث کی گئی ہے



ترسیم 3.6.4:رینڈم فاریسٹ الگورتھم کا کام [156]

3.7ماڈل کے تشخیص کی تکنیک

ماڈل کی کارکردگی کو جانچنے کے لئے کسی کو یہ اندازہ کرنے کے لئے منظم طریقے کی ضرورت ہے کہ data mining کی تکنیک کس طرح کام کرتی ہے۔ درجہ بندی کی دشواریوں کے لئے ، کنفیوژن میٹرکس ،

کراس- توثیق، غلطی کی شرح، حساسیت، وضاحتی، درستگی، صحت سے متعلق اور آر او سی ROC منحنی خطوط کے لحاظ سے درجہ بندی کرنے والے کی کارکردگی کی پیمائش کرنا فطری بات ہے، جس پر ذیل میں بحث کی گئی ہے۔

3.7.1 کنفیوژن میٹرکس (Confusion Matrix)

کنفیوژن میٹرکس بنیادی طور پر ٹیسٹ ڈیٹا پر مشین لرننگ کلاسیفائیرز کی کارکردگی کی پیمائش کے لئے استعمال کیا جاتا ہے جن کی اصل اقدار پہلے ہی معلوم ہیں[157] ۔ جدول 3.7.1 میں دو طبقاتی الجھن(confusion) کا میٹرکس دکھایا گیا ہے جو درجہ بندی کرنے والے کے ذریعہ ہونے والی غلطیوں کی اقسام کے بارے میں بصیرت فراہم کرتا ہے۔

جدول 3.7.1: دو طبقاتی درجہ بندی کے لئے کنفیوژن میٹرکس[158]

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

مثبت ٹپلس جو الگورتھم کے ذریعہ صحیح طور پر لیبل لگائے جاتے ہیں انہیں ٹروپوزیٹیو (TP) کہا جاتا ہے۔ جبکہ درجہ بندی کرنے والے کے ذریعہ منفی طور پر منفی لیبل لگانے والے منفی ٹپلس کو ٹرو نیگیٹیو (TN) کہا جاتا ہے، جھوٹے مثبت (FP) کو ٹائپ-1 کی غلطی (ایرر) کے نام سے بھی جانا جاتا ہے، یہ وہ منفی ٹیوپلس ہیں جن کی غلط طور پر مثبت درجہ بندی کی گئی ہے، جھوٹی منفی (FN) کو ٹائپ-II کی غلطی (ایرر) بھی کہا جاتا ہے، یہ وہ مثبت ٹپلس ہیں جن کی مثبت طور پر غلط درجہ بندی کی گئی ہے[159] ۔

i. **senstivity**: حساسیت کو حقیقی مثبت شرح/پہچان/یاد آوری کے نام سے بھی پکارا جاتا ہے جب حقیقی مثبت مثال واقعی مثبت کے طور پر پہچانی جاتی ہے، حساسیت کا موازنہ کیا گیا ہے جیسا کہ ذیل میں دیئے گئے مساوات 3.11 میں دکھایا گیا ہے۔

$$Sensitivity = \frac{TruePositive}{TruePositive + False\ Negative} \qquad (3.11)$$

ii. **Specificity**: اس کی وضاحت کو سچ منفی شرح بھی کہا جاتا ہے جیسا کہ ذیل میں دی گئی مساوات 3.12 میں دکھایا گیا ہے جس کی وضاحت کی گئی ہے کہ درجہ بندی کتنی مرتبہ منفی مثال کی پیشن گوئی کرتی ہے۔

$$Specificity = \frac{TrueNegative}{\text{TrueNegative} + False\ Positive} \qquad (3.12)$$

iii. **Accuracy:** درستگی ایسے معاملات کی مجموعی فیصد ہے جو الگورتھم (یعنی مجموعی طور پر کتنا درجہ بندی درست ہے) کے ذریعہ درست طریقے سے درجہ بند ہیں۔

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

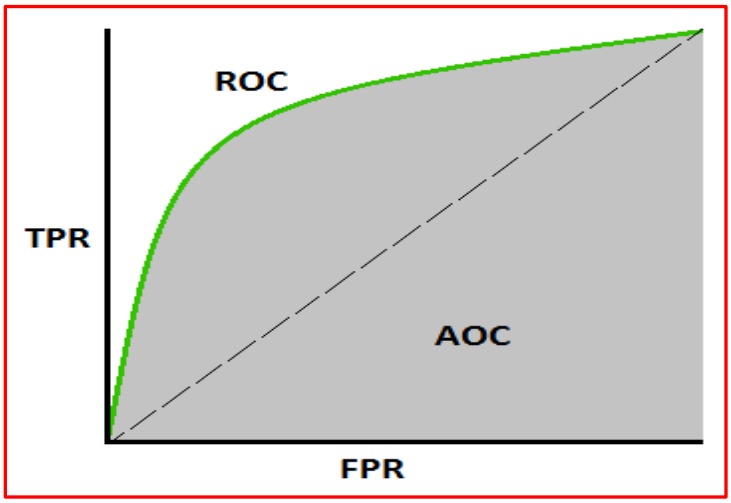
Iv **Precision:** صحت سے متعلق جب درجہ بندی کرنے والی پیش گوئی کرتی ہے کہ ہاں، یہ کتنی بار صحیح پیش گوئی کرتی ہے (یعنی اس تک کہ کتنے فیصد لیبل لگا ہوا ٹیوپس کا تناسب حقیقت میں صحیح ہے)۔

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

3.7.2 AUROC (Area under the Receiver Operating Characteristics)

(آر او سی ROC) منحنی خطوط پر مشین لرننگ کلاسیفائیر کی کارکردگی کو حقیقی مثبت (ٹرو پوزیٹیو TP) اور جھوٹے مثبت (FP) کی شرح کے درمیان پیمائش کرنے کا ایک طریقہ ہے۔[160] آر او سی ROC ایک امکانی منحنی خط ہے اور اے یو سی AUC علیحدگی کی ڈگری یا پیمائش کی نمائندگی کرتا ہے [161]۔ یہ بائنری

درجہ بندی کی صورت میں دو طبقوں کے درمیان فرق کرنے میں درجہ بندی کی کارکردگی کی وضاحت کرتا ہے جیسا کہ ذیل میں دیئے گئے ترسیم[162] 3.7.2 میں دکھایا گیا ہے۔، جھوٹی مثبت شرح کے مقابلہ میں آراوسی ROC منحنی خط کو صحیح مثبت شرح کے ساتھ بنایا گیا ہے جہاں حقیقی مثبت(ٹروپوزیٹیوTP) شرح Y- محور پر ہے جھوٹے مثبت(فالس پوزیٹیوFP) شرح X- محور پر ہے



ترسیم3.7.2: AUROC منحنی خط خاکہ

ایک بہترین ماڈل 1 کے قریب اے یو سیAUC رکھتا ہے، جس کا مطلب یہ ہے کہ ماڈل مثبت اور منفی کے درمیان کلاس کو بہت اچھی طرح سے فرق کرتا ہے،جب ماڈل کے پاس اے او سی 0 کے قریب ہوتا ہے تو اس کا مطلب ہے کہ پھر وہ نتائج کی تکرار کر رہا ہے،ایسے حالات میں ماڈل0s کو 1s اور1s کو0s کی

طرح پیشن گوئی کر رہا ہے ، جب ماڈل کی اے یو سی AUC تقریباً 0.5 ہو تو ماڈل میں مثبت طبقے اور منفی طبقے کو فرق کرنے میں امتیازی صلاحیت نہیں رہتی ہے۔

3.7.3 کراس-ویلیڈیشن (Cross-Validation)

کراس ویلیڈیشن بنیادی طور پر ایک شماریاتی تکنیک ہے جو کسی خاص ڈیٹاسیٹ پر غلطی کی شرحوں کی اصطلاح میں درجہ بندی کرنے والوں کی مہارت کی پیمائش کے لئے استعمال ہوتی ہے۔، تربیت ڈیٹاسیٹ ڈیٹا مائننگ کی تکنیک کو اس ڈیٹا سے سیکھنے کی اجازت دیتا ہے، ٹیسٹنگ ڈیٹاسیٹ کو ٹریننگ ڈیٹاسیٹ سے سیکھی گئی چیز کے سلسلے میں ڈیٹا مائننگ تکنیک کی کارکردگی کی جانچ کرنے کے لئے استعمال کیا جاتا ہے[163] ۔

3.7.4 غلط درجہ بندی کی شرح (Misclassification Rate)

درجہ بندی ماڈل کے ذریعہ کی جانے والی غلطیاں عام طور پر دو اقسام میں تقسیم ہوتی ہیں: تربیت کی غلطیاں اور عام کرنے کی غلطیاں۔ تربیت کی غلطیاں جسے دوبارہ ردوبدل کی غلطی یا بظاہر غلطی کے نام سے جانا جاتا ہے جو کہ تربیت کے ریکارڈوں پر عائد غلط فہمی کی غلطیوں کی تعداد ہے، جبکہ عام طور پر غلطی پچھلے

غائب ریکارڈوں پر ماڈل کی متوقع غلطی ہے۔ اچھی درجہ بندی کے ماڈل میں تربیت کی غلطی کے کم ہونے کے ساتھ ساتھ عام کرنے کی غلطی بھی کم ہونی چاہئے، درجہ بندی کرنے والا ہر مثال کی کلاس کی پیشن گوئی کرتا ہے: اگر صحیح ہے تو، اسے کامیابی کے طور پر شمار کیا جاتا ہے اور اگر نہیں تو یہ ایک غلطی ہے، غلطی کی شرح نمونہ (instances) کے پورے سیٹ پر ہونے والی غلطیوں کا تناسب ہے اور یہ درجہ بندی کی مجموعی کارکردگی کی پیمائش کرتی ہے[164]۔

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

3.8 اسمبل تکنیکس (Ensemble Techniques)

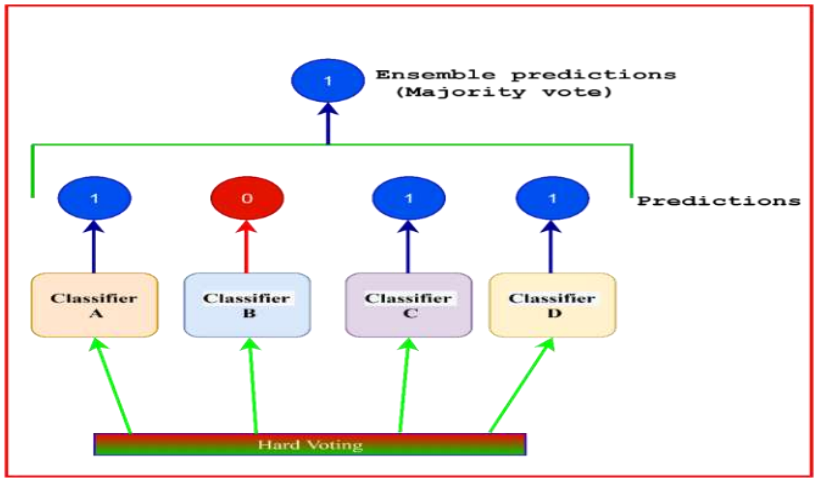
اسمبل میتھڈ بہتر درستگی اور کم تغیر کے حصول کے لئے متعدد مشین لرننگ کلاسیفائی کو ایک پیشن گوئی درجہ بندی میں شامل کرتے ہیں[165] ۔ یہ معروف ہے کہ اسمبل میتھڈ زیادہ درستگی کے ساتھ بہ نسبت سنگل کلاسیفائیر کے زیادہ درست نتائج پیدا کرتے ہیں ، اور اس کی وجہ سے کمپیوٹر سائنس کے مختلف شعبوں میں جوڑنے والے طریقوں کا استعمال ہوا ہے، جوڑنے والی تکنیکیں دو طرح کی ہیں۔ hard voting ensembling اور soft voting ensembling تکنیک۔

Hard voting جہاں ماڈل کا انتخاب اکثریتی ووٹ کے ذریعہ کیا جاتا ہے۔ مثال کے طور پر اگر ہمارے پاس بالترتیب 1، 1، 0، 1 آؤٹ پٹس کے ساتھ چار درجہ بندیاں ہیں تو ہم اکثریت کے قاعدہ کی وجہ سے 1 کو آؤٹ پٹ کے طور پر منتخب کرتے ہیں۔ ہارڈ ووٹنگ کی کارکردگی کو اعداد و شمار 3.8.1 میں دکھایا گیا ہے، ہم وزن wj اور درجہ بندی کرنے والے cj کو ساتھ ملا کر وزنی اکثریتی ووٹ کی گنتی کرتے ہیں[166] ۔

$$y^{\wedge}=\operatorname{argmax}_i \sum_{j=1}^{m} w_j \chi_A(C_j(x)=i) \qquad (3.6.1)$$

جہاں χ_A خصوصیت کا کام ہوتا ہے $[C_j(x) = i \in A]$، اور A منفرد کلاس لیبلوں کا مجموعہ ہے۔

ترسیم 3.8.1 کی تفصیل ذیل میں بیان کی گئی ہے



ترسیم3.8.1:ہارڈووٹنگ [167]

درجہ بندیAآوٹ پٹ 1 کے نتائج کی پیش گوئیاں ، درجہ بندیBآوٹ پٹ 0 کے نتائج کی پیشگوئیاں ، درجہ بندی C آوٹ پٹ 1 کے نتائج کی پیش گوئیاں درجہ بندی کرنے والا(کلاسیفائیرD) آوٹ پٹ 1 کے نتائج کی پیش گوئیاں۔ جیسا کہ ہم نے دیکھا ہے کہ 3/4درجہ بندی نے 1 کو بطور آؤٹ پٹ پیشن گوئی کی ہے، لہذا1 سخت رائے دہی یا جوڑ فیصلہ ہے۔

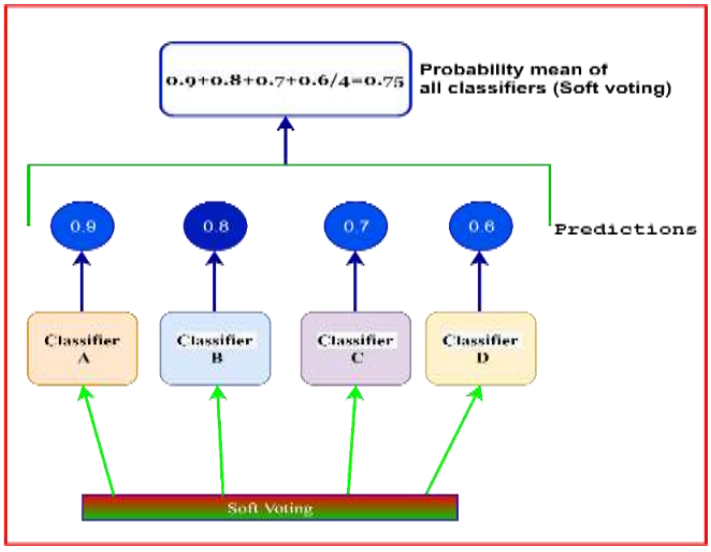
سافٹ ووٹنگ نتائج کے لئے استعمال ہونے والے تمام درجہ بندی کے اوسط امکانات کا حساب کتاب کرکے کی گئی ہے، سافٹ ووٹنگ انفرادی الگورتھم کے حساب سے احتمالات کا اوسط نکال کر بہترین نتائج پر پہنچتی ہے۔[168] سافٹ ووٹنگ میں ہم درجہ بندی کرنے والوں کے لئے پیشن گوئی کی

گئی امکانات P کی بنیاد پر کلاس لیبل کی پیشن گوئی کرتے ہیں جیسا کہ نیچے دیئے گئے ترسیم 3.12 میں دکھایا گیا

ہے۔

$$y^{\wedge}=\text{argmax}_i \sum_{j=1}^{m} w_j p_{ij} \qquad (3.6.2)$$

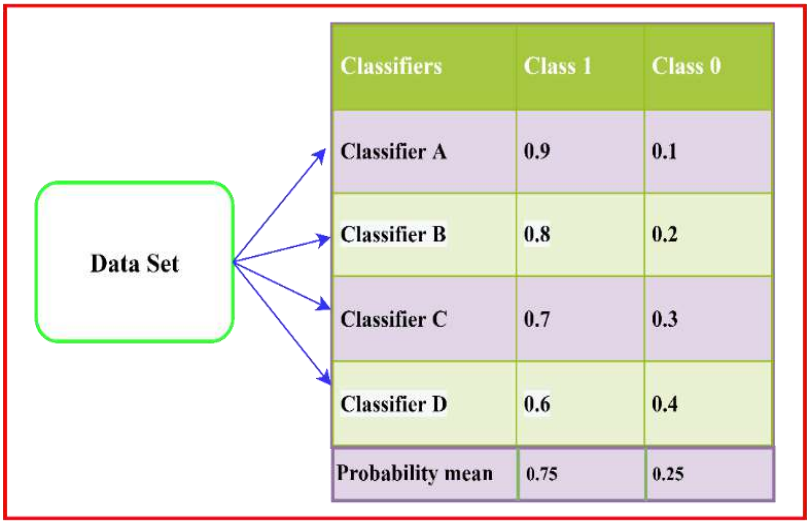
جہاں wj وزن ہے جس میں jth درجہ بندی کرنے والے کو تفویض کیا جا سکتا ہے۔



ترسیم 3.8.2: سافٹ ووٹنگ [169]

سافٹ ووٹنگ میں اوسط امکانات کا مطلب حتمی نتائج بنانے کے لئے تمام درجہ بندی کرنے والوں کو زیر غور رکھنا ہے، سافٹ ووٹنگ میں تمام درجہ بندی کی آؤٹ پٹ کو کلاس 1 اور کلاس 0

کے درمیان تقسیم کیا جاتا ہے، پھر جس کلاس میں سب سے زیادہ آؤٹ پٹس ہوتے ہیں وہ حتمی کلاس سمجھا جاتا ہے جیسا کہ نیچے ترسیم 3.8.3 میں دکھایا گیا ہے، اس طرح سافٹ ووٹنگ کلاسیفائیر ہارڈ ووٹنگ سے بہتر کارکردگی کا مظاہرہ کرتا ہے کیونکہ یہ انتہائی پر اعتماد ووٹوں کو زیادہ وزن دیتا ہے۔



ترسیم 3.8.3: سافٹ ووٹنگ کلاسیفائیرس

مذکورہ بالا ترسیم 3.8.3 میں ہم نے چار درجہ بندیاں کیں اور دو طبقے (یعنی کلاس 1 اور کلاس 0) کے امکان کے معنی کا حساب لگایا اور ہم نے محسوس کیا کہ کلاس 1 میں سب سے زیادہ احتمال ہے کلاس 0 کی بہ نسبت، اس طرح ہم کلاس 1 کو حتمی نتائج بنانے کے لئے حتمی کلاس کے طور پر منتخب

کرتے ہیں،ensemble learning سے حاصل کردہ انتخابی پیشن گوئی کے نتائج پر باب 5 میں تبادلہ خیال کیا گیا ہے۔

3.9 شماریاتی ٹیسٹ (Statistical Test)

شماریاتی ٹیسٹ یہ معلوم کرنے کے لئے کئے جاتے ہیں کہ آیا ماڈل اہم ہے یا نہیں p ویلیو کا استعمال کر کہ، جیسے کہ اگر $p<0.05$ ہے پھر ہم اس کا لعدم مفروضے کو مسترد کرتے ہیں اور اس سے یہ ثابت ہوتا ہے کہ موازنہ کرنے والے ماڈل مختلف کارکردگی کا مظاہرہ کر رہے ہیں اور جس ماڈل پر موازنہ کیا گیا ہے وہ اہم ہے، اس تحقیقی کام میں ہم اعداد و شمار ٹی جوڑی ٹیسٹ کرتے ہیں جو دو لرننگ الگورتھم کا موازنہ کرنے کے لئے استعمال کیا جانے والا ایک مشہور جوڑی میں فرق کا ٹیسٹ ہے۔

T-paired test اس ٹیسٹ میں P-ویلیو کا اندازہ کرنے کے لئے t تقسیم کا استعمال کیا گیا ہے جس کی نمائندگی اس امکان کی نمائندگی کرتی ہے کہ دیکھے گئے اختلافات کا مطلب بے ترتیب طور پر ہوا ہے، اگر نتیجے میں p ویلیو بہت کم ہے (عام طور پر 0.05 سے نیچے) تو یہ نتیجہ اخذ کیا جا سکتا ہے کہ مشاہدہ کیا گیا فرق غیر مرتب موقع کے ذریعہ بیان کیے جانے والے مقابلے میں کہیں زیادہ ہے،

اور اس وجہ سے اعداد و شمار اہم ہیں، مشین لرننگ کے تناظر میں اس کا یہ موازنہ کرنے کے مترادف ہے کہ الگورتھم A بمقابلہ الگورتھم B کسی خاص سیکھنے کی دشواری میں کتنا اچھا ہے جو ایک ڈیٹا سیٹ کی جانب سے ترتیب شدہ ہے۔ اگر ڈیٹا سیٹ کا استعمال کرتے ہوئے الگورتھم A کی کارکردگی اور الگورتھم B کی کارکردگی کے درمیان وسطی فرق اعداد و شمار کے لحاظ سے اہمیت کا حامل ہے تو، اس خاص جھکاؤ مسئلے کے لئے الگورتھم A کے استعمال کو ترجیح دینے کی حمایت کی جا سکتی ہے[170]، اس وجہ سے دو یا دو سے زیادہ الگورتھم کا موازنہ کرنے والی مشینیں مشین لرننگ الگورتھم کی افادیت کی توثیق کرنے اور مختلف دشواری ڈومینز میں ایک دوسرے سے موازنہ کرنے میں اہم ہیں۔

3.10 الیکشن پیشن گوئی ماڈل کی ترقی کے لئے ڈیٹا اکٹھا کرنے اور ریسرچ کرنے کے طریقے:

ڈیٹا سیٹ اکٹھا کرنے کیلئے قابل فخر چیلنج معیار اور متعلقہ ڈیٹا کو حاصل کرنا ہے، بنیادی اعداد و شمار مختلف متفاوت اعداد و شمار کے ذرائع جیسے ہندوستان کے الیکشن کمیشن اور جموں و کشمیر کے الیکشن کمیشن کی ویب سائٹوں اور ہر ضلعی انتخابی سیل کے کچھ اعداد و شمار سے حاصل کیا گیا ہے،

اس انتخابی ڈیٹاسیٹ میں بارہ اوصاف کے ساتھ 3770 ریکارڈ شامل ہیں جیسا کہ ذیل میں دیا گیا جدول 3.10 میں بیان کیا گیا ہے،

اس تحقیقی کام میں وضاحتی تحقیقی طریقہ کار کی پیشن گوئی کی گئی ہے تاکہ وہ سیاسی پیشن گوئی کے ماڈل کو تیار کر سکیں، انتخاب کی پیشن گوئی کا نمونہ پائتھان پروگرامنگ لینگویج اور ڈیٹا کلیننگ ، شماریاتی ماڈلنگ ، ڈیٹا ویزولائزیشن ، اور مشین لرننگ کے عمل میں تیار کیا گیا ہے جو پیٹر ویب ایپلی کیشن میں کیا گیا ہے۔ ہم نے اعداد و شمار سے مختلف بصیرت حاصل کرنے کے لئے جموں اور کشمیر کے انتخابی ڈیٹاسیٹ پر ایک تحقیقاتی ڈیٹا تجزیہ (EDA) کیا۔ اعداد و شمار پر اس طرح کی بصیرت حاصل کرنے کے بعد کے تجزیے میں مدد ملے گی جیسے زیادہ سے زیادہ آؤٹ پٹ حاصل کرنے کے لئے اور ناپسندیدہ تغیرات کو قابو کرنے کیلئے اس تکنیک کو ڈیٹا پر استعمال کی جا سکتا ہے۔

تحقیقاتی ڈیٹا تجزیہ EDA کے بعد یہ پتہ چلا ہے کہ انتخابی ڈیٹاسیٹ پُر شور ہے اور متعدد گمشدہ اوصاف والی اقدار پر مشتمل ہے جن کی نمائندگی پوچھ گچھ (؟) علامت کے ساتھ کی گئی ہے، اعداد و شمار کا انقباض انجام دیا گیا ہے (غیر قانونی طور پر گمشدہ اقدار کا مطلب ہے کہ ڈیٹا کلیننگ کی تکنیک کو لاپتہ

ہونے والے وصف سے متعلق اقدار کو پُر کرنے کے لئے استعمال کیا جاتا ہے اور واضح صفات کے لئے ، گمشدہ اقدار کو پُر کرنے کا موڈ طریقہ استعمال کیا جاتا ہے)۔انتخابی ڈیٹاسیٹ کی صفات کو دو قسموں میں بنام اور ہندسوں میں ترتیب دیا گیا ہے۔ مثال کے طور پر ، صنف وصف کی حیثیت سے مرد اور عورت برائے نام وصف کی نمائندگی کرتے ہیں اور عمر کی خصوصیت 50 سال عددی وصف کی نمائندگی کرتی ہے، مزید برآں ، برائے نام اعداد و شمار بائنری اور عام متغیر کے مطابق ہوتے ہیں اور اعداد و شمار کے اوصاف عدد ، وقفہ سے پیمانے اور تناسب سے متعلق متغیر کے مطابق ہوتے ہیں۔

جدول3.10:انتخابی پیرامیٹرزاوران کی تفصیل۔

پیرامیٹرز	وضاحت
مرکزی اثرورسوخ	بھارت میں حکومت بنانے والی پارٹی کے جموں و کشمیر میں حکومت بنانے کازیادہ سے زیادہ امکان موجود ہے جیسا کہ 2008 میں کانگریس نے کولیشن حکومت تشکیل دی، جموں و کشمیر میں جموں و کشمیر نیشنل کانفرنس کے ساتھ بھی، 2014 میں بی جے پی نے کولیشن گورنمنٹ تشکیل دی، جموں و کشمیر میں جموں و کشمیر پیپلز ڈیموکریٹک پارٹی کے ساتھ۔

مذہب کے پیروکار	جموں و کشمیر کی سیاست میں مذہبی عنصر کا غلبہ تھا، مذہب بی جے پی کی بنیادی وجہ ہے کیونکہ وہ گزشتہ تین انتخابات میں صوبہ کشمیر میں ایک بھی نشست نہیں جیت سکی کیونکہ مذہب کے پیروکاروں کی اکثریت مسلمان ہے۔
پارٹی لہر	اس کا مطلب ہے کہ پارٹی جس نے انتخابات کے اوقات میں مثبت لہریں حاصل کیں، 2014 کے ہندوستان کے پارلیمانی انتخابات کے دوران بی جے پی کی لہر جیسے انتخابات میں بیشتر نشستوں پر کامیابی کا زیادہ سے زیادہ امکان ہے۔
پارٹی خلاصے	اس میں انڈین نیشنل کانگریس، جموں و کشمیر نیشنل کانفرنس، مرکزی و ریاستی پارٹی یا آزاد جیسے جماعتوں کے نام اور قسم کی وضاحت کی گئی ہے، اگر کوئی امیدوار کسی جماعت سے الیکشن لڑ رہا ہے تو آزاد امیدوار کی حیثیت سے انتخاب لڑنے والے امیدوار کے مقابلے میں انتخابات جیتنے کا زیادہ امکان موجود ہے۔
حساس علاقے	یہ واقعہ صوبہ کشمیر میں خاص طور پر انتخابی اوقات میں بہت زیادہ پایا جاتا ہے، لوگ رائے شماری کا بائیکاٹ کرتے ہیں اور اپنا ووٹ نہیں ڈالتے ہیں جس سے جموں و کشمیر میں انتخابی عمل متاثر ہوتا ہے۔

ووٹ بینک	ووٹ بینک کا مطلب ہے یا تو لوگ پارٹی کو یا امیدواروں کو ووٹ دیتے ہیں۔
موروثی	موروثی مطلب یہ ہے کہ اگر کسی فرد کے مقابلہ انتخابات ہوں اور کسی شخص نے پہلے ہی اس کے گھر والوں یا اس کے وارث سے الیکشن لڑا ہو، تو اس شخص کا جیتنے کا امکان اس کے حریف کے امیدواروں کے مقابلے میں زیادہ ہوتا ہے۔ جیسے عبداللہ کنبہ، مفتی خاندان وغیرہ۔
مابعد پارٹی	یہ ہر حلقے کی فاتح پارٹی کو ملازمت دیتی ہے، نشست پارٹی کے پاس ووٹروں کی طرف مائل ہونے کی وجہ سے بار بار انتخاب جیتنے کا روشن امکان نہیں ہو سکتا ہے۔
ذات کا فیکٹر	جموں و کشمیر کے کچھ علاقوں جیسے ذات پونچھ، راجوری، کپواڑہ، گول ارناس اور پورے لداخ خطے میں ذات پات کا ایک اہم کردار ہے، لوگوں نے پارٹی یا مذہب کے بجائے اپنی ذات پات کے امیدواروں کو ووٹ دیا۔

3.11 باب کا خلاصہ اور نتیجہ

یہ تحقیق ڈیٹا مائننگ تجزیہ کا استعمال کرتے ہوئے انتخابی پیشن گوئی ماڈل تیار کرنے کی تحقیقات کرتی ہے، بنیادی اعداد و شمار کو مختلف وابستہ اعداد و شمار کے ذرائع سے جمع کیا گیا ہے جیسے ہند کے انتخابی کمیشن اور جموں و کشمیر کے الیکشن کمیشن کی ویب سائٹوں سے مقداری ڈیٹا اکٹھا کرنے کے طریقوں کے

ذریعے جمع کیا گیا ہے جبکہ کچھ اعداد و شمار ہر ضلع کے انتخابی سیل کے ذریعہ جمع کیے گئے ہیں، یہ تحقیقی کام وضاحتی تحقیقی طریقہ کار کی پیروی کرتا ہے اور انتخابی پیشن گوئی کے ماڈل تیار کرنے کے لئے پائتھان پروگرامنگ لینگویج اور جوپیٹر ویب ایپلی کیشن کا استعمال کرتا ہے، اس تحقیقی کام میں جموں و کشمیر کے انتخابی ڈیٹا سیٹ پر انتخابی نتائج کی جلد پیشن گوئی کے لئے نمایاں ہونے والے نمایاں ذیلی سیٹ کو منتخب کرنے کے لئے مختلف خصوصیت کے انتخاب کی تکنیک کا استعمال کیا گیا ہے، اس تفتیش میں ڈیٹا مائننگ کی مختلف تکنیکوں کا اطلاق کیا گیا ہے جیسے ڈیسیزن ٹری، کے این این، ایس وی ایم اور رینڈم فارسٹ وغیرہ۔ انتخابی پیشن گوئی کے جدید ماڈل کی کارکردگی کی پیمائش کرنے کے لئے مختلف ماڈل کی جانچ کی تکنیک اور شماریاتی طریقوں کا استعمال کیا گیا ہے، آخر میں باب کا اختتام جموں و کشمیر کے لئے انتخابی پیشن گوئی کی صفات کی اہمیت پر گفتگو کر کے کیا گیا ہے۔

باب-4

4. مجوزہ طریقہ کار

اس باب میں، ہم انتخاب کے نتائج کی پیشن گوئی کے لئے ایک موثر ماڈل بنانے کے لئے ڈیٹا مائننگ (KDD) کے طریقہ کار میں علمی انکشاف کا استعمال کریں گے۔ مجوزہ طریقہ کار کو مختلف منظم تحقیقی مراحل اور درجوں میں جموں و کشمیر کے انتخاب کی پیشن گوئی کے بارے میں پیچیدہ معلومات حاصل کرنے کے لیے لاگو کیا گیا ہے اور حاصل شدہ ڈیٹا کو مائن کیا جائے گا۔ اس باب میں، ہم تحقیقی سرگرمیوں کو آسان بنانے اور ریسرچ مقاصد کے بیان سے تحقیق کو نتیجہ خیز بنانے کے لئے ایک تحقیقی ڈیزائن مرتب کریں گے۔ یہ تحقیق فطرتاً ایکسپلورٹری ہے جس کا مطلب یہ ہے کہ یہ انتخابی پیشن گوئی کے ماڈل کو بنانے میں اہم اجزاء، رکاوٹوں، مسائل اور ضرورتوں کا انکشاف کرنے کی کوشش کرتی ہے۔ تاکہ یہ صحیح پیشن گوئی کرنے کے لئے سیاسی پیشن گوئی کرنے والوں کی مدد کر سکے، آخر میں، ایک انتخابی پیشن گوئی کا ماڈل تیار کیا گیا ہے جو جموں و کشمیر کے لئے کسی بھی سیاسی جماعت کی انتخابی حلقہ کی سطح پر جیت یا ہار کی نشاندہی کرنے میں مدد کرتا ہے۔

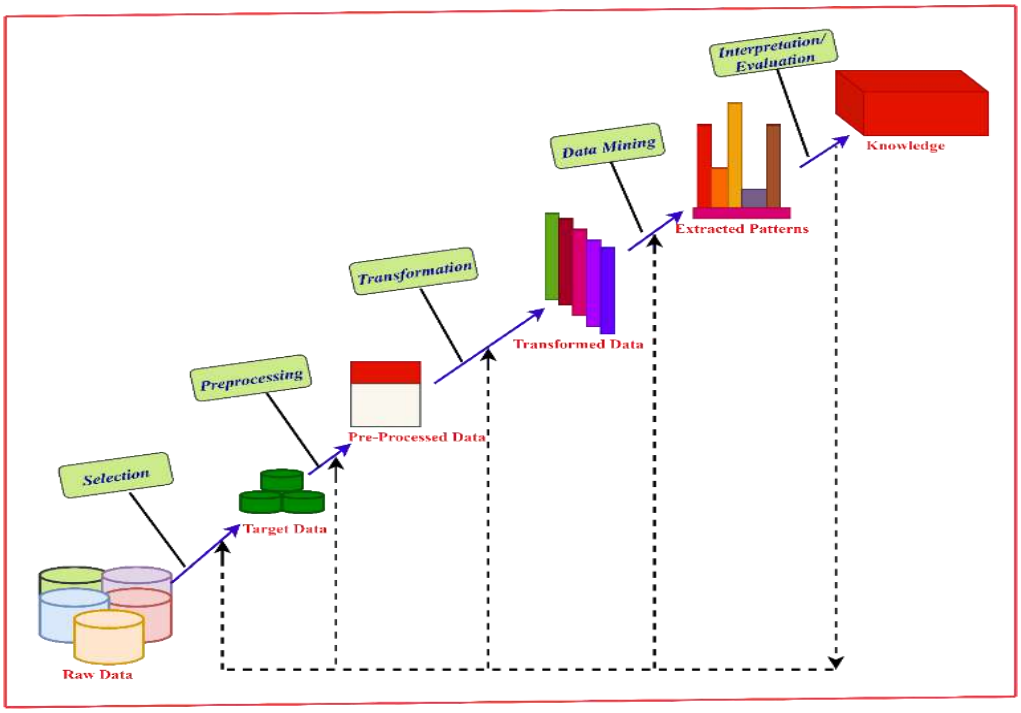
یہ ماڈل اپنے استعمال کرنے والے کے ساتھ دوستانہ تال میل بٹھاتا ہے اسے ایک پیشہ ورانہ صارف کے

ساتھ ساتھ ایک عام انسان بھی آسانی سے استعمال کر سکتا ہے اور سمجھ سکتا ہے۔

4.1 انتخابات کی پیشن گوئی کے لئے ڈیٹا مائننگ کا طریقہ کار

ڈیٹا مائننگ کا طریقہ کار خام ڈیٹا لینے اور اسے قابل فہم شکل میں تبدیل کرنے کے لیے متبادل طریقوں کو استعمال کرنے کی ایک تکنیک ہے تاکہ صارفین کے لئے معلومات فراہم کی جا سکے۔ڈیٹا سے علم کی دریافت کے لئے ڈیٹا مائننگ کے دو مشہور طریقہ کار موجود ہیں RISP-DM اور SEMMA [171] ۔ سی آر آئی ایس پی-ڈی ایم (کراس انڈسٹری اسٹینڈرڈ پروسیس ماڈل برائے ڈیٹا مائننگ) کو انڈسٹری کے زیر قیادت کنسورشیم نے سن 1996 میں تیار کیا تھا [172]۔ سی آر آئی ایس پی-ڈی ایم کو ایک پروسس ماڈل کے طور پر بیان کیا گیا ہے جو ڈیٹا کے منصوبوں کو انجام دینے کے لئے ایک ساخت مہیا کرتا ہے جو صنعت کے شعبے اور اس میں استعمال ہونے والی ٹیکنالوجی دونوں سے آزاد ہے [173]۔SEMMA (سیمپل ایکسپلور موڈیفائی ماڈل کا ایکسیس) ڈیٹا مائننگ کا ایک طریقہ کار ہے جو شماریاتی تجزیہ سافٹ ویئر ادارہ (SAS، 2008) سے اخذ کیا گیا ہے [171]۔ یہ دونوں طریقے ہمارے تحقیقی کام کے لئے موزوں نہیں ہیں کیونکہ استعمال میں یہ بہت بڑے اور بہت پیچیدہ ہیں۔ لہذا اس

تحقیقی کام کے لئے، کے ڈی ڈی کے طریقہ کار کی پیروی کی گئی ہے جو ذیل میں دی گئی ترسیم 4.1 میں دکھائی گئی ہے۔



ترسیم 4.1: انتخابی پیشن گوئی ماڈل کا طریقہ کار

ڈیٹا کا انتخاب: اس مرحلے میں، مختلف متضاد ذرائع سے مناسب انتخابی ڈیٹا منتخب کیا جاتا ہے اور پھر اسے معیاری ڈیٹا بیس میں محفوظ کیا جاتا ہے۔

ڈیٹا کی تیاری: ڈیٹا کی تیاری کے مرحلے میں، انتخابی ڈیٹا سیٹ کا تجزیہ کیا جاتا ہے اور اس سے بامقصد بصیرت حاصل کرنے اور زیادہ سے زیادہ آؤٹ پٹ حاصل کرنے کے لئے ڈیٹا مائننگ

الگورتھم کے لئے ایک مناسب شکل میں تیار کیا جاتا ہے۔

ڈیٹا ٹاسک فلٹر: اس مرحلے میں بعد میں آنے والے مراحل میں انتخابی پیش گوئی کے متوقع نتائج کا تعین کرنے کے لئے انکشافی نتائج کے قواعد استعمال کیے جاتے ہیں۔ اس کے بعد منتخب ڈیٹا سیٹ کو ڈیٹا مائننگ ٹاسک اسٹو میں محفوظ کیا جاتا ہے۔

ڈیٹا مائننگ کی تکنیک: اس مرحلہ کے لیے تیسرے مرحلہ میں مجوزہ کام کے لئے موزوں ڈیٹا سیٹ کے ساتھ ایک مناسب الگورتھم منتخب کیا جاتا ہے۔

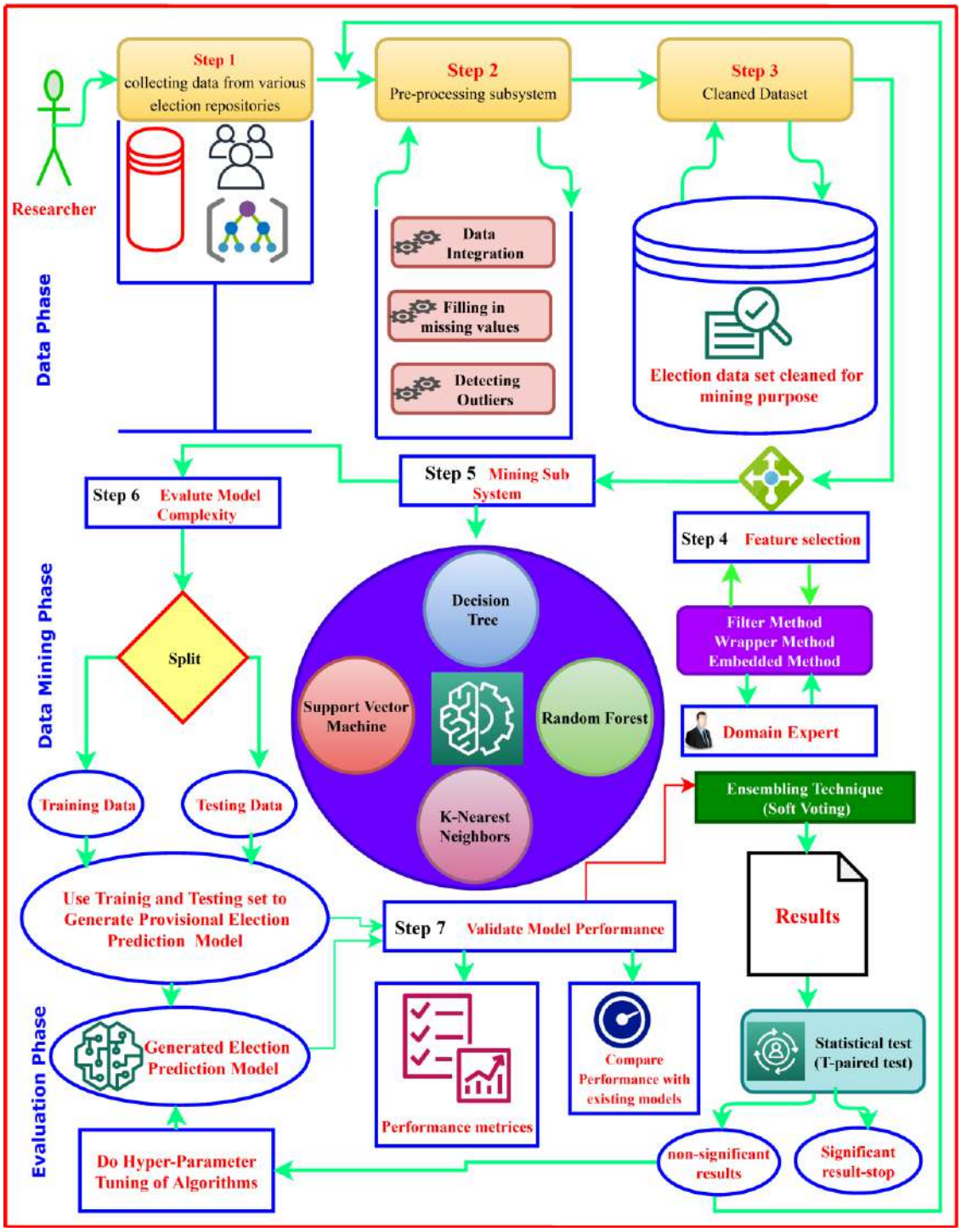
موازنہ اور تشخیص: اس مرحلے میں، ڈیٹا مائننگ کی مختلف تشخیصی کسوٹی کی بنا پر درجہ بند نتائج کو باہم موازنہ کیا جاتا ہے اور تخمینہ لگایا جاتا ہے۔

نئے ماڈلز کی تعمیر: اس مرحلے میں، بڑی احتیاط سے تیار درجہ بندی کے ماڈل اگلی پیشن گوئی کی پریشانیوں کے حل کے لیے ڈیٹا مائننگ اسٹور میں محفوظ کیے جاتے ہیں۔ پیشن گوئی کے نئے منصوبوں کے لئے مختلف عمل تیسرے مرحلے سے لے کر پانچویں مرحلہ تک دہرائے جاتے ہیں۔ اس تحقیقی کام میں اس مخصوص طریقہ کار کو استعمال کرنے کی وجہ یہ ہے کہ وہ ہمارے تحقیقی مقاصد کو واضح کرتا ہے۔

4.2 انتخابی پیشن گوئی ماڈل کے لئے ریسرچ ڈیزائن

اس تحقیقی کام میں ہم مقاصد کے بیان سے تحقیقی سرگرمیوں کو آسان بنانے اور تحقیق کو نتیجہ خیز بنانے کے لئے تحقیقی ڈیزائن کی پیروی کرتے ہیں۔ یہ تحقیقی ڈیزائن تحقیقی مسائل کو حل کرنے کے لئے، مطلوبہ اعداد و شمار اور تجزیے میں استعمال کیے جانے والے طریقوں کی جانکاری فراہم کرتا ہے۔ مجوزہ تحقیقی ڈیزائن ذیل میں دیئے ترسیم 4.2 میں دکھایا گیا ہے جس کی جموں و کشمیر (انتخابی حلقہ وار) کے لئے انتخابی پیشن گوئی کے ماڈل کی تشکیل کے لئے مرحلہ وار پیروی کی جاتی ہے۔ تحقیقی ڈیزائن میں تین اہم دور ہوتے ہیں جن میں آٹھ مراحل ہوتے ہیں۔ تحقیقی ڈیزائن کے ادوار مع مراحل کی وضاحت مندرجہ ذیل ہے۔

i. **ڈیٹا فیز:** ڈیٹا فیز میں ڈیٹا اکٹھا کرنے سے لے کر اس میں موجود خصوصیت کا باریکی سے مطالعہ کرکے اسے کارآمد شکل میں لانے تک کا سارا عمل ہوتا ہے۔ اس مرحلے میں کوالٹیٹو ڈیٹا اکٹھا کرنے کا عمل، تحتی نظام کا قبل از پروسینگ مرحلہ، چھوٹے ہوئے اعداد و شمار کی زخیرہ اندوزی، اور آخر میں فیچر سلیکشن کا مرحلہ شامل ہے۔



ترسیم 4.2: انتخابی پیش گوئی کے لئے ریسرچ ڈیزائن

ii. **ڈیٹا مائننگ کا دور:** ڈیٹا مائننگ کے دور میں مشین لرننگ کلاسیفائیرز جیسے کے سب سے قریبی پڑوسی، ڈسیزن ٹری، سپورٹ ویکٹر مشین، رینڈم فاریسٹ شامل ہیں اور آخر میں انھیں ایک درجہ بندی میں شامل کیا جاتا ہے جو انتخابی پیش گوئی کے ماڈل کو تیار کرنے کے لئے بنایا جاتا ہے۔

iii. **ماڈل کی تشخیص اور توثیق کا مرحلہ:** ماڈل کی تشخیص اور توثیق کا مرحلہ ڈیٹا مائننگ کے مختلف تکنیکوں کا استعمال کرتے ہوئے تشخیصی ماڈل کی کارکردگی کا حساب کتاب اور توثیق کرتا ہے۔ انتخابی پیشن گوئی کا ماڈل انتخابی اعداد و شمار کی ٹرین ٹیسٹ تقسیم کے ذریعے 10 فولڈ کراس ویلیڈیشن کا استعمال کرتے ہوئے تشخیص کی جاتی ہے۔ اس کے بعد ماڈل کی توثیقی کارکردگی کے مختلف پیمانوں کا استعمال کرکے موجودہ ماڈل اور مختلف میٹرکس موڈل (حساسیت، وضاحتی، درستگی، غلط طبقے کی شرح، اور آر او سی اسکور) کے ساتھ نتائج کا موازنہ کرکے کی جاتی ہے۔ پھر آخر میں، اہمیت کی سطح کو جانچنے کے لئے ایک اعداد و شمار ٹیسٹ جیسے ٹی ٹیسٹ کو عمل میں لایا جاتا ہے۔

iv. **علمی بنیاد کا دور:** علمی بنیاد پر مبنی مرحلے میں انتخاب کی پیشن گوئی کے بارے میں معلومات کا ذخیرہ کرنے اور بازیافت کرنے کے اقدامات شامل ہیں۔ رائے شماری کی پیشن گوئیوں کے لئے ضروری مہارت کے

مطابق انتخابی پیشن گوئی ماڈل کے تیار کردہ اصول و قواعد نالج بیس رکھے جائیں گے اور پول پسین گوئی کے لیے معروف تجربہ کاروں کے مطابق باہم جانچے جائیں گے۔

4.3 ایکسپلوریٹری ڈیٹا انیلیسیس (ای ڈی اے) عمل (Exploratory Data Analysis (EDA)

:(Process

یہ بنیادی شماریاتی تفصیل جموں و کشمیر کے انتخابی ڈیٹا سیٹ کے ہر ایک وصف کی اہمیت کے بارے میں جاننے کے لئے عمل میں لائی گئی ہے۔ ہر ایک وصف کے بارے میں اس طرح کے بنیادی اعداد و شمار کو جاننے سے بڑی بڑی اقدار کو ہلکا کرنے، حدود کا پتا لگانے، گمشدہ اقدار کو پر کرنے میں مدد ملتی ہے۔ انتخابی پیشن گوئی کے اعداد و شمار نامی اور ہندسی وصف کے امتزاج پر مشتمل ہوتے ہیں۔ لاپتہ عددی اقدار کو سیدھے سادے مین امپیوٹیشن (mean imputation) طریقہ کار کے ذریعہ ختم کیا جاتا ہے، اور درجتا گم شدہ اقدار کو موڈا مپیوٹیشن (mode imputition) کے ذریعہ پُر کیا جاتا ہے۔

4.3.1 ڈیٹاسیٹ میں درجہ کے عدم توازن اور ڈیٹا کی تقسیم کے مسائل کی تحقیقات

(Checking Class Imbalance and Data Distribution Problems in Dataset):

انتخابی ڈیٹاسیٹ پر کسی بھی کارروائی کو انجام دینے سے پہلے طبقاتی توازن کی تحقیقات کرنے کی اشد ضرورت ہے۔کیونکہ انتہائی عدم توازن والے ڈیٹا مشین لرننگ الگورتھم کو جانب دار بنا دیتے ہیں۔ ڈیٹاسیٹ میں مختلف سیاسی جماعتوں کے 3776 ریکارڈ موجود ہیں، جنہوں نے سال 2002 سے 2014 تک جموں و کشمیر میں الیکشن لڑا تھا۔ لہذا، 2002 کے انتخابات سے لے کر 2014 کے ریاستی اسمبلی انتخابات کے اسمبلی انتخابات کے نتائج کے پیش نظر، محقق جموں و کشمیر کے لئے انتخابی پیش گوئی کا نمونہ تیار کر رہا ہے تاکہ وہ متعصب نتائج کی بجائے درست نتائج کی پیش گوئی کر سکے۔ محقق نے 2002 سے 2014 کے نتائج کے پچھلے تین شرائط کے لئے ہندوستان کے انتخابی کمیشن اور جموں و کشمیر کے الیکشن کمیشن سے حاصل کردہ معلومات کو جمع کیا۔ کچھ ڈیٹاسیٹ جموں و کشمیر کے مختلف حلقوں کے فیلڈ سروے کے طور پر اکٹھا کیا گیا ہے۔ڈیٹاسیٹ میں مختلف خصوصیات کے ساتھ جمع کیا گیا تھا State, constituency, Account No, Gender, Caste, Party wave, Religion

followers, Caste factor, Hereditary, Vote-bank, Central Government Influence, Party Abbreviations, Sensitive areas, Votes polled, Votes -majority and finally Party Won ڈیٹاسیٹ جو https://eco.Gov.in سے حاصل کیا گیا ہے

ہم 2002 کے اسمبلی انتخابات سے لے کر 2014 کے اسمبلی انتخابات تک اس تحقیقی کام میں انتخابی ڈیٹاسیٹ کو استعمال کر رہے ہیں کیونکہ 2002 کے اسمبلی انتخابات سے قبل جموں اور کشمیر نیشنل کانفرنس (جے کے این) نے جموں و کشمیر میں زیادہ تر اسمبلی انتخابات جیتے تھے[174] اور اسی وجہ سے وہ ماڈل کو جانب دار نتائج پیش کرنے کی طرف راغب کرتا ہے لیکن 2002 کے بعد دیگر پارٹیاں جیسے جموں و کشمیر پیپلز ڈیموکریٹک پارٹی (جے کے پی ڈی پی) اور بھارتیہ جنتا پارٹی (بی جے پی) نے جموں و کشمیر میں حکومتیں تشکیل دینا شروع کر دی، ساتھ ہی جموں و کشمیر نیشنل کانفرنس (جے کے این) اور انڈین نیشنل کانگریس (آئی این سی) نے بھی جیسا کہ جدول 4.3.1 میں دکھایا گیا ہے۔

جدول 1.4.3.1 2002 سے 2014 تک اہم سیاسی جماعتوں کی کارکردگی

S.NO.	Years	Name of Party	Seats won	Governments formed
1	2002	JKN	28	INC+ JKPDP
2	2002	JKPDP	16	
3	2002	INC	20	
4	2002	BJP	01	
5	2002	Other	22	
6	2008	JKN	28	INC+ JKN
7	2008	JKPDP	21	
8	2008	INC	17	
9	2008	BJP	11	
10	2008	Other	10	
11	2014	JKN	15	JKPDP+BJP
12	2014	JKPDP	28	
13	2014	INC	12	
14	2014	BJP	25	
15	2014	Other	07	

باب 5

5. نفاذ اور نتائج

اس باب میں، ہم انتخابات کی پیشن گوئی ماڈل کے نتائج اور اس کے نفاذ پر بحث کریں گے۔انتخاب کے مختلف متغیر پیمانوں کے مابین نسبت تلاش کرنے کے لئے پیئرسن کے باہمی تعلق کی مکرر قدر کا استعمال کیا جاتا ہے۔فیچر سلیکشن کی تکنیک کے ذریعہ حاصل کردہ نتائج کا موجودہ نتائج کے ساتھ نقل اور توثیق کیا جاتا ہے۔

5.1 مختلف انتخابی پیمانوں کے مابین باہمی تعلق کا پتہ لگانا

کسی بھی ڈیٹاسیٹ میں، صفات کے مابین کثیر الجہتی اور عجیب وغریب تعلقات ہو سکتے ہیں۔ لہذا، اعداد و شمار میں ایک دوسرے سے وابستہ ہونے والی مقدار کا تعین اور اس کی پیمائش ضروری ہے۔ڈیٹاسیٹ پیمانوں کے مابین تعلقات کی مقدار تلاش کرنے کے اس عمل کو ارتباط کے نام سے جانا جاتا ہے۔ متغیر پیمانوں کے مابین ارتباط کا علم علوم آلات الگورتھم کی توقعات کو پورا کرنے کے لئے ڈیٹا تیار کرنے میں مدد کرتا ہے۔پیئرسن کا باہمی تعلق انتخابی پیشن گوئی کے پیمانوں کے مابین باہمی

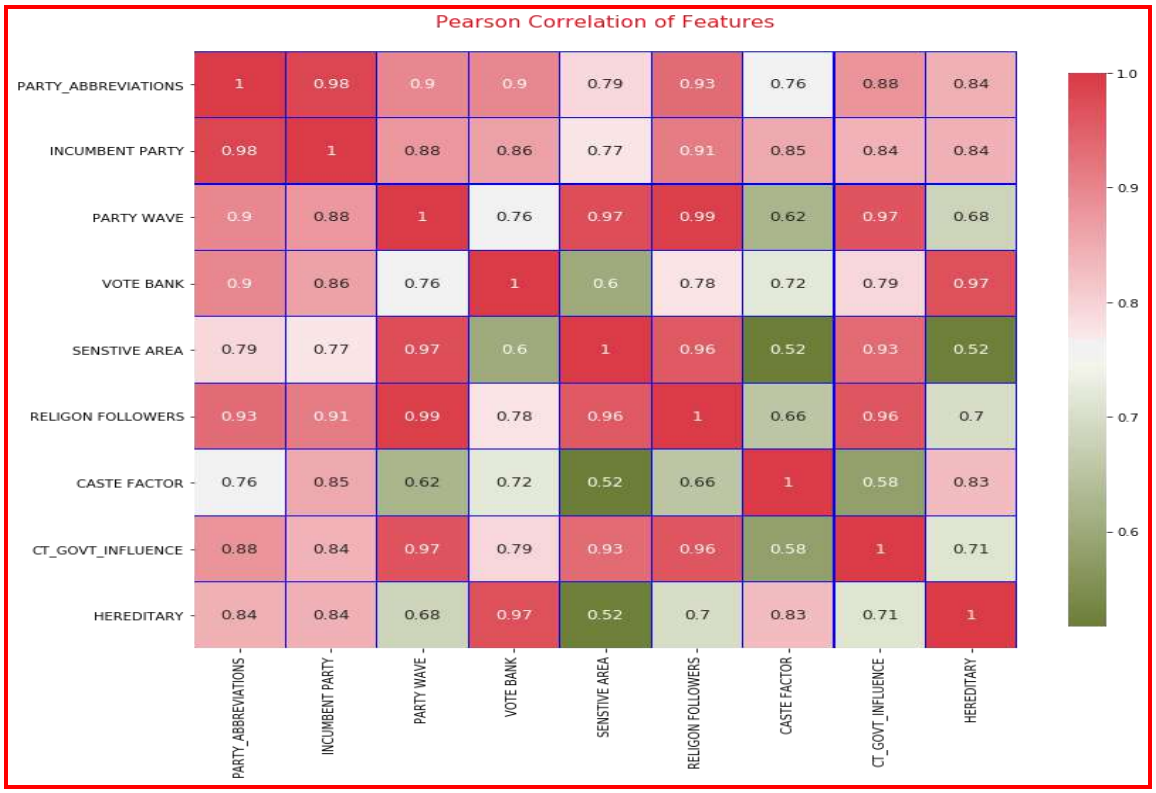
تعلقات کو جانچنے کے لئے لاگو ہوتا ہے۔ ایک ارتباط مثبت بھی ہو سکتا ہے (جس کا مطلب ہے کہ تمام متعلقہ متغیر پیمانے ایک ہی سمت میں چلتے ہیں)، منفی (جس کا مطلب ہے کہ تمام متعلقہ متغیر پیمانے مخالف سمتوں میں منتقل ہوتے ہیں) یا غیر جانبدار (جس کا مطلب ہے کہ متغیر پیمانے ایک دوسرے سے متعلق نہیں ہیں)۔ انتخابی پیمانوں میں لاگو پیئرسن کے باہمی تعلق کے نتائج کو ہاٹ میپ حناکہ کی شکل میں نیچے دیئے گئے ترسیم 5.1 میں دکھایا گیا ہے۔

ہیٹ میپ حناکہ انتخابی پیمانوں کے مابین ان کے متعلقہ مکرر قدر کے ساتھ ارتباط کی نمائندگی کرتا ہے۔ متوازی ہیٹ میپ سانچہ تمام جوڑے کی خصوصیات کے مابین ارتباط دینے کے لئے اوپر اور نیچے کی سمت تمام پیمانوں کی نمائندگی کرتی ہے۔ نیچے دائیں کونے سے اوپر بائیں تک میٹرکس کے اس خط قتری ہر پیمانوں کا خود سے کامل ارتباط کی نمائندگی کرتی ہے۔ قدر 1 کا مطلب پیمانوں کے درمیان ایک کامل مثبت ارتباط ہے اور قدر -1 جس کا مطلب انتخابی پیشن گوئی کے متغیر پیمانوں کے درمیان ایک کامل منفی تعلق ہے۔ صفر کے قریب ارتباطی مکرر قدر انتخابی پیشن گوئی کے متغیر پیمانوں میں کمزور انحصار کی نشاندہی کرتا ہے۔ ہیٹ میپ سے وابستگی کے نتائج کا تجزیہ کرنے کے بعد، یہ پتہ

چلا ہے کہ جموں و کشمیر کے انتخابی ڈیٹا سیٹ کے آزاد متغیر پیمانے ایک دوسرے کے ساتھ مضبوطی

سے وابستہ ہیں۔تاہم، اگر کسی ڈیٹا سیٹ میں متغیر پیمانے مضبوطی سے منسلک ہوتے ہیں تو پھر ایک

متغیر قدر میں تبدیلی سے دوسرے میں تبدیل ہو سکتی ہے۔



ترسیم 5.1: ہیٹ میپ نمائندگی کے ذریعہ انتخابی پیشن گوئی کے متغیر پیمانوں میں باہمی تعلقات

پیمانوں کے درمیان باہمی تعلق کا مطلب تسبیب نہیں ہے، پیمانوں کے مابین مضبوط تعلقات کو نمایاں

طور پر جانچنا چاہئے۔ زیادہ تر، پیمانوں کے مابین تعلق کچھ فرو گزاشت عوامل کی وجہ سے مضبوط ارتباط کے ذریعے کار باعث سبب لگ سکتے ہیں۔

5.2 انتخابات کی پیش گوئی کے لئے فیچر سلیکشن کی تکنیک

انتخابی نتائج کی درست پیش گوئی کے لئے انتخابی حلقوں کی سطح پر اہم اور مناسب خصوصیات کا انتخاب کرنے کے لئے خصوصیت کے انتخاب کی تکنیک کا استعمال کیا جاتا ہے۔۔ فیچر سلیکشن نامناسب اور بیکار اوصاف کو کم کرنے میں مدد کرتا ہے، جو اکثر ماڈل کی کارکردگی کو گھٹا دیتے ہیں۔ اس تحقیقی کام میں، فیچر سلیکشن کے مختلف طریقوں کا اطلاق کیا جاتا ہے جو الگ کرنے کا طریقہ، ستہ کرنے کا طریقہ اور سرایت شدہ طریقہ ہیں، تاکہ انتخاب کی پیشن گوئی کے ماڈل کے لے ایک مناسب خصوصیت کا ذیلی سیٹ حاصل کیا جا سکے۔ یہ فیچر سلیکشن کی تکنیک انتخاب کی پیشن گوئی میں ان کے کردار کے مطابق ہر خاصیت میں وزن میں ڈالتی ہیں۔ اطلاقی فیچر سلیکشن کی تکنیک انتخابی پیشن گوئی کے ہر پیمانوں کے لئے 0 سے 1 کے پیمانے کے درمیان وزن ڈالتی ہے۔ علیحدہ فیچر سلیکشن تکنیک کے ذریعہ ہر پیمانے کو وزن تفویض کرنے کے بعد، ان فیچر سلیکشن کی تکنیک کے ذریعہ ہر پیمانے

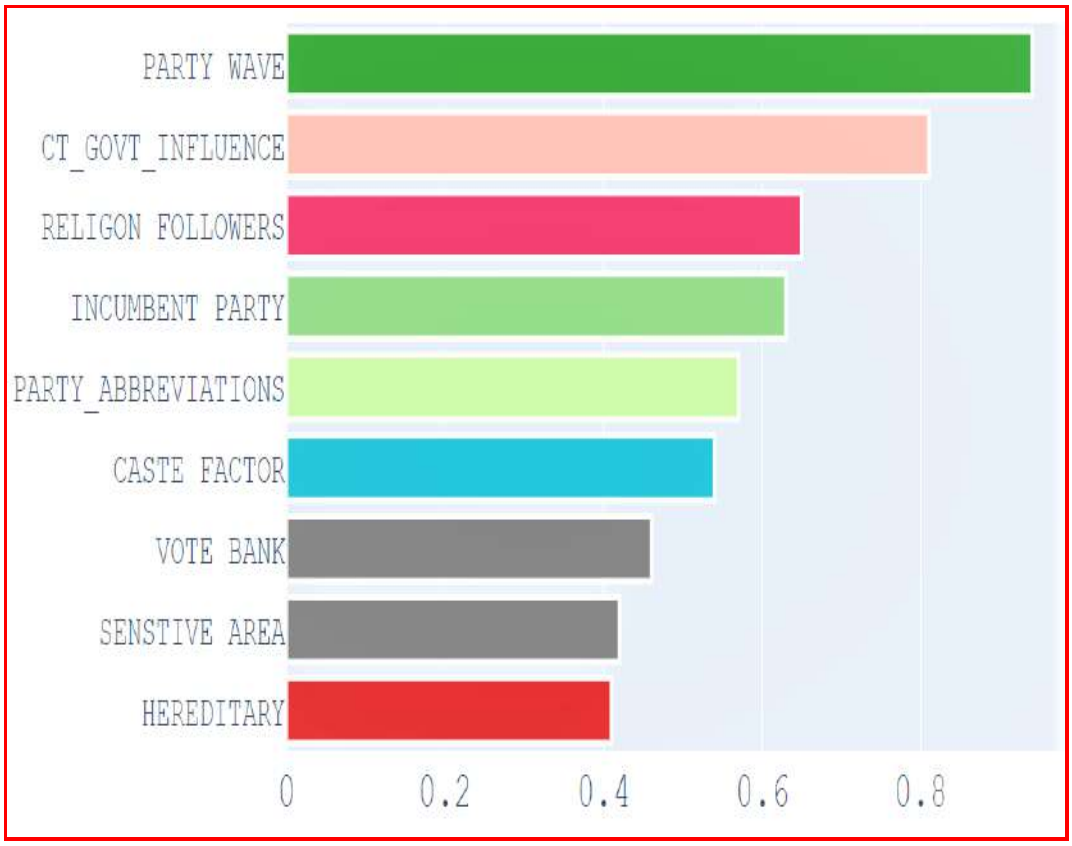
کے تمام لاگو وزن کا مجموعی مطلب آخری وزن سمجھا جاتا ہے۔ انتخابی نتائج کی پیشن گوئی کرنے میں 1 کے قریب پیمانوں کو اہم سمجھا جاتا ہے، اور وہ پیمانے جن کی وابستہ اقدار 0 کے قریب ہیں انتخابی نتائج کی پیشن گوئی کرنے میں کم اہم سمجھے جاتے ہیں جموں و کشمیر حلقہ وار کے انتخابی نتائج کی پیش گوئیاں کرنے میں۔ ذیل میں دیئے گئے جدول 5.2.1 میں انتخاب کی پیشن گوئی کی مختلف خصوصیات ان کے متعلقہ وزن کے ساتھ ظاہر کی گئیں ہیں جو مختلف فیچر سلیکشن کی تکنیکوں کے ذریعے تفویض کی گئی ہیں اور جدول 5.1 میں آخری کالم تمام تکنیکوں کا مجموعی مطلب ظاہر کرتا ہے۔

جدول 5.2.1 فیچر سلیکشن کی تکنیک ہر خصوصیت کو وزن مہیا کراتی ہے

	Filter_Method	Embedded_Method	Wrapper_Method	MEAN
PARTY_ABBREVIATIONS	0.72	0.50	0.50	0.57
INCUMBENT PARTY	0.23	0.77	0.88	0.63
PARTY WAVE	0.98	0.83	1.00	0.94
VOTE BANK	0.00	0.75	0.62	0.46
SENSTIVE AREA	1.00	0.00	0.25	0.42
RELIGON FOLLOWERS	0.99	0.84	0.12	0.65
CASTE FACTOR	0.39	0.84	0.38	0.54
CT_GOVT_INFLUENCE	0.69	1.00	0.75	0.81
HEREDITARY	0.36	0.86	0.00	0.41

ان انتخابی پیشن گوئی کی صفات کی نشاندہی بہت سے انتخابی تجزیہ نگاروں کے ذریعہ کی گی ہے۔ جیسے رنجیت سنگھ کارلا (اسسٹنٹ پروفیسر)، اسد نعمانی (سماجی کارکن) اور بہت سارے دوسرے عام ماہرین، جو مختلف سطح پر ہندوستان بھر میں رائے شماری کی پیشن گوئی میں کام کر رہے ہیں۔ ہر انتخابی پیشن گوئی کے پیمانوں کو تفویض شدہ وزن کی مختلف حلقہ کے ماہرین جیسے ڈاکٹر افروز عالم (صدر شعبہ سیاسیات مانو حیدر آباد انڈیا) نے توثیق کی ہے اور اس کی منظوری دی ہے۔ اسی اثنا میں ان انتخابی پیشن گوئی کے حلقہ کے ماہرین نے انتخابات کی پیشن گوئی کے لیے کچھ اہم پیمانوں جیسے ترقیاتی ایجنڈا اور صنف عنصر کو شامل کرنے کے لئے اپنی متعلقہ آراء پیش کی ہیں۔

نتائج کے تجزیے سے پتہ چلتا ہے کہ یہ خصوصیات [پارٹی لہر، مذہب کے پیروکار، عہدیدار پارٹی، ووٹ بینک، مرکزی حکومت کا اثر و رسوخ، وراثت، پارٹی کی تخفیف کا عمل، حاس علاقے اور ذات پات] ان کی اعلیٰ متعلقہ عددی اقدار کی وجہ سے انتخابی نتائج کی ابتدائی پیشن گوئی کے لئے سب سے اہم پیمانے ہیں۔ ان نتائج کو رائے دہندگی کے مختلف پیشن گوئوں سے بھی توثیق اور منظوری ملی ہے۔ متعلقہ وزن کے ساتھ ان کی انتسابی درجہ بندی کی مصورانہ نمائندگی ذیل کی ترسیم 5.2 میں دکھائی گئی ہے۔



ترسیم 5.2: فیچر سلیکشن کی تکنیک کے ذریعہ انتخاب کی پیشن گوئی کی خصوصیت کی درجہ بندی

ذیل میں دیئے گئے جدول 5.2.2 میں انتخاب کی پیشن گوئی کے پیمانوں کی نزولی ترتیب کو ان کی اوسط اقدار کے مطابق دکھایا گیا ہے جو خصوصیت کے انتخاب کی تین مختلف تکنیکوں کے ذریعہ تفویض کی گئی ہیں۔ سب سے زیادہ وزن والے پیمانے انتہائی اہم ہیں، اور انتخابی نتائج کی پیشن گوئی کرنے میں کم اقدار والی صفات کم اہم ہیں۔ انتخابی پیشن گوئی کے ماڈل کو تیار کرنے کے لئے منتخب پیمانوں میں سے انتہائی وزن والا اہم ذیلی سیٹ استعمال کیا جاتا ہے۔

جدول5.2.2:فیچر سلیکشن کی تکنیک کے ذریعہ انتخابی پیشن گوئی کے اوسطوں کی درجہ بندی

	Feature	Mean Ranking
1	PARTY WAVE	0.94
2	CT_GOVT_INFLUENCE	0.81
3	RELIGON FOLLOWERS	0.65
4	INCUMBENT PARTY	0.63
5	PARTY_ABBREVIATIONS	0.57
6	CASTE FACTOR	0.54
7	VOTE BANK	0.46
8	SENSTIVE AREA	0.42
9	HEREDITARY	0.41

5.3 مجوزہ ڈیٹا مائننگ کی تراکیب کے تجرباتی نتائج

یہ پایا گیا ہے کہ مروجہ انتخابی پیشن گوئی ماڈل نقص سے پاک نہیں ہیں کیونکہ وہ مختلف ڈیٹا سیٹوں میں مختلف نتائج دکھاتے ہیں جو نظام کی تاثیر کو بہت حد تک کم کردیتے ہیں۔اس تحقیق میں، جموں و کشمیر کے انتخابی ڈیٹاسیٹ کی ڈسیزن ٹری، کے نیرسٹ نائبر، رینڈم فاریسٹ، اور سپورٹ ویکٹر مشین الگورتھم کے ذریعہ تحقیق کی گئی ہے۔پیشن گوئی ماڈل کی کارکردگی کو بہتر بنانے کے لئے سافٹ ووٹنگ اسیمبلنگ تکنیک استعمال کی جاتی ہے۔حساسیت،وضاحت،درستگی، باقاعدگی،AUROCاسکور،

اور غلط طبقے کی شرح اور ماڈل پیمانے جیسے شماریاتی تکنیک کا اعداد و شمار اقدامات کی اہمیت کی سطح کو جانچنے کے لئے حساب کیا جاتا ہے۔ مختلف ذیلی حصوں کے نیچے انتخابی پیشن گوئی کے مختلف ماڈلز کے ذریعہ حاصل کردہ تجرباتی نتائج کی وضاحت کی گئی ہے۔

5.3.1 ڈسیزن ٹری ماڈل کے تجرباتی نتائج (Decision Tree Model Experimental Results)

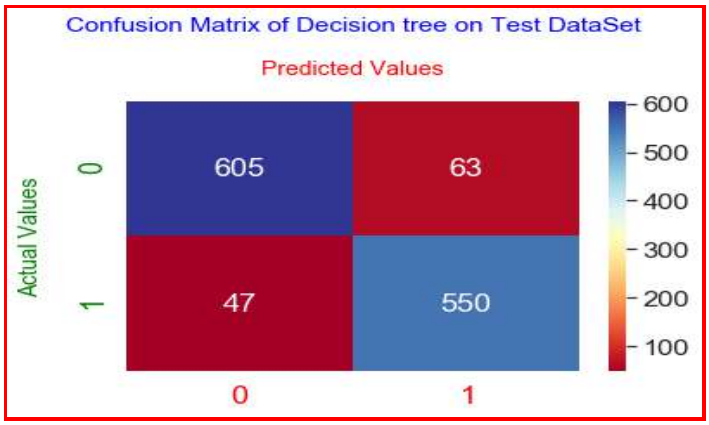
ڈسیزن ٹری کو استعمال کرنے کی دلیل انتخابی پیشن گوئی کا ماڈل تیار کرنا ہے جو تربیتی ڈیٹا سیٹ سے کٹوتی فیصلے کے قواعد سیکھ کر انتخابی نتائج کی پیش گوئی کر سکتا ہے۔۔ٹریننگ ڈیٹا سیٹ پر کراس ویلیڈیشن غیر جانبدارانہ نتائج حاصل کرنے کے لئے استعمال کی جاتی ہے۔ڈسیزن ٹری ماڈل کی کارکردگی کے نتائج کنفیوژن میٹرکس کے ترسیم 5.3.1.1 میں دکھائے گئے ہیں۔ڈسیزن ٹری ماڈل کے کنفیوژن میٹرکس (ترسیم 3.7.1) سے حساسیت، صراحت، درستگی، صحت سے متعلق، اور غلط درجہ بندی کی شرحیں اخذ کی گئی ہیں جن کو بیان کیا گیا ہے۔

سیاسی جماعتوں اور آزاد امیدواروں کی فیصد جو صحیح طور پر انتخاب میں کامیابی حاصل کرنے

کے لئے تسلیم کی گئی تھی (حقیقی مثبت) سیاسی جماعتوں اور آزاد امیدواروں کی کل تعداد پر انتخاب ہوا جس نے حقیقت میں کامیابی حاصل کی ہے۔ اسے حساسیت کے نام سے جانا جاتا ہے۔ کنفیوژن میٹرکس ترسیم 5.3.1.1 کی ماخوذ حساسیت کی اقدار کو مساوات 3.11 میں ڈالنا 92٪ کی حساسیت حاصل کی جاتی ہے۔ جتنا ہی اس اقدام کی قدر 1 کے قریب ہے، اتنا ہی سیاسی جماعتوں یا آزاد امیدواروں کی نشاندہی کرنے میں قواعد بہتر ہوں گے جنہوں نے انتخابات میں کامیابی حاصل کی ہے۔

اسی طرح، سیاسی جماعتوں اور آزاد امیدواروں کی فیصد جو صحیح طور پر تسلیم کی گئی تھی کہ وہ انتخابات میں ہاریں گے (حقیقی منفی) جن سیاسی جماعتوں اور آزاد امیدواروں نے انتخابات میں کامیابی حاصل نہیں کی ہے ان کی کل تعداد پر، وضاحت کے نام سے جانا جاتا ہے۔ کنفیوژن میٹرکس کے ترسیم کی ماخوذ اقدار کو مساوات 5.3.1.1 میں ڈالنا 90 فیصد کی خصوصیت حاصل کی جاتی ہے جس کا مطلب ہے کہ ڈسیزن ٹری ماڈل ان سیاسی جماعتوں یا آزاد امیدواروں کو پہچان سکتا ہے جو 90 فیصد کی درستگی کے ساتھ انتخابات میں کامیابی حاصل نہیں کرتے ہیں۔ جتنا اس پیمائش کی قدر

1 سے قریب ہے اتنے ہی بہترین قواعد بغیر نقص کے ان سیاسی پارٹیوں یا آزاد امیدواروں کی نشاندہی کرنے میں ہیں ۔ ڈسیزن ٹری ماڈل کی مجموعی درستگی ترسیم 5.3 اور مساوات 3.13 کو استعمال کرکے حاصل کی گئی ہے جو 91 فیصد کے برابر ہے جو نمائندگی کرتی ہے کہ جیتنے والے اور انتخابات ہارے ہوئے دونوں کی پیشن گوئی کرنے میں ڈسیزن ٹری انتخابی پیشن گوئی ماڈل کی مجموعی کارکردگی 91 فیصد ہے۔ جتنی زیادہ فیصد کی درستگی ہوگی، ماڈل اتنا ہی درست ہوگا۔



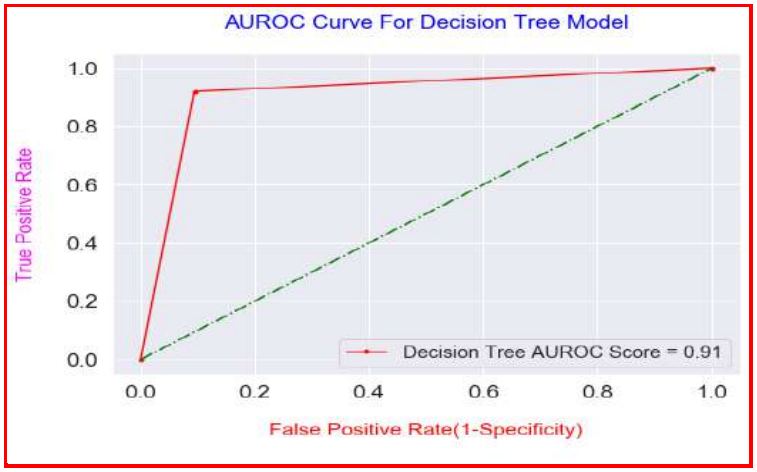
ترسیم 5.3.1.1 ڈسیزن ٹری ماڈل کنفیوژن میٹرکس

اسی طرح، کنفیوژن میٹرکس کے ترسیم 5.3.1.1 کو مساوات 3.14 میں ڈالنے سے، 89 فیصد کی صحت حاصل کی جاتی ہے۔ جتنی اس پیمائش کی قدر 1 سے قریب ہے، اس کا امکان اتنا زیادہ ہوگا کہ

مثبت نتائج والوں نے حقیقت میں انتخابات میں کامیابی حاصل کی ہوگی۔ اگر ڈسیزن ٹری ماڈل کی اعلیٰ صحت سے متعلق شرح حاصل کی جاتی ہے، تو اس کا مطلب یہ ہے کہ ماڈل کم غلط مثبت شرح وصول کرے گا۔ ترقی یافتہ فیصلہ ڈسیزن ٹری ماڈل کی غلطی کی شرح کنفیوژن میٹرکس کے ترسیم 5.3.1.1 کی مساوات 3.15 میں ڈال کر حاصل کی جاتی ہے، جو 8 فیصد کے برابر ہے۔ ماڈل کی غلط درجہ بندی کی شرح جتنی کم ہے، جموں و کشمیر کے ہر حلقے کے انتخابات جیتنے یا ہارنے کی نشاندہی کرنے میں ماڈل اتنا ہی درست ہے۔

AUROC کارکردگی کی پیمائش کا استعمال ڈسیزن ٹری الگورتھم کے ذریعہ حاصل کردہ امتیازی کرو اور علیحدگی کی پیمائش کی جانچ کے لئے کیا جاتا ہے۔ AUROC واضح کرتا ہے کہ ماڈل کتنے موثر انداز میں سیاسی جماعتوں یا آزاد امیدواروں کے انتخابی حلقے سے انتخاب جیتنے اور ہارنے میں فرق کر سکتا ہے۔ AUROC کرو 0.0 اور 1.0 کے درمیان مختلف امیدواروں کی حد اقدار کے لئے غلط مثبت شرح (X-axis) کے خلاف حقیقی مثبت شرح (Y-axis) کا پلاٹ ہے۔ نیچے دیئے گئے ترسیم 5.3.1.2 ڈسیزن ٹری ماڈل کی AUROC ہے۔ AUROC اسکور = 0.91 فیصد کے ساتھ۔ غیر حقیقی مثبت کے خلاف

حقیقی مثبت پلاٹ لگانے والے ارتباط 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 موجودہ تحقیق کے ساتھ تیار شدہ ڈسیزن ٹری انتخابی پیشن گوئی ماڈل کے کامیاب تجرباتی نتائج کی نقالی کرتے ہیں۔ حاصل کردہ نتائج ہمارے نزدیک بہترین علم کے حامل ہیں جو لٹریچر میں شائع شدہ نتائج سے کہیں زیادہ بہتر ہیں۔

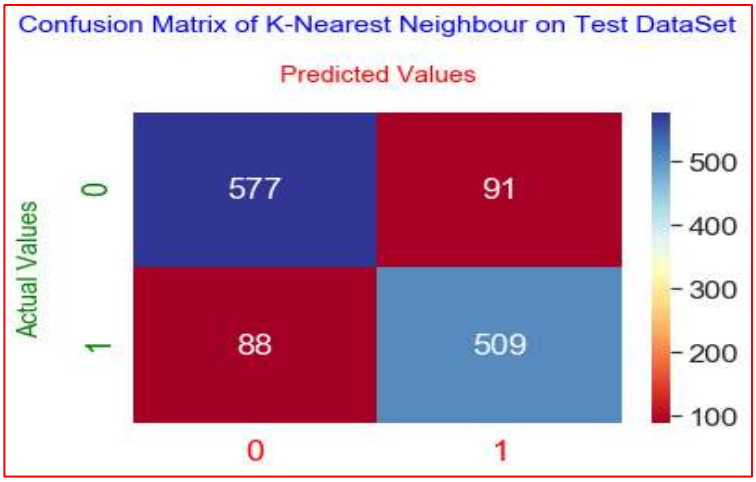


ترسیم 5.3.1.2: ڈسیزن ٹری ماڈل کی AUROC کرو

5.3.2 کے نیرسٹ نائبر ماڈل تجرباتی نتائج (K-NN Model Experimental Results):

کے نیرسٹ نائبر الگورتھم کو استعمال کرنے کا مقصد انتخابی پیشن گوئی کا ماڈل بنانا ہے جو انتخابی نتائج کی

جلد سے جلد پیشن گوئی کر سکتا ہے۔ ٹریننگ ڈیٹا پر 10 فولڈ کراس ویلیڈیشن کا استعمال زیادہ سے زیادہ اور غیر جانبدارانہ نتائج حاصل کرنے کے لئے کیا جاتا ہے۔ حاسیت، وضاحتی، درستگی، صحت سے متعلق اور K-NN ماڈل کے غلط درجہ بندی کی شرح جیسے کارکردگی کے نتائج کنفیوژن میٹرکس کے ترسیم 5.3.2.1 سے اخذ کیے گئے ہیں۔ ۸۵ فیصد کی حاسیت ترسیم 5.5 میں مساوات 3.11 کا استعمال کرکے حاصل کی گئی ہے، جس کا مطلب ہے کہ کے نیرسٹ نائبر ماڈل جیتنے والی سیاسی جماعتوں اور آزاد امیدواروں کو 85 فیصد کی درستگی کے ساتھ پہچان سکتا ہے۔ اسی طرح، مختلف سیاسی جماعتوں اور آزاد امیدواروں کے ذریعہ ہارے گئے انتخابات کی مقدار جو ترسیم 5.3.2.1 میں وضاحتی مساوات 3.12 کا استعمال کرکے 86 فیصد ہے جسے شکست کے طور پر تسلیم کیا گیا ہے۔ اس کا مطلب ہے کہ کے نیرسٹ نائبر ماڈل 86 فیصد کی درستگی کے ساتھ انتخاب ہار جانے کو پہچان سکتا ہے۔

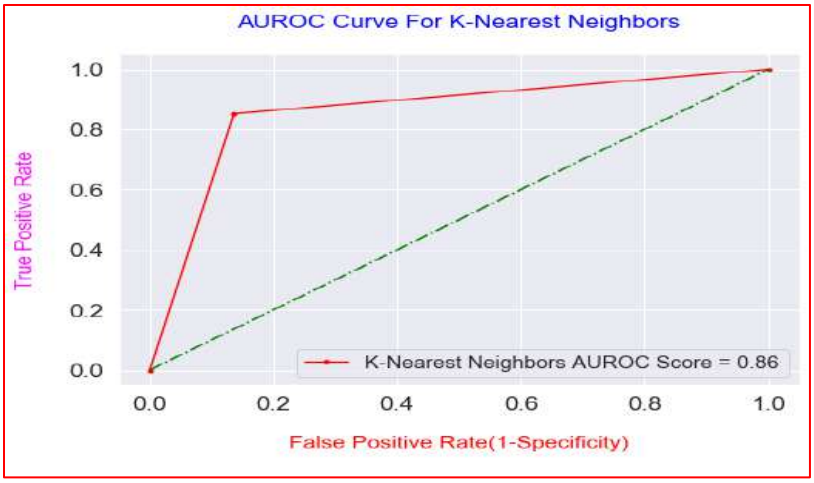


ترسیم 5.3.2.1: کے نیئرسٹ نائبر کنفیوژن میٹرکس ٹیسٹ ڈیٹاسیٹ پر

کے نیئرسٹ نائبر ماڈل کی مجموعی طور پر درستگی کنفیوژن میٹرکس میں ترسیم 5.5 کو مساوات 3.13 میں ڈالنے سے حاصل کیا جاتا ہے جو 85 فیصد کے برابر ہے اس کا مطلب یہ ہے کہ جیتنے اور ہارنے دونوں کے معاملات کا تعین کرنے میں کے نیئرسٹ نائبر انتخابی پیشن گوئی ماڈل کی مجموعی کارکردگی کی درستگی 85 فیصد ہے۔ اسی طرح، کے نیئرسٹ نائبر ماڈل کی مساوات 3.14 درستگی کا استعمال کرتے ہوئے اس کا حساب لگایا جاتا ہے جو 84 فیصد کے برابر ہے۔ اگر کے نیئرسٹ نائبر ماڈل کی اعلی صحت سے متعلق شرح حاصل کی گئی ہے تو، اس کا مطلب یہ ہے کہ ماڈل کم حقیقی-مثبت شرح

حاصل کرے گا۔ ترقی پذیر کے نیئرسٹ نائبر ماڈل کی غلطی کی شرح کنفیوژن میٹرکس کے ترسیم 5.5 میں مساوات 3.15 کا استعمال کرکے حاصل کی گئی ہے، جو 14 فیصد کے برابر ہے۔

AUROC کی کارکردگی کی پیمائش کو یہ دیکھنے کے لئے استعمال کیا جاتا ہے کہ آیا پیشن گوئی کی درجہ بندی کا ماڈل جیتے اور ہارے ہوئے معاملات میں درست طور پر فرق کر سکتا ہے۔ تاہم، ناقص ماڈلز کو دونوں طبقوں میں فرق کرنے میں مشکلات پیش آئیں گی۔ ذیل میں دیئے گئے ترسیم 5.3.2.2 میں AUROC کرونیئرسٹ نائبر الگورتھم سے حاصل کیا گیا ہے جس میں AUROC اسکور 86 فیصد ہے۔ ہم مروجہ تحقیق کے ساتھ تیار شدہ کے نیئرسٹ نائبر انتخابی پیشن گوئی ماڈل کے کامیاب تجرباتی نتائج کی نقالی کرتے ہیں۔ نتائج سے پتہ چلتا ہے کہ کے نیئرسٹ نائبر ماڈل انتخابی پیشن گوئی کے لئے بہترین ہے کیوں کہ غلط استعمال کی شرح کم ہے۔



ترسیم 5.3.2.2: کے نیرسٹ نائبر ماڈل کا AUROC کرو۔

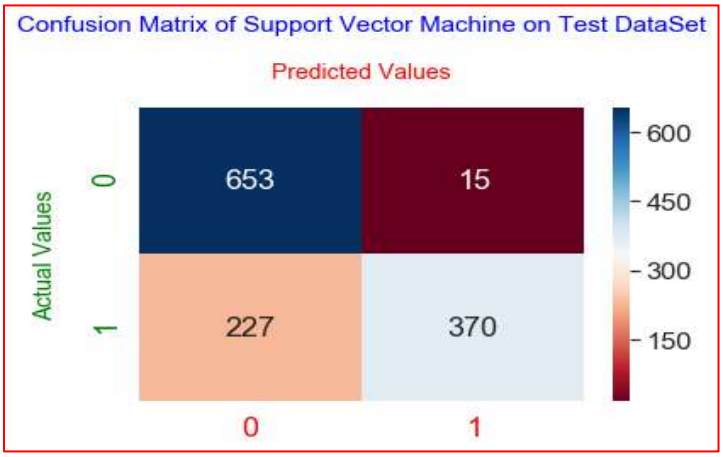
5.3.3 سپورٹ ویکٹر مشین ماڈل تجرباتی نتائج (SVM Model Experimental Results)

سپورٹ ویکٹر مشین اعداد و شمار کو زیادہ سے زیادہ مارجن والی ہائپر پلیون کی بنیاد پر کلاسوں میں الگ کرتی ہے اور سپورٹ ویکٹر بنانے کے کلاسوں کے مابین زیادہ سے زیادہ ہائپر پلان تلاش کرتا ہے-[176] [175]۔اس تحقیق میں، سپورٹ ویکٹر مشین انتخابی پیشن گوئی ماڈل تیار کرنے کے لئے استعمال کی جاتی ہے جو ابتدائی مرحلے میں انتخابی نتائج کی پیشن گوئی کر سکتی ہے۔جموں و کشمیر کے انتخابی ڈیٹاسیٹ پر سپورٹ ویکٹر مشین ماڈل کی کارکردگی کے نتائج کنفیوژن میٹرکس کے ترسیم 5.7 میں دکھائے

گئے ہیں اور اس سے حساسیت، صراحت، درستگی، صحت سے متعلق اور غلط طبقے کی شرح اخذ کی گئی ہے جس کی تفصیل یہ ہے:

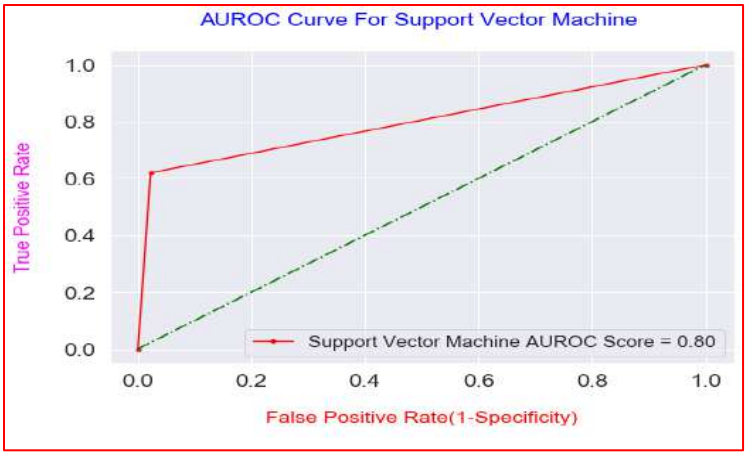
کنفیوژن میٹرکس 5.7 میں مساوات 3.11 کا استعمال کرتے ہوئے سپورٹ ویکٹر مشین ماڈل کی حساسیت کو 61 فیصد کے حساب سے سمجھا جاتا ہے۔ لہذا ماڈل جیتنے والے معاملات کو 61 فیصد کی درستگی کے ساتھ پہچان سکتا ہے۔ اسی طرح، ترسیم 5.3.3.1 میں مساوات 3.12 کا استعمال کر کے 97 فیصد کی درستگی حاصل کی گئی ہے، جس کا مطلب ہے کہ ایس وی ایم ماڈل ہار جانے والی جماعتوں یا آزاد امیدواروں کے معاملات کو 97 فیصد کی درستگی کے ساتھ پہچان سکتا ہے۔ اسی طرح، مجموعی طور پر درستگی 80 فیصد، ترسیم 5.3.3.1 کی قدروں کو مساوات 3.13 میں ڈال کر حاصل کی گئی ہے۔ اس کا مطلب یہ ہے کہ حلقہ وار انتخابات میں جیتنے اور ہارنے دونوں کی پیشن گوئی کرنے میں ایس وی ایم انتخابی پیشن گوئی ماڈل کی مجموعی کارکردگی 80 فیصد ہے۔ اسی طرح، مساوات 3.14 کی مدد سے ترسیم 5.7 سے 96 فیصد کی درستگی صحت سے متعلق حاصل کی گئی ہے۔ ترقی یافتہ ایس وی ایم ماڈل کی غلط درجہ

بندی کی شرح ترسیم 5.3.3.1 میں مساوات 3.15 کا استعمال کرکے حاصل کی گئی ہے، جو 19 فیصد کے برابر ہے۔



ترسیم 5.3.3.1: ٹیسٹ ڈیٹاسیٹ پر ایس وی ایم کنفیوژن میٹرکس

AUROC کارکردگی کی پیمائش کا استعمال ایس وی ایم ماڈل کے ذریعے حاصل کردہ امتیازی کرو اور علیحدگی کی پیمائش کی جانچ کرنے کے لئے کیا جاتا ہے۔ نیچے دیئے گئے اعداد و شمار 5.3.3.2 دکھایا گیا ہے جس میں AUROC، ایس وی ایم ماڈل سے اخذ کیا گیا ہے جس کا AUROC اسکور 80 فیصد ہے۔ ہم موجودہ تحقیق کے ساتھ تیار کردہ سپورٹ ویکٹر مشین انتخابی پیشن گوئی ماڈل کے کامیاب تجرباتی نتائج کی نقالی کرتے ہیں۔ حاصل کردہ نتائج ہماری بہترین معلومات کے ہیں۔



ترسیم 5.3.3.2: سپورٹ ویکٹر مشین ماڈل کی AUROC کرو

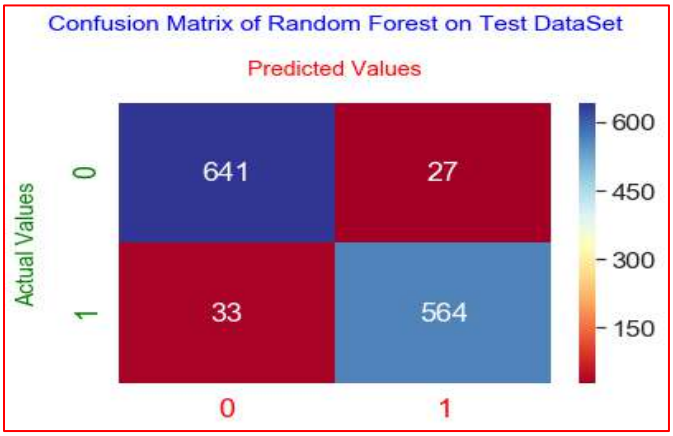
لہذا ترقی یافتہ ایس وی ایم ماڈل کو اس کے عملی نفاذ کے لئے استعمال کیا جا سکتا ہے۔ تاہم، ایس وی ایم ماڈل میں مزید بہتری کی ضرورت ہے جو ہائپر پیرامیٹر ٹننگ کا استعمال کرتے ہوئے باب 6 میں انجام پائے ہیں۔

5.3.4 رینڈم فاریسٹ ماڈل کے تجرباتی نتائج (Experimental Results)

Random Forest Model

جموں و کشمیر اسمبلی ڈیٹاسیٹ پر رینڈم فاریسٹ ماڈل کے پیشن گوئی کے نتائج کنفیوژن میٹرکس کے ترسیم 5.3.4.1 میں دکھائے گئے ہیں۔

اس اعداد و شمار سے اخذ کردہ حساسیت، صراحت، درستگی، صحت سے متعلق اور غلط درجہ بندی کی شرحوں کی وضاحت اس طرح کی گئی ہے:



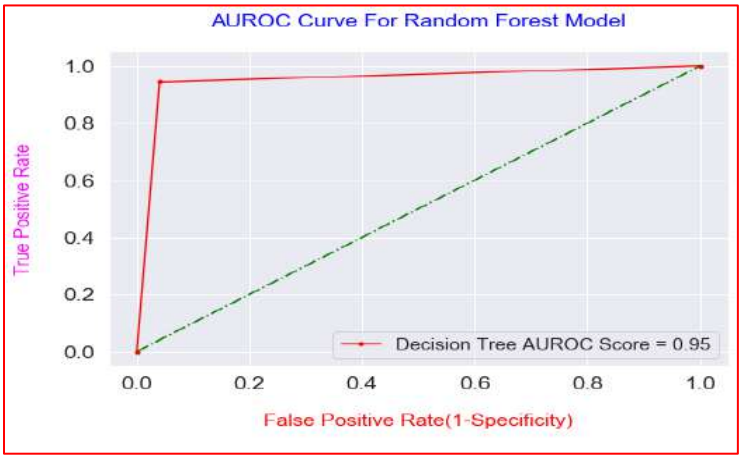
ترسیم 5.3.4.1: ٹیسٹ ڈیٹاسیٹ پر رینڈم فاریسٹ ماڈل کنفیوژن میٹرکس

رینڈم فاریسٹ ماڈل مساوات 3.11 کی ترسیم 5.3.4.1 میں رکھ کر 94 فیصد کی حساسیت کے ساتھ جیتنے والے معاملات کو پہچان سکتا ہے۔ اسی طرح، سیاسی جماعتوں اور آزاد امیدواروں کی تعداد جنہوں نے حقیقت میں انتخابات میں کامیابی حاصل نہیں کی تھی وہ 95 فیصد کے برابر ہے جو ترسیم 5.3.4.1 میں مساوات 3.12 کو استعمال کرکے حاصل کی گئی ہے۔ جتنا اس اقدام کی قدر 1 کے قریب ہے، انتخابی نتائج کی نشاندہی کرنے پر قواعد بہتر ہوں گے۔

مجموعی طور پر 95 فیصد درستگی ترسیم 5.3.4.1 میں مساوات 3.13 کا استعمال کر کے حاصل کی گئی ہے اس کا مطلب یہ ہے کہ رینڈم فاریسٹ انتخابی پیشن گوئی ماڈل کی مجموعی درستگی (انتخابی مقدمات کی جیتنے اور ہارنے دونوں کی پیشن گوئی کرنے میں) 95 فیصد ہے، جس کا درستگی فیصد زیادہ ہے، یہ ایک بہترین ماڈل ہے۔ اسی طرح، 95 فیصد کی صحت سے متعلق ترسیم 5.3.4.1 میں مساوات 3.14 کا استعمال کرتے ہوئے حاصل کی گئی ہے۔ اس پیمائش کی قدر جتنی زیادہ 1 کے قریب ہے، اس کے امکانات اتنے ہی زیادہ ہوں گے کہ مثبت نتائج والے انتخابات میں جیت پائیں گے۔ 4 فیصد کی خرابی کی شرح ترسیم 5.3.4.1 کی اقدار کو مساوات 3.15 میں ڈال کر حاصل کی جاتی ہے۔ ماڈل کی غلط درجہ بندی کی شرح میں جو تناسب کم ہے، انتخابی نتائج کی نشاندہی کرنے میں ماڈل اتنا ہی درست ہے۔

AUROC اسکور کا تخمینہ امکان کَرو اور رینڈم فارسٹ کے ماڈل کے ذریعہ حاصل ہونے والی علیحدگی کی پیمائش کی جانچ کرنے کے لئے کیا جاتا ہے۔ AUROC یہ بتاتا ہے کہ انتخابی حلقہ کی سطح پر اسمبلی انتخابات کے جیتنے اور ہار جانے والے معاملات میں ماڈل کتنا اچھا فرق کر سکتا ہے۔ نیچے

دیئے گئے ترسیم5.3.4.2میں دکھایا گیا ہے کہAUROCرینڈم فاریسٹ ماڈل سے حاصل کیا گیا ہے جس کاAUROCاسکور95فیصد ہے۔



ترسیم:5.3.4.2رینڈم فاریسٹ ماڈل کاAUROCکرو

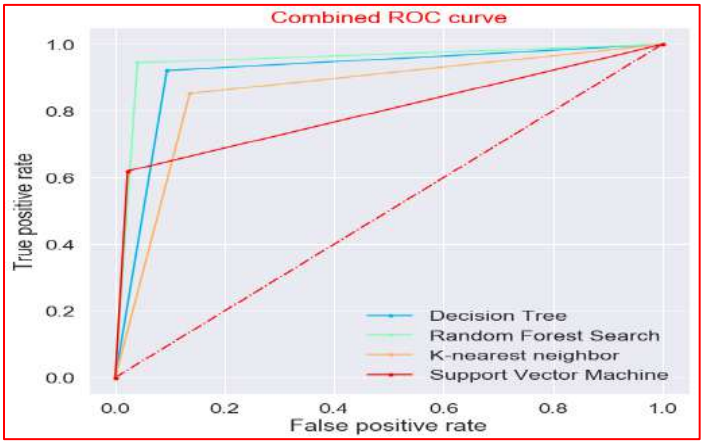
ہم مروجہ تحقیق کے ساتھ تیار شدہ رینڈم فاریسٹ انتخابی پیشن گوئی ماڈل کے کامیاب تجرباتی نتائج کی نقالی کرتے ہیں۔ حاصل کردہ نتائج ہمارے علم کے مطابق بہترین ہیں۔ لہذا مجوزہ رینڈم فارسٹ ماڈل کو جموں وکشمیر کے انتخابی نتائج کی جلد پیشن گوئی کے لئے استعمال کیا جاتا ہے۔تاہم رینڈم فاریسٹ ماڈل میں مزید بہتری کی ضرورت ہے جو ہائپر پرمیٹر ٹننگ کا استعمال کرتے ہوئے باب 6میں شامل ہیں۔

5.4 ترقی یافتہ انتخابی پیشن گوئی ماڈلز کی کارکردگی کا موازنہ

یہ سیکشن مختلف اقدامات کے ذریعہ ڈسیزن ٹری، کے نیئرسٹ نائبر، سپورٹ ویکٹر مشین، اور رینڈم فارسٹ انتخابی پیشن گوئی ماڈل کی کارکردگی اور موازنہ پیش کرتا ہے جیسا کہ ذیل میں دیئے گئے جدول 5.4 میں بیان کیا گیا ہے۔

جدول نمبر 5.4 انتخابی پیشن گوئی ماڈلز کی کارکردگی کی پیمائش۔

	Performance Measures of The Model					
Models	AUROC Score	F1 Score	Classifiers Accuracy	Recall Score	Precision Score	Miss-Classification Score
Decision Tree	0.9134%	0.9090%	0.9130%	0.9212%	0.8972%	0.0869%
Random Forest	0.9521%	0.9494%	0.9525%	0.9447%	0.9543%	0.0474%
K - Nearest Neighbors	0.8581%	0.8504%	0.8584%	0.8525%	0.8483%	0.1415%
Support Vector Machine	0.7986%	0.7535%	0.8086%	0.6197%	0.9610%	0.1913%



ترسیم 5.4 انتخابات کی پیشن گوئی ماڈلز کی مشترکہ AUROC کرو

نتائج سے پتہ چلتا ہے کہ رینڈم فاریسٹ ماڈل انتخابی پیشن گوئی کے دیگر ماڈلز کو 95 فیصد کی زیادہ سے زیادہ درستگی، 95 فیصد کی صراحت، 94 فیصد کی حساسیت، 95 فیصد کی صحت سے متعلق، AUROC اسکور 95 فیصد اور غلط شرح بندی کے 4 فیصد کے ساتھ بہتر کارکردگی کا مظاہرہ کرتا ہے۔ انتخابات کی پیشن گوئی کے لئے رینڈم فاریسٹ ماڈل کے ذریعہ حاصل کردہ درستگی سب سے زیادہ ہے مذکورہ بالا ترسیم 5.4 مختلف ترقی یافتہ انتخابی پیشن گوئی ماڈل کے مشترکہ AUROC منحنی خطوط کو ظاہر کرتا ہے۔ ترسیم 5.4 سے یہ بات واضح ہے کہ رینڈم فاریسٹ ماڈل میں 95 فیصد کا سب سے زیادہ AUROC اسکور ہے، جس کا مطلب ہے کہ انتخابی نتائج کی پیشن گوئی کرنے میں یہ ماڈل انتہائی ہنر مند ہے۔

5.5 باب کا خلاصہ

اس تحقیقی کام میں، ڈیوس کے ڈیٹا مائننگ کے طریقہ کار کی پیروی انتخابی پیشن گوئی ماڈل کی ترقی کے لئے کی گئی ہے۔ اس باب میں، تحقیقی سرگرمیوں کو آسان بنانے اور مقاصد کے بیان سے تحقیق کو نتیجہ خیز بنانے کے لئے تحقیقی ڈیزائن تیار کیا گیا ہے۔ جموں و کشمیر اسمبلی انتخابات کے ڈیٹاسیٹ میں نمایاں صفات کے انتخاب کے لئے مختلف فیچر سلیکشن کی تکنیک کا استعمال کیا گیا ہے۔ انتخابی نتائج کی پیشن گوئی حلقہ وار پیشن گوئی کے لئے ان اہم صفات کو ڈیٹا مائننگ کی تکنیک کے لئے استعمال کیا جاتا ہے۔ ڈیٹا مائننگ کی درجہ بندی کی تکنیک جیسے ڈیسیزن ٹری، سپورٹ ویکٹر مشین، کے نیئرسٹ نائبر اور رینڈم فارسٹ، انتخابی نتائج کی جلد پیشن گوئی کے لئے استعمال کی جاتی ہے۔ تجرباتی نتائج سے ظاہر ہوتا ہے کہ رینڈم فارسٹ ماڈل دوسرے ماڈل سے سب سے زیادہ درستگی، کم شرح غلط درجہ بندی کی شرح کے ساتھ بہتر کارکردگی کا مظاہرہ کرتا ہے۔ لیکن استعمال شدہ درجہ بندی میں ہمیں جن اہم پریشانیوں کا سامنا کرنا پڑا ہے وہ اوور فٹ ہیں، اور ہم کوشش کریں گے کہ باب پنجم میں زیادہ مناسبات کو ختم کرنے کے مسئلے کو کم کیا جا سکے، تاکہ انتخابی پیشن گوئی کا

صحیح نمونہ تیار کیا جا سکے۔ تب ترقی یافتہ ماڈل صارف کو انتخابی نتائج کی جلد پیشن گوئی کرنے

میں مدد دے گا لہذا شدید پیچیدگیوں کی پیشرفت میں کمی کر دے گا۔

باب 6

6. نتائج پر تبادلہ خیال اور توثیق

6.1 تعارف

انتخابی پیشن گوئی اس کی بنیادی پیچیدگیوں کی وجہ سے ایک چیلنجنگ کام ہے۔ کم سے کم غلطیوں کی شرح کے ساتھ انتخابی نتائج کا زیادہ درست اور موثر اندازہ لگانے کے لئے، ہم مجوزہ ماڈلز کے ہائپر پرامیٹرز کو بہتر بناتے ہیں۔ 'ہائپر پیرامیٹر آپٹیمائزیشن' ایک لرننگ ماڈل کے لئے بہترین ہائپر پیرامیٹرس کی اصلاح کا عمل ہے [177]۔ اس کا بنیادی مطلب یہ ہے کہ ہائپر پیرامیٹرس کے ممکنہ امتزاج کی ایک بہت بڑی کائنات کو اس سیٹ کے لئے تلاش کرنا ہے جو میرٹ کے مطلوبہ اعداد و شمار کو بہتر بناتا ہے [178]۔ تربیت کے مرحلے کے دوران ماڈل پیرامیٹرز سیکھے جاتے ہیں تاہم ہائپر پیرامیٹرس درجہ بندی کرنے والے یا تخمینہ لگانے والے کے پیرامیٹرز ہیں جو تربیت کے اعداد و شمار سے مشین لرننگ مرحلے میں براہ راست نہیں سیکھتے ہیں بلکہ علیحدہ طور پر بہتر بنائے جاتے ہیں [179]۔ ہائپر پیرامیٹرس کی اصلاح کے فن میں عبارت حاصل کرنے کے لئے نہ صرف مشین

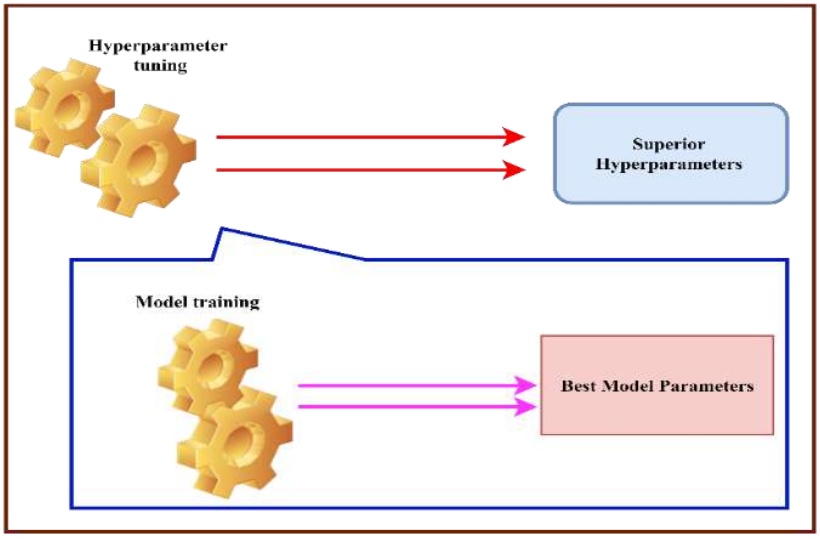
لرننگ الخوارزمی میں ایک ٹھوس پس منظر کی ضرورت ہوتی ہے بلکہ حقیقی دنیا کے ڈیٹا سیٹس کے ساتھ کام کرنے کے لئے وسیع تجربے کی بھی ضرورت ہے۔

اس باب میں بتایا گیا ہے کہ باب 5 میں مجوزہ ماڈلز کے ہائپر پیرامیٹرس کو کس طرح بہتر بنایا جائے۔ اس باب میں، ہم مختلف قسم کے ہائپر پیرامیٹرس آپٹیمائزیشن تراکیب اور ان کے جوڑ کے ماڈل کے ساتھ موازنہ کی بھی وضاحت کرتے ہیں۔ آخر میں، ہم انتخابی پیشن گوئی کے تجزیہ کے مختلف امتزاجوں، پیدا کردہ انتخابی نتائج کے قواعد، انتخابی پیشن گوئی کے تشخیص ماڈل کے اجزاء، اور تیار کردہ انتخابی پیشن گوئی ماڈل پر بھی تبادلہ خیال کریں گے اور باب کا اختتام خلاصے اور نتیجے پر کریں گے۔

6.2 ہائپر پیرامیٹر آپٹیمائزیشن تکنیکس Hyperparameter Optimization Techniques

ہائپر پیرامیٹرز وہ نوبس (گرہ) اور سطح ہیں جسے ہم مشین لرننگ ماڈل تیار کرتے وقت کھینچتے اور موڑتے ہیں [180]۔ ہائپر پرمیٹرس کو مختلف ترتیب کی چھان بین کے ذریعے بہتر بنایا گیا ہے تاکہ یہ معلوم کیا جاسکے کہ کون سی اقدار اعلی ترین درجے کے نتائج مہیا کرتی ہیں [181]۔ ہائپر پارامیٹر کی اصلاح میں اکثر fine ٹوننگ پیرامیٹرز شامل ہوتے ہیں جو ماڈل کے باہر ہی رہتے ہیں، لیکن اس سے اس کے

طرز عمل پر گہرا اثر پڑ سکتا ہے [182],[178] کارکردگی کو بہتر بنانے کے لئے ہائپر پارامیٹر ٹننگ ایک فن ہے اور مناسب ہائپر پر میٹرز کا انتخاب انتہائی درست نتائج پیدا کرے گا اور ہمارے اعداد و شمار میں ہمیں انتہائی قیمتی بصیرت عطا کرے گا۔[184]۔کارکردگی کو بہتر بنانے اور تعصب اور تغیر کے مابین جب صحیح توازن ڈھونڈتے ہیں تو ہائپر پر میٹرز مشین لرننگ ماڈلز کے سلوک کو کنٹرول کرنے میں ہماری مدد کرتے ہیں۔[185]۔ہائپر پیرامیٹر آپٹیمائزیشن کا تصویری خاکہ ذیل میں دیئے گئے ترسیم 6.2 میں دکھایا گیا ہے۔



ترسیم 6.2 ہائپر پیرامیٹر آپٹمائزیشن کا ماڈل اور خاکہ

ہائپر پیرامیٹر آپٹیمائزیشن ایک مساوات کی شکل میں اس طرح پیش کیا جاتا ہے:

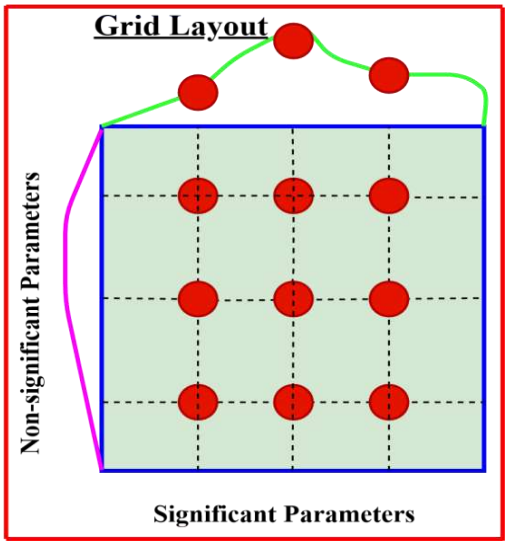
$$x^* = \underset{x \varepsilon X}{arg\,min}\, f(x) \qquad (6.1)$$

یہاں $f(x)$ توثیق سیٹ پر تشخیص شدہ خرابی کی شرح کو کم سے کم کرنے کے لئے آبجیکٹیو اسکور کی نمائندگی کرتا ہے۔ x^* ہائپر پیرامیٹرس کا سیٹ ہے جو اسکور کی سب سے کم قیمت حاصل کرتا ہے ، اور x ڈومین X میں کسی بھی قیمت کو لے سکتا ہے۔ آسان الفاظ میں ، ہم ہائپر پیرامیٹرس تلاش کرنا چاہتے ہیں جو توثیق سیٹ میٹرک پر بہترین اسکور حاصل کرتا ہے۔ ادب میں مختلف ہائپر پیرامیٹرس تکنیکوں کی وضاحت کی گئی ہے ، تاہم اس تحقیقی کام میں ہم صرف انتہائی موثر تدابیر پر تبادلہ خیال کرتے ہیں:

6.2.1 گرڈ سرچ ہائپر پیرامیٹر آپٹمائزیشن Grid Search Hyperparameter Optimization

گرڈ سرچ ہائپر پیرامیٹر کی اصلاح مطلوبہ امیدوار کے پیرامیٹر کی اقدار کی ایک وسیع تلاش ہے جو طے شدہ تلاش کی جگہ میں تمام ممکنہ اقدار سے زیادہ ہے۔[186]-[187]- کسی ماڈل کے لئے ہر ممکنہ پیرامیٹر کے امتزاجوں کا اندازہ کرنے کے بعد، بہترین مجموعہ برقرار رکھا جائے گا۔ گرڈ سرچ

ہائپر پر میٹرز (سیکھنے کی شرح اور متعدد تہوں) کے دو سیٹ کا استعمال کر کے تمام مجموعوں کے لئے الگورتھم کی تربیت کرتا ہے اور کراس–ویلیڈیشن کی تکنیک کا استعمال کرتے ہوئے کارکردگی کو ناپتا ہے [187]۔



ترسیم 6.2.1: گرڈ سرچ لے آؤٹ

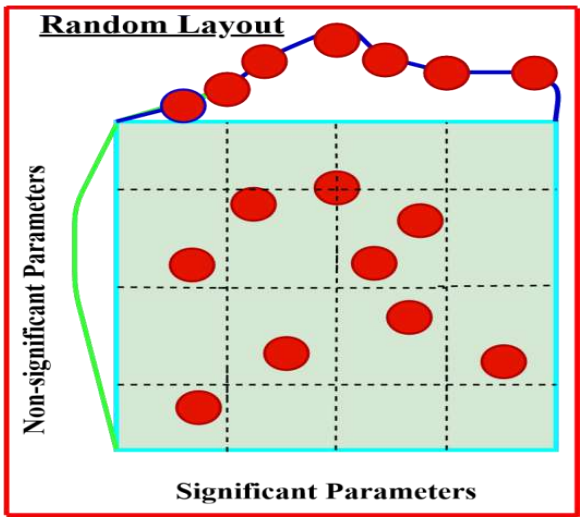
اس ویلیڈیشن کی تکنیک سے یہ یقین دہانی ہو سکتی ہے کہ ہمارے تربیت یافتہ ماڈل کو ڈیٹا سیٹ سے بیشتر نمونے ملے ہیں۔ ترسیم 6.2.1 [188] میں گرڈ سرچ کے طریقہ کار کے خاکہ کو ظاہر کیا گیا ہے۔ گرڈ سرچ میتھڈ استعمال کرنے کے لئے ایک آسان طریقہ ہے لیکن یہ ایک مہنگا طریقہ ہے اور اگر اعداد و شمار میں اعلی جہتی جگہ (جہتی کی لعنت) ہوتی ہے تو وہ متاثر ہوتا ہے۔ فرض کریں کہ

ہمارے پاس n ہائپر پیرامیٹرس ہیں اور ہر ہائپر پیرمیٹر کی دو اقدار ہیں، پھر ہیئت کی کل تعداد 2^n ہے ، لہذا یہ ہیئت کی بہت کم تعداد پر ہی گرڈ سرچ کرنا ممکن ہے۔

6.2.2 رینڈم سرچ ہائپر پیرامیٹر آپٹیمائزیشن Random Search Hyperparameter Optimization

رینڈم سرچ میں، ہائپر پیرامیٹر کی قدریں متعین سرچ اسپیس سے اتفاقی طور پر منتخب کی جاتی ہیں [188]۔ رینڈم سیمپلنگ سرچ اسپیس کو مجرد اور مسلسل ہائپر پیرامیٹرس دونوں کو شامل کرنے کی اجازت دیتا ہے۔ رینڈم سرچ متوازی ہے اور نمونہ کی تقسیم کو مخصوص کرتے ہوئے پیشگی معلومات کو شامل کرنے کی اجازت دیتا ہے[189]۔

ترسیم [188]6.2.2 رینڈم سرچ آپٹیمائزیشن کے خاکہ کو ظاہر کرتے ہیں۔ گرڈ اور رینڈم سرچ ہائپر پیرامیٹر آپٹیمائزیشن ماضی کی تشخیصوں سے مکمل طور پر بے خبر ہیں، اور اس کے نتیجے میں، اکثر 'خراب' ہائپر پیرامیٹرس کی تشخیص کرنے میں کافی وقت خرچ کرتے ہیں[178]۔



ترسیم 6.2.2: رینڈم سرچ لے آؤٹ

ان تکنیک میں ، اعلی جہتی اعداد و شمار کی غلط تشہیر کی وجہ سے انسانی مہارت ہائپر پیرامیٹرس کی قریب ترین ترتیب حاصل نہیں کر سکتی ہے اور جب متعدد ہائپر پیرامیٹرس کو ہم آہنگ کرنے کی کوشش کی جاتی ہے تو آسانی سے غلط تشریح کی جا سکتی ہے[142]۔ گرڈ اور رینڈم سرچ ہائپر پیرامیٹر آپٹیمائزیشن کی تکنیک سے وابستہ ان خرابیوں کی وجہ سے ، ہم جموں و کشمیر کے انتخابی پیشن گوئی کے ماڈل کی ترقی کے لئے بایسیئن ہائپر پیرامیٹر آپٹیمائزیشن کو استعمال کرنے کو ترجیح دیتے ہیں۔

6.2.3 بائیسین ہائپر پیرامیٹر آپٹیمائزیشن Bayesian Hyperparameter Optimization

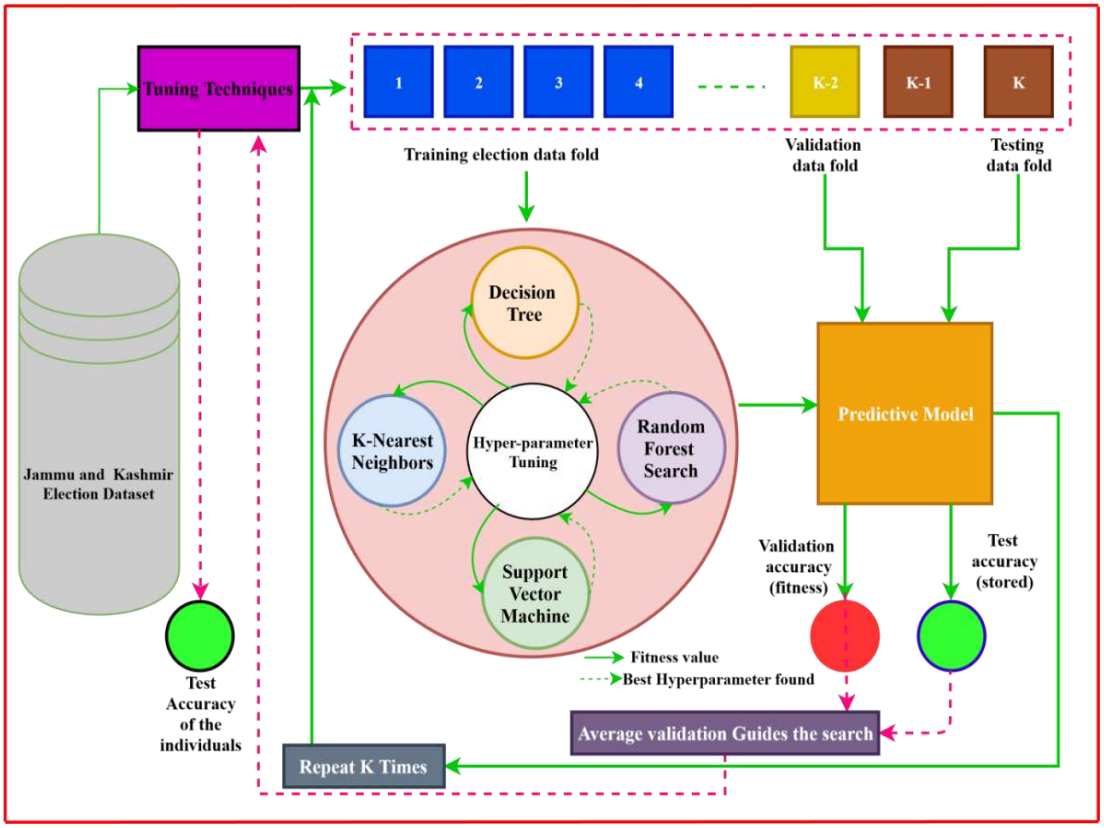
بائیسین آپٹیمائزیشن فنکشن میپنگ کا ایک امکانی نمونہ تیار کرتی ہے جس میں ہائپر پیرامیٹر اقدار سے لے کر ایک ویلیڈیشن سیٹ کی تشخیص کی جاتی ہے [182]۔ گرڈ سرچ اور رینڈم سرچ میں، ہم بغیر کسی قاعدے اور بغیر دیکھے ترتیب آزماتے ہیں اور نیا تجربہ پہلے ہونے والے تمام تجربوں سے آزاد ہے [191]۔ اس کے برعکس، خودکار طریقے سے ہائپر پیرامیٹر اگلے پیرامیٹر کی سیٹنگ کے لئے زیادہ بہتر انتخاب کرنے کے لئے ہائپر پیرامیٹر کی سیٹنگ اور ماڈل کی کارکردگی کے مابین تعلقات کے بارے میں علم کی تشکیل کرتی ہے [192]۔ خاص طور پر، یہ پہلے کئی ترتیبوں پر کارکردگی کو اکٹھا کرے گا، پھر کچھ نتیجہ اخذ کرے گا اور فیصلہ کرے گا کہ اگلے تجربے میں کس ترتیب کا استعمال کرنا ہے۔ اس کا مقصد یہ ہے کہ خوب سے خوب تر کی جستجو میں تجربات کی تعداد کو کم سے کم کیا جائے۔ بائیسین آپٹمایزیشن کو انجام دیتے وقت دو اہم انتخابات کرنے چاہیے۔

i. افعال کے مقدم کا انتخاب کریں جو عمل کے بہتر ہونے کے بارے میں مفروضوں کا اظہار کریں گے۔ اس کے لئے، ہم پہلے "گاؤسی پروسیز" کا انتخاب کرتے ہیں۔

ii. اگلا، ہمیں ایک حصول فنکشن کا انتخاب کرنا چاہئے جو ماڈل کے بعد سے افادیت فنکشن کی تیاری کے لئے استعمال ہوتا ہے، جس سے ہمیں اگلے نقطہ کا اندازہ کرنے کا موقع مل سکے۔

ہر ٹیوننگ تکنیک کے ذریعہ بایسیئن آپٹمائزیشن میں، ہم ترسیم 6.2.3 میں دکھائے گئے سنگل کراس ویلیڈیشن (S-CV) کے طریقہ کار کا استعمال کرتے ہیں۔ جب ایک ہائپر پیرامیٹر ٹیوننگ عمل میں لائی جاتی ہے تو، انتخابی ڈیٹاسیٹ کو k پرتوں میں تقسیم کیا جاتا ہے۔ ڈیسیزن ٹری، رینڈم فارسٹ سرچ، K-NN اور SVM جیسے الگورتھم کو تکنیک کے ذریعہ پائے جانے والے ہر امیدوار حل کے لئے 2 – k پارٹیشنز (ٹریننگ فولڈ) کی تربیت دی جاتی ہے۔ ایک حصہ، ماڈل کی توثیق کرنے کے لئے الگ ہے (ویلیڈیشن فولڈ) اور باقی حصے کو ٹیسٹ کرنے کے لئے الگ کر دیا گیا ہے (ٹیسٹ فولڈ)۔ جانچ اور توثیق کے درست ہونے کا اندازہ تربیتی پارٹیشنوں اور آپٹیمائزیشن تکنیک کے ذریعہ پائے جانے والے ہائپر پیرامیٹر کی اقدار کے ساتھ حوصلہ افزائی ماڈل کے ذریعے کیا جاتا ہے۔ یہ عمل سنگل کراس ویلیڈیشن میں تمام متبادل K کے لئے دہرایا جاتا ہے۔ اوسط توثیق کی درستگی کو فٹنس ویلو کے بطور استعمال کیا جاتا ہے، جو تلاش کے عمل میں رہنمائی کرے گا۔ آخر میں، سب سے زیادہ توثیق کی درستگی والا فرد واپس ہو جاتا ہے

(اس کے ہائپر پیرامیٹر اقدار کے ساتھ)،اور تکنیکی کارکردگی فرد کی اوسط ٹیسٹ کی درستگی سمجھی جاتی ہے۔



ترسیم 6.2.3:ہائپر پیرامیٹر ٹیوننگ کے لئے سنگل کراس ویلیڈیشن تجرباتی طریقہ کار

6.3 انتخابی پیشن گوئی کے ماڈل کو بہتر بنانا

ہائپر پیرامیٹر ماڈل پیرامیٹرز کا اندازہ لگانے کے لئے استعمال ہوتے ہیں اور پیشن گوئی کرنے کے لئے تربیت یافتہ ماڈل استعمال نہیں ہوتے ہیں۔ مشین لرننگ الگورتھم اکثر ایسے بہت سارے ہائپر پیرامیٹرس

پر مشتمل ہوتے ہیں جن کی اقدار پیچیدہ طریقوں سے حوصلہ افزا ماڈلز کی پیشن گوئی کارکردگی کو متاثر کرتی ہیں [185]۔ ان ہائپر پیرامیٹر کی ہیئت کے اعلیٰ امکانات کی وجہ سے، ہمارے پاس اس ضمن میں بصیرت کا فقدان ہے کہ ہیئت کی اس وسیع جگہ کو موثر انداز میں کیسے استعمال کیا جائے۔ ہم نے ماڈل کی تعمیر میں استعمال ہونے والے تمام مشین لرننگ کی درجہ بندی کرنے والوں میں رینڈم اسٹیٹ کو "42" کے طور پر مرتب کیا، کیونکہ اگر ہم رینڈم اسٹیٹ کو سیٹ نہیں کرتے ہیں تو ہر بار، جب ہم ماڈل کو چلاتے ہیں تو ایک بے ترتیب قدر پیدا ہو جاتی ہے اور ٹیسٹ اور ہمارے ماڈل کے تربیت یافتہ ڈیٹا سیٹ میں ہر بار مختلف اقدار ہوتی ہیں۔ لہذا رینڈم اسٹیٹ کو '42' سے ظاہر کرتے ہوئے جب بھی ہم ماڈل چلاتے ہیں تو ہم ایک جیسے نتائج حاصل کرتے ہیں (یعنی تربیت یافتہ اور ٹیسٹ ڈیٹا سیٹس میں ایک ہی قدر)۔ ہر ماڈل پر روشنی ڈالنے کے لئے یہاں بہت سارے امکانی پیرامیٹرز موجود ہیں اور تمام پیرامیٹرز قابل قدر ہیں لیکن ہمیں پیرانٹرز کا اہم سبسیٹ منتخب کرنے کی ضرورت ہے۔

6.3.1 ڈیسیزن ٹری ہائپر پرامیٹر آپٹیمائزیشن ماڈل Decision Tree Hyperparameter Optimization Model

زیادہ سے زیادہ درست نتیجہ حاصل کرنے کے لئے ہم انتہائی اہم ہائپر پیرامیٹر ماڈل ڈیسیزن ٹری کا استعمال کرتے ہیں تاہم موزوں ہونے سے بچنے کے لئے ہمیں ٹیسٹ ڈاٹافولڈ پر ان کی توثیق کرتے ہوئے محتاط رہنا چاہئے۔

ڈیسیزن ٹری ماڈل کی کارکردگی کو بہتر بنانے والے اہم ہائپر پیرامیٹر ماڈل کے ذیلی حصے مندرجہ ذیل ہیں:

i. **زیادہ سے زیادہ گہرائی**: زیادہ سے زیادہ گہرائی کا ہائپر پیرامیٹر اشارہ کرتا ہے کہ ٹری کتنا گہرا ہو سکتا ہے۔ ٹری جتنا گہرا ہوتا ہے، اور اس میں جتنے ٹکڑے ہوں گے تو وہ ڈیٹا سے زیادہ سے زیادہ معلومات اکٹھا کر سکتا ہے [144]۔ ہم نے 1 سے 32 تک کی گہرائیوں کے ساتھ ڈیسیزن ٹری کو فٹ کیا اور تربیت کا منصوبہ تیار کر کے، اے یو آر او سی اسکور کی جانچ کرتے ہیں۔ ہم دیکھتے ہیں کہ ہمارا ماڈل بڑی گہرائی والے اقدار کے لئے فائدہ مند ہے۔ ڈیسیزن ٹری تمام مرتب اعداد و شمار کی بالکل صحیح پیشن گوئی کرتا ہے۔ تاہم، نئے اعداد و شمار کی تلاش کو عام کرنے میں ناکام ہے۔

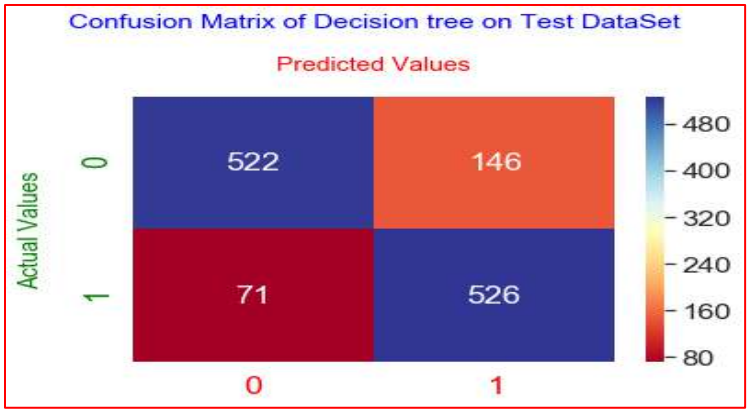
ii. کم از کم نمونے کا لیف: کم از کم نمونے کا لیف، لیف کے نوڈ درکار نمونوں کی کم از کم تعداد کی نمائندگی کرتا ہے، تاہم قیمت کے نتائج میں غیر موزوں اضافہ کرتا ہے[193]۔

iii. کم از کم نمونہ تقسیم: کم سے کم نمونہ تقسیم داخلی نوڈ کے لیف کو درکار نمونوں کی کم از کم تعداد کی نمائندگی کرتا ہے [194]۔ ہر نوڈ میں کم از کم ایک نمونہ پر غور کرنے اور ہر نوڈ کے تمام نمونوں پر غور کرنے کے درمیان یہ مختلف ہو سکتا ہے۔ جب ہم اس پیرامیٹر میں اضافہ کرتے ہیں تو، ٹری زیادہ مجبور ہو جاتا ہے کیونکہ اسے ہر نوڈ پر مزید نمونوں پر غور کرنا پڑتا ہے۔ یہاں ہم نمونے کے 10 فیصد سے 100 فیصد تک پیرامیٹر کو تبدیل کرتے ہیں اور نتائج سے یہ واضح ہوتا ہے کہ جب ہم ہر نوڈ پر نمونے میں سے 100 فیصد غور کرتے ہیں تو ماڈل انڈر فٹ کی طرف جاتا ہے۔ ڈیسیزن ٹری ماڈل کے ہائپر پیرامیٹرز کو ٹیوننگ کرنے کے بعد ہم ذیل میں جدول 6.3.1 میں بیان کردہ نتائج حاصل کرتے ہیں۔ ہائپر پیرامیٹرائزڈ ڈسیزن ٹری ماڈل کا تبادلہ اور امتزاج مختلف نتائج دکھاتے ہیں تاہم، ہم صرف ان امتزاج کو بیان کرتے ہیں جو قریب ترین نتائج فراہم کرتے ہیں۔

جدول 6.3.1: ڈسیزن ٹری ماڈل کے ہائپر پیرامیٹر آپٹمائزیشن کے نتائج

Max depth	Min samples leaf	Min samples split	Random state	Accuracy
10	55	30	42	76%
10	55	30	42	82%
10	130	200	42	80%
5	20	42	42	64%
10	20	42	42	73%
8	5	15	42	78%
9	55	90	42	81%
9	30	183	42	82%
9	30	183	42	73%
6	55	30	42	74%
7	55	30	42	75%
7	30	50	42	77%
8	30	50	42	78%
8	30	150	42	77%
8	90	150	42	79%
8	70	150	42	78%
8	70	110	42	78%
8	10	150	42	77%
8	10	180	42	76%
9	10	183	42	82%
10	20	70	42	82%

ہم نے انتخابی نتائج کی پیشن گوئی کے لئے آپٹیمائزڈ ڈسیزن ٹری الگورتھم کا اطلاق کیا اور ماڈل کی کارکردگی کے نتائج کنفیوژن میٹرکس کے ترسیم 6.3.1.1 میں دکھائے گئے ہیں۔



ترسیم 6.3.1.1: ہائپر پرامیٹر ڈسیزن ٹری ماڈل کا کنفیوژن میٹرکس

ترسیم 6.3.1.1 سے ہم حقیقی مثبت شرح (ری کال)، حقیقی منفی شرح (مخصوصیت)، شناخت کی شرح، صحت سے متعلق اور مختلف درجہ بندی کی شرح جس کو مندرجہ ذیل میں بیان کیا گیا ہے۔

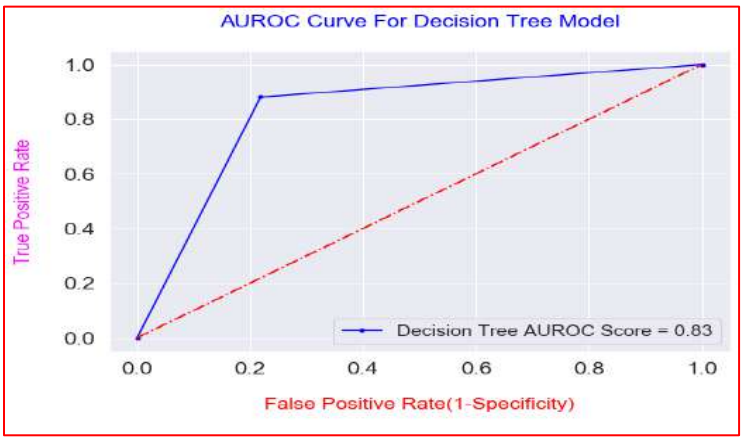
مساوات 3.11 کا استعمال کرتے ہوئے، ہم ڈسیزن ٹری ماڈل کی حقیقی مثبت شرح کو حاصل کرتے ہیں جیسے $\frac{(526)}{(71+526)}$ جو کہ 0.8810 فیصد کے برابر ہے جس کا مطلب ہے کہ ہمارے ڈسیزن ٹری ماڈل انتخابات جیتنے والے معاملات کو 88 فیصد کی درستگی کے ساتھ مثبت تسلیم کر سکتے ہیں۔ اسی طرح، مساوات 3.12 کا استعمال کرکے ہم ڈسیزن ٹری ماڈل کی حقیقی منفی شرح کو حاصل کرتے ہیں جیسے $\frac{(522)}{(522+146)}$ جو کہ 0.7814 فیصد کے برابر ہے اس کا مطلب ہے کہ ہمارے ڈسیزن ٹری ماڈل انتخابات ہارنے والے معاملات کو 78 فیصد کی درستگی کے ساتھ تسلیم کر سکتے ہیں۔ ڈسیزن ٹری ماڈل کی درستگی مساوات 3.13

کے ذریعہ حاصل کی گئی ہے جو کہ $\frac{(522 + 526)}{(522+146+71+526)}$ ہے۔جو 0.8284 فیصد کے برابر ہے جس کا مطلب ہے کہ ڈسیزن ٹری ماڈل کے فیصلے کی انتخابی نتائج کی جیت اور ہار دونوں کی پیشن گوئی کرنے میں مجموعی کارکردگی 82 فیصد ہے۔اسی طرح، بہتر ڈسیزن ٹری آپٹمائزڈ ماڈل کی درستگی حاصل کرنے کے لئے مساوات 3.14 کو $\frac{(526)}{(526+146)}$ کے طور پر استعمال کیا جاتا ہے جوکہ تشخیص کے بعد 0.7827 فیصد ہے اس کا مطلب یہ ہے کہ ہمارے ڈسیزن ٹری آپٹمائزڈ ماڈل میں بہت کم غلط-مثبت شرح ہے۔مجوزہ ڈسیزن ٹری ماڈل کی مختلف درجہ

بندی کی شرح مساوات 3.15 $\frac{(146+ 71)}{(522+146+71+526)}$ کا استعمال کرکے حاصل کی جاتی ہے جو 17 فیصد کے برابر ہے۔

محقق AUROC کی کارکردگی کی پیمائش کا استعمال بھی کرتے ہیں تاکہ یہ دیکھیں کہ ماڈل انتخابی نتائج کی جیت اور ہار کے معاملوں میں کتنا اچھا فرق کر سکتا ہے۔بہتر ماڈلز دونوں کے درمیان درست طور پر تمیز کر سکتے ہیں (یعنی جیت اور ہار کے معاملات کے درمیان) تاہم ، ناقص ماڈلز کو دونوں کے درمیان تمیز کرنے میں مشکلات پیش آئیں گی۔

نیچے دیئے گئے ترسیم 6.3.1.2 یہ ظاہر کر رہا ہے کہ AUROC Curve 83 فیصد AUROC اسکور کے ساتھ ڈسیزن ٹری ماڈل کے ذریعہ حاصل کیا گیا ہے۔ ہم موجودہ تحقیق کے ساتھ آپٹیمائزڈ ڈسیزن ٹری کے انتخابی پیشن گوئی ماڈل کے کامیاب تجرباتی نتائج کو نقل کرتے ہیں۔ حاصل کردہ نتائج ہمارے بہترین مشاہدے ہیں جو ادب میں شائع شدہ اقدار سے کہیں زیادہ ہیں۔ لہذا، ہم جموں و کشمیر کے حلقہ وار انتخابی نتائج کی پیشن گوئی کے لئے مجوزہ آپٹیمائزڈ ڈسیزن ٹری ماڈل کا استعمال کر سکتے ہیں۔ تاہم، ماڈل کی کارکردگی میں مزید بہتری کی ضرورت ہے۔



ترسیم 6.3.1.2: ہائپر پیرامیٹرائزڈ ڈسیزن ٹری ماڈل کے ذریعہ AUROC کرو

6.3.2 کے -نیرسٹ نائیبر ہائپر پیرامیٹر آپٹیمائزیشن ماڈل K-Nearest Neighbor)

Hyperparameter Optimization Model)

KNN اکثریتی ووٹوں تصور پر مبنی ایک نامعلوم شے کی درجہ بندی کرتی ہے[195]۔ ہر نائیبر کو یا تو مساوی وزن دیا جا سکتا ہے یا ووٹ فاصلے پر مبنی ہو سکتا ہے [196]۔ مماثلت کی پیمائش ڈیٹا کی قسم پر منحصر ہے۔ ہم KNN کے لئے ایک ہائپر پیرامیٹر ٹیوننگ جاب چلاتے ہیں تاکہ اس کے لیے بہترین ماڈل تلاش کر سکیں جو جانچ کے انتخابی ڈیٹاسیٹ میں سب سے زیادہ درستگی اور کم ترین غلطی کرتا ہے۔ ہم KNN کے اہم ترین پیرامیٹرز کی چھان بین کرتے ہیں اور معلوم کرتے ہیں کہ وہ کس طرح اوور فٹننگ اور انڈر فٹنگ کی اصطلاح میں ہمارے ماڈل پر اثرانداز ہوتے ہیں۔ یہاں تین بنیادی ہائپر پیرامیٹر ہیں:

i. نائبرس کی تعداد 'k' ۔
ii. فاصلہ میٹرک/مماثلت کا فنکشن۔
iii. پیرامیٹر وزن، نائیبرس کے وزن کو بیان کرنے کے لئے استعمال ہوتا ہے۔

یہ سب قدریں ڈرامائی انداز میں K-NN درجہ بندی کی درستگی کو متاثر کرتی ہیں۔ ظاہر ہے، بہترین K وہی ہے جو سب سے کم غلطی کی شرح سے مطابقت رکھتا ہے، لہذا ہم K کی مختلف اقدار کے لئے بار بار جانچ کی غلطی کی پیمائش کرتے ہیں۔ فٹنگ کا عمل اس ذیلی سیٹ، جسے ویلیڈیشن سیٹ کہا جاتا ہے، ہمارے الگورتھم کی مناسب لچکدار سطح کو منتخب کرنے کے لئے استعمال کیا جا سکتا ہے۔ عملی طور پر ویلیڈیشن کے مختلف طریقوں کا استعمال کیا جاتا ہے، لیکن ہم 10-فولڈ کراس ویلیڈیشن کی تحقیق کرتے ہیں۔ ایک اور ضروری بات نوٹ کرنے کی یہ ہے کہ K-NN کے وقت کی پیچیدگی بہت زیادہ ہے کیونکہ یہ ٹیسٹ سیٹ میں ہر ایک نقطہ کے لئے انفرادی فاصلے طے کرتا ہے۔ آپٹیمائزیشن سرچ سی وی متعدد دفعہ ہمارے تربیت یافتہ ماڈل کی تربیت کے ذریعہ کام کرتی ہے، پیرامیٹرز کی اس رینج پر جسے ہم نے مخصوص کیا ہے۔ اس طرح، ہم ہر پیرامیٹر کے ساتھ اپنے ماڈل کی جانچ کر سکتے ہیں اور بالکل درست نتائج حاصل کرنے کے لئے بہترین قدروں کا پتہ لگا سکتے ہیں۔

آپٹمائزڈ K-NN ماڈل کے تجرباتی نتائج ذیل میں دیئے گئے جدول 6.3.2 میں دکھائے گئے ہیں۔ ہم زیادہ سے زیادہ درستگی حاصل کرنے کے لئے K-NN پیرامیٹر کے مختلف تبادلہ اور امتزاج کا استعمال کرتے

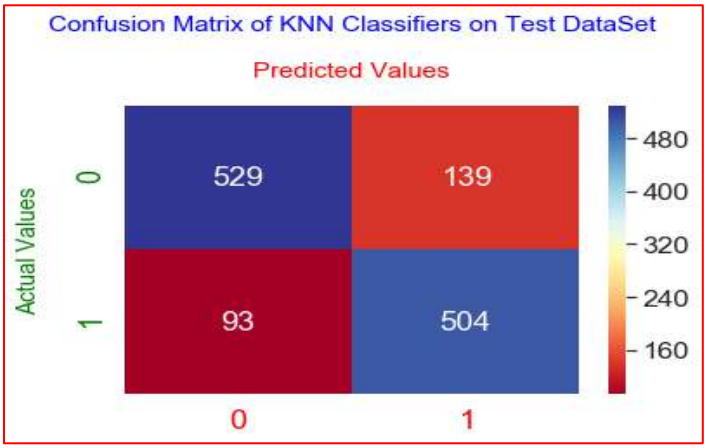
ہیں۔ پیرامیٹر کے امتزاج جس کا نتیجہ سب سے زیادہ درست ہوتا ہے ، ذیل میں دیئے گئے جدول 6.3.2 میں بیان کیا گیا ہے۔ ہم اپنے ماڈل کی درستگی کو جانچنے کے لئے 'بہترین اسکور' فنکشن کا استعمال کر سکتے ہیں کیونکہ 'بہترین اسکور' کراس ویلیڈیشن کے ذریعہ حاصل کردہ اسکور کی اوسط درستگی ہے۔ جب K کی قدر چھوٹی ہوتی ہے تو پھر ہمیں کم تعصب اور اعلی تغیر ملتا ہے، تاہم جب K کی قدر بڑی ہے تو زیادہ تعصب اور کم تغیر پایا جاتا ہے۔ لہذا، ہم نے K پیرامیٹر کی قدر کو سویٹ پوزیشن میں رکھا ہے اور آپٹیمائزیشن سرچ کی مدد سے ہم نے اپنی ماڈل کی درستگی میں 81% فیصد تک بہتری لائی ہے۔

جدول 6.3.2 K-NN ماڈل کے ہائپر پرامیٹر آپٹمائیزیشن کے نتائج

Leaf size	N neighbours	weights	Metric	Random State	Accuracy
30	11	uniform	Minkowski	42	81%
50	13	uniform	Minkowski	42	81%
30	13	uniform	Minkowski	42	80%
20	13	uniform	Minkowski	42	81%
20	13	uniform	Euclidean	42	81%
20	19	uniform	None	42	80%
20	15	uniform	Euclidean	42	80%
20	19	None	Minkowski	42	80%
25	17	uniform	Minkowski	42	79%
20	19	uniform	None	42	80%
20	19	uniform	Euclidean	42	80%
25	17	uniform	Euclidean	42	79%
20	19	None	Minkowski	42	80%
10	11	uniform	Minkowski	42	81%
15	11	uniform	Minkowski	42	81%
40	11	uniform	Minkowski	42	81%
50	11	uniform	Euclidean	42	81%
25	11	uniform	Euclidean	42	81%

25	15	uniform	Euclidean	42	80%
25	15	uniform	Minkowski	42	80%
20	19	uniform	Minkowski	42	80%
30	19	uniform	None	42	80%
25	11	Uniform	Minkowski	42	81%

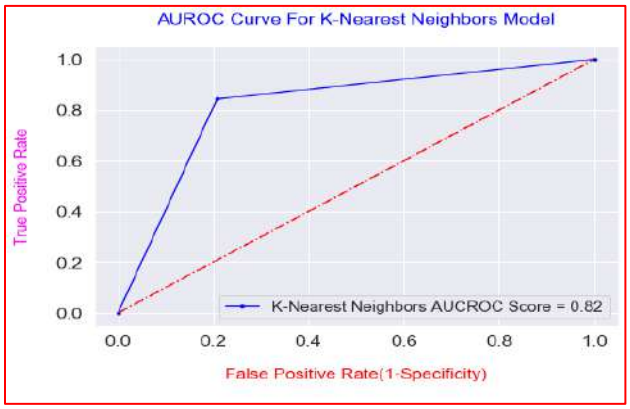
تجرباتی نتائج سے پتہ چلتا ہے کہ جب ہائپر پیرامیٹر کے امتزاج ہوتے ہیں تو [لیف کا سائز=25، این نائیبر =11، وزن = یکساں، میٹرکس = منکووسکی، رینڈم اسٹیٹ =42] سب سے زیادہ درستگی 81 فیصد حاصل کی جاتی ہے۔ مجوزہ K-NN ماڈل کی کارکردگی کے نتائج کنفیوژن میٹرکس کے ترسیم 6.3.2.1 میں دکھائے گئے ہیں۔



ترسیم 6.3.2.1 ہائپر پرامیٹر KNN کنفیوژن میٹرکس

ترسیم 6.3.2.1 سے، ہم حقیقی مثبت شرح (ری کال)، حقیقی منفی شرح (وضاحت)، شناخت کی شرح، صحت سے متعلق اور مختلف درجہ بندی کی شرح اخذ کرتے ہیں جس کو بیان کیا گیا ہے۔ مساوات 3.11 کا استعمال کرتے ہوئے ہم K-NN ماڈل کی حقیقی مثبت شرح $\frac{(504)}{(504+93)}$ حاصل کرتے ہیں جو 0.84 فیصد کے برابر ہے لہذا ہمارا K-NN ماڈل فاتح پارٹی یا آزاد امیدواروں کو پہچان سکتا ہے 84 فیصد کی درستگی کے ساتھ۔ اسی طرح، مساوات 3.12 کے استعمال سے ہمیں K-NN ماڈل کی حقیقی منفی شرح موصول ہوتی ہے جیسا کہ $\frac{(529)}{(529+139)}$ جو 0.79 فیصد کے برابر ہے۔ اس کا مطلب ہے کہ ہمارے K-NN ماڈل 79 فیصد کی درستگی کے ساتھ پارٹی یا آزاد امیدوار جو انتخابات نہیں جیت پائے ان کو پہچان سکتے ہیں۔ K-NN ماڈل کی مجموعی درستگی مساوات 3.13 کا استعمال کرکے حاصل کی گئی جو کہ $\frac{(529 + 504)}{(529+139+93+504)}$ ہے جو حسابات کے بعد 0.81 فیصد کے برابر ہے۔ جس کا مطلب ہے کہ انتخابات جیتنے اور ہارنے والے معاملات کی پیشن گوئی میں K-NN ماڈل کی مجموعی کارکردگی 81 فیصد ہے۔ مطلوبہ K-NN ماڈل کی درستگی حاصل کرنے کے لئے مساوات 3.14 استعمال کی جاتی ہے اور نتائج کا اندازہ لگایا جاتا ہے جیسے $\frac{(504)}{(504+139)}$ جو کہ 0.78 فیصد کے برابر ہے۔ اس کا مطلب یہ

ہے کہ آپٹمائزڈ K-NN ماڈل کی شرح کم منفی مثبت ہے۔ مجوزہ K-NN ماڈل کی غلط درجہ بندی کی شرح مساوات $3.15\frac{(139+\ 93)}{(529+139+93+504)}$ کا استعمال کرکے حاصل کی گئی ہے جو کہ 18 فیصد کے برابر ہے۔ ہم آپٹمائزڈ KNN ماڈل کے ذریعہ حاصل کردہ احتمال کے منحنی خطوط اور جداکاری کے پیمائش کو دیکھنے کے لئے AUROC کارکردگی کی پیمائش کا بھی استعمال کرتے ہیں۔ AUROC کرو ہمیں بتاتا ہے کہ حلقہ کی سطح پر پارٹی یا امیدواروں کے انتخابات جیتنے اور ہارنے کے درمیان ماڈل کتنا اچھا فرق کر سکتا ہے۔ بہتر ماڈل انتخابی جیت یا ہار کے معاملے میں درست طور پر تمیز کر سکتے ہیں تاہم ناقص ماڈلز کو ان میں فرق کرنے میں مشکلات پیش آئیں گی۔



ترسیم 6.3.2.2: ہائپر پیرامیٹرائزڈ KNN ماڈل کا AUROC کرو

جو کہ 6.3.2.2 آپٹمائزڈ KNN ماڈل کا AUROC کرو کو ظاہر کر رہا ہے AUROC کے 82 فیصد اسکور کے ساتھ۔ ہم موجودہ تحقیق کے ساتھ آپٹیمائزڈ KNN ماڈل کے انتخابی پیشن گوئی ماڈل کے کامیاب تجرباتی نتائج کو نقل کرتے ہیں۔

حاصل کردہ نتائج ہمارے بہترین مشاہدے ہیں جو ادب میں شائع شدہ اقدار سے کہیں زیادہ ہیں۔ لہذا، ہم جموں و کشمیر کے حلقہ وار انتخابی نتائج کی پیشن گوئی کے لئے مجوزہ آپٹیمائزڈ KNN ماڈل کا استعمال کر سکتے ہیں۔ تاہم، ماڈل کی کارکردگی میں مزید بہتری کی ضرورت ہے۔

6.3.3 سپورٹ ویکٹر مشین ہائپر پیرامیٹر آپٹمائزیشن ماڈل Support Vector Machine Hyperparameter Optimization Model

انتخابی نتائج کی جلد پیشن گوئی کے لئے اعلی ترین درستگی حاصل کرنے کے لئے، ہم ایس وی ایم ماڈل کے انتہائی اہم ہائپر پیرامیٹر کا استعمال کرتے ہیں۔ تاہم، ہمیں ٹیسٹ ڈیٹاسیٹ پر ان کی توثیق کرنے میں محتاط رہنا چاہئے۔ SVM ماڈل کے ہائپر پیرامیٹرز جو ہم نے بہتر بنائے ہیں وہ مندرجہ ذیل ہیں:

i. **کرنل (دانا):** کرنل پیرامیٹر اعداد و شمار کو الگ کرنے کے لئے استعمال ہونے والے ہائپرپلان کی قسم کا انتخاب کرتا ہے[185]۔ کرنل کا مرکزی کام دیئے گئے انپٹ ڈیٹا کو مطلوبہ شکل میں تبدیل کرنا ہے۔ یہاں مختلف اقسام کے افعال ہوتے ہیں جیسے قطاری، کثیرالاسمی، سگمائی اور شعاعوں پر مبنی فنکشن (آرٹی ایف)۔ کثیرالاسمی اور آرٹی ایف غیر قطاری ہائپرپلان کے لئے مفید ہیں جو اعلی پیمائش میں علیحدہ لائن کو شمار کرتے ہیں[198]۔

ii. **ریگولرائزیشن:** ریگولرائزیشن پیرامیٹر (پائتھن کے اسکائٹ لرن 'این یو' پیرامیٹر میں) باقاعدگی کو برقرار رکھنے کے لئے استعمال کیا جاتا ہے۔ یہاں 'این یو' پینالٹی پیرامیٹر ہے، جو غلط درجہ بندی یا غلط اصطلاح کی نمائندگی کرتا ہے[199]۔ غلط درجہ بندی یا غلط اصطلاح SVM آپٹمائزیشن کو بتاتی ہے کہ کتنی غلطی قابل برداشت ہے۔ فیصلے کی حد اور غلط درجہ بندی کی اصطلاح کے مابین تجارتی تعلقات کو آپ اس طرح کنٹرول کر سکتے ہیں۔ 'این یو' کی ایک چھوٹی سی قیمت ایک چھوٹے مارجن ہائپرپلان کی تخلیق کرتی ہے اور 'این یو' کی ایک بڑی قدر بڑے مارجن ہائپرپلان پیدا کرتی ہے[200]۔

iii. **گاما:** گاما غیر قطاری ہائپر پلان کے لئے ایک پیرامیٹر ہے۔ گاما کی ایک چھوٹی قدر ٹریننگ ڈیٹا سیٹ کو آسانی سے فٹ کرے گی، جب کہ گاما کی اعلیٰ قدر ٹریننگ ڈیٹا سیٹ کے عین مطابق فٹ بیٹھتی ہے، جس کی وجہ سے اوور فٹنگ ہونے کا سبب بنتا ہے[201]۔ دوسرے لفظوں میں، آپ کہہ سکتے ہیں کہ گاما کی ایک چھوٹی قدر علیحدگی کی لکیر کا حساب لگانے میں صرف قریبی نکات پر ہی غور کرتی ہے، جبکہ گاما کی اعلی قدر علیحدگی کی لکیر کے حساب کتاب میں موجود تمام ڈیٹا پوائنٹس پر غور کرتی ہے۔ ہم دیکھ سکتے ہیں کہ بڑھتے ہوئے گاما کی وجہ سے زیادہ مناسب ہوتا ہے کیونکہ درجہ بند ٹریننگ ڈیٹا سیٹ کو مکمل طور پر فٹ کرنے کی کوشش کرتا ہے۔

iv. **امکان** (پروبیبلیٹی): ماڈل کا طرز عمل گاما پیرامیٹر کے ساتھ بہت حساس ہے۔ اگر ہم گاما کی قدر کو بہت زیادہ طے کرتے ہیں تو اس کا نتیجہ زیادہ مناسب ہو جاتا ہے اور جب گاما کی قدر چھوٹی ہوتی ہے تو، ماڈل بہت محدود ہوتا ہے اور ڈیٹا کی پیچیدگی یا 'شکل' پر گرفت نہیں کر سکتا ہے[202]۔ انٹرمیڈیٹ اقدار کے لئے، ہم دیکھ سکتے ہیں کہ ایک اچھا ماڈل 'این یو' اور گاما کے کونے پر پایا جا سکتا ہے۔ آخر میں، کوئی یہ بھی مشاہدہ کر سکتا ہے کہ گاما کی کچھ انٹرمیڈیٹ اقدار کے لئے، ہم برابر کارکردگی کا مظاہرہ کرنے والے ماڈل حاصل کرتے ہیں۔ نیچے دیئے

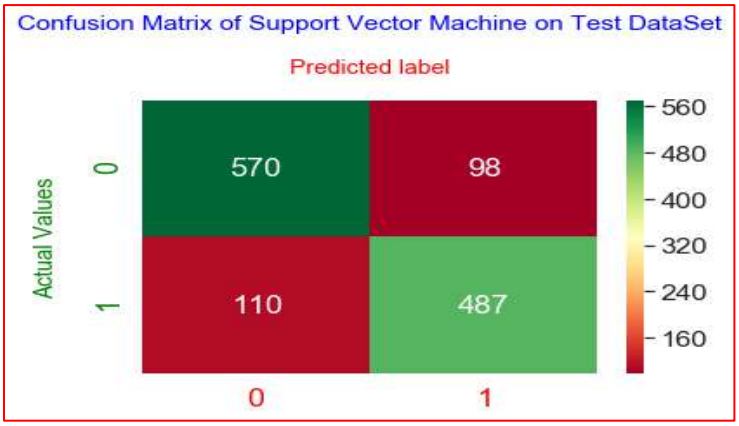
گئے جدول 6.3.3 کے ایس وی ایم ماڈل کے مختلف ہائپر پر میٹرز کو مد نظر رکھنے کے بعد حاصل کی گئی مختلف درستگی دکھاتی ہیں۔ایس وی ایم ماڈل کے 'کرنل (دانا)'، 'این یو'، 'گاما'، اور امکان (پرو بیبلیٹی) ہائپر پیرامیٹر کا استعمال زیادہ سے زیادہ درستگی کے لئے کیا جاتا ہے۔

جدول 6.3.3: درستگی کے ساتھ ایس وی ایم ہایپر پیرامیٹر آپٹمائزیشن

nu	Gamma	kernel	Probability	Accuracy
0.1	0.1	rbf	True	81%
0.6	0.01	rbf	True	78%
0.2	0.01	rbf	True	83%
0.8	0.01	rbf	True	76%
0.6	0.01	rbf	True	78%
0.7	0.01	rbf	True	76%
0.7	0.001	rbf	True	78%
0.5	0.0001	rbf	True	83%
0.6	0.0001	rbf	True	81%
0.7	0.0001	rbf	True	79%
0.8	0.0001	rbf	True	78%
0.9	0.0001	rbf	True	77%
0.9	0.001	rbf	True	75%
0.8	0.001	rbf	True	77%
0.7	0.001	rbf	True	78%
0.6	0.001	rbf	True	79%
0.5	0.001	rbf	True	83%
0.5	0.0001	rbf	True	83%
0.6	0.1	rbf	True	78%
0.7	0.1	rbf	True	76%
0.8	0.1	rbf	True	76%

ہم نے پایا کہ جب جب کرنل ہائپر پیرامیٹر قدریں (لکیری یا سگموئڈ) سیٹ کی جاتی ہیں تو ماڈل کے وقت کی پیچیدگی بڑھ جاتی ہے۔ تجرباتی نتائج سے پتہ چلتا ہے کہ جب SVM ہائپر پیرامیٹر کے امتزاج ہوتے ہیں تو

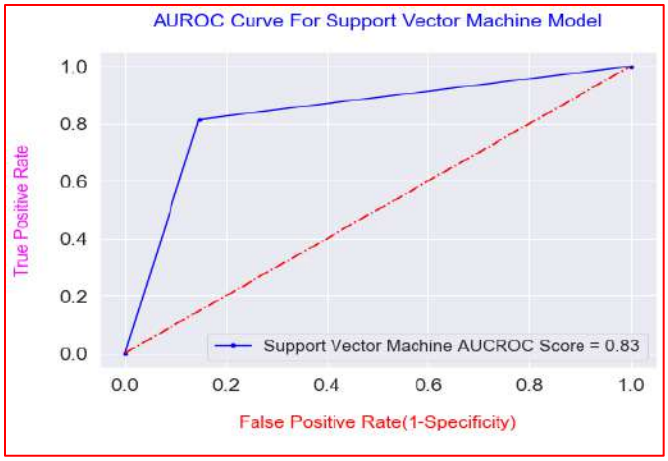
[nu=0.5, Kernel=rbf, Gamma=0.0001, Probability =True]اور83فیصد کی اعلیٰ ترین درستگی حاصل کی جاتی ہے۔ ماڈل کی کارکردگی کے نتائج کنفیوژن میٹرکس کے ترسیم 6.3.3.1 میں دکھائے گئے ہیں۔ مندرجہ بالا ڈیٹاسیٹ 6.3.3.1 سے، ہم یاد(ری کال)، خاصیت، پہچان کی شرح، صحت سے متعلق اور مختلف درجہ بندی کی شرح اخذ کرتے ہیں جس کو مندرجہ ذیل میں بیان کیا گیا ہے۔



ترسیم 6.3.3.1 ہائپر پرامیٹر ایس وی ایم کنفیوژن میٹرکس

مساوات 3.11 کا استعمال کرتے ہوئے ہم ایس وی ایم ماڈل کی حقیقی مثبت شرح حاصل کرتے ہیں جیسے $\frac{(487)}{(487+110)}$ جو کہ 0.81 فیصد کے برابر ہے لہذا ایس وی ایم ماڈل مثبت انتخابات جیتنے والے معاملات کو

81 فیصد کی درستگی کے ساتھ پہچان سکتا ہے۔اسی طرح، مساوات 3.12 کو استعمال کر کے ہمیں ایس وی ایم ماڈل کی حقیقی منفی شرح حاصل ہو جاتی ہے جیسے $\frac{(570)}{(570+98)}$ جو کہ 0.85 فیصد کے برابر ہے، جس کا مطلب ہے کہ آپٹمائزڈ ایس وی ایم ماڈل انتخاب ہارے ہوئے معاملوں 85 فیصد کی درستگی کے ساتھ شناخت کر سکتا ہے۔ایس وی ایم ماڈل کی درستگی مساوات 3.13 یعنی $\frac{(570 + 487)}{(570+98+110+487)}$ کا استعمال کر کے حاصل کی گئی ہے جو کہ 0.83 فیصد کے برابر ہے جس کا مطلب ہے کہ ایس وی ایم ماڈل کی مجموعی کارکردگی میں الیکشن جیتنے اور الیکشن ہار جانے والے معاملات کی پیشن گوئی 83 فیصد ہے۔آپٹمائزڈ ایس وی ایم ماڈل کی درستگی حاصل کرنے کے لئے مساوات 3.14 استعمال کی گئی ہے اور جس کی اقدار کا حساب لگایا جاتا ہے جیسے $\frac{(487)}{(487+98)}$ جو نتائج کو 0.83 فیصد دکھاتے ہیں، اس کا مطلب یہ ہے کہ آپٹمائزڈ ایس وی ایم ماڈل کم غلط مثبت شرح کا حامل ہے۔ مجوزہ ایس وی ایم ماڈل کی غلط درجہ بندی کی شرح مساوات 3.15 $\frac{(110+ 98)}{(570+98+110+487)}$ کا استعمال کر کے حاصل کی گئی ہے جو کہ 16 فیصد کے برابر ہے۔



ترسیم 6.3.3.2: ہائپر پیرامیٹر ایس وی ایم ماڈل کا AUROC کرو

ہم یہ دیکھنے کے لئے بھی AUROC کارکردگی کی پیمائش کا استعمال کرتے ہیں کہ ماڈل دو چیزوں میں کتنا اچھا فرق کر سکتا ہے (جیسے اگر کسی پارٹی یا آزاد امیدوار نے الیکشن جیت لیا یا ہار گیا)۔ ترسیم 6.3.3.2 یہ ظاہر کرتا ہے کہ AUROC ایس وی ایم ماڈل سے 83 فیصد اسکور کے ساتھ حاصل کیا۔

ہم موجودہ تحقیق کے ساتھ آپٹمائزڈ ایس وی ایم انتخابی پیشن گوئی ماڈل کے کامیاب تجرباتی نتائج کو نقل کرتے ہیں۔ ایس وی ایم ماڈل کے تجرباتی نتائج انتخابی نتائج کی پیشن گوئی کے لئے بہترین ہیں۔

6.3.4 رینڈم فاریسٹ ہائپر پیرامیٹر آپٹمائیزیشن Random Forest

Hyperparameter Optimization

رینڈم فاریسٹ میٹا کا تخمینہ لگانے والا ہے جو ڈیٹا سیٹ کے مختلف ذیلی نمونوں پر کئی ایک کی درجہ بندیوں کو فٹ کرتا ہے اور پیشن گوئی کی درستگی کو بہتر بنانے اور اوور فٹنگ کو کنٹرول کرنے کے لئے اوسط استعمال کرتا ہے [203] ہم رینڈم فارسٹ الگورتھم کے انتہائی اہم پیرامیٹرز کی چھان بین کرتے ہیں جن پر ذیل میں تبادلہ خیال کیا گیا ہے۔

i. **بوٹ اسٹریپ:** یہ ڈیٹا پوائنٹس (متبادل کے ساتھ یا بغیر) نمونے لینے کا ایک طریقہ ہے۔

ii. **زیادہ سے زیادہ گہرائی:** زیادہ سے زیادہ گہرائی کا پیرامیٹر ہر ایک درخت کی زیادہ سے زیادہ گہرائی کی وضاحت کرتا ہے، زیادہ سے زیادہ گہرائی کی پہلے سے طے شدہ قیمت ''کوئی نہیں'' ہے، اس کا مطلب یہ ہے کہ ہر ایک درخت اس وقت تک وسعت پذیر ہو جائے گا جب تک کہ ہر پتی پاک نہیں ہوتا ہے، خالص پتی وہ ہے جہاں پتی کے تمام اعداد و شمار ایک ہی طبقے سے آتے ہیں۔

iii. **N اسٹیمیٹر:** N اسٹیمیٹر جنگل میں درختوں کی تعداد کی نمائندگی کرتے ہیں، عام طور پر درختوں کی تعداد جتنی زیادہ ہوتی ہے اس سے اعداد و شمار کو سیکھنے میں بہتر ہوتا ہے۔ تاہم، بہت سارے درخت شامل کرنے سے تربیت کا عمل کافی سست ہو سکتا ہے، لہذا ہم بہترین جگہ تلاش کرنے کے لئے پیرامیٹر کی تلاش کرتے ہیں، ہم دیکھ سکتے ہیں کہ ہمارے اعداد و شمار کے لئے، ہم 32 درختوں پر رک سکتے ہیں کیونکہ درختوں کی تعداد میں اضافے سے ٹیسٹ کی کارکردگی میں کمی واقع ہوتی ہے۔

iv. **کم از کم نمونے تقسیم:** کم از کم نمونے اسپلٹ، اسپلٹ نمونوں کی کم از کم تعداد کی نمائندگی کرتا ہے جو داخلی نوڈ کو تقسیم کرنے کے لئے درکار ہوتا ہے۔ ہر نوڈ میں کم از کم ایک نمونہ پر غور کرنے اور ہر نوڈ کے تمام نمونوں پر غور کرنے کے درمیان یہ مختلف ہو سکتا ہے، جب ہم اس پیرامیٹر میں اضافہ کرتے ہیں تو، جنگل میں ہر ایک درخت زیادہ مجبور ہو جاتا ہے کیونکہ اسے ہر نوڈ پر مزید نمونوں پر غور کرنا پڑتا ہے۔ یہاں ہم نمونے کے 10 فیصد سے 100 فیصد تک پیرامیٹر کو مختلف کریں گے، یہ معلوم ہے کہ جب ہر نوڈ پر تمام نمونے لینے کی ضرورت ہوتی ہے تو، ماڈل اعداد و شمار کے بارے میں کافی کچھ نہیں سیکھ سکتا ہے اور انڈر فٹنگ کیس کا باعث بنتا ہے۔

v. **کم سے کم نمونوں کا پتا:** کم از کم نمونے لیف، پتی کے نوڈ پر نمونے کی کم از کم تعداد ہوتی ہے، یہ پیرامیٹر

کم از کم نمونے اسپلٹ کی طرح ہے، تاہم اس میں درخت کی بنیاد، پتی پر نمونوں کی کم از کم تعداد کی وضاحت کی گئی ہے۔ اس قدر کو بڑھانا انڈر فٹنگ کا سبب بن سکتا ہے۔

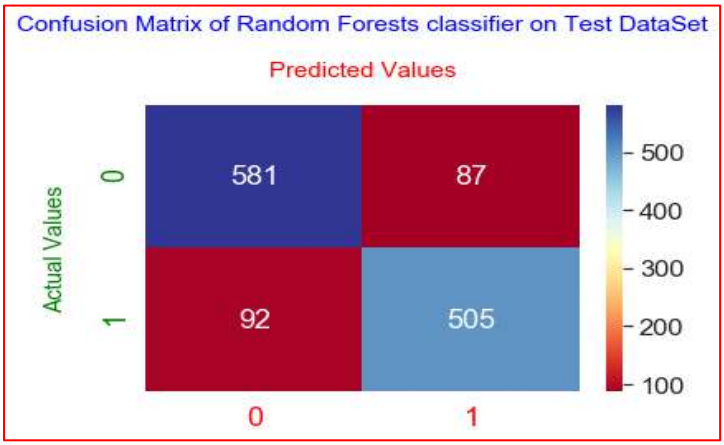
رینڈم فارسٹ ماڈل کے ہائپر پیرامیٹروں کو ٹیوننگ کرنے کے بعد، ہم نتائج حاصل کرتے ہیں، جیسا کہ نیچے دیئے گئے جدول 6.3.4 میں دکھائے گئے ہیں، مرضی کے مطابق جنگلاتی ماڈل کے تقویت اور امتزاج مختلف نتائج دکھاتے ہیں تاہم ہم صرف وہی پیرامیٹر کے امتزاج بیان کرتے ہیں جو سب سے زیادہ درستگی فراہم کرتے ہیں۔ تجرباتی نتائج سے پتہ چلتا ہے کہ جب ہائپر پیرامیٹر کے امتزاج ہوتے ہیں تو [بوٹ اسٹرپ = صحیح، زیادہ سے زیادہ گہرائی=100، کم از کم نمونوں کی لیف=50، کم از کم نمونے تقسیم=110، N estimators=1000، رینڈم اسٹیٹ=42،] سب سے زیادہ درستگی 85 فیصد حاصل کر لیا گیا۔

جدول 6.3.4: رینڈم فاریسٹ ہائپر پیرامیٹر آپٹمائزیشن اپنی درستگی کے ساتھ

Boot-strap	Max depth	Min samples leaf	Min samples split	N estimators	Random state	Accuracy
True	200	10	150	1000	42	85%
True	500	10	150	1000	42	85%
True	100	10	150	1000	42	85%
True	100	5	150	1000	42	85%
True	100	20	150	1000	42	84%
True	100	20	170	1000	42	83%
True	100	20	170	100	42	83%
False	100	20	270	100	42	83%
False	100	40	270	100	42	83%
False	100	50	270	100	42	82%

True	100	50	150	1000	42	83%
True	100	50	130	1000	42	84%
True	100	50	110	1000	42	85%

ہم نے جموں و کشمیر کے انتخابی پیشن گوئی کے لئے رینڈم فاریسٹ ماڈل کا انتخاب کیا، ماڈل کی کارکردگی کے نتائج کنفیوژن میٹرکس کے ترسیم 6.11 میں دکھائے گئے ہیں۔



ترسیم 6.3.4.1: ہائپر پرامیٹر رینڈم فاریسٹ کنفیوژن میٹرکس

مذکورہ بالا ترسیم 6.3.4.1 سے، ہم حقیقی مثبت شرح، حقیقی منفی شرح (خصوصیت)، شناخت کی شرح، صحت سے متعلق اور متفرق درجہ کی شرح اخذ کرتے ہیں جس کو مندرجہ ذیل بیان کیا گیا ہے۔

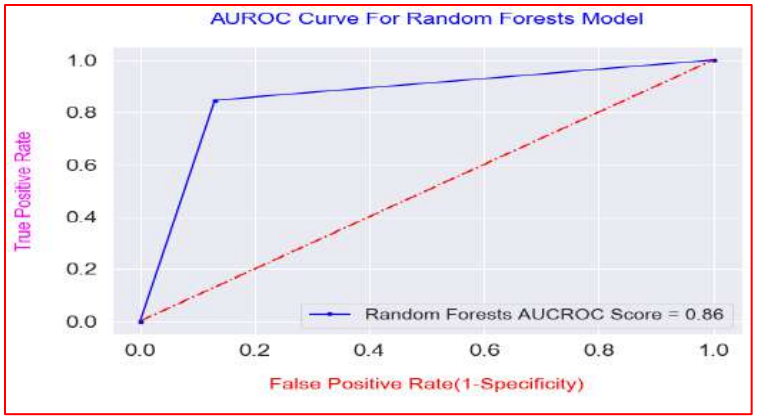
مساوات 3.11 کا استعمال کرتے ہوئے ہم نے رینڈم فاریسٹ ماڈل کی حقیقی مثبت شرح حاصل کی جیسے $\frac{(505)}{(505+92)}$ جو کہ 0.84 فیصد کے برابر ہے لہذا ہمارے رینڈم فاریسٹ ماڈل انتخاب کی جیت کے

مثبت معاملات 84 فیصد کی درستگی کے ساتھ پہچان سکتے ہیں۔اسی طرح، مساوات 3.12 کے ذریعہ ہمیں رینڈم فاریسٹ ماڈل کی حقیقی منفی شرح $\frac{(581)}{(581+87)}$ مل جاتی ہے جو کہ 0.86 فیصد کے برابر ہے اس کا مطلب ہے کہ رینڈم فاریسٹ ماڈل انتخاب کے کھوئے ہوئے معاملوں کو 86 فیصد کی درستگی کے ساتھ پہچان سکتا ہے، رینڈم فارسٹ ماڈل کی درستگی کو مساوات 3.13 کا استعمال کر کے حاصل کی گئی ہے جو کہ $\frac{(581 + 505)}{(581+87+92+505)}$ ہے حساب کے بعد جو 0.85 فیصد کے برابر ہے، جس کا مطلب ہے کہ آپٹمائزڈ رینڈم فاریسٹ ماڈل انتخابات جیتنے اور ہارے ہوئے دونوں واقعات کی پیشن گوئی کرنے میں اس کی مجموعی کارکردگی 85 فیصد ہے۔ مطلوبہ رینڈم فاریسٹ ماڈل کی صحت سے متعلق حاصل کرنے کے لئے مساوات 3.14 استعمال کی جاتی ہے اور مساوات کی جانچ کے بعد نتائج $\frac{(526)}{(526+146)}$ کے طور پر حاصل کیے جاتے ہیں نتیجہ 0.85 فیصد ہے، اس کا مطلب یہ ہے کہ ہمارا مطلوبہ رینڈم فاریسٹ ماڈل میں غلط شرح کم ہے، مجوزہ رینڈم فارسٹ ماڈل کی غلط درجہ بندی کی شرح مساوات 3.15 $\frac{(87+ 92)}{(581+87+92+505)}$ کا استعمال کر کے حاصل کی جاتی ہے جو 14 فیصد کے برابر ہے۔ ہم احتمال کے منحنی خطوط اور مرضی کے مطابق رینڈم فارسٹ ماڈل سے

حاصل کردہ علیحدگی کی پیمائش کو دیکھنے کے لئے AUROC کارکردگی کی پیمائش کا بھی استعمال کرتے ہیں، AUROC کرو ہمیں بتاتا ہے کہ پارٹی یا آزاد امیدواروں کے ذریعہ الیکشن جیت یا ہار کے معاملے میں ماڈل کتنا اچھا فرق کر سکتا ہے۔

بہتر ماڈلز دونوں کے درمیان درست طور پر تمیز کر سکتے ہیں تاہم ناقص ماڈلز کو ان میں تمیز کرنے میں مشکلات پیش آئیں گی، ذیل میں دیئے گئے ترسیم 6.12 کے مطابق، AUROC اسکور کے ساتھ رینڈم فارسٹ ماڈل سے حاصل کردہ AUROC کرو دکھاتا ہے۔ ہم مروجہ تحقیق کے ساتھ آپٹمائزڈ رینڈم فارسٹ انتخابی پیشن گوئی ماڈل کے کامیاب تجرباتی نتائج کی نقالی کرتے ہیں۔

حاصل کردہ نتائج ہمارے بہترین مشاہدے کے ہیں جو ادب میں شائع شدہ اقدار سے کہیں زیادہ ہیں تاہم ماڈل کی کارکردگی میں مزید بہتری کی ضرورت ہے، لہذا ہم جموں و کشمیر کے انتخابی نتائج کی مؤثر طریقے سے پیشن گوئی کرنے کے لئے مجوزہ مطلوبہ رینڈم فارسٹ ماڈل کا استعمال کرتے ہیں۔



ترسیم 6.3.4.2: ہائپر پرامیٹر رینڈم فارسٹ ماڈل کے ذریعے AUROC کرو

6.4 آپٹمائزڈ ماڈل میں کارکردگی کا موازنہ Performance Comparison among Hyperparameterized Models

اس حصے میں ہم ترقی یافتہ انتخابی پیشن گوئی کے ماڈلز کی تشخیص اور موازنہ پیش کر رہے ہیں۔ کارکردگی کو جانچنے کے لئے ہم مختلف پرفارمنس میٹرکس استعمال کرتے ہیں جس کو نیچے دیئے گئے جدول 6.4 میں بیان کیا گیا ہے۔ تجرباتی نتائج سے پتہ چلتا ہے کہ مطلوبہ رینڈم فاریسٹ ماڈل نے دوسرے ماڈلز کو پیچھے چھوڑ دیا، مجوزہ انتخابی پیشن گوئی ماڈل کی کارکردگی کا تجربہ موجودہ ماڈلز کے ساتھ کیا جاتا ہے جو انکشاف کرتے ہیں کہ نتائج درست پیشن گوئی کرنے والی طاقت کے ساتھ بہت قریب ہیں۔ تجرباتی نتائج کی تنقیدی جانچ پڑتال کے بعد ہمیں معلوم ہوا کہ قیمتی معلومات کو نکالنے اور غیر متزلزل ماڈل تیار

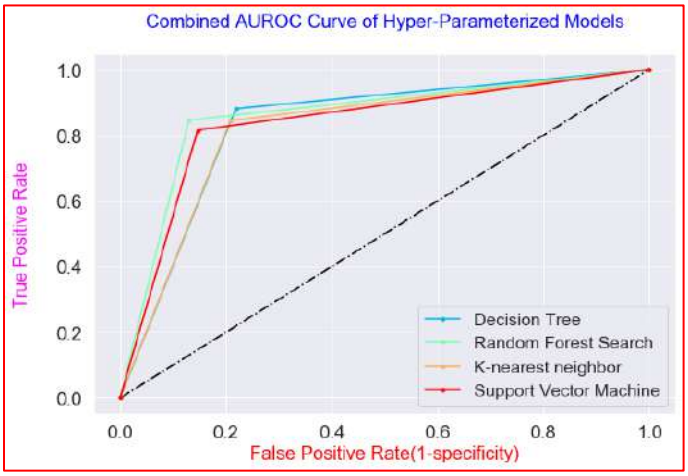
کرنے کے لئے اعداد و شمار کو احتیاط سے جانچنا ضروری ہے۔نتائج سے پتہ چلتا ہے کہ رینڈم فارسٹ ماڈل نے کم سے کم غلطی کی شرح صرف 14 فیصد کے ساتھ 85 فیصد کی زیادہ سے زیادہ درستگی حاصل کی ہے۔

جدول 6.4 مختلف انتخابی پیشن گوئی کے ماڈلز کی کارکردگی میٹرکس

	Performance Measures of The Models					
Models	AUROC Score	F1Score	Classifier Accuracy	Recall Score	Precisio Score	Miss-Classification Score
Decision Tree	0.8312%	0.8289%	0.8284%	0.8810%	0.7827%	0.1715%
Random Forest	0.8578%	0.8494%	0.8584%	0.8458%	0.8530%	0.1415%
K-Nearest Neighbors	0.8180%	0.8129%	0.8166%	0.8442%	0.7838%	0.1833%
Support Vector Machine	0.8345%	0.8240%	0.8355%	0.8157%	0.8324%	0.1644%

ترسیم 6.4 کے مختلف مطلوبہ انتخابی پیشن گوئی ماڈلز کے مشترکہ AUROC منحنی خطوط دکھائے گئے ہیں۔ترسیم 6.4 سے یہ واضح ہے کہ رینڈم فارسٹ انتخابی پیشن گوئی ماڈل میں 85 فیصد کا سب سے زیادہ اسکور ہے جس کا مطلب یہ ہے کہ ماڈل انتخابی حلقوں کی سطح پر سیاسی جماعتوں یا آزاد امیدواروں کے جیتنے

یا ہار جانے والے انتخابات میں فرق کرنے کی بہترین صلاحیت رکھتا ہے، تاہم اس میں مزید بہتری کی ضرورت ہے۔

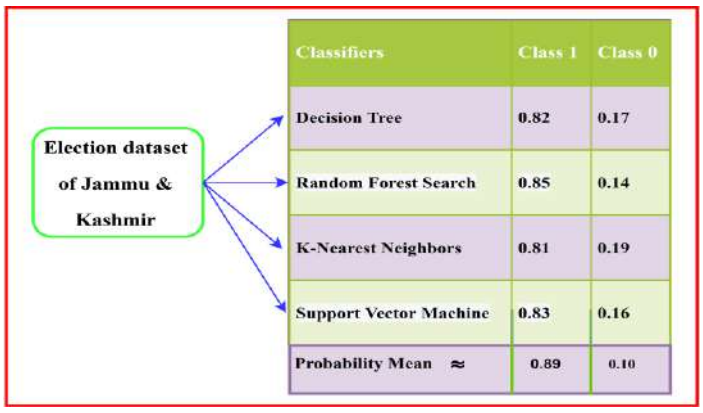


ترسیم 6.4: مجوزہ مطلوبہ انتخابی پیشن گوئی ماڈلز کے مشترکہ AUROC منحنی خطوط

6.5 اسمبل میتھڈ Ensemble Method

جیسا کہ اسمبلنگ تکنیک تکنیک (یعنی سخت ووٹنگ اور سافٹ ووٹنگ) کے بارے میں (باب تیسرا) میں زیر بحث آیا، اس تحقیقی کام میں سافٹ ووٹنگ اسمبلنگ تکنیک کا استعمال کیا گیا ہے کیونکہ سافٹ ووٹنگ سے استعمال ہونے والے تمام درجہ بندی کے امکانات کا مطلب ہوتا ہے۔ اس تحقیقی کام میں محقق نے چار مختلف مشین لرننگ کلاسیفائیرز جیسے decision tree, random

forest, K-NN اور SVM استعمال کی ہے۔اب سوال یہ پیدا ہوتا ہے کہ انتخابی پیش گوئی ماڈل کے حتمی نتائج بنانے میں کس درجہ بندی کار محقق نے استعمال کیا، زیادہ تر محقق نے صرف وہ درجہ بندی کا اطلاق کیا جو زیادہ سے زیادہ درستگی دیتا ہے۔ لیکن اس تحقیقی کام میں محققین اسمبلنگ تکنیک والے طریقے (سافٹ ووٹنگ) کا استعمال کر کے چاروں مشین لرننگ کلاسیفائرز کی اوسط امکان کا مطلب لے رہے ہیں۔



Classifiers	Class 1	Class 0
Decision Tree	0.82	0.17
Random Forest Search	0.85	0.14
K-Nearest Neighbors	0.81	0.19
Support Vector Machine	0.83	0.16
Probability Mean ≈	0.89	0.10

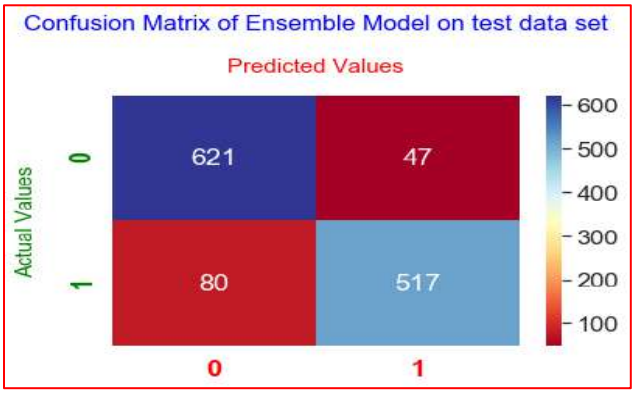
ترسیم 6.5: سافٹ ووٹنگ اسمبلنگ انتخابی پیشن گوئی ماڈل

۔اسمبل ماڈل کی مجموعی درستگی ترسیم 6.5 میں مساوات 3.6.2 کا استعمال کر کے حاصل کی گئی ہے جو کہ 0.89 فیصد کے برابر ہے جو نمائندگی کرتا ہے کہ اسمبل ماڈل ہیں سافٹ ووٹنگ کا استعمال کرتے ہوئے،

انتخابی پیشن گوئی ماڈل کی مجموعی کارکردگی میں پارٹی یا آزاد امیدوار کی کامیابی اور ہار دونوں کی پیشن گوئی کرنے میں اضافہ ہوتا ہے۔

6.5.1 مجوزہ اسمبل انتخابی پیشن گوئی ماڈل کے تجرباتی نتائج

جموں وکشمیر اسمبلی ڈیٹاسیٹ پر اسمبل ماڈل کی پیشن گوئی کے نتائج کنفیوژن میٹرکس کے ترسیم 6.5.1.1 میں دکھائے گئے ہیں۔ ترسیم 6.5.1.1 حاصل کردہ حساسیت، صراحت، درستگی، صحتسے متعلق اور غلط درجہ بندی کی وضاحت ذیل میں کی گئی ہے۔



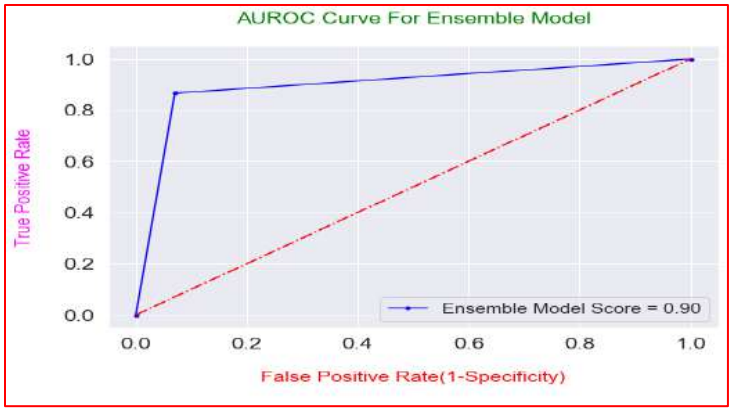
ترسیم 6.5.1.1: ٹیسٹ ڈیٹاسیٹ پر اسمبل ماڈل کا کنفیوژن میٹرکس

ترسیم 6.5.1.1 سے، ہم اسمبل ماڈل کی حقیقی مثبت شرح (یاد)، حقیقی منفی شرح (مخصوصیت)، شناخت کی

شرح، صحت سے متعلق اور غلط طبقے کی شرح اخذ کرتے ہیں جس کو ذیل میں بیان کیا گیا ہے۔

مساوات 3.11 کا استعمال کرتے ہوئے ہم اسمبل ماڈل کی حقیقی مثبت شرح جیسا کہ $\frac{(517)}{(517+80)}$ حاصل کرتے ہیں جو کہ 0.86 فیصد کے برابر ہے - لہذا ہمارا اسمبل ماڈل انتخابی جیت کے مثبت معاملات کو 86 فیصد درستگی کے ساتھ پہچان سکتا ہے، اسی طرح مساوات 3.12 کے ذریعہ ہمیں اسمبل ماڈل کی حقیقی منفی شرح جیسے $\frac{(621)}{(621+47)}$ موصول ہوتی ہے جو 0.92 فیصد کے برابر ہے اس کا مطلب ہے کہ اسمبل ماڈل انتخاب کے ہارے ہوئے معاملات کو 92 فیصد درستگی کے ساتھ پہچان سکتا ہے، اسمبل ماڈل کی درستگی مساوات 3.13 کا استعمال کرکے حاصل کی جاتی ہے جو $\frac{(621 + 517)}{(621+47+80+517)}$ ہے جو حساب کے بعد 0.89 فیصد کے برابر ہے، جس کا مطلب ہے کہ اسمبل ماڈل کی مجموعی کارکردگی انتخابات کی جیت اور ہار دونوں معاملوں میں 89 فیصد ہے۔ اسمبل ماڈل کی درستگی حاصل کرنے کے لئے مساوات 3.14 کا استعمال کیا جاتا ہے اور نتائج $\frac{(517)}{(517+47)}$ کے طور پر حاصل کیے جاتے ہیں، مساوات کی جانچ پڑتال کے بعد نتیجہ 0.91 فیصد ہے اس کا مطلب یہ ہے کہ ہمارے اسمبل ماڈل میں کم غلط شرح ہے، مجوزہ اسمبل ماڈل کی غلط درجہ بندی کی شرح مساوات 3.15 $\frac{(47+80)}{(621+47+80+517)}$

کا استعمال کر کے حاصل کی گئی ہے جو 10 فیصد کے برابر ہے۔ ہم اسمبل ماڈل کے ذریعہ حاصل کردہ امتیازی کَرو اور علیحدگی کی پیمائش کو دیکھنے کے لئے بھی AUROC کارکردگی کی پیمائش کا استعمال کرتے ہیں۔ AUROC کَرو ہمیں بتاتا ہے کہ پارٹی یا آزاد امیدواروں کے ذریعہ انتخابات جیتنے یا ہارنے دونوں معاملوں میں ماڈل کتنا اچھا فرق کر سکتا ہے، بہتر ماڈل دونوں میں درست طریقے سے تمیز کر سکتے ہیں۔ تاہم ناقص ماڈلز کو دونوں کے درمیان فرق کرنے میں مشکلات پیش آئیں گی، ذیل میں دیئے گئے ترسیم 6.5.1.2 میں AUROC اسکور 90 فیصد کے ساتھ ملحقہ ماڈل سے حاصل کردہ AUROC کَرو کو ظاہر کرتا ہے۔ ہم مروجہ تحقیق کے ساتھ تیار کردہ اسمبل انتخابی پیشن گوئی ماڈل کے کامیاب تجرباتی نتائج کی نقالی کرتے ہیں۔ حاصل کردہ نتائج ہمارے بہترین علم کے ہیں جو لٹریچر میں شائع شدہ نتائج سے کہیں زیادہ ہیں، لہذا مجوزہ اسمبل ماڈل جموں و کشمیر کے انتخابی نتائج کی حلد پیشن گوئی کے لئے استعمال کیا جاتا ہے۔



ترسیم 6.5.1.2: اسمبل ماڈل کے ذریعے AUROC کرو

6.6 مختلف مجوزہ انتخابی پیشن گوئی ماڈلز کی کارکردگی کا موازنہ:

حتمی بہترین پیشن گوئی ماڈل کی نشاندہی کرنے کے لئے، اسمبل ماڈل کے مقابلے میں ہر استعمال شدہ طریقہ کار کا بہترین کارکردگی ماڈل منتخب کیا گیا ہے، جدول 6.6 ہر طریقہ کے بہترین پرفارمنس ماڈل کے ساتھ وابستہ مختلف حسابات کی جانچ کے نتائج کو ظاہر کرتا ہے، منتخب کردہ مشین لرننگ کے دیگر چار ماڈلز کے مقابلے میں اسمبل ماڈل زیادہ درستگی اور کم غلط درجہ بندی کی کمی کی اقدار کو دکھاتا ہے۔

جدول 6.6: مختلف مجوزہ انتخابی پیشن گوئی ماڈل کی کارکردگی کا موازنہ

	Performance Measures of The Models					
Models	AUROC Score	F1 Score	Classifiers Accuracy	Recall Score	Precision Score	Miss-Classification Score

Decision Tree	0.8312%	0.8289%	0.8284%	0.8810%	0.7827%	0.1715%
Random Forest	0.8578%	0.8494%	0.8584%	0.8458%	0.8530%	0.1415%
K -Nearest Neighbors	0.8180%	0.8129%	0.8166%	0.8442%	0.7838%	0.1833%
Support Vector Machine	0.8345%	0.8240%	0.8355%	0.8157%	0.8324%	0.1644%
Ensemble Methods	0.8978%	0.8906%	0.8996%	0.8659%	0.9166%	0.1003%

اس سے یہ اشارہ ہوتا ہے کہ اسمبل میتھڈ دوسرے چار منتخب مشین لرننگ ماڈلز کی کارکردگی کو بہتر سمجھتے ہیں، اس کے علاوہ حتمی بہترین ماڈل کی شناخت اسمبل ماڈل کی حیثیت سے کی گئی ہے، چونکہ اس کی گنتی کی درستگی سب سے کم غلط شرح کے ساتھ ہے۔ ترسیم 6.5.1.2 نے اس نقطہ نظر کے پیشن گوئی شدہ نتائج کا مظاہرہ کیا جس سے یہ ظاہر ہوتا ہے کہ گزشتہ تین انتخابات کے لئے اسمبلی انتخابات کے نتائج کی پیشن گوئی کرنے کے لئے استعمال شدہ اسمبل پیشن گوئی کا طریقہ زیادہ بہتر ہے۔

6.7 ٹی پیئرڈ ٹیسٹ T-Paired Test

اس شماریاتی ٹی پیئرڈ ٹیسٹ کا استعمال کرتے ہوئے ہم تمام ترقی یافتہ انتخابات کی پیشن گوئی ماڈلز کا موازنہ ایک ایک کرکے اسمبل ماڈل کے ساتھ کرتے ہیں پی ویلیو معیار کا استعمال کرتے ہوئے، اگر ماڈل کی پی ویلیو 0.05 سے کم ہے، تو پھر اسمبل ماڈل نمایاں ہے اور یہ اس بات کا ثبوت ہوگا کہ اسمبل ماڈل منتخب

ماڈل سے مختلف انداز میں کارکردگی کا مظاہرہ کرتا ہے ، لہذا ہم منسوخ نظریے کو مسترد کرتے ہیں۔چونکہ کالعدم قیاس آرائی یہ ہے کہ یہاں کچھ نہیں چل رہا ہے یا مشین لرننگ کے دو ماڈلز کی کارکردگی میں کوئی فرق نہیں ہے، سمجھی جانے والی اہمیت کی سطح سے چھوٹی P ویلیو نے متبادل مفروضے کے حق میں غلط مفروضے کو مسترد کر دیا، جس سے یہ فرض کیا جاتا ہے کہ اسمبل ماڈل انتخابی پیشن گوئی کے جدید ماڈل کے لئے مختلف طریقے سے کارکردگی کا مظاہرہ کرتا ہے، اس کے علاوہ اہمیت کی سطح سے زیادہ ایک p-value سے پتہ چلتا ہے کہ ہم کالعدم نظریے کو مسترد کرنے میں ناکام ہیں۔ہم نے اس تحقیقی کام کے لئے اسمبل ماڈل کے ساتھ ڈسیزن ٹری رینڈم فارسٹ، K-NN اور SVM ماڈل کی کارکردگی کا موازنہ کرنے کے لئے مذکورہ طریقہ کار کا استعمال کیا، اور اخذ کردہ نتائج نیچے دیئے گئے جدول نمبر 6.7 میں درج ہیں۔

جدول نمبر 6.7: ٹی پیئرڈ ٹیسٹ کا استعمال کرتے ہوئے اسمبل ماڈل کے ساتھ مشین لرننگ ماڈلز کا موازنہ

S.No	Comparison of models with ensemble model	p value
1	Decision tree	0.006
2	Random Forest	0.001
3	K-Nearest Neighbors	0.009

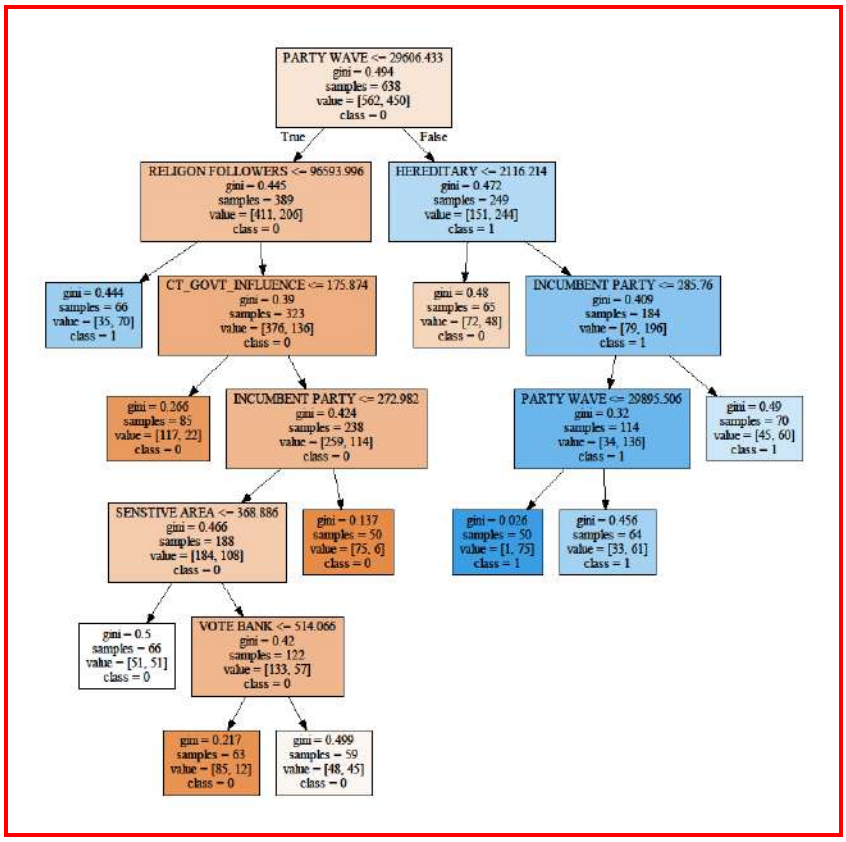
4	Support Vector Machine	0.014

اس تحقیقی کام میں استعمال شدہ سب سے بہتر انتخابی پیشن گوئی کے نمونوں کی پی ویلیو سمجھی جانے والی اہمیت کی سطح سے 0.05 فیصد چھوٹی ہے چنانچہ ہم اس ناقابل قیاس آرائی کو مسترد کرتے ہیں۔ لہذا نتائج اعداد و شمار کے مطابق اس بات کے قائل ثبوت فراہم کرتے ہیں کہ آپٹمائزڈ ڈسیزن ٹری ،رینڈم فاریسٹ، کے نیئرسٹ نائیبر اور ایس وی ایم ماڈل اسمبل ماڈل کے مقابلے میں مختلف کارکردگی کا مظاہرہ کرتے ہیں، اسمبل ماڈل کی اوسطاً درستگی تمام مجوزہ انتخابی پیشن گوئی ماڈلز کی نسبت زیادہ ہے، لہذا اسمبل ماڈل کی کارکردگی اور درستگی دوسرے مجوزہ انتخابی پیشن گوئی ماڈلز کے مقابلے میں اہم ہے، لہذا یہ اسمبل ماڈل جموں و کشمیر کے

6.8 انتخابات کی پیشن گوئی کی تشخیص کے لئے قواعد تیار کرنا:

انتخابی پیشن گوئی کو ایک نان لینیئر مسئلہ سمجھا جاتا ہے جو متغیر کے مابین پیچیدہ تعلقات کو ظاہر کرتا ہے، محققین انتخابی نتائج کی پیشن گوئی کرنے میں، رائے شماری کی پیشن گوئی کرنے والے کی مدد کے لئے ڈیٹا مائننگ کی متعدد تکنیک استعمال کر رہے ہیں، ہمارے کام میں، ڈسیزن ٹری، کے نیئرسٹ نائیبر، رینڈم

فاریسٹ ، سپورٹ ویکٹر مشین کی درجہ بندی اور اسمبل ماڈل کو پیشن گوئی کے لئے اہم انتخابی پیرامیٹرز کا تعین کرنے کے لیے تجویز کیا گیا ہے۔ سیاسی پیشن گوئی کرنے والے وہ عنصر ہیں جو پیشن گوئی کی شرح پر اثر انداز ہوتے ہیں، اور اسی وجہ سے پیشن گوئی کی جاتی ہے کہ یا تو سیاسی جماعتیں یا آزاد امیدوار مختلف اسمبلی حلقوں میں انتخاب جیت یا ہار سکیں۔ پارٹی کی لہر، مرکزی حکومت کے اثر ورسوخ، مذہب کے پیروکار، پارٹی خلاصے، حاس علاقوں، ووٹ بینک، وراثتی، مابعد پارٹی اور ذات پات کے عوامل انتخابی پیشن گوئی کے سب سے نمایاں عوامل۔انتخابی پیشن گوئی کے قواعد پیشہ ورانہ اور نابالغ صارف کو سمجھنے میں مدد فراہم کرنے کے لئے نکالا جاتا ہے کہ انتخابی نتائج کی پیشن گوئی کرنے میں ان مجوزہ ماڈلز کے ذریعہ ان اوصاف کے امتزاج کا استعمال کیا جاتا ہے- گیا ہے، تیار کردہ قواعد کو مختلف انتخابی پیشن گوئی کے معروف ماہرین کے ذریعہ منتخب، کٹے ہوئے، جانچنے اور توثیق کئے ہوئے ہیں۔ ڈسیزن ٹری کے اس خلاصہ کو سیاسی پیشن گوئی کرنے والے ماحول میں استعمال کیا جاسکتا ہے تاکہ عام سروے کے تجزیہ کار کم انتخاب ، قابل اعتماد اور موثر اندازہ لگانے والے ماڈل کے ساتھ نو خصوصیات کا استعمال کرتے ہوئے انتخابی نتائج کی ابتدائی پیشن گوئی قائم کر سکیں۔

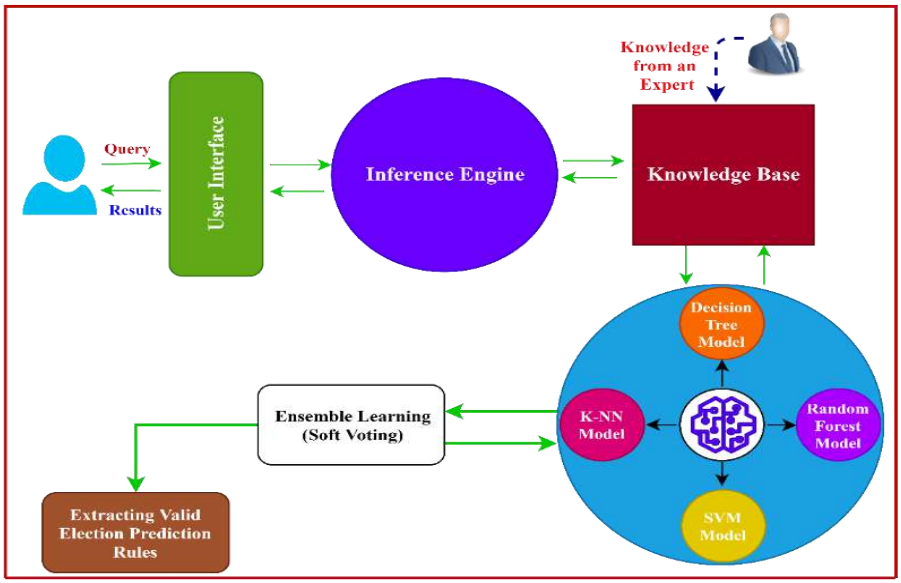


ترسیم 6.8: نو خصوصیات کا استعمال کرتے ہوئے انتخابی پیشن گوئی ڈسیزن ٹری

6.9 الیکشن پیشن گوئی ماہر سسٹم کی تشخیص ماڈل کے اجزاء:

مجوزہ انتخابی پیشن گوئی ماڈل جدید ہے کیونکہ اس میں کسی سیاسی جماعتوں یا آزاد امیدواروں کے انتخابی کامیابی یا صرف انتخابی اعداد و شمار کے اوصاف کا استعمال کرتے ہوئے ہارنے کے امکانات کی نشاندہی کی جاتی ہے، اس طرح اس کی عوامی اسکریننگ جانچ کے طور پر اس کی درخواست کی حمایت ہوتی ہے، عام زبان میں ہم اس ماڈل کو جے کے ای پی ایم (جموں و کشمیر الیکشن پریڈکشن ماڈل) کہتے ہیں۔ نیچے دیئے گئے

ترسیم 6.9 جے کے ای پی ایم کے تین اہم اجزاء کو دکھا رہے ہیں، علم کی بنیاد، انجن اور انٹرفیس۔ ماہرین نظام کے اصولوں کو نکالنے کے لئے علمی اساس جموں و کشمیر کے انتخابی اعداد و شمار کی خصوصیات پر مجوزہ ماڈل کا اطلاق کرتا ہے، اس کی خصوصیات مرکزی حکومت کا اثر و رسوخ، مذہب کے پیروکار، پارٹی کی لہر، پارٹی خلاصہ، حساس علاقوں، ووٹ بینک، وراثتی، پارٹی، ذات پات ہیں۔ انفیرینس انجن نکلے ہوئے اصولوں اور صارفین کے اِن پٹ کا استعمال کرتے ہیں تاکہ علم کی بنیاد سے نتائج اخذ کریں اور صارف انٹرفیس کے ذریعے صارف کے سامنے پیش کریں۔



ترسیم 6.9 انتخابات کی پیشن گوئی تشخیص کے آلے کے اُجزاء

صارف انٹرفیس 'مواصلات' اسکرینوں کی اجازت دیتا ہے جہاں صارف اِن پٹ ڈیٹا

میں داخل ہوتا ہے اور ماہر سسٹم انتخاب جیتنے یا ہارنے کا موقع واپس کرتا ہے جیسا کہ انفیرینس انجن کے

حساب سے ہوتا ہے۔

6.10 جموں و کشمیر انتخابی پیشن گوئی ماڈل (جے کے ای پی ایم):

ہم جموں و کشمیر کے لئے انتخابی پیشن گوئی ماڈل بناتے ہیں جو انتخابی نتائج کی پیشن گوئی کے لئے انتخابی حلقے کے

مطابق استعمال ہو سکتی ہے، باب 5 اور باب 6 میں، ترقی یافتہ ماڈل انتخابی پیشن گوئی ماڈل کی تعمیر میں

مختلف اوصاف کا استعمال کرتے ہیں۔ نتائج سے یہ ظاہر ہوتا ہے کہ پارٹی لہر، مرکزی حکومت کے اثر و

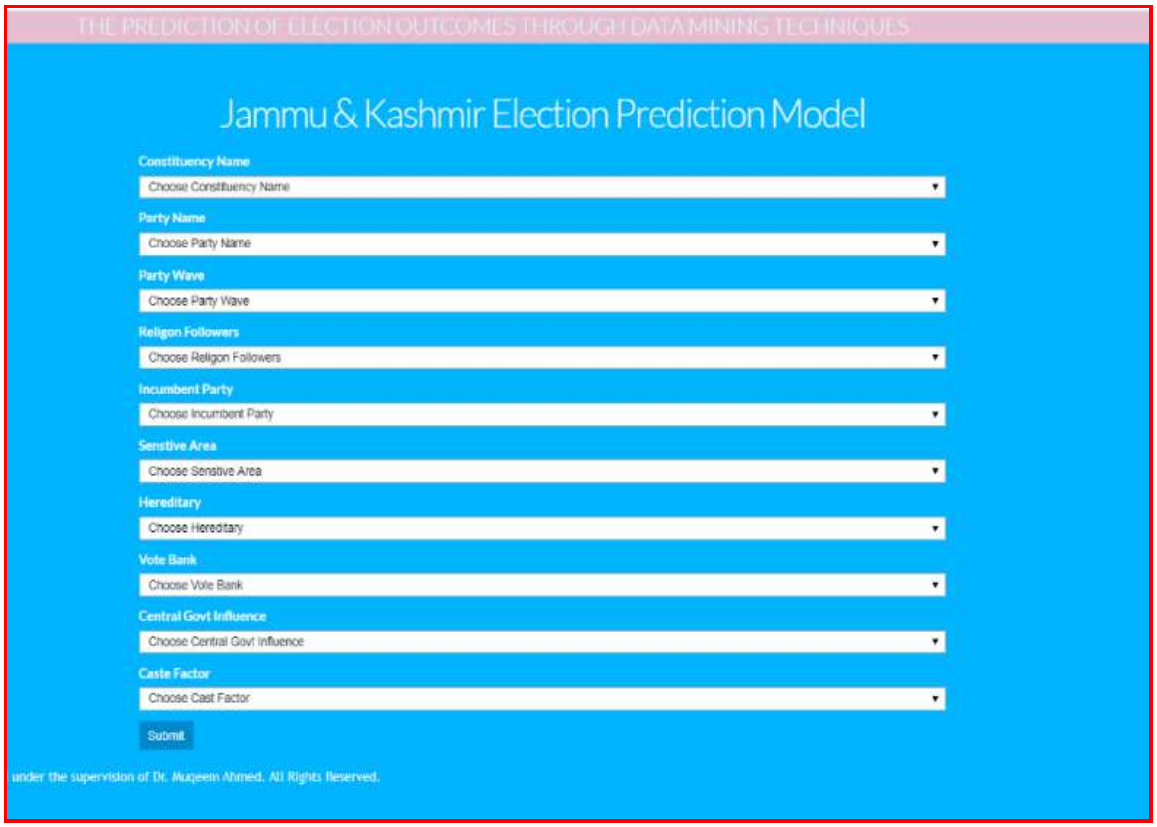
رسوخ، مذہب کے پیروکار، پارٹی کے مخففات، حساس علاقوں، ووٹ بینک، ذات پات، موجودہ پارٹی

اور موروثی کا امتزاج بہترین نتائج فراہم کرتا ہے۔ ان نتائج کو کافی حد تک زیادہ محسوس ہوتا ہے کہ اسیمبل ماڈل جو

کہ (ڈیسیزن ٹری، رینڈم فاریسٹ، کے نیرسٹ نائیبر اور سپورٹ ویکٹر مشین) پر مشتمل ہے، انتخابی

نتائج کی تشخیص کے لئے اسکریننگ ٹیسٹ بنانے کے لئے استعمال کیا جا سکتا ہے جس کو سمجھا اور استعمال کیا جا سکتا ہے۔ نیز پیشہ ور صارفین کے ذریعہ۔

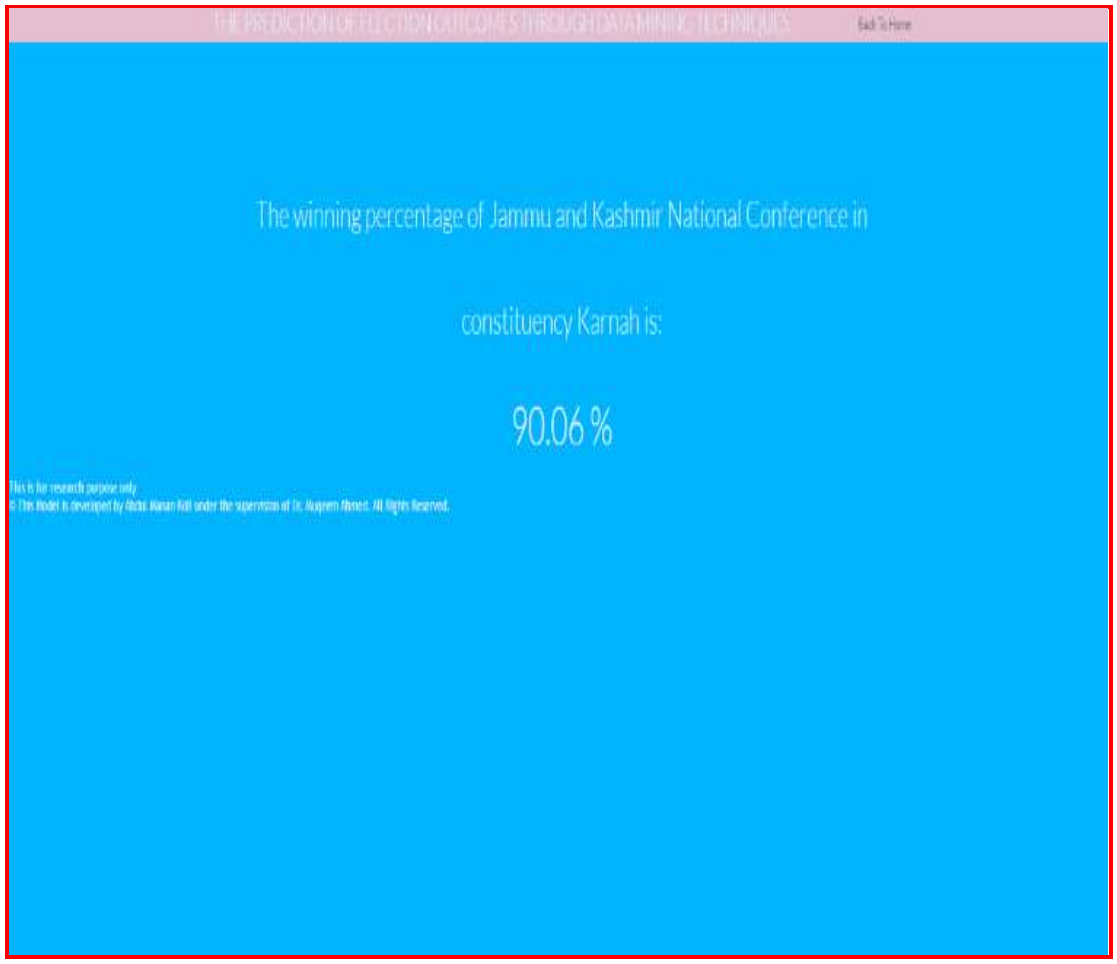
جے کے ای پی ایم کا ترقیاتی منصوبہ دو اہم مراحل پر مشتمل ہے، پہلے مرحلے میں اوصاف کو لوڈ کرنا اور مجوزہ ماڈلز کو جموں و کشمیر کے انتخابی ڈیٹا سیٹ کی صفات میں لاگو کرنا اور پھر پیشن گوئی کے قواعد کو نکالا اور محفوظ کیا جاتا ہے۔ دوسرے مرحلے میں ، صارف ماڈل اسکرین پر دکھائے گئے مختلف اوصاف پر کلک کرتا ہے اور ذخیرہ شدہ اقدار میں داخل ہوتا ہے، ان اوصاف کا استعمال ذخیرہ شدہ پیشن گوئی قواعد کے ذریعہ استعمال کنندہ کی ڈگری کو انتخاب کی پیشن گوئی کرنے کے لئے جمع کیا جاتا ہے جو کہ صارف کو ظاہر ہوتا ہے۔ جے کے ای پی ایم کا نفاذ پائتھن جوپیٹر نوٹ بک کا استعمال کرتے ہوئے کیا جاتا ہے۔



ترسیم 6.10.1:انتخابی پیشن گوئی ماڈل انٹرفیس

ترسیم 6.10.1 جے کے ای پی ایم کی اسٹارٹ اسکرین کو دکھاتا ہے جہاں صارف مختلف پیرامیٹرز پر ڈیٹا داخل کرتا ہے اور الیکشن جیتنے یا ہار جانے کے امکانات کا حساب کتاب اور ظاہر کیا جاتا ہے۔

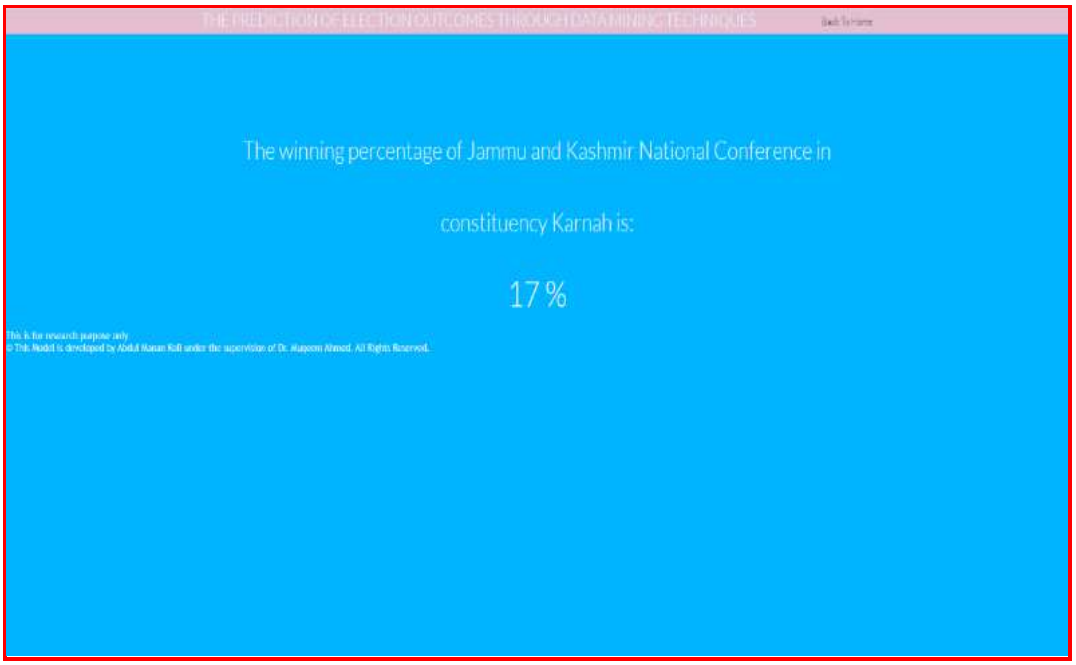
ترسیم 6.10.2 صارف کی طرف سے اعداد و شمار کا نتیجہ ہے جو شروعاتی اسکرین کے مختلف وابستہ قدروں میں داخل ہوتا ہے، یہاں جے کے ای پی ایم نے درج کردہ اعداد و شمار کے لئے حلقہ وار انتخابی جماعت کے لئے انتخابی نتائج کی اعلی ڈگری کا حساب کتاب کیا ہے۔



ترسیم6.10.2:انتخابات کی پیشن گوئی کی تشخیص کی مثال

ہم نے قیمت کی حد5.=< مقرر کیا ہے جس کا مطلب 50فیصد ہے۔اس کے اوپر سیاسی

جماعتیں یاآزاد امیدوار الیکشن جیت گئے اور اس کے نیچے وہ الیکشن ہار گئے ۔



ترسیم 6.10.3: انتخاب جیتنے کے کم امکانات کی مثال

ترسیم 6.10.3 جے کے ای پی ایم ماڈل کی دوسری مثال ہے، اس معاملے میں، انتخاب جیتنے کے کم مواقع کا حساب ابتدائی اسکرین پر درج اعداد و شمار کی بنیاد پر کیا جاتا ہے۔

ان مثالوں سے ظاہر ہوتا ہے کہ جے کے ای پی ایم عوامی سطح کی اسکریننگ ٹیسٹ کے طور پر کام کر سکتا ہے۔ صارف کے انٹرفیس کی سادگی سیاسی یا غیر سیاسی پیشہ ور افراد کو سیاسی جماعتوں یا آزاد امیدواروں کے انتخابی حلقے کے لحاظ سے جیتنے یا ہارنے کے امکانات کی نشاندہی کرنے کی اجازت دیتی ہے۔ جے کے ای پی ایم کو موبائل کے ساتھ ساتھ ڈیسک ٹاپ ایپلیکیشنز پر بھی لاگو کیا جاتا ہے۔

6.11 باب کا خلاصہ اور نتیجہ:

اس باب میں، ہم نے ہائپر پیرامیٹر آپٹمائزڈ تکنیک اور ان کی مختلف اقسام کا تعارف کرایا ہے۔ ہم انتخابی نتائج کی درست پیشن گوئی کرنے کے لئے سیاسی یا غیر سیاسی صارفین کی مدد کرنے کے لئے مجوزہ ماڈل کو بہتر بناتے ہیں۔ ہم آپٹمائزڈ الیکشن پریڈکشن ماڈل اور اسیمبل ماڈل کے درمیان موازنہ بھی کرتے ہیں۔ مجوزہ ماڈل جو پیٹر نوٹ بک ویب ایپلی کیشن پر تیار کیا گیا ہے اور ہر ترقی یافتہ ماڈل کے لئے ٹی پی آر، ٹی این آر، درستگی، پریشانی، متفرق شرح اور AUROC کروا سکور کی کمپیوٹنگ کے ذریعہ کارکردگی کے مختلف اقدامات کئے جاتے ہیں اور آخر میں سافٹ ووٹنگ کا استعمال کرکے چاروں ماڈلز کو ایک ماڈل میں جوڑ دیتے ہیں۔ ووٹ ڈالنے کی تکنیک شماریاتی ٹی پیئرڈ ٹیسٹ کا اطلاق اس بات کے لئے کیا جاتا ہے کہ مجوزہ انتخابی پیشن گوئی کا ماڈل نمایاں ہے یا نہیں۔ تجرباتی نتائج سے پتہ چلتا ہے کہ اسیمبل ماڈل اہم ہے کیونکہ اس نے دوسرے مجوزہ ماڈلز کو مات دیدیا ہے، جے کے ای پی ایم (جموں و کشمیر انتخابی پیشگوئی ماڈل) ماڈل کی ترقی کو بھی بیان کیا گیا ہے جو عوامی سطح کے اسکریننگ ماڈل کے طور پر استعمال ہوتا ہے جو جموں و کشمیر کے انتہائی اہم پیرامیٹرز کی بنیاد پر الیکشن ہارنے یا جیتنے کی ڈگری کی نشاندہی کرتا ہے، اسی طرح عام لوگوں کے

ساتھ ساتھ تربیت یافتہ رائے شماری کی پیشن گوئی ممکنہ انتخابی نتائج کی سستی، قابل اعتماد اسکریننگ کے ساتھ فراہم کرے گی، آخر میں، انتخابی نتائج کے قواعد کا زیادہ سے زیادہ سیٹ ان ماڈلز کے ذریعہ تیار کیا جاتا ہے اور سروے کی پیشن گوئی کرنے والے معروف ماہرین کے ذریعہ ان کی جانچ اور توثیق کی جاتی ہے۔

باب 7

7. اختیتام اور مستقبل کا کام

7.1 مقالے کا خاتمہ

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۔ جموں و کشمیر کے لئے اعداد و شمار کی (data mining) کی تکنیک کا استعمال کرتے ہوئے انتخابی نتائج کی پیش گوئیاں جو پیرامیٹرک نقطہ نظر پر مبنی انتخابی حلقہ وار بھی ایک مشکل کام ہے کیونکہ مختلف حلقوں کی ثقافت، رسم و رواج اور رہائش گاہ مختلف ہیں۔ نیز، جموں ڈویژن کے عوام کا سیاسی تناظر مختلف ہے اور پھر کشمیر ڈویژن کے عوام۔ اگرچہ ٹویٹر اور فیس بک کے ڈیٹا یا ایگزٹ پول پر مبنی پیش گوئی کے نئے طریقے اب پیشن گوئی کا معیار بن چکے ہیں لیکن یہ طریقہ کار مہنگا اور پیچیدہ عمل ہے جو جموں و کشمیر کے دیہی علاقوں میں ان کے استعمال کو روکتا ہے۔ یہ تواتر سے انٹرنیٹ بند ہونے، کم رابطے کی وجہ سے ہے، لہذا آبادی کا صرف ایک چھوٹا حصہ پول سروے میں حصہ لے کر سوشل میڈیا پر انتخابات کے بارے میں اپنے خیالات کا اظہار کرتا

ہے۔ سروے کے اخراجات ، بار بار ہونے والے ایگزٹ پول اور غلط پیش گوئوں نے خاص طور پر ٹویٹر اور فیس بک کے ڈیٹا سے ہونے والی انتخابی پیش گوئوں کو دنیا بھر میں ایک اضطراب میں بدل دیا ہے۔ لہذا، محقق خاص طور پر جموں و کشمیر کے لئے ایک مناسب ماڈل تیار کرنے کی کوشش کرتے ہیں جو حلقہ وار انتخابی نتائج کی جلد پیش گوئی کو آسان بنا دے۔

اس مقالے میں ، محقق نے پیرامیٹرک نقطہ نظر پر مبنی انتخابی پیش گوئی کے ماڈل تیار کیے ، خاص طور پر جموں و کشمیر کے لئے ،جو انتخابات کے نتائج کی پیش گوئی کرنے کے لئے صرف سروے کی پیش گوئی کرنے میں مدد نہیں کرے گا، لیکن سیاسی جماعتوں یا آزاد امیدواروں کو الیکشن لڑنے سے پہلے مخصوص پیرامیٹرز (اس ماڈل عمارت کے لئے استعمال ہونے والے) پر سخت محنت کرنے کی وارننگ بھی دیتا ہے۔ یہ ماڈل پیشہ ورانہ اور عام عوام کے لئے انتخابی نتائج کی ابتدائی اور درست پیش گوئوں کے لئے حقیقی نتائج کے اعلان سے قبل مستعمل ہے۔

اس تحقیقی کام میں ، محقق جپٹر نوٹ بک ویب ایپلی کیشنز کا استعمال کرتے ہوئے پیرامیٹرک نقطہ نظر کی بنیاد پر انتخابی پیش گوئی کا ماڈل بناتا ہے۔ محقق نے (data mining) کی تکنیکوں کو بنیادی طور پر درجہ

بندی کی تکنیک جیسے(Decision Tree)، رینڈم فاریسٹ، (K-NN)، سپورٹ ویکٹر مشین کو جمع کردہ انتخابی اعداد و شمار کے سیٹ پر لاگو کیا اور آخر میں (Soft voting) کا استعمال کرتے ہوئے انہیں ایک ماڈل میں جوڑ دیا۔ تاکہ سیاسی جماعتیں یا آزاد امیدوار انتخابات میں کامیابی حاصل کریں یا نہ ہوں۔

ترقی یافتہ انتخاب کی پیش گوئی ماڈل کی درستگی اور وشوسنیتا کی جانچ کرنے کے لئے ہر فرد کی تکنیک پر مبنی ماڈل اور نمونہ ٹیسٹنگ کے لئے AUROC recognition rate, precision, accuracy, sensitivity, misclassification اسکور اور اسکور کا استعمال کرتا ہے۔ تجرباتی نتائج سے پتہ چلتا ہے کہ ضمنی انتخابی پیش گوئی ماڈل 86 فیصد کی sensitivity کے ساتھ ہائپر پرمیٹر کی ترتیبات پر درجہ بندی پر مبنی انتخاب کی پیشن گوئی کے ماڈل کو بہتر بنا دیتا ہے، 92 کی specificity، of89 کی accuracy، of91 کی precision، miss-classification rate 10.1، AUROC score 90، جیسا کہ باب 6 جدول نمبر 6.6 میں دکھایا گیا ہے۔ آخر میں، محقق نے ماڈل کی اہمیت کی سطح حاصل کرنے کے لئے شماریاتی (t-paired test) چلایا اور تکنیک پر مبنی تمام ماڈلز (Decision Tree، K-NN، رینڈم

فاریسٹ ، سپورٹ ویکٹر مشین) کے لئے (0.05) p ویلیو سے کم value حاصل کی۔ اس statistical t-paired ٹیسٹ کو ہمارے ماڈل پر لاگو کرنے کے نتیجے میں، null hypothesis کو مسترد کر دیا جاتا ہے۔ کیونکہ (ensemble) ماڈل اور دیگر مجوزہ انتخابی پیش گوئی ماڈل کے مابین شماریاتی فرق ہے۔

7.2 تحقیقی حدود

اس تحقیق کی کچھ حدود ہیں:

i. صرف بنیادی ڈیٹا مائننگ تکنیک جیسے ڈیسیزن ٹری ، کے نیئرسٹ نائیبر، رینڈم فارسٹ ، سپورٹ ویکٹر مشین کا استعمال۔

ii. ڈویلپمنٹ فیکٹر، ووٹ سوئنگ ، امیدوار کی اہلیت ، اور جی ڈی پی جیسے دیگر اوصاف کا استعمال نہ کرتے ہوئے صرف کچھ اوصاف کا استعمال۔

iii. جموں و کشمیر کے انتخابی ڈیٹا سیٹ پر ڈیٹا مائننگ تکنیک کی جانچ کرنا جس میں بہت کم ریکارڈ موجود ہیں۔

iv. انتخابی پیشن گوئی کی مختلف اقام یا سوشل میڈیا ڈیٹا بطور خاص ٹویٹر یا فیس بک ڈیٹا جیسے طریقوں پر توجہ نہ دینا۔

7.3 مستقبل کا کام

مستقبل میں اضافے کے لئے، ہم مندرجہ ذیل سمتوں میں مجوزہ کام کو بہتر بنانے کو ترجیح دیں گے۔ موافق کارکردگی کے نتائج تیار کرنے کے لئے دیگر مضبوط ڈیٹا مائننگ تکنیک جیسے ڈیپ لرننگ وغیرہ کی کارکردگی کی جانچ پڑتال۔

i. انتخابی نتائج کی پیشن گوئی حلقہ وار سطح کے بجائے امیدوار کی سطح پر کی جائے۔

ii. انتخابی نتائج کی پیشن گوئی میں ڈیٹا مائننگ کی مختلف تکنیکوں کی کارکردگی (جیسے دیگر ترقیاتی عنصر، ووٹ سوئنگ اور امیدوار کی اہلیت) کو شامل کرنے کی اہمیت کا مطالعہ کرنا۔

iii. دیگر بڑے اصلی ڈیٹا سیٹس کا استعمال کرنا جس میں مختلف خصوصیات، مختلف ریاستوں کے انتخابی ڈیٹا سیٹ اور ریکارڈوں کی ایک بڑی تعداد شامل ہیں۔

iv. پول کے دیگر پیشن گوئی کرنے والوں اور ایگزٹ پول فراہم کنندگان کے مابین JKEPM کی حقیقت پسندی کے استعمال اور قابل قبولیت کی جانچ کرنا۔