Prediction of Consanguinity Marriages in various regions of Pakistan using Machine Learning Algorithms

By

Muhammad Akhlaq

Department of Statistics

Faculty of Natural Sciences

Quaid-i-Azam University, Islamabad

2023

بسم الله الرحمن الرحيم

In the Name of Allah The Most Merciful and The Most Beneficent

Prediction of Consanguinity Marriages in various regions of Pakistan using Machine Learning Algorithms

By

Muhammad Akhlaq

A THESIS SUBMITTED IN THE PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF PHILOSOPHY IN STATISTICS

Supervised By

Prof. Dr. Ijaz Hussain

Department of Statistics

Faculty of Natural Sciences

Quaid-i-Azam University, Islamabad

2023

CERTIFICATE

Prediction of consanguinity marriages in various regions of Pakistan using machine learning algorithms

By

Muhammad Akhlaq

(Reg. No. 02222113015)

A THESIS SUBMITTED IN THE PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF M.PHIL. IN

STATISTICS

We accept this thesis as conforming to the required standards

1. ______________ Prof. Dr. Ijaz Hussain (Supervisor)	2. ______________ Dr. Zahid Iqbal (External Examiner)

3 ______________
Prof. Dr. Ijaz Hussain
(Chairman)

DEPARTMENT OF STATISTICS
QUAID-I-AZAM UNIVERSITY
ISLAMABAD, PAKISTAN
2023

Declaration

I “Muhammad Akhlaq” hereby solemnly declare that this thesis titled, “Prediction of Consanguinity Marriages in various regions of Pakistan using Machine Learning Algorithms”.

- I declare that this MPhil thesis has not previously been submitted, in part or in full, for any other academic degree at any other institution. I affirm the accuracy of the data and information given in this MPhil thesis, conforming to ethiI declare that this MPhil thesis has not been published previously.
- Under the supervision of Prof. Dr. Ijaz Hussain, I worked on my MPhil thesis, which I am proud of. All sources utilised are properly acknowledged.
- In my MPhil thesis, I guarantee for the accuracy of all data and material presented, following to ethical norms and academic requirements throughout the data gathering and analysis procedures.
- I followed all grammatical rules and academic standards during the data collecting and analysis processes, properly citing and referencing any published works or other peoples’ intellectual property used in this MPhil thesis.
- Any published works or other third-party intellectual property that I used in this MPhil thesis has been properly assigned and referenced.

Dated:________________________ Signature:________________________

I am feeling Wonderful and pleasure to dedicate this research work to

My Believed Parents and Friends

Whose endless affection, prayers and wishes have been a great source of comfort for me during my whole education period and my life

Acknowledgment

First and foremost, I thank to **Allah Almighty**, the Most Compassionate and Merciful, for bestowing upon me the spirit and enthusiasm essential for this undertaking, and then I want to express my heartfelt gratitude to my supervisor, **Professor Dr. Ijaz Hussain**. His energy, support, and guidance contributed to my search for knowledge. Professor Dr. Ijaz Hussain's knowledge and direct views improved the quality of this thesis, and as a result, I was inspired to pursue academic interests in new ways in my research area. Professor Dr. Ijaz Hussain has been an advisor, friend, and inspiration to me. I also want to a spatial thank and express my gratitude to faculty member staff, especially Dr. Ismail Shah, Dr. Manzor Khan, Dr. Abdul Haq, Dr. Sajid Ali, and Dr. Maryam Asim. Their dedication to encouraging academic excellence had a significant influence on my development as a scholar. A special thanks goes out to **Asad Ellahi**, whose friendliness, friendly nature, and desire to help have been an ongoing source of support. His insightful advice and help were really helpful in resolving a number of obstacles that came up during the study process. I would want to thank every member of my family for their support throughout my life, especially my parents. It would not have been possible without their prayers, love, encouragement, and support. I would also like to express my heartfelt gratitude to **Sir Farukh Kamran** for his financial support at this time. His kind assistance was crucial in making it possible to finish this thesis. Last but not least, I would want to thanks my classmates, particularly **Hamza Amin**, who has been really helpful in this work and whose friendship and shared experiences have made this academic journey more enjoyable and meaningful. This thesis completion has been a difficult but rewarding experience, and I acknowledge that it would not have been possible without the combined efforts of these people. They have been the pillars that have kept me standing during this academic endeavour because of their encouragement and support. I would like to conclude by expressing my sincere gratitude to Professor Dr. Ijaz Hussain, Asad Ellahi, the teachers, staff, my family, Sir Farokh Kamran, and my classmates for their essential contributions. My academic experience has been significantly impacted by your advice and encouragement, and I am very appreciative of your presence in my life. I appreciate each and every one of you for contributing to this journey and making a difference in my life by being kind and helpful.

Abstract

Abstract Consanguinity is the genetic connection between individuals through shared ancestry from a common parent. Consanguineous unions can lead to an increased risk of inherited genetic disorders due to the sharing of a higher proportion of common ancestors. This study aims to identify factors influencing consanguineous marriages, employing machine learning algorithms to develop and validate predictive models. For this purpose, the data sourced from a prior study by (Jabeen and Malik, 2014a) focuses on consanguinity in marriages, with 16 variables including age, language, caste, and education for both women and husbands in the Bhimber district. External validation is conducted using a separate dataset from Mardan, Pakistan Tufail et al. (2017), with 13 variables, including age, caste, income, and education, to assess the generalizability of findings on consanguinity. Initially, the risk factors are investigated using association analysis and odds ratios for each category of the covariates. The Boruta algorithm is also employed to capture the relative importance of each factor concerning consanguineous marriage. The prediction of consanguineous marriages is carried out using various machine learning algorithms (i.e., logistic regression, decision trees, random forests, ensemble stacked meta-models, and support vector machines) based on the selected most important factors. Furthermore, the generalizability, reliability, and robustness of the models were assessed by external validation in the Mardan district of Pakistan. The study identified several statistically significant factors associated with consanguineous marriages, including the age, caste, language, and marriage year of wives, as well as the age, caste, and language of husbands. Additionally, variables such as marriage type, family type, and specific tehsils within the Bhimber district were found to significantly impact consanguineous unions. The bagging and ensemble method of the stacked meta-model demonstrated superior performance with AUC values of 0.59 for the Bhimber district, 0.67 for the Mardan district, and 0.58 when the Bhimber district served as the training dataset and Mardan district as the test dataset, outperforming other models, with logistic regression being the second-best performing. The study concluded that consanguineous marriage is significantly influenced by the factors, i.e., age, caste, language, family type, and marriage types of both parents. The findings highlighted the sensitive interplay of diverse factors in shaping consanguinity trends and emphasized the effectiveness of ensemble methods in predictive modeling for such complex phenomena. The study provides significant insights to improve healthcare policies and may help in executing targeted measures.

Contents

List of Tables

List of Figures

List of Abbreviation

CU	Consanguineous Union
WHO	World Health Organization
PDHS	Pakistan Demographic and Health Survey
PSLM	Pakistan Social and Living Standards Measurement
ES	Exome Sequencing
AR	Autosomal-Recessive
GIS	Geographical Information System
DT	Decision Tree
UNICEF	United Nations Children's Fund
RF	Random Forest
OOB	Out-of-bag error rate
SVM	Support Vector Machine
BORUTA	Bootstrap aggregating with randomized trees
AUC	Area Under the Normal Curve
ROC	Receiver Operating Characteristic curve

Chapter 1

Introduction

Consanguinity is described as a contract between two biologically related people (cousins) that connects to blood. Multiple investigations have shown that congenital abnormalities, autosomal diseases, a high newborn mortality rate, underweight offspring, etc. are all more likely to occur in consanguineous marriages. Parents who are biologically related to one another, such as cousins or siblings, are more likely to have consanguineous offspring. The relationship refers to the fact that the parents share a common ancestor (Bennett et al., 2002). This is known as the consanguineous risk from biological relationships. There is a noticeable difference between consanguineous marriages and non-consanguineous marriages in cousins, according to several studies. There are several reasons to study consanguineous marriages and why it is important (Dronamraju and Khan, 1963) (Modell and Darr, 2002) (Abdulrazzaq et al., 1997). Some studies have revealed that consanguineous marriages can raise the risk of specific genetic disorders and diseases, as well as the strengths and weaknesses of any society's public health system and information about the decision-makers who can address this issue (Little et al., 2009). The consanguineous-related genetic disorder is dangerous for every individual in society; therefore, an investigation is needed to overcome this disease (Al Aqeel, 2007). For basic research purposes, researchers can better understand the underlying mechanisms of certain events and potentially develop new treatments for any specific problem related to such events. It is an invaluable site for researching consanguineous marriages because of its cultural significance and historical background (Zayed, 2016).

1.1 Percentage of consanguineous marriages in World

The prevalence of consanguineous marriages ranges from 51–58% in Jordan, 54% in Kuwait, 50% in the United Arab Emirates, and 52% in Qatar, but Pakistan has been demonstrating the highest prevalence rate due to social, religious, and economic benefits (Bener and Mohammad, 2017). Approximately 1% of consanguinity marriages occur in the United States, Russia, Australia, and some parts of Latin America and Europe. In South Asia, approximately 1 to 10% of consanguinity marriages are in China, North India, South Europe, and Canada. 10 to 50% are present consanguinity marriages. In Muslim countries like Pakistan, Arab countries, Turkey, Iran, Afghanistan, and South India, a very small percentage of the world Amish population is present in India (Hamamy, 2012). There are social differences in stillbirth risk all over the world, and it has been proposed that the shown ethnic discrepancy is a result of the socioeconomic disadvantage that most migrants face. Consanguinity has been proposed as another reason for the increased incidence of stillbirth and congenital abnormalities reported in several migratory groups, and based on Pakistani and German medical authorities, consanguineous marriages cause half of all child deaths in Pakistan (Nybo Andersen et al., 2016). According to the 2020 year, roughly 6700 newborn infants passed away every day, 2.4 million children worldwide died in their first month of life, and 47% of children under five died in a given year (De Andrade et al., 2022). Every year, 30 million newborns are born, yet four million of them die during the first 28 days, with some dying within the first 24 hours. The United Nations International Children's Emergency Fund (UNICEF) seeks to protect children's rights while also attempting to provide for their fundamental needs. (Oestergaard et al., 2011).

1.2 Prohibited of marriages in Quran

The Qur'an mentions this in verses 4:22–24. You are not allowed to marry your mom, your daughters, your sisters, your father's sisters, your mother's sisters, your brother's daughters, or your sister's daughters. According to Surah 4:23 of the Qur'an, first cousins are not on the list of ineligible spouses. Zaynab bint Jahsh, Muhammad's (PBUH) first cousin, was married to him. Relatives through Marriage: In the case of a second, third, or fourth marriage, you aren't allowed to marry your wife's mother or grandmother, your wife's daughter or grandchild, or your son's wife. Relatives through Fosterage: Foster relationships are also forbidden, as are relationships that are based on blood ties or marriage. For instance, a man cannot marry the daughter of his foster mother.

1.3 Consanguineous marriages in China

The effectiveness of clinical exome sequencing (ES)-based carrier screening in identifying potential genetic risks in Chinese consanguineous couples was the goal of the study conducted in China (He et al., 2021). To screen consanguineous couples for autosomal-recessive (AR) diseases, based on the clinical ES of 5000 genes associated with human diseases. Consanguineous marriage was not prohibited in earlier times and was a common practice in every society. More than 10% of respondents prefer first- and second-cousin marriages inside their own families (Warsy et al., 2014). Over time, such marriages were deemed illegal in various jurisdictions. Previously, Australia, North America, and South America all practiced cosine marriages. It was prohibited in certain countries, including Taiwan, South and North Korea, China's mainland, and others. It is critical for working women to take a 56-day break before and after giving birth. During this time, female employees should continue to receive their full pay from their employers. This is because pregnant women need this time to rest and recover, and if there were twin-child issues or other issues, they would be given an extra 14 days to leave work with full pay (Robinson, 1985). In China According to the 1980 Law, there were completely banned marriages between cousins. The law of 1980 replaced the law of 1950. In the 1950 law, more necessities could be provided to women during and after pregnancy (Ocko, 1991). There were fewer than 1% consanguineous marriages in the United States, Russia, Australia, and several parts of Latin America and Europe, 1 to 10% in China, Latin America, North India, Japan, South Europe, and Canada, and 10 to 50% in the Arab world, Turkey, Iran, Pakistan, Afghanistan, and South India (Hamamy, 2012).

1.4 Trend of consanguineous marriages in Pakistan.

According to a World Health Organization (WHO) estimate from 2008, 99% of maternal deaths occur in underdeveloped countries. Between 1990 and 2018, studies revealed that nearly two-thirds of women of reproductive age were married to relatives, more frequently paternal than maternal. The findings revealed that consanguineous marriage has typically remained stable in occurrence over the last thirty years (Iqbal et al., 2022). According to the Pakistan Demographic and Health Survey (PDHS) (2020) survey, the PDHS 2017-18 survey, and the Pakistan Social and Living Standards Measurement (PSLM) 2018-19 survey, Pakistan's infant mortality rate is dropping; however, it is significantly higher in rural areas than in urban areas (Haque et al., 2016). In urban areas, it was 59 per thousand, and in rural areas, it was 50 per thousand. According to gender, male infant mortality is 58, which is higher than the female mortality rate of 50 per thousand in the same year in Pakistan (Ayaz and Saleem, 2010). In the UK, some immigrant Asian people show highly inherited disorders, and some are rare cases (Anwar et al., 2014). From the different experiments we have studied, we have found that there is a strong relationship between consanguineous disorders and some genetic and health problems such as immunodeficiency disorders, low birth weight, down syndrome, protein-S deficiency, etc (Shawky et al., 2013). Consanguineous marriage is useful in determining the image of social, biological, and health systems and disease patterns in the structure of populations (Bittles, 1994). It is associated with higher incidences of different diseases like congenital anomalies, child mortality, and adult morbidity, as well as a decreased fertility rate in the population. Here, we divide into different categories to understand the percentage of consanguineous marriages present in Pakistan. If we divide it into different categories like 20%, 21:30%, 41:60%, and unknown, then Pakistan falls into the third category because Pakistan has more consanguineous diseases as compared to other countries. Here we were more studying according to different interesting places in Pakistan, like Punjab, Khyber Pakhtunkhwa, Baluchistan, and Azad Jammu Kashmir, and district-wise also in literature. (Hamamy, 2012) Consanguineous marriage related to genetic disorders is very dangerous for every individual in society, particularly in the genetic field, so therefore, an investigation is needed to overcome this disease. The Bhimber district of Kashmir, Pakistan, is known for having diverse populations and a social structure that is distinct from Pakistan's other social cultures.

1.5 Modeling of consanguineous marriages

Investigating the occurrence of consanguineous marriages and creating a reliable prediction model unique to the Bhimber district are the main goals of this research work. Furthermore, by comparing the model's predictions with actual data on consanguineous marriages in this area, this work aims to externally validate machine-learning approaches used based on independent data (Li et al., 2021). By doing this study, researchers will be able to ensure the prediction models' application in real-world circumstances and enhance the validity and accuracy of the models. The techniques of machine-learning used here like logistic regression, decision tree, random forest, support vector machine, bagging, and ensemble method stacked model were applied to develop the prediction of the model using R-programming for this investigation. Furthermore, the model will be subjected to rigorous validation using independent datasets of different places to assess its predictive power and generalizability. Here used Classification technique of machine-learning predicts the new events because it identifies the insight of the risk factors that may be dangerous for every individual or family regarding consanguineous marriage. Therefore, we are going to deal with this problem by using existing models to predict new events. Because it is important to validate the model for new events to evaluate its generalizability and ensure its reliability in the real-world population. According to the Ramspek et al. (2021) External validation prognostic prediction models when it is appreciated to validate the existing models for our study, and why is it important for such type of study? External validation is very important because sometimes the existing model is better performed on the given testing and training data but not better performed on an independent set, therefore external validation is very important, especially in such types of study as genetic study, medical study, etc.

1.6 Factors that affect the consanguineous rate and what factors may decrease the consanguineous rate?

Several important factors have an impact on the consanguineous union rate or the practice of marrying close relatives. First, both urban and rural areas are important. Respondents from rural areas were more likely than those from urban areas to be positive about consanguinity (62% vs 56%). Respondents' knowledge of congenital malformations, as well as their understanding of the association between consanguineous marriage and congenital anomalies, have substantial differences in their attitudes toward consanguineous marriages. (Mazharul Islam, 2017) Those who are unaware of congenital malformations are more likely to agree that consanguineous marriage is beneficial than those who are aware of congenital anomalies (70% vs 51%). Additionally, the prevalence of consanguineous marriages is significantly influenced by socioeconomic conditions. A person's decision to marry a close cousin might be influenced by the economic and social circumstances of their family and society. Consanguineous marriage is considered in some communities to maintain social standing and family riches while also keeping a stable income (Charsley, 2005).

According to the different variables like age, culture, and religion Joseph et al. (2015). Higher levels of female education are important because educated women and girls are more willing to look outside their families for partners and are more aware of the potential risks involved with such unions. Consanguinity is less likely when people are more migratory between rural and urban areas and have access to a wider range of possible partners. Better economic circumstances also allow families to base decisions on criteria other than money, which lessens the necessity for in-law unions to maintain social standing or financial resources. Nouri et al. (2017) As people become more aware of the genetic risks connected with consanguinity, they have a greater capacity to make decisions when they have access to appropriate medical facilities and treatment in every area. Improved maternal and infant health results from caring for mothers before, during, and after childbirth, which reduces the idea that consanguineous unions are necessary to obtain family support in the event of a medical emergency. By supporting these elements, we can create a culture that values studying, variety, and the health of the next generation, which will improve marriage customs and general well-being.

1.7 Objectives of the study

The main objectives of the studies are:

- To predict consanguineous marriages in two regions of Pakistan using machine learning algorithms.
- To obtain generalization through external validation utilizing the Mardan district.
- To examine the effects of influential factors on consanguinity marriages.

1.8 Structure of thesis

This thesis covers seven chapters of the study, each of which serves a specific function to advance the overall research investigation. Chapter 1 outlines the issue being studied, gives a straightforward and convincing explanation of its importance, and highlights the insight into this issue. A thorough evaluation and summary of the relevant research that has already been conducted on the subject is provided in Chapter 2's substantial literature survey. It gives the reader a background understanding of the subject under study and points out any differences between the current work and earlier work. In Chapter 3, we prepared the data for our analysis and defined the structure of the data since, in the actual world, the raw data is incomplete and noisy, and we defined the variables for our study, which is extremely important before we deal with it. In Chapter 4, we used advanced machine-learning algorithms such as decision trees, support vector machines, and bagging to understand the overall picture of consanguineous marriages in the Bhimber district of Kashmir, Pakistan. In Chapter 5, we conducted an external validation comparative analysis for Pakistan's Mardan district, using a variety of machine-learning algorithms such as logistic regression, random forest, and ensemble method staked meta-models for external validation from the Mardan district. Furthermore, in Chapter 6, we did comparative research for external validation based on the training set considering the Bhimber district and the testing set considering the Mardan district, applying several machine-learning algorithms as presented in Chapters 4 and 5. The researcher provides an overview of the findings and explains how these techniques were useful for data analysis. To determine which approach performed better than the other. Through the representation of the performance metric, they also investigate the models in greater detail to comprehend how they function. These chapters are crucial since they clarify for readers what the researcher found and their methodology. The discussion and conclusion of the research problem are provided in Chapter 7, which also contains a full evaluation of the most important conclusions and insights from the entire study.

Chapter 2

Literature Review

In Pakistan's Jammu and Kashmir Bhimber district, participant studies on a total of 1731 girls, aged 12-75, were conducted to determine the cause of illness and death caused by congenital and non-congenital defects. Because of this, it was impossible to control the detrimental consequences of consanguinity and abnormalities, especially in rural areas. They used various demographic independent variables to assess the variability in congenital anomalies and noncommunicable diseases using statistical approaches such as odds ratios and confidence intervals. The most frequent causes of mortality and morbidity were determined to be hereditary, communicable, and non-communicable disorders (Jabeen and Malik, 2014b). Furthermore, because the majority of individuals lived in rural areas, research was undertaken on 1800 women in the Sargodha district population in Punjab, Pakistan, to estimate consanguinity rates and inbreeding coefficients. The study discovered that at the time, the prevalent rate in the district of Sargodha was high, with consanguinity marriages, and that there were also high inbreeding coefficients. According to the participant, cultural and societal factors played a larger role in this condition, and the unfavorable health outcomes were connected with consanguinity. It is also stated that there is a significant difference between Sargodha and other districts in Punjab in terms of consanguinity marriages and reducing consanguinity frequency through education, among other things. Further research in that area focuses on the probable need to eliminate the risks associated with consanguinity unions and improve the reproductive health care system (Hina and Malik, 2015). In Okara, Pakistan, 1521 married women were studied using a cross-sectional sampling design between 2016 and 2017. According to the sample (n = 763), the first marriage had the highest consanguineous marriage. They also observed that consanguinity unions were statistically significant, particularly in rural areas. Descriptive statistics and multivariate logistic regression were used in the Okara population. The participant discovered that the percentage of consanguinity was higher in the Okara population (Nawaz et al., 2021). Consanguineous marriages in India and the connection between marriage among blood relations and having an unfavorable pregnancy outcome This problem is being investigated using bivariate and trivariate estimations, as well as the Cox proportional hazard regression model. The variance of consanguinity marriage was studied in southern India using different demography variables, and the bad pregnancy effect is highly dependent on consanguineous marriage in India (Kuntla et al., 2013). Researchers continued to investigate the study's findings by examining additional conditions like hypertension, dilated cardiomyopathy (DCM), and cardiovascular disease that are linked to consanguinity marriage. They asserted that consanguinity marriages, cardiomyopathy, and hypertension were related. The family was divided into two groups, maternal members and non-maternal members, for comparison using the statistical method. Clinical investigation reveals that the majority of people have hereditary cardiomyopathy and hypertension (Zhu et al., 2016). According to a Beirut population study, the likelihood of consanguineous marriage increased as the husband's education and occupational rank decreased. The researcher discovered, using a multivariate analysis technique, that two key variables were highly connected with consanguinity marriage sickness when they belonged to a Muslim faith country and had a low occupational class (Khlat, 1988). The researcher was looking into why there were so many marriages in Muslim countries, and he discovered that consanguineous marriage was negatively impacting people's health, so he declared it a major public health issue. As a result, the primary and crucial determinants of consanguineous marriage disease in Arab countries were economic, cultural, and rural women's low education (El Goundali et al., 2022). Using local and national data, researchers examined the trend incidence and medical implications of consanguinity in the Turkish south coast population over the preceding 11 years. The cross-sectional study was carried out in the Manavgat region, which is a major tourism center on Turkey's Mediterranean coast. They conducted a 1500-person random survey of married couples on the Mediterranean coast of Manavgat province, Turkey, and observed that first-cousin weddings were the most common type of marriage, with consanguineous marriages remaining common among Turkey's Mediterranean population (Alper et al., 2004). Consanguinity had long been acknowledged as a societal norm by certain Iranians. The primary goal was to determine the function of consanguineous in congenital malformations and the relationship between the inbreeding coefficient and defects in Iranians. The cross-sectional study included all neonates delivered between April and December 2008 at the Shahid Sadoughi hospital in Yazd, Iran ($n = 1195$). As a result, the researchers established a relationship between parental marriages and the occurrence of abnormalities (Alper et al., 2004).

From January to June 2014, the data was collected from the patients at the start of the trial at Khyber Teaching Hospital in Peshawar. Women with at least one birth problem were questioned, and 1062 deliveries were documented; around 2.9 percent of newborns had various congenital defects. The most prevalent contributing defects observed in the investigation were anencephaly and hydrocephalus. Furthermore, inadequate intake and a high consanguineous rate were the most linked risk factors for congenital malformations. It can be decreased by increasing knowledge and avoiding consanguineous marriage. Similarly, congenital and inherited abnormalities (CA), a primary cause of child death and morbidity, are common in Pakistan (Khan et al., 2015).

To understand the burden and biodemographic correlates of CA, the study was designed to describe the prevalence-pattern and phenotypic characteristics of CA in the Hazara population of Khyber Pakhtunkhwa, Pakistan. For a retrospective cross-sectional study, CA patients and families were enlisted from district hospitals and community centers, and descriptive data were gathered. The CA trend and high frequency of sporadic cases observed in this cohort imply that nongenetic factors may contribute significantly to the origin of these conditions, which may be diminished by enhancing the healthcare system (Bibi et al., 2022). Utilizing a cross-sectional sample strategy in Pakistan's Okara society. Based on a sample size of $n = 763$, the first union had the highest percentage of consanguineous marriages. After discovering that this link was statistically significant, it was discovered that it was most prevalent in Pakistan's rural Okara district. It involved determining the social, cultural, and demographic elements that contribute to consanguineous marriage among the inhabitants of Pakistan's Okara area (Nawaz et al., 2021).

In 2004, the results from the study were examined by the Indian Journal of Community Medicine. They estimated the prevalence of consanguineous marriages and carried out chi-square tests to compare the frequency of unfavorable pregnancy outcomes between consanguineous and non-consanguineous marriages. They used logistic regression analysis to determine the association between consanguinity and unfavorable pregnancy outcomes while adjusting for potential confounding factors such as maternal age, education, and parity. The study aimed to investigate the prevalence of consanguineous unions in rural areas and the effects of consanguinity on the outcome of pregnancies (Nath et al., 2004). The mother's sociodemographics, consanguinity, previous pregnancies' outcomes, and the pregnancy's final results were all obtained. The outcomes were then examined to see if consanguinity was linked to unsuccessful pregnancies. To summarize the information gathered from the pregnant women, the investigators of the study used the descriptive statistics mean and standard deviation for numerical results and the chi-square test for categorical data. Consanguinity was shown to be highly prevalent in the research population and to be directly related to low birth weight, stillbirth, and neonatal death (Metgud et al., 2012).

In the north of Jordan, the major objective was to assess the connection between consanguineous marriages and unfavorable pregnancy outcomes, including stillbirths, low birth weight, preterm births, and congenital abnormalities. Thus, the relationship between consanguineous marriages and unfavorable pregnancy outcomes was investigated. The procedures used in this study included a cross-sectional design, a structured questionnaire administered to women within 24 hours of delivery, data collection, and statistical analysis using the Statistical Package for Social Sciences (SPSS, version 15) software. The sociodemographic and obstetric characteristics of the participants were described using frequency distribution, and the proportions of women were compared using the chi-square test (Obeidat et al., 2010).

The description of new autosomal recessive primary immunodeficiencies (PIDs) has greatly benefited from research on inbred populations. Here, we look into the PID pattern in North African populations and evaluate the effects of the extremely common consanguinity. In Egypt, Morocco, and Tunisia, countries where PID knowledge is still relatively new, this review details the current situation regarding pediatricians' awareness of PIDs. The phenotypic distribution of PIDs is described, and comparisons with other populations and the three nations are made. According to data analysis, there are more mixed immunodeficiencies than antibody diseases, and autosomal recessive types are distributed oddly. Molecular diagnosis is essential in these endogamous populations for creating a genetically based preventive strategy. In these environments with few resources, there are several challenges to organizing diagnosis and care services (Barbouche et al., 2011).

Ramspek et al. (2021) state that internal and external validation is crucial for the sake of generalization, especially when we are dealing with any hereditary condition. The importance of external validation was studied by the researcher, who used various prognostic models in a statistically sound and thoughtful approach. For the generalization purpose, simulation data was used for external validation, and machine-learning methods including cross-validation, bootstrapping, holdout, and others were applied to compare the results of internal and external validation to verify the performance of the model. The researcher discovered that these methods perform poorly if the test set is smaller. When the test set size was increased, the results were more exact and had a lower standard deviation. These data were used for prediction purposes based on logistic regression (Eertink et al., 2022).

Based on the previous literature analysis, a variety of statistical methods were used to investigate the relationship between consanguineous marriage and various demographical, cultural, and socioeconomic characteristics, as well as congenital malformations and adverse pregnancy outcomes. Most statistical tools, such as odds ratio, chi-square, confidence interval, cross-sectional designs, structured questionnaires, and logistic regression, were utilized to examine the significance of associations in the previous literature, particularly those between consanguinity and adverse pregnancy results. Based on the previous literature review, we proceed to the advanced statistical analysis to gain an understanding of the current study's insight picture, which may be a risk factor in the future in various districts of Pakistan. The major purpose of this investigation is to evaluate external validation studies and generalize consanguineous marriages in various areas of Pakistan using machine-learning algorithms. We then move on to more advanced machine-learning algorithms such as decision trees, random forests, logistic regression, SVM, bagging, and ensemble methods for predicting consanguineous marriages, as well as external validation of our work in various real-world scenarios.

Chapter 3

Data description and measurements

3.1 Preprocessing of the data

During the data mining process, we can see that the raw data is likely to be incomplete, contain useless information, and include some noisy and missing observations. As a result, preprocessing is an essential strategy for comprehending raw data. Many machine-learning algorithms only function with numerical data and do not work with such data. As a result, categorical data preprocessing is required in this investigation. Algorithms can make use of categorical variables' information by converting it to a numerical representation. Furthermore, pretreatment strategies help avoid the introduction of unexpected biases or ordinality into the data, ensuring algorithm compatibility, and improving overall model performance and interpretability.

3.2 Data inspection

Examining and comprehending the features of the categorical variables in a dataset is necessary for data inspection, particularly for categorical analysis in machine-learning. Here's why it's crucial. To understand the variable types, you can separate categorical variables in your dataset from numerical variables by determining which variables are categorical. Your ability to apply relevant statistical methods and algorithms created especially for categorical data is enhanced by this understanding, and you can also check to see whether there is a major imbalance in the distribution of the categories, as this can affect the performance of the model and require specific management, such as oversampling or undersampling methods. Visualization is also important for data inspection. For visualization of the categorical variables using plots such as bar charts, pie charts, or stacked bar charts, visualizations provide a visual representation of the distribution, particularly in research work.

3.3 Missing values

Missing values could be a problem since they might affect the accuracy of analyses. During this phase, you identify any missing data and decide how to handle it, such as by imputing the missing values or by removing the corresponding rows or columns. Categorical data with missing observations provides several options for substituting the values. Here are a few common techniques. Mode imputation: With mode imputation, missing categorical values are replaced with the mode, or the category that occurs the most frequently in the variable. The mode is considered the most likely category in this method for the missing values. Using machine-learning techniques, the missing categorical values can be predicted using machine-learning models, such as decision trees or random forests, depending on the other variables. Utilizing the observed data, create a model that will be utilized for predicting the missing values.

3.4 Outlier detection

Outliers are frequently related to numerical data rather than category data. Since categorical variables have distinct labels or categories, they are less likely to be outliers in the usual way. Rare or unusual categories, also referred to as ”outliers” in the context of categorical data, are still possible. The following should be considered when handling such circumstances: Rare categories: In categorical data, rare categories are those that happen very infrequently. Even if uncommon categories aren’t typically outliers, they might always present challenges for analysts. Rare categories must be recognized and treated carefully since they may generate absence difficulties or diminish the performance of machine-learning models. Consider grouping all the distinct, unique categories into an ”other” or ”rare” category if there are many of them.

3.5 Encoding the data

In dealing with categorical data in statistics, we often need to convert the data into coding using any numerical value. The coding process allows us to transform the raw data into a format that can easily be analyzed; therefore, the coding process is crucial for categorical data. With label encoding, each category of a categorical variable is given a distinct numerical label. It works well with ordinal data where the categories are arranged naturally. Algorithms can use categorical data as numeric values due to label encoding. If it is binary coding data, then each category is represented as a binary code through binary encoding like 0 or 1, as shown in Table 3.1 and in Table 3.2. Each distinct category is given a binary code, and these binary codes are used as features. While maintaining categorical data, binary encoding reduces the dimensionality in comparison to one-hot encoding. It achieves equilibrium by achieving a balance between dimensionality reduction and false ordinal avoidance.

3.6 Overview of the research work

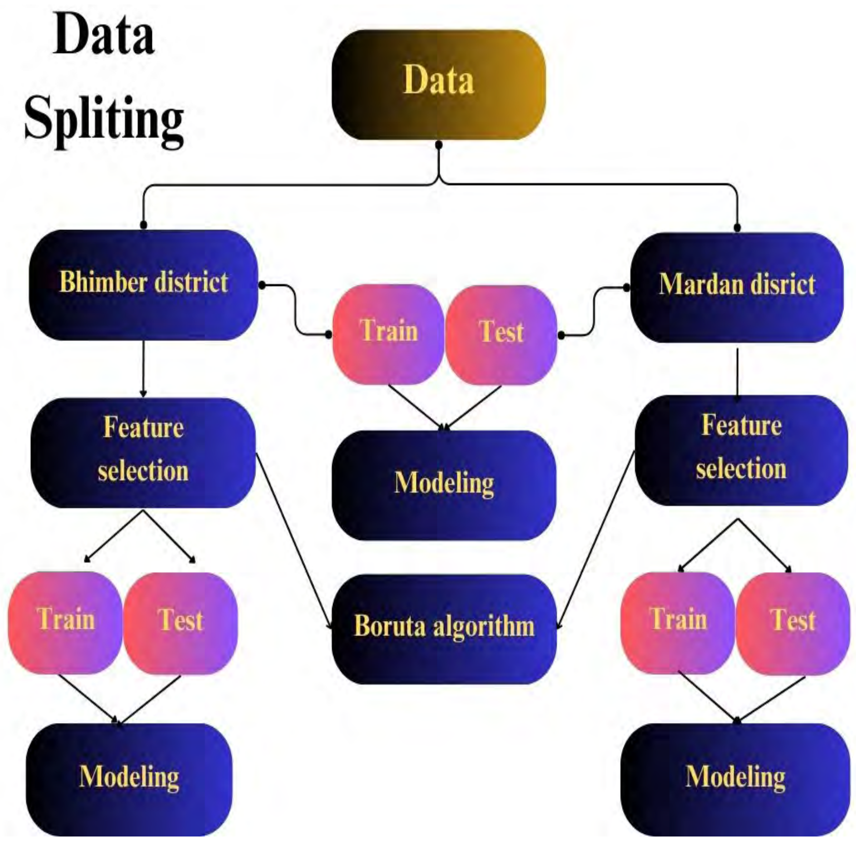


Figure 3.1: Data segmentation for model building into training and testing sets for effective machine-learning models.

3.6.1 Dependent variable

Consanguineous marriage is the dependent variable that is categorized as ”Yes” or ”No” to reflect the presence or absence of consanguineous relationships. There are both categorical and continuous independent variables in the dataset. All parameters, however, have been turned into categorical variables for predictive analysis to maintain uniformity and comparability among districts. This transformation ensures that variables are handled consistently throughout both districts.

3.6.2 Independent variables

The independent variables, or covariates, may consist of socioeconomic statistics, cultural or religious characteristics, or any other pertinent variables that are known to affect consanguineous marriage. The machine-learning models applied in the prediction process depend heavily on these factors as inputs. The dataset is divided into two sections for the prediction analysis of consanguineous marriage: one for predicting and modeling consanguineous in the Bhimber district, and another for external validation in the Mardan district. By using the Bhimber district dataset as training data, machine-learning models can discover patterns and connections between variables and consanguineous marriage. On the other hand, the performance and generalizability of the trained models are assessed using the Mardan district dataset. The dataset is separated into many sets before the prediction analysis is carried out, as shown in Fig. 3.1. We organized the data into three tables, Table 3.1 and Table 3.2, which contain information about wives and husbands of wives, respectively. These tables contain descriptions and measurements of the variables utilized in the study. We aim to get insights into the factors that may influence consanguineous marriages in the Bhimber district by using encoded data and demographic covariates. In a similar way to the previous study in the Bhimber district, we are now focusing on the Mardan district to predict consanguineous marriages. For this study, we are using some demographic factors as predictors, and these factors are listed in Table 3.3. This table provides detailed descriptions and measurements of the variables used in our analysis. By using the data in Table 3.3, we aim to gain more insights and ensure the reliability of our predictions for this specific region.

Table 3.1: Coded the demographic data to analyze the insight discoveries according to wife.

Variable	**Measurement**	**Coded**
Consanguineous	Yes & No	1 & 2
Age	18-30, 31-40, 41-50 & Above(50)	1,2,3 & 4
Education	Metric, Undergraduate & Graduate	1,2 & 3
Occupation	House Wife & Others	1 & 2
Rural-Urban	Rural, Urban & Peri-Urban	1,2 & 3
Caste	Choudhary, Rajput, Jat & Others	1,2,3 & 4
Language	Punjabi & Pahari	1 & 2
Marriage year	Below(1990) & Above(1990)	1 & 2

Table 3.2: Coded the demographic data to analyze the insight discoveries according to the husband.

Variable	**Measurement**	**Coded**
Age	18-30, 31-40, 41-50 & above(50)	1,2,3 & 4
Education	Metric, Undergraduate & Graduate	1,2 & 3
Occupation	Well job & Others	1 & 2
Rural-Urban	Rural, Urban & Peri-Urban	1,2 & 3
Caste	Choudhary, Rajput, Jat & Others	1,2,3 & 4
Language	Punjabi & Pahari	1 & 2
Family type	Nuclear, Grndprnts and one cple, Extnded Fmly & More than one cple	1,2,3 & 4
Marriage type	Arrange marriage & Self arrange	1 & 2
Tehsil	Bernala, Samahni & Bhimber	1,2 & 3

Table 3.3: Coded the demographic data to analyze the insight discoveries in the Mardan district for external validation.

Variable	**Measurements**	**Coded**
Consanguineous	Yes & No	1 & 2
Age	18-30, 31-40, 40-50 & Above(50)	1,2,3 & 4
Income (Thousand)	Nill, 1 to 30, 31 to 60 & 60>	1,2,3 & 4
Education	Metric, Undergraduate & Graduate	1,2 & 3
Origin	Mardan & Others	1 & 2
Occupation	Housewife & Others	1 & 2
Rural-Urban	Rural & Urban	1 & 2
Family type	Nuclear & Extended	1 & 2
Caste	Khatak & Others	1 & 2
Marriage type	Arrange marriage & Self-arrangement	1 & 2
Marriage year	1959 to 1990 & 1991 to 2015	1 & 2
Age at marriage	1 to 15, 16 to 30 & 30>	1 & 2
Tehsil	Mardan, katlang & takht bie	1, 2 & 3

3.7 For imbalance classification task assessing class distribution.

An imbalanced dataset occurs when one class considerably exceeds the others, creating problems for machine-learning models. Addressing class imbalance is critical because it can lead to biased predictions. To balance class distribution, several strategies, such as oversampling and undersampling, are applied. Hybrid approaches combine various methodologies for better results. Ensemble approaches and visual aids like as bar charts can also help alleviate class imbalance and understand underrepresented classes in a study (Ganganwar, 2012; Napierala and Stefanowski, 2016). In the present study, we are assessing Consanguineous class distribution by using bar charts to show the distribution of consanguinity classes, we can see the frequencies or percentage of each class of the consanguineous marriage depended variable, shown in both Fig 4.5 and in Fig 5.4. This visual depiction gives us the ability to identify any differences in how various classes are represented and gives us a thorough overview of their distribution patterns. We can identify potential imbalances and learn more about the inadequate representation of particular consanguinity groups by looking at the bar charts in the result section.

Chapter 4

Application of machinelearning algorithms for Bhimber district of Kashmir, Pakistan.

4.1 Introduction

To investigate consanguineous marriages in the Bhimber district, we used a comprehensive methodology that included association analysis, feature selection using the BORUTA technique, predictive modeling, and robust evaluation metrics. Initially, we identify relevant variables utilizing the BORUTA technique, a technology that assists us in identifying the primary elements driving these marriages. Our study employs three models for forecasting consanguineous marriages based on demographic variables: a decision tree for binary classification, a support vector machine for discriminating between related and unrelated relatives, and the bagging technique. The AUC value for overall discriminative power, sensitivity and specificity for proper classification, and prevalence rate for the proportion of consanguineous marriages in the district are performance evaluation measures. Furthermore, the (ROC) curve visually depicts the trade-offs between true and false positive rates in our models.

4.2 Methodology

4.2.1 Selection of study area and data source

The data for this research was sourced from a previous study conducted by (Jabeen and Malik, 2014a). Azad Jammu and Kashmir, sometimes referred to as Azad Kashmir, is an administratively autonomous region within Pakistan (Shahzad et al., 2016). It is part of the larger Kashmir region, which India and Pakistan have fought for since 1947. Azad Kashmir covers 13,297 square kilometers and is divided into 10 regions and three divisions (Shafique et al., 2023). It is administered by Pakistan. So, Bhimber district (32.58°N, 74.04°E) of Mirpur division is one of eight districts in Pakistan's Azad Jammu and Kashmir (AJK) region. Jabeen and Malik (2014a) state that it is known as a "door to Kashmir" since it is a historical corridor to the north. In their study, a total of 1,584 married females originating from three tehsils and 24 sampling sites in Bhimber district were randomly recruited. We utilized a dataset with 16 variables in this study. The primary objective of this research is to examine consanguineous marriage as the dependent variable, with other covariates such as age, language, caste, education, etc., considered for both wife and husband. Fig. 2 shows the study area, focusing on consanguineous marriages in Bhimber district, Pakistan. A GIS-created map has been used to determine the district's boundaries. The map provides a visual picture of the district's layout, giving a thorough comprehension of the region being researched for consanguineous marriages.

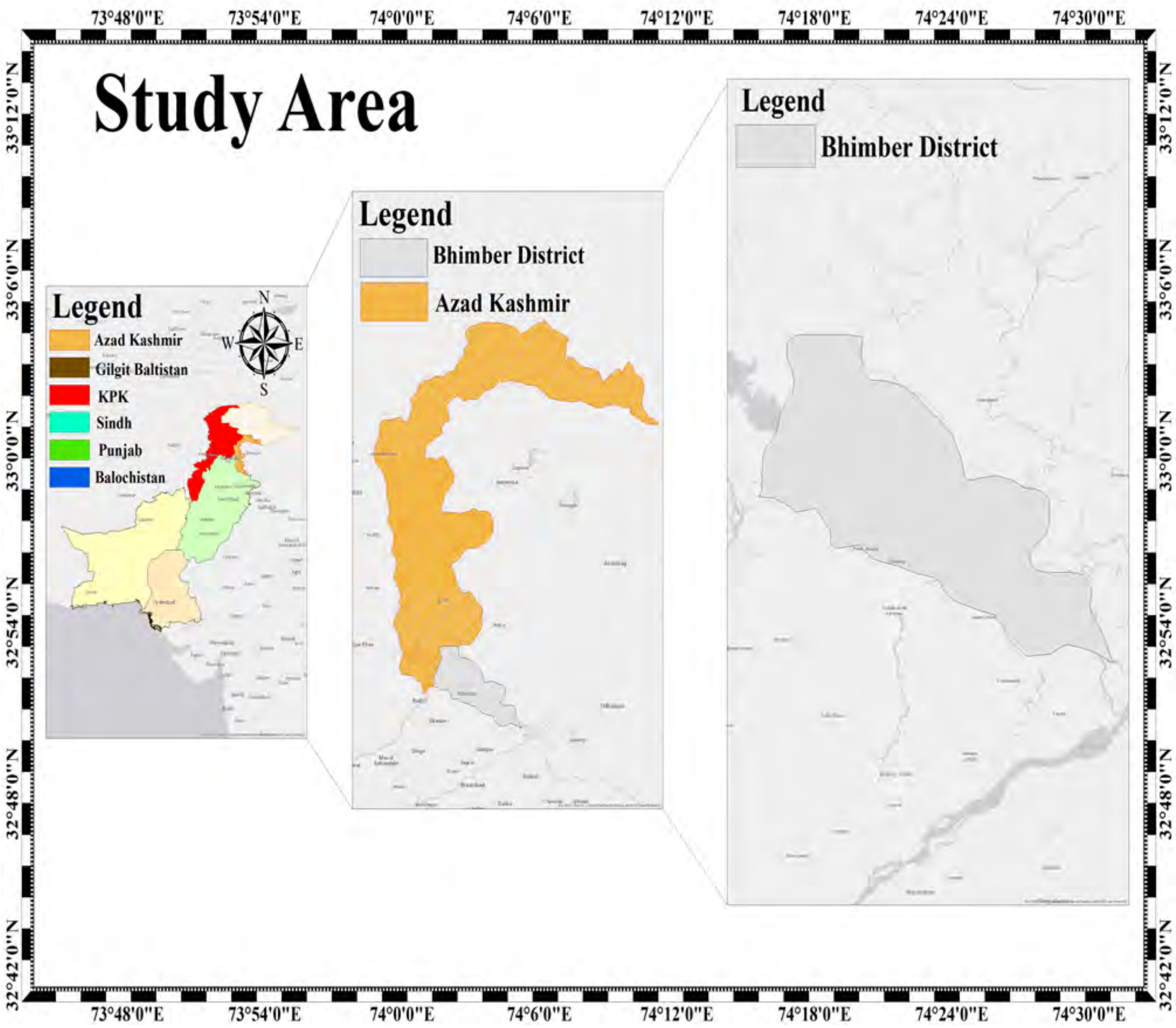


Figure 4.1: Study area map for Bhimber district, Pakistan.

4.2.2 Relevant features based on a BORUTA-algorithm

When developing our data model, variable selection is a crucial strategy before fitting the models because it eliminates irrelevant data from the study and takes out strongly correlated variables. (Bootstrap Aggregating with Random trees) BORUTA graphs illustrate the significant characteristics in our study, aid predictions based on input features, and provide a detailed discussion of the BORUTA, shown in the given article (Bagherzadeh-Khiabani et al., 2016). BORUTA iteratively performs a randomization process before ranking the significance of each feature in the data. All the green box plots depict the variables that are significant in the sets of data and statistically more significant than the corresponding characteristics in the study. The BORUTA method's initial step is to create a shadow feature set, which is a modified version of the original features, according to the general description. Once the combined set of original features and shadow features is ready, a random forest model is trained. The method evaluates the highest importance of each related shadow characteristic and compares it to the importance of each original feature. It is considered important and given the "Confirmed" label if a feature's relevance is noticeably greater than the maximum importance of its shadow counterparts. It is marked as "Tentative" and further assessed in subsequent iterations if the significance of a trait doesn't change noticeably. Features consistently lacking significant importance are marked as 'Rejected.'" This repeated process is carried out using the BORUTA algorithm until all features are accepted or rejected. Additionally, a term known as "Z-score" is introduced to describe the relevance of feature importance. The Z-score measures the distance between a feature's maximum relevance and the maximum importance of its shadow features. BORUTA excels in feature selection for its ability to handle complex datasets with a large number of features. It solves the drawbacks of conventional feature selection strategies, such as step-wise selection or univariate methods, which may neglect important features or miss interactions between features.

Steps for BORUTA algorithm

- The way that the method works is by copying features from the original dataset. Each column's values on this copy are randomly shuffled to create randomization.

- The term "Shadow Features" refers to these arranged characteristics. The original features are combined with the shadow features to create a new feature space whose dimension is double that of the first data.

- The formula Z-Score original > Z-Score Maximum, shadow tests to determine if the actual (original) feature is more important than the maximum importance of the shadow features.

- If that was the case the feature remains since it is significant, alternatively if it is insignificant, it is removed from the dataset.

- The dataset used in the second iteration was produced utilizing the features that fulfilled the requirements in the first iteration. These characteristics are used again to construct shadow features, and the algorithm assesses their importance as it did in the first iteration.

- While certain features remain, others are deleted. This continues until a certain number of iterations have been completed, all features have been confirmed, or all features have been dropped.

4.2.3 Decision tree

The decision tree is a crucial machine-learning technique utilized for both classification and regression analysis. Its visual representation, resembling a flowchart, facilitates easy interpretation, allowing individuals to comprehend the decision-making process visually. It enables you to decide on any specific point. In classification problems, it predicts labels based on features or attributes, while in regression problems, it predicts numerical values based on input features or attributes. It traverses from the root node to a leaf node, making the final decision based on the provided data. The decision tree model is constructed by partitioning the data into a training set and a testing set. The model is then fitted to the training data and evaluated on the test data. The main goal is to find the most informative features in the data to split into distinct groups or classes.

The nodes of the decision tree include: Root Nodes—these are the starting nodes that split the data variables. Intermediate nodes: Intermediate nodes split the data but do not provide predictive values for prediction; they display multiple outcomes in the decision tree. End nodes or Final nodes: It gives us the final decision of the outcome or the class. The decision tree model for predicting consanguineous marriages was trained and evaluated using cross-validation with the specified parameter configuration. Our objective was to enhance the robustness and generalizability of the decision tree model for predicting consanguineous marriages in the Bhimber district district. This was achieved by employing cross-validation on the same training and testing dataset with consistent parameters. Through cross-validation, we assessed the model's performance across multiple subsets of the training data, allowing us to identify potential bias or overfitting. The shape of the decision tree is like in Fig 4.2.

Gini-Index:

The Gini-Index is a metric algorithm in machine-learning. It is used for classification analysis

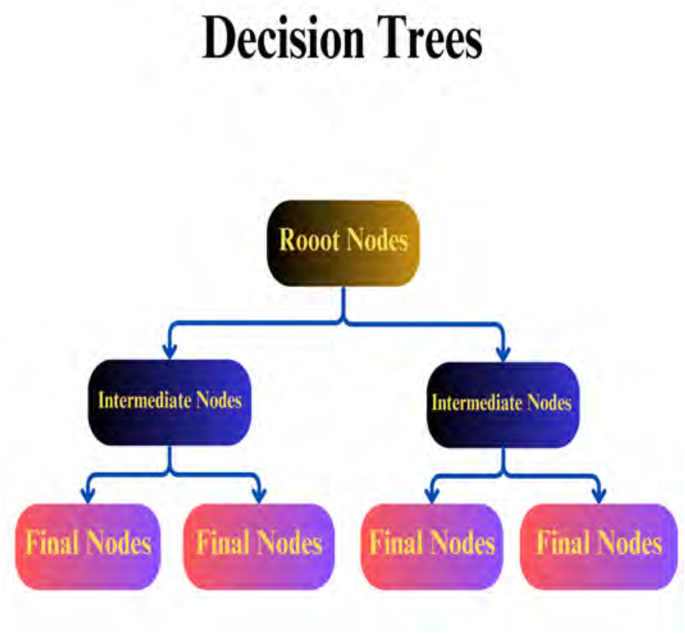


Figure 4.2: A graphical illustration of the decision tree node structure.

in machine-learning and for evaluating the performance of models. It is the sum of the square of the probability of each class. The Gini coefficient, sometimes known as the Gini index, is the most widely used indicator of inequality. By (Ceriani and Verme, 2012) (Berndt et al., 2003) Corrado Gini (1884–1965), an Italian statistician, proposed it. The following is the Gini-index formula:

GI(Y)= $1 - \sum_{i}^{k}(p_i^2)$

In the above equation, k is the total number of classes and p_i is the relative frequency of class i in Y. If the data set on attributes into $class1$ and $class2$ and $Y \in (class1, class2)$ with size N_1 and N_2, respectively, then Gini-index is calculated as:

1–$(P(class1)^2 + P(class2)^2)$

where $P(class1)$ = proportion of class 1/number of total class.

The given two steps proceeded to make the decision tree: First step: Find the Gini index for each independent variable in the data. Second Step: In the second step, select the attribute that has the lowest Gini-Index, and that will be the root node in your data if you want to build the decision tree. The given procedure continues for the second root node or intermediate node until it reaches the final nodes or end nodes.

4.2.4 Support vector machine

Another commonly used method in machine-learning is the support vector machine, particularly known for its applications in classification. SVM finds applications in classification, regression analysis, image classification, bioinformatics, and text organization. However, it is most commonly employed for binary classification. In classification, SVM aims to find a decision boundary that effectively separates classes and generalizes well to an independent set. It handles both linear and non-linear datasets by using different kernel functions, and SVM can handle high-dimensional data. In the SVM algorithm, we try to separate each point by using hyperplanes that differentiate two classes of data points. Here is one question that arises in our minds if we deal with SVM: which hyperplane will be best? So, that hyperplane will be optimal, which maximizes the distance between the nearest data points, which means it should be the maximum width between support vectors; that distance is called the margin. SVM employs various kernels such as linear, polynomial, radial basis functions, etc., and its performance is significantly influenced by the choice of kernels and their parameters. SVM is used for small data in machine-learning, and it is a supervised machine-learning algorithm. If the data can be easily split by a hyperplane, a linear kernel is used; otherwise, other kernels such as Gaussian, polynomial, radial, etc., are employed.

Linear support vector machine

Linear support vector machine is commonly used in binary classification tasks, especially when the data is separable linearly. The goal is to find the optimal hyperplane that best separates the data, resulting in robust and precise predictions for the particular problem at hand. In cases of linear separability based on sample information (x_1,y_1),(x_2,y_2),..., (x_n,y_n), new points can be classified using the formula $f(y) = w \cdot x + b$. Here, (w) represents weights, (b) is the linear coefficient estimated from the training data, and x is the data matrix for classification. The following graphical representation illustrates the decision boundary generated by the linear support vector machine. If $f(y)$ is greater than 0, then the sample belongs to the positive class of the dependent variable, while if $f(y)$ is less than 0, then the sample belongs to the negative class of dependent variable.

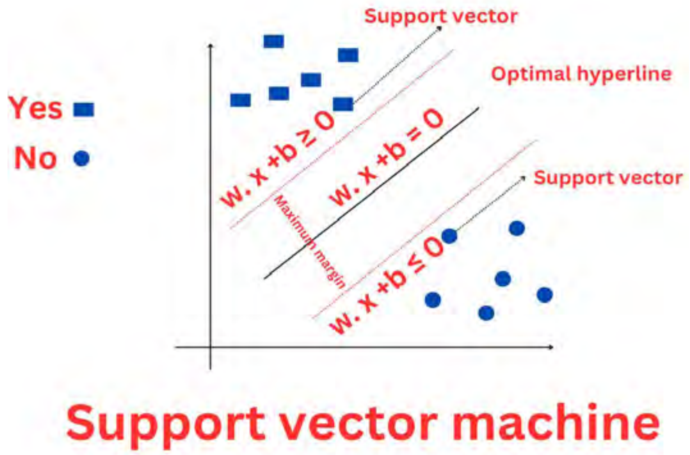


Figure 4.3: A graphical illustration of the support vector machine.

4.2.5 Bagging or Bootstrapping algorithm

Bagging, or Bootstrap aggregating, is an ensemble learning technique that combines multiple base models to create a robust prediction model for input data. It involves generating multiple bootstrap samples from the original dataset, with each sample used to train a base model. From the original dataset, multiple bootstrap samples are produced, and each sample is used to train a base model. The final prediction is generated by averaging or voting the predictions from each base model, a process designed to enhance the overall accuracy and reliability of the prediction. The following Fig 4.4 shows the bagging technique and how to do this work. In

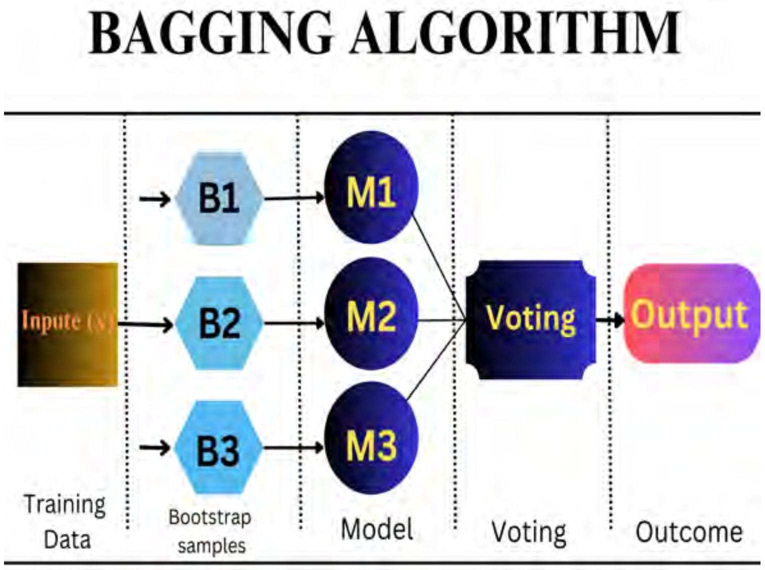


Figure 4.4: The bagging approach illustrated for classification output.

the context of binary classification problems, bagging involves partitioning the initial training data into subsets through bootstrap sampling. Various classifiers are trained on these subsets, and their predictions are combined to make a final classification determination. To improve prediction accuracy, cross-validation is also added to the bagging model. Let's denote the training data set by S, which consists of N samples of the training data. Each sample is represented as a pair (x, y), where x is the feature vector and y is the corresponding target value. Bagging is the process of creating multiple bootstrap samples by randomly selecting N samples and replacing them.

A base model is trained to produce a corresponding prediction model for each bootstrap sample. Let's the base model denoted by $M(b)$. Various base models yield the final prediction once all of the base models $M(b)$ have been trained. The combination depends on the problem type because the goal of this research is to predict consanguinity using a binary classifier. For classification problems, the most common method is voting. Each base model predicts the class label for a specific input sample, and the base models vote to decide the final prediction. To assess how effectively the bagging model will perform on untested data, cross-validation is employed. It involves dividing the provided dataset into several folds or subsets. A portion of the data is used to train the bagging model, while the remaining data is used to evaluate it. This procedure is performed multiple times, with the validation set changing each time. An estimate of the bagging model's overall performance is then provided by averaging or combining the findings from each fold.

4.3 Evaluation metrics for evaluating model performance.

Area Under the Curve (AUC)

Accuracy is commonly used for prediction purposes in everyday life, in classification problems used AUC value is because it is the measure of binary classifier. It is the plot of the true positive rate (TPR) against the false positive rate (FPR). AUC is used for classification problems with probability and more deeply analyzing the prediction of outcomes. The range of AUC values is between 0 to 1, if the value of AUC is close to zero, it indicates that the classifier is random, and if the value of AUC is close to 1, it indicates that the classifier is best performed.

Receiver Operating Characteristic Curve (ROC)

The ROC curve is commonly used to illustrate the trade-offs between sensitivity and specificity for a binary classifier. The majority of machine-learning classifiers generate real-valued scores that are correlated with how strongly an instance is predicted to be positive. Setting a threshold is required to turn these real-valued ratings into ”yes” or ”no” predictions; circumstances with scores above the threshold are considered positive, while situations with scores below the threshold are considered negative. Different threshold levels give varying degrees of sensitivity and specificity. While a high threshold is more careful in identifying a case as positive, it increases the likelihood that positive cases will be missed (lower rate of true positives). Because positive labels are created more frequently, a low threshold is less exact (more false positives), but also more sensitive (more true positives). The ROC curve shows the entire range of such options by displaying the true positive rate against the false positive rate.

Sensitivity

Sensitivity is the ability to show a positive rate after the screening test. It tells about a person who is suffering from a disease. If sensitivity is low, then it predicts that your model may predict similar predictions for different input data. If the model has low sensitivity, then it needs further improvement by using different techniques like feature selection, ensemble methods, cross-validation, etc. To calculate sensitivity, use the following formula.

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity

Specificity is the ability of a test to tell about the true negative or to declare about the healthy person, it means no suffering from the disease. It is another metric performance in machine-learning to evaluate the performance of a model to correctly identify negative cases.

When the model has high specificity, it means that the model is more likely to be correct. If the model is predictive of negative cases, then you have more confidence in the accuracy of the model. The given formulate is used to calculate Specificity.

$$Specificity = \frac{TN}{TN + FP}$$

Confusion Matrix

Confusion matrix is an evaluation tool used to display model performance; it includes summary information on classification models such as true positive, true negative, accuracy, as well as prevalence rate, etc. It can be used for two-by-two binary classification as well as an extended form of multi-class classification. It provides a detailed analysis of the model's actual and predicted classifications because binary classification has two classes, one positive and one negative.

4.4 Results and discussion

Utilizing various statistical analyses to assess consanguineous marriages in Bhimber district, Kashmir, Pakistan, an association study was conducted, as illustrated in Table 4.1. This analysis aimed to explore the relationship between consanguinity and several variables, employing statistical measures such as odds ratio, chi-square, and confidence interval. Covariates identified as significant during this analysis included age, caste, wife' language, and years of marriage.

Results from the analysis indicated that, when comparing women aged 18-25 to those aged 26-30, 31-40, and above 40, the odds ratios for consanguineous marriages were 0.96, 0.92, and 0.69, respectively. Similarly, wives from the Choudary and Rajpot castes, when compared to the Jat reference group, showed an identical odds ratio of 0.67 for consanguineous marriages. This implies that Choudary and Rajpot individuals have 33% lower odds of consanguineous marriages compared to those from the Jat caste, while individuals from other castes exhibit slightly lower odds with an odds ratio of 0.96. Table 4.1 presents comprehensive information on the associations between consanguineous marriages and various factors, including odds ratios, chi-square values, and confidence intervals.

Similar association analyses were conducted for the husband, as depicted in Table 4.2, examining demographic variables related to consanguineous marriages. Odds ratios for consanguineous marriages were significantly identified when comparing husbands aged 18-30 to those aged 31-40, 40-50, and above 50 (0.72, 0.77, and 0.45, respectively). Additionally, odds ratios for consanguineous marriages were 1.29 for husbands with metric education and 1.05 for husbands with a graduation degree, compared to the reference group of husbands with an undergraduate education. This indicates that individuals with metric education have 29% higher probabilities of consanguineous marriage, and those with a graduate degree have 5% higher odds.

Table 4.2 provides additional insights into odds ratios, chi-square values, and confidence intervals for each category of husband-related variables. Furthermore, bar charts were created for each relevant variable associated with consanguineous marriages, as depicted in Figure 4.10.

Table 4.1: Association analysis of wife consanguineous marriage with demographic factors using chi-square test, confidence intervals, and odds ratio.

Variable	Categories	OR	χ^2,df	P-value	Confidence interval
Age					
	18-25	Reference			
	26-30	0.96			0.70-1.31
	31-40	0.92			0.69-1.23
	above(40)	0.69			0.51-0.93
			7.81, 3	P=0.04, **sig**	
Education					
	Metric	0.70			0.42 - 1.18
	Undergraduate	0.70			0.41 - 1.21
	Graduate	Reference			
			1.88, 2	P=0.38, Non-sig	
Occupation					
	House Wife	1.08			0.79 - 1.51
	Others	Reference			
			0.17, 1	P=0.68, Non-sig	
Rural/Urban					
	Rural	0.91			0.58 - 1.46
	Urban	1.18			0.67 - 2.05
	Peri-Urban	Reference			
			0.33, 2	P=0.85, Non-sig	
Castes					
	Choudhary	0.67			0.48 - 0.93
	Rajput	0.67			0.45 - 0.95
	Others	0.96			0.71 - 1.29
	Jat	Reference			
			12.97, 3	P=0.00, **sig**	
Language					
	Punjabi	0.71			0.57 - 0.88
	Pahari	Reference			
			9.08, 1	P=0.00, **sig**	
Marriage year					
	Below 1990	1.35			1.08-1.70
	Above 1990	Reference			
			6.46, 1	P=0.01, **sig**	

In this analysis, we utilize bar charts to examine the distribution of classes within the data. The bar chart provides valuable insights into the prevalence of consanguineous marriages in the study district.

Table 4.2: Association analysis of husband consanguineous marriage with demographic factors using chi-square test, confidence intervals, and odds ratio.

Variables	Categories	OR	χ^2,df	P-value	Confidence interval
Age					
	18-30	Reference			
	31-40	0.72			0.54 - 0.95
	40-50	0.77			0.56 - 1.05
	above(50)	0.45			0.33 - 0.62
			25.64, 3	P=0.00, **sig**	
Education					
	Metric	1.29			0.85 - 0.99
	Graduate	1.05			0.75 - 0.51
	Undergraduate	Reference			
			2.03, 2	P=0.36, Non-sig	
Occupation					
	Well Job	0.90			0.73 - 1.11
	Others	Reference			
			0.86, 1	P=0.35, Non-sig	
Castes					
	Choudhary	0.67			0.48 - 0.94
	Jat	0.65			0.45 - 0.94
	Others	0.97			0.71 - 1.31
	Rajput	Reference			
			13.45, 3	P=0.00, **sig**	
Language					
	Pahari	0.69			0.56 - 0.87
	Punjabi	Reference			
			9.09, 1	P=0.00, **sig**	
Family type					
	Nuclear	1.06			0.51 - 0.26
	Grand parents and one couple	0.85			0.83 - 1.97
	Extended family	0.67			0.32 - 1.43
	More than One Couple	Reference			
			18.059, 3	P=0.00, **sig**	
Marriage Type					
	Arrange marriage	1.77			1.25 - 2.55
	Self Arrange	Reference			
			9.54, 1	P=0.00, **sig**	
District					
	Bernala	0.71			0.55 - 0.90
	Samahni	0.57			0.42 -0.74
	Bhimber	Reference			
			18.19, 2	P=0.00, **sig**	

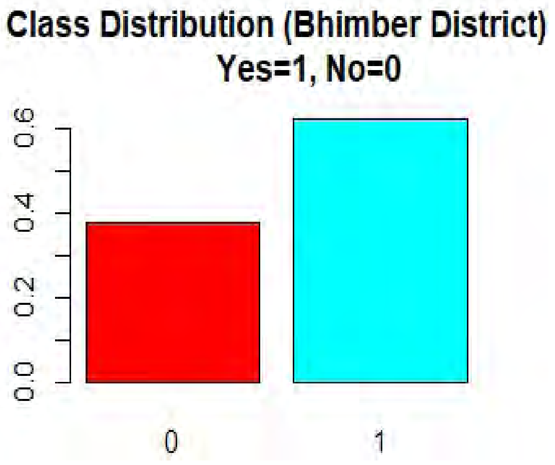


Figure 4.5: Class distribution of consanguineous marriages in Bhimber district Kashmir, Pakistan.

Upon analyzing the bar chart, as shown in Figure 4.5, we observe that approximately 60% of the data points are classified as 'Yes' for consanguineous marriage, while the remaining 40% are labeled as 'No.' This distribution sheds light on the prevalence of consanguineous instances in the two categories. The information derived from the bar charts forms the basis of our study, allowing us to enhance the accuracy and reliability of our classification results. Addressing class imbalance ensures that all classes are adequately represented in the predictive models, contributing to a more robust analysis. Based on the correlation plot depicted in Figure 4.6, a distinct correlation analysis specific to the Bhimber district was conducted to examine the relationships among various independent variables. The plot reveals a substantial correlation between the castes of wife and husbands, as well as correlations between the languages spoken by both parents. Additionally, some variables exhibit highly negative correlations, such as the language of wife and tehsil. These findings offer valuable insights into the intricate interactions of these variables within each district, highlighting underlying patterns and relationships. The robust correlations identified indicate the presence of multicollinearity, a phenomenon that can pose challenges in regression analysis, leading

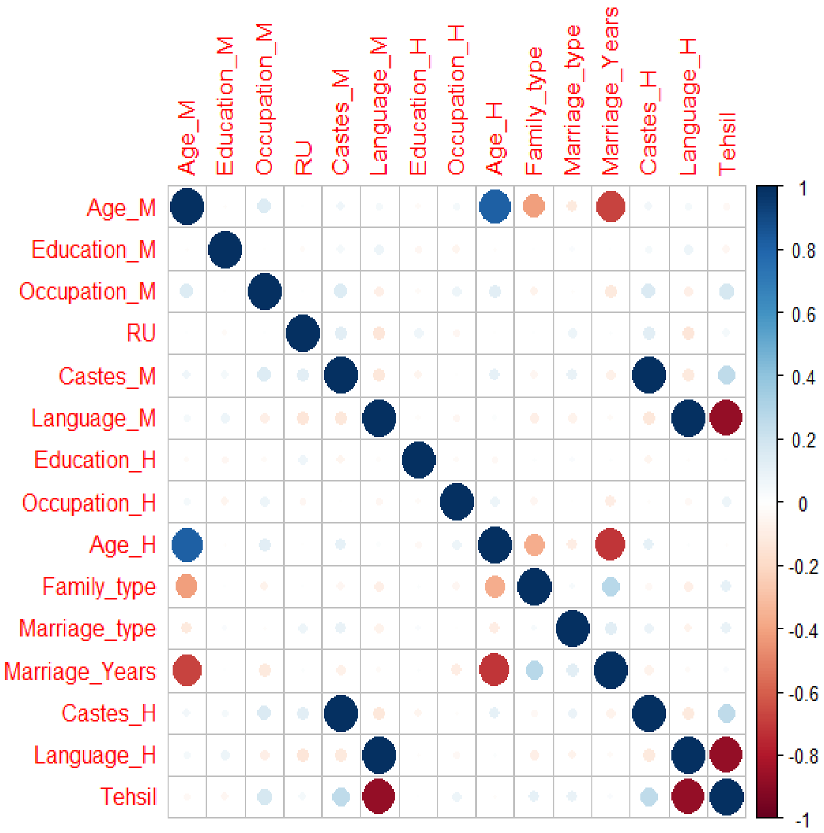


Figure 4.6: Plotting connections between variables using Correlation plot for Bhimber district.

to overestimated standard errors and difficulties in evaluating individual predictor effects. To address this issue, we may consider exploring the removal of highly correlated variables through feature selection techniques. Effectively managing multicollinearity will enhance the reliability of our findings, allowing us to draw stronger conclusions and make informed decisions based on the data.

Table 4.3: Results of the BORUTA feature selection for the Bhimber district of Kashmir.

Variable	meanImp	medianImp	minImp	maxImp	normHits	decision
Age_M	4.66	4.52	0.62	8.92	0.70	Confirmed
Education_M	0.93	1.57	-2.06	2.93	0.00	Rejected
Occupation_M	0.94	0.69	-1.65	3.97	0.01	Rejected
RU	-0.70	-1.31	-2.90	2.04	0.00	Rejected
Castes_M	5.70	5.61	2.55	9.08	0.91	Confirmed
Language_M	8.13	8.10	5.78	11.52	1.00	Confirmed
Education_H	1.24	1.28	-0.22	2.63	0.01	Rejected
Occupation_H	-0.10	0.54	-2.91	2.38	0.00	Rejected
Age_H	8.19	7.95	4.59	12.39	1.00	Confirmed
Family_type	5.24	5.18	1.51	10.33	0.77	Confirmed
Marriage_Type	6.50	6.40	2.09	10.28	0.92	Confirmed
Marriage_Years	2.16	1.81	-0.64	6.78	0.02	Rejected
Castes_H	6.48	6.52	3.62	12.59	0.97	Confirmed
Language_H	7.51	7.39	5.60	10.29	0.97	Confirmed
Tehsil	7.80	7.90	3.65	11.12	0.98	Confirmed

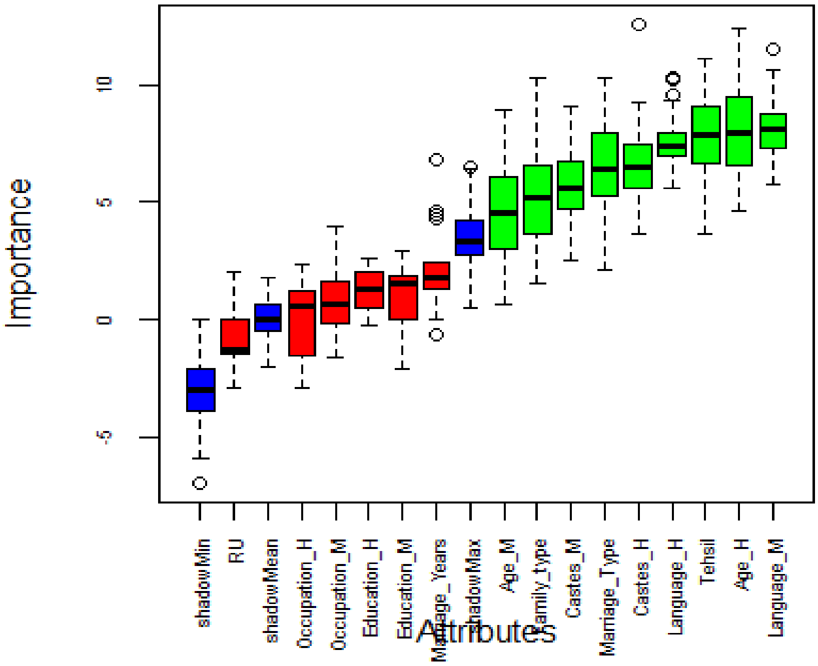


Figure 4.7: Using BORUTA technique to extract the important variables for the Bhimber district.

Fig 4.7 illustrates the importance of variables in our study as determined by the BORUTA function. In the graph, the green box signifies all the important variables identified in the present study. The concept of Shadow Features is employed in BORUTA, where the values of each feature are randomly shuffled and added to the original dataset. These shadow features serve as a baseline for comparison with the actual features, acting as representations of noise. Relevant Features, in contrast, exhibit box plots that are shifted to the right (towards greater values) when compared to the shadow features. The analysis identifies nine crucial characteristics, as depicted in Fig 4.7 and detailed in Table 4.3, which are highly associated with consanguineous marriage in the Kashmir district of Bhimber, Pakistan. Other variables in the study are not deemed as highly relevant.

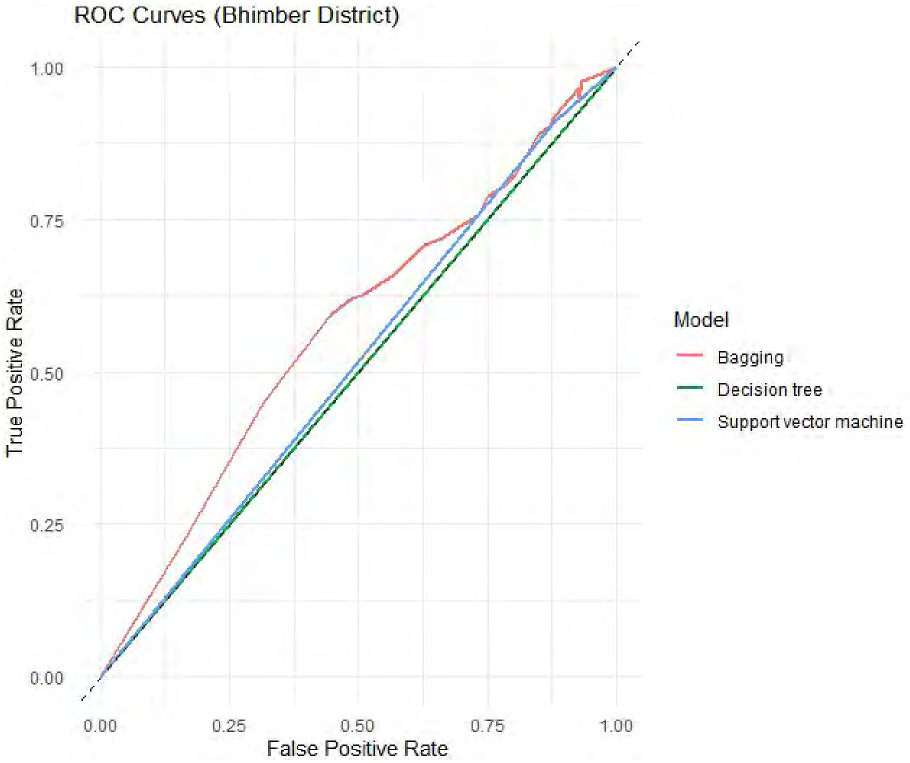


Figure 4.8: ROC comparison of models performance for Bhimber district.

Table 4.4: Comparison of models prediction using Bhimber district Kashmir, Pakistan.

Models	AUC	Sensitivity	Specificity	Prevalence rate
Decision tree	0.46	0.24	0.83	0.4
Support vector machine	0.52	0.12	0.91	0.4
Bagging	0.57	0.25	0.79	0.4

Comparing the sensitivity values of three models—decision tree, support vector machine

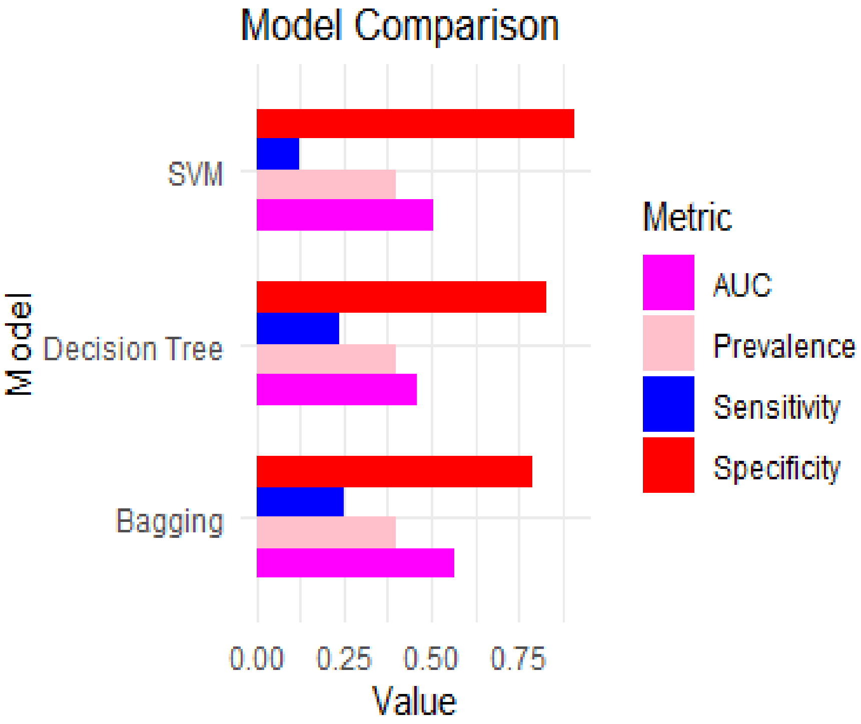
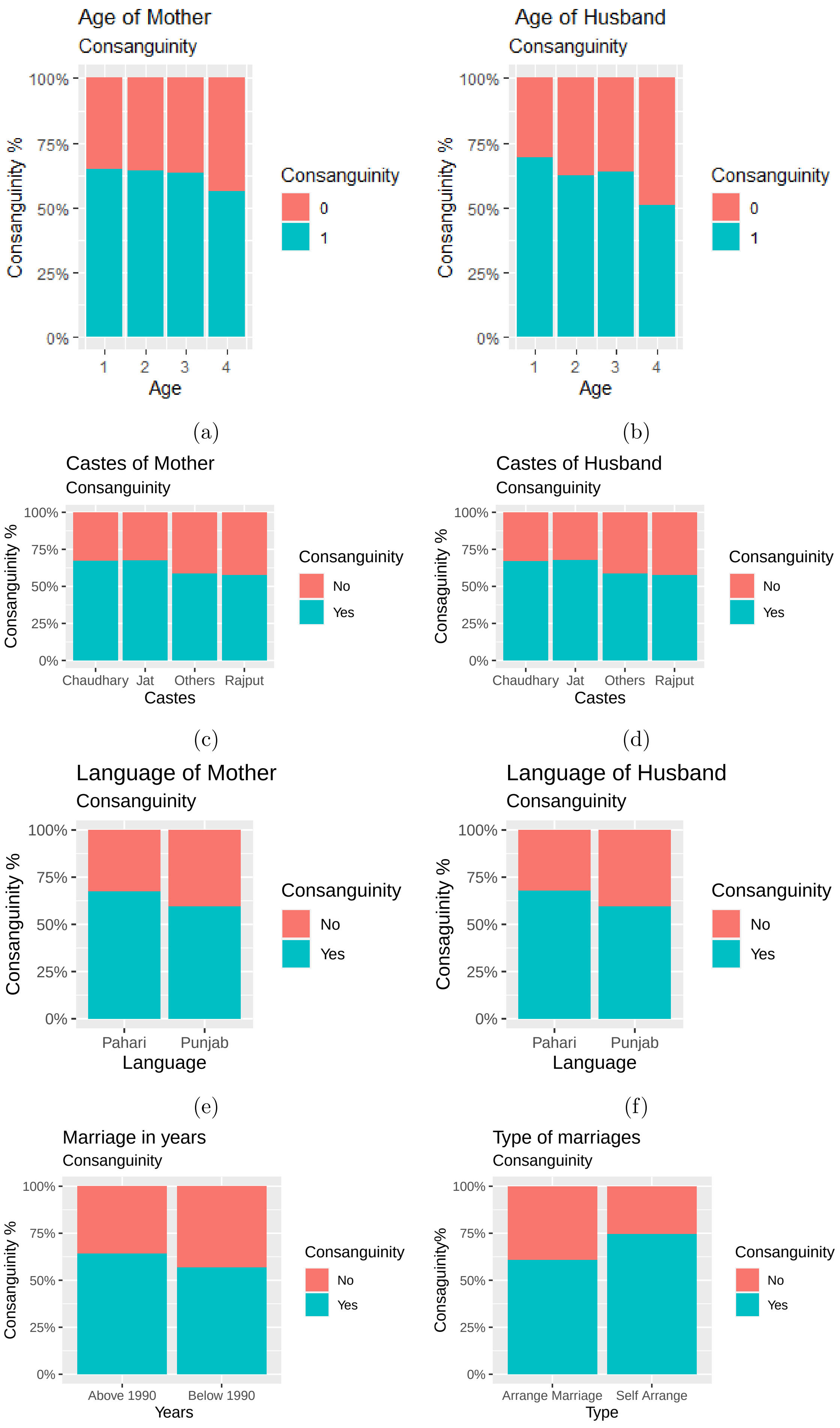


Figure 4.9: Comparison of model performance using various metrics for Bhimber district.

(SVM), and bagging—provides valuable insights into their effectiveness in accurately detecting positive cases. The decision tree model exhibited a sensitivity of 0.13, correctly identifying 13% of actual positive cases but missing the remaining 87%. Similarly, the SVM model had a sensitivity of 0.12, correctly identifying only 12% of true positive cases. The bagging model demonstrated a sensitivity of 0.25, indicating that, like the decision tree, it correctly recognized only one-quarter of positive situations, as shown 4.9.

Turning attention to specificity, which identifies correctly classified negative cases, the SVM model stood out with a high specificity of 0.91, showcasing its superiority in identifying true negative cases. The bagging model, with a specificity of 0.79, performed well in identifying true negatives but might encounter challenges with false positives. In summary, the SVM outperformed the decision tree and bagging models in their respective ability to identify negative cases. Table 4.4 presents the efficacy of several models for consanguineous marriages, with AUC values ranging from 0.46 for the decision tree, 0.51 for SVM, to 0.57 for bagging. The AUC values provide a single scalar measure to evaluate the performance of different models. The ROC curve, or AUC values, assess how well the models distinguish between two groups. While all models moderately performed for the given data, bagging exhibited superior performance with an AUC value of 0.57, indicating higher predictive ability compared to the other models. This conclusion is supported by the visual representation of model performance in Fig 4.8 through the ROC curve, and provides additional information about the investigated models, including sensitivity, specificity, prevalence rate, and AUC values, offering a comprehensive overview of their performance, as shown in Table 4.4.


Age of Mother
Consanguinity
Consanguinity %
100%
75%
50%
25%
0%
1
2
3
4
Age
Consanguinity
0
1
Age of Husband
Consanguinity
Consanguinity %
Age
Consanguinity
0
1
Castes of Mother
Consanguinity
Consanguinity %
Chaudhary
Jat
Others
Rajput
Castes
Consanguinity
No
Yes
Castes of Husband
Consanguinity
Consaguinity %
Chaudhary
Jat
Others
Rajput
Castes
Consanguinity
No
Yes
Language of Mother
Consanguinity
Consanguinity %
Pahari
Punjab
Language
Consanguinity
No
Yes
Language of Husband
Consanguinity
Consaguinity %
Pahari
Punjab
Language
Consanguinity
No
Yes
Marriage in years
Consanguinity
Consanguinity %
Above 1990
Below 1990
Years
Consanguinity
No
Yes
Type of marriages
Consanguinity
Consaguinity%
Arrange Marriage
Self Arrange
Type
Consanguinity
No
Yes

(a)

(b)

(c)

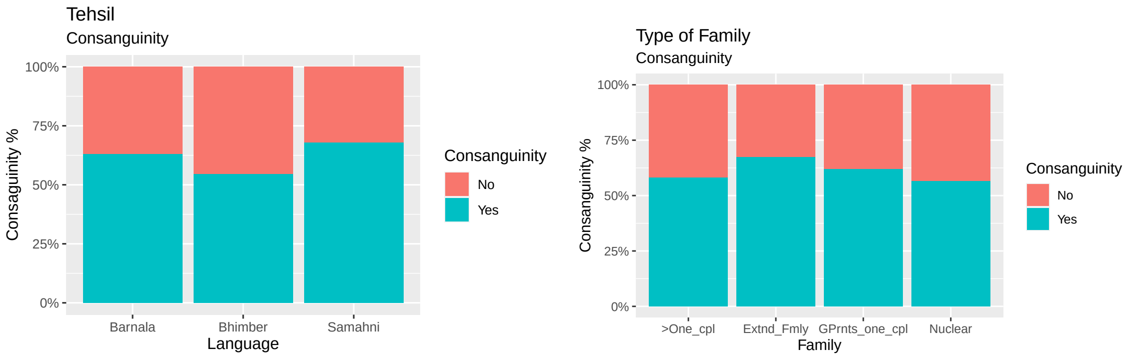
(d)

(e)

(f)

(g)

(h)



(i)

(j)

Figure 4.10: Percentage of consanguineous marriage in Bhimber district Kashmir Pakistan using significant variables.

Chapter 5

External validation of machine-learning algorithms for Mardan district.

5.1 Introduction

Various binary classification approaches are employed in this study to predict consanguineous marriages in the Mardan district. To achieve this goal, a combination of strategies is utilized, including feature selection using BORUTA before model creation. The implemented models encompass the Random Forest model, Logistic Regression, and a mixed ensemble method referred to as the Stacked Model. Each method is chosen for its distinct advantages and potential drawbacks, recognizing that performance may vary based on the characteristics of the data. To ensure the precision of predictions, the BORUTA approach selectively identifies critical marriage attributes. The dataset is divided into training and testing sets, with the former used for model training and the latter for evaluating model performance. The effectiveness of the models is evaluated using essential metrics such as AUC, sensitivity, specificity, prevalence rate, and the ROC curve, which play a crucial role in assessing the models' ability to distinguish between different marriage forms. The ultimate goal is to externally validate consanguineous marriages in the Mardan district based on the obtained findings. While details of BORUTA feature selection are presented in the first chapter, this section provides an overview of the outcomes and graphical representations of BORUTA performance specific to the Mardan district. Further model details can be found in the associated methodology and materials section.

5.2 Methodology

5.3 Selection of study area and data source

To validate the external applicability and generalizability of our findings, we utilize a distinct dataset from another region of Pakistan, known as the 'land of brave men.' The district, situated between 34° 05' and 34° 32' north latitudes and 71° 48' and 72° 25' east longitudes (Saeed et al., 2012). It is bordered to the north by the Buner district and the Malakand protected region, to the east by the Swabi and Buner districts, to the south by the Nowshera district, and to the west by the Charsadda district and the Malakand protected area. Covering a total area of 1632 square kilometers (Urooj and Ali, 2016). The Mardan district is separated into two sections: the north-eastern hilly terrain and the south-western plain. The district's southwestern part is largely made up of fertile plain with scattered minor hills (Khan et al., 2014). This dataset was derived from a prior study conducted by (Tufail et al., 2017) involving a population of 1202 individuals for external validation. In this validation, a set of 13 variables is employed, with the consanguineous variable serving as the dependent variable and others acting as covariates, including age, castes, income, occupation, and education (refer to Table 3.3). A Geographical Information System (GIS) is used to define a specific research zone, with a primary focus on Pakistan's Mardan area. This strategic decision promotes external validation research by defining the study region and ensuring a thorough understanding, as shown in Fig 5.1.

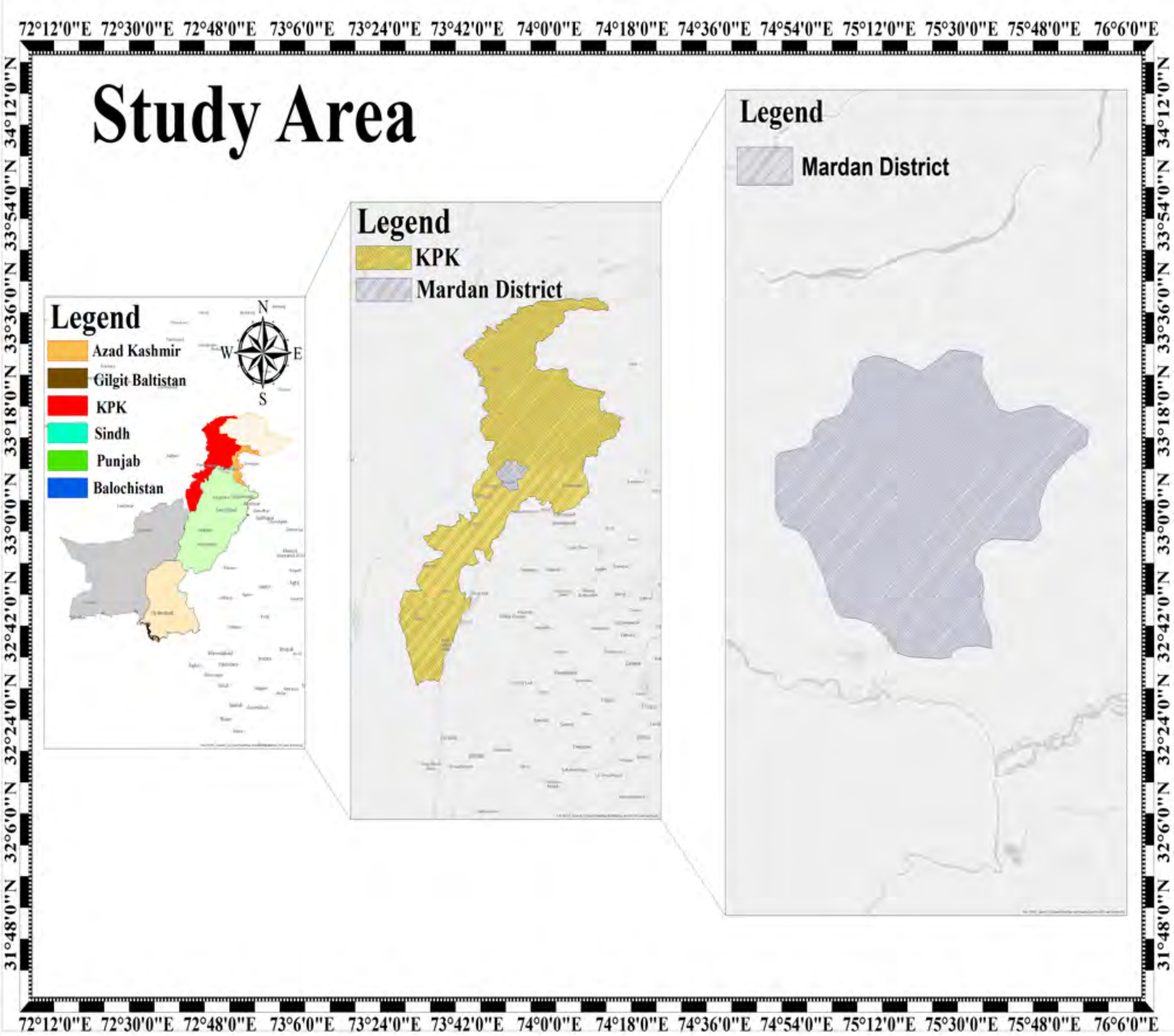

Study Area
Legend
Azad Kashmir
Gilgit Baltistan
KPK
Sindh
Punjab
Balochistan
N
W
E
S
Legend
KPK
Mardan District
Legend
Mardan District
72°12'0"E 72°30'0"E 72°48'0"E 73°6'0"E 73°24'0"E 73°42'0"E 74°0'0"E 74°18'0"E 74°36'0"E 74°54'0"E 75°12'0"E 75°30'0"E 75°48'0"E 76°6'0"E
31°48'0"N 32°6'0"N 32°24'0"N 32°42'0"N 33°0'0"N 33°18'0"N 33°36'0"N 33°54'0"N 34°12'0"N

Figure 5.1: Study area map for Mardan district, Pakistan.

5.3.1 Random forest algorithm

The random forest model is a widely used machine-learning technique, particularly popular for both classification and regression purposes. It operates by employing multiple decision trees and making decisions based on the average accuracy of the dataset. The model's precision improves with an increased number of trees. Compared to a single decision tree, a random forest generally provides more accurate results. Notably, it effectively handles missing observations in the data using robust methods.

The primary objective of the random forest model is to minimize overfitting, as it is trained on different parts of the data with the same training data. Known as an Ensemble Method in machine-learning, the random forest utilizes numerous decision trees during training. While the interpretation of a random forest model is more complex than a single decision tree, it performs exceptionally well, often ranking as a top model in competitions such as Kaggle.

The choice of a random forest model over other models is motivated by its ability to predict outcomes with high accuracy, handle large datasets efficiently, manage a high proportion of missing data, and determine the importance of ranking variables in classification or regression analyses.

In our study, to maximize accuracy, specific parameters were set for consanguinity prediction during external validation: $cost = 0.01$, $mtry = 3$, and $ntree = 500$. These parameter choices aimed to enhance the precision and robustness of the predictions in our study.

To fit a random forest model, the first step is to set the mtry value, where mtry is the square root of the predictor, and then set the ntrees. After this process, record the out-of-bag (OOB) error rate for each tree. Identify the model where the OOB error rate is at its minimum; this model will be the best-fitting model during training

In a dataset where m is the total number of input variables, only x_i attributes are randomly chosen for each tree, where m is less than x_i. The best possible split from these attributes is determined using the Gini index to develop the decision tree model. This process continues until the termination condition is met. The "meanDecrease Accuracy table" interprets how removing each variable from the data reduces the accuracy of the model. There are two main parameters in random forest models: mtry and ntrees.

where mtry= $\sqrt{ncol(x_i)}$ and x_i are the attributes in the data. ntrees: The number of trees you want to build in the data. Hyperparameter tuning is crucial in enhancing the overall performance of the machine-learning model. These parameters are commonly set before the model is fitted. Minimizing the error during the model fitting process is achieved by using hyperparameters. The goal is to find optimal hyperparameter estimates that maximize the model's performance and minimize the loss function error, resulting in better output

Start by inputting the dataset, ensuring the binary target variable is labeled and organized. Clean the data by addressing categorical variables, outliers, and missing values. Divide the dataset into training and test sets. The test set is used to assess the random forest model's performance after training on the training set To find the most pertinent characteristics for the classification problem, use feature selection approaches like correlation analysis, BORUTA algorithm, or recursive feature elimination, if necessary. This action may reduce overfitting and enhance model performance. Create a random forest classifier and import the required libraries, such as ”Randomforest” in R. Set the hyperparameters if you want, such as the number of trees in the forest, the maximum depth of each tree, and any other parameters if you want to modify your model.

Model training involves fitting the random forest model to your training set of data. In this step, many decision trees are built, each using a different random subset of the training data. Use the test set to evaluate the model's performance after training and then use the predict function to predict the test set's target variable, then compare the result to what happened. AUC-ROC, accuracy, precision, recall, and F1-score are illustrations of common evaluation metrics for binary classification.

5.3.2 Logistic Regression

In regression analysis, the prediction of the mean value is based on independent variables, unlike logistic regression, where the dependent variable is binary. Logistic Regression is employed for binary classification tasks, predicting the probability of outcomes such as a student passing or failing, or a person having a disease or not. It is a supervised machine-learning method grounded in mathematical probability to predict whether an instance belongs to a specific category. Logistic regression is also referred to as the Sigmoid function, and its curve is always S-shaped. While dealing with logistic regression, it is important to note that we predict categorical variables like true and false, right and wrong, where traditional regression analysis is not applicable. A training/testing data split was used in logistic regression in our study, which was consistent with other models. The likelihood of a binary result is estimated based on input variables, capturing correlations between these qualities and the chance of consanguineous marriage. The training data is utilized to train the model, identifying the best coefficients and thresholds for classifying cases as consanguineous marriages. Following that, the logistic regression model is evaluated on the remaining data to determine its performance.

The following equation of the logistic regression is given by.

$$P(y=1|x) = \frac{1}{1 + exp - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)} + \epsilon$$

Here y is the dependent variable and y is a function of $x_1, x_2, ..., x_k$, and $\beta_0, \beta_1, \beta_2, ..., \beta_k$ are the estimates parameter values by maximum likelihood using the following function.

Logit Model

Used to predict the logit of the dependent variable is given as:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

Exponentiate both sides and take their multiplicative inverses,

$$\frac{1-p}{p} = \frac{1}{exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)}$$

Remove the fraction on the left-hand side of the equation and add 1 to both sides of the equation,

$$\frac{1}{p} = 1 + \frac{1}{exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)}$$

$$\frac{1}{p} = \frac{exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k) + 1}{exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)}$$

Finally, we are taking the multiplicative inverse again to obtain the formula for the probability,

$$P(y=1) = p = \frac{exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)}{1 + exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)}$$

OR

The above equation multiplying and dividing by $exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)$, we get the other form of logistic equation, which is given below.

$$P(y = 1) = p = \frac{1}{1 + exp - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)}$$

where p is the probability of event, $\beta_0, \beta_1, \beta_2, \beta_3, ..., \beta_k$ are the regression coefficients, and x_i is the explanatory variables. The coefficient of β_0 is the logit of the dependent variable when all other independent variables have zero value, and the $exp(\beta)$ is the quantity of odds ratio when all other independent variables have zero value.

5.3.3 To improve the accuracy through a meta-model in ensemble learning.

machine-learning approaches called ensemble methods combine the predictions of various models to get predictions that are more reliable and accurate. By combining multiple viewpoints from many models, they draw into the insight of the public and frequently achieve better results than when utilizing just one model. Due to its ability to manage complicated patterns, minimize overfitting, and improve prediction accuracy across a variety of domains. The details discussion of the ensemble technique is available visit (Rajadurai and Gandhi, 2020). The predictions of these models can be combined using ensemble methods to generate a final prediction that is more trustworthy and accurate than the predictions of any one of the individual models. Ensemble techniques can include bagging, boosting, or Stacking. Bootstrap sampling is used in bagging approaches to construct subsets of the training data from which individual models are trained. These models' predictions are then integrated, frequently by voting or averaging. On the other hand, Boosting strategies train weak models iteratively while concentrating on fixing the errors generated by the prior models. In Stacking, various models are trained and their output features are used by a Meta-Model, which eventually learns to produce the final prediction. We combined a number of basic models, including logistic regression, support vector machine, and random forest, to create an ensemble stacking model for binary classification in this investigation. These foundation models were chosen due to their flexibility in modeling techniques and their usefulness for binary classification problems. The graphical representation of the ensemble technique is given below in Fig 5.2. We combined the predictions made by the basis models after they had been trained to serve as input features for the Meta-Model, in this instance, logistic regression is used in the Meta-Model. The original attributes from the training set are used to train the Meta-Model along with the combined predictions of the basic models. The testing set, which is made up of the data, is used to assess the ensemble stacking model's performance once it has been trained. The testing set is used as a separate sample to judge how well the model could generalize and predict outcomes from new event data. The efficiency of the model in binary classification for consanguineous marriage in the Mardan district of Pakistan is assessed using a variety of performance indicators, including AUC value, sensitivity, specificity, and ROC-AUC. In the

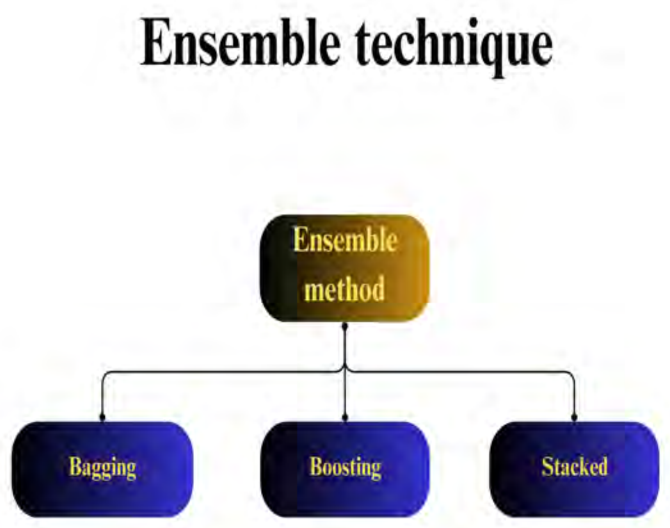


Figure 5.2: A visual representation of the various kinds of ensemble methods used in machine-learning.

procedure for external validation, ensemble approaches might be quite important. Ensemble approaches attempt to represent different perspectives and take into account differences in the data by integrating the predictions of various models. Due to the fact that they produce an aggregate project based on multiple models that have been trained on various subsets of data or using various algorithms, ensemble methods offer a more stable and dependable strategy when used for external validation. In external validation, independent data that was not used in the initial training process is used to assess the ensemble models. This data offers a good opportunity to evaluate the effectiveness and generalizability of the ensemble models and is frequently indicative of real-world circumstances. Insights into the model's capacity to handle new cases, previously observed trends, or potential biases can be obtained by examining the ensemble's predictions on the independent data.

Stacked Meta-Model

In this investigation, we are applying an ensemble technique with a staking model of machine-learning. The graphical representation of the stacked meta-model of the ensemble technique is given below in Fig 5.3. Combining the predictions of multiple models by training a Meta-

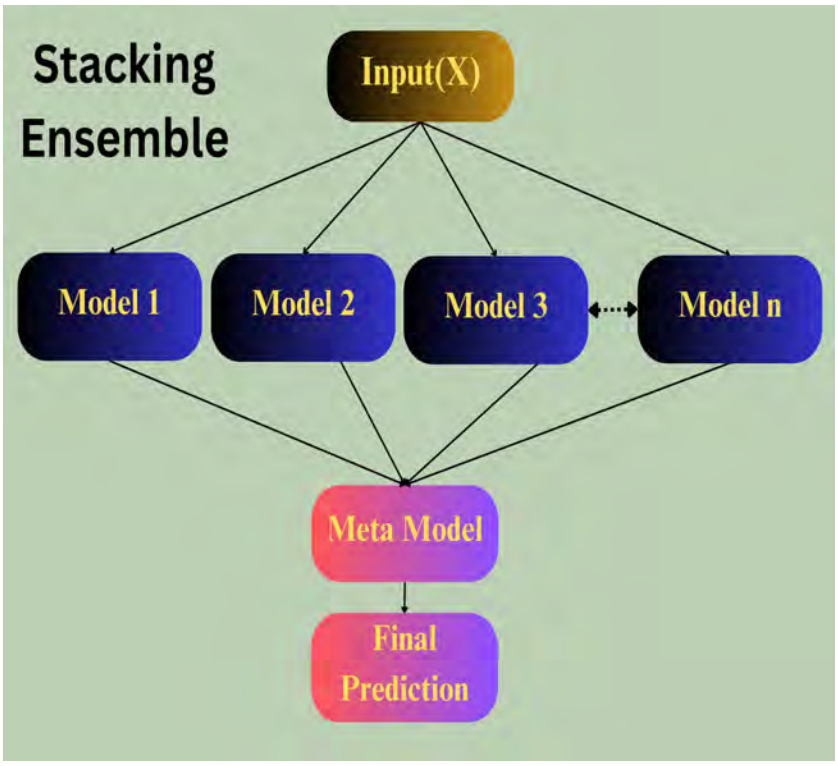


Figure 5.3: A visual representation of the Stacked-Meta-Model of ensemble methods in machine-learning.

Model on their outputs. The input attributes and base models of the Meta-Model are the predictions from the various models. Stacking improves prediction accuracy by focusing on the benefits of different models. Training a model to incorporate the predictions of various learning algorithms is known as Stacking, sometimes it is known as stacked generalization. A final estimator is trained to make a final prediction using all of the predictions of the base estimators as additional inputs or using cross-validated predictions from the base estimators, which can prevent overfitting. First, all of the other algorithms are trained using the available data. Although in practice a logistic regression model and random forest are frequently implied as the combiner, a detailed discussion of the stacked model is given in the given article (Rajadurai and Gandhi, 2020). In most cases, Stacking produces performance that is superior to any single trained model. It has been applied successfully to supervised learning problems like regression, classification, and distance learning as well as unsupervised learning like density estimator. It's essential to understand that the selection of an ensemble technique depends on the characteristics of the dataset, the performance of the individual models, and the issue domain. Prediction accuracy can be considerably increased by using ensemble techniques, especially when the individual models have different strengths and weaknesses. For each training example, the base models are predicted using the formula $h_i(x)$, where i is a number between 1 and M and x is the input instance. The predictions are stored in a matrix H with N x M dimensions, where N is the total number of training samples. A Meta-Model (such as logistic regression, Random Forest, etc.) is trained using the real labels of the training instances as well as the predictions of the basic models. The predictions of the Meta-Model are written as $g(x)$. For each test instance, the base models predict using the notation $h'_i(x)$, where i is a number between 1 and M and x is the input instance. The predictions are kept in a matrix H' with T x M dimensions, where T denotes the number of test examples to illustrate. Meta-model: The trained Meta-model generates the final ensemble prediction using the predictions of the basic models for the test instances as input. The predictions of the Meta-Model are written as $g'(x)$. The following formula can be used to compute the ensemble prediction, denoted as $Y(x)$: $Y(x) = 1$ if $g'(x)$ >threshold, else 0. In this formula, the term threshold refers to a judgment threshold that establishes the class assignment based on the predicted probability or confidence. The ensemble prediction is 1 if

$g'(x)$ is greater than the threshold and 0 otherwise.

5.4 Results and discussion

In this chapter, a variety of machine-learning tools is employed to conduct a comparative analysis of the Mardan district. The investigation encompasses critical procedures to enable accurate prediction and robust model evaluation. The method begins by examining class distribution using bar charts, as shown in Fig 5.4, offering insight into the balance of classes in the dataset. Following that, a correlation plot is created to identify any multicollinearity among independent variables, confirming the dependability of predictors. Certain variables in the Mardan district dataset, such as ”age” and ”age at marriage,” exhibit a positive association, indicating that as one variable increases, the other tends to increase as well. In contrast, factors such as ”marriage time” and ”age at marriage” display a negative association, implying that if one variable rises, the other tends to fall, as shown in Fig 5.5.

The BORUTA method is used to select features for improving model performance for consanguineous marriage prediction for the independent set of the Mardan district. This method enables the identification and retention of only variables that have strong relationships with consanguineous outcomes, thereby improving model efficiency and interpretability. Subsequently, several machine-learning models, such as logistic regression, random forest, and an ensemble technique, are built based on these fundamental features.

Table 5.1: Results of the BORUTA feature selection for the Mardan district.

Variables	meanImp	medianImp	minImp	maxImp	normHits	decision
Tehsil	16.99	16.88	11.66	22.05	1	Confirmed
Rural and urban	13.29	13.29	8.74	17.86	1	Confirmed
Age	2.65	2.63	-2.37	7.02	0.45	Tentative
Education	3.98	3.98	-1.02	9.97	0.69	Confirmed
Origin	2.36	2.39	-1.21	6.37	0.28	Rejected
Occupation	2.27	2.31	-2.36	6.04	0.2	Rejected
Origin	4.76	4.73	-2.57	10.03	0.79	Confirmed
Caste	4.64	4.53	-0.32	10.83	0.8	Confirmed
Family type	-0.63	-0.8	-2.02	1.8	0	Rejected
Marriage type	19.91	19.82	13.74	26.21	1	Confirmed
Marriage time	0.08	0.24	-2.04	1.91	0	Rejected
Age at marriage	4.16	4.17	0.51	10.13	0.73	Confirmed

Once more, we are using the BORUTA algorithm to comprehend the significance of each variable for an independent set in the Pakistani district of Mardan. The results of

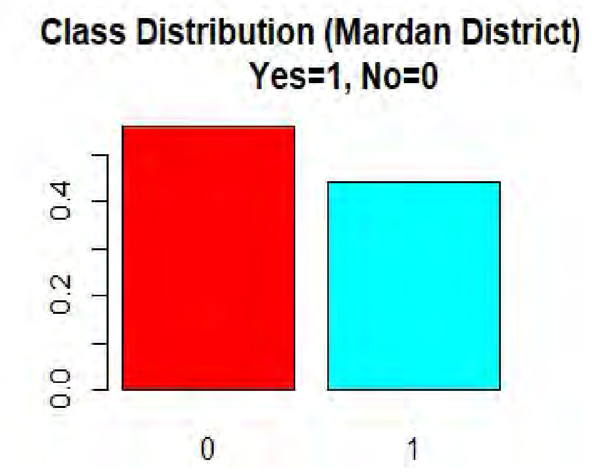


Figure 5.4: Class distribution of consanguineous marriages in Mardan district, Pakistan.

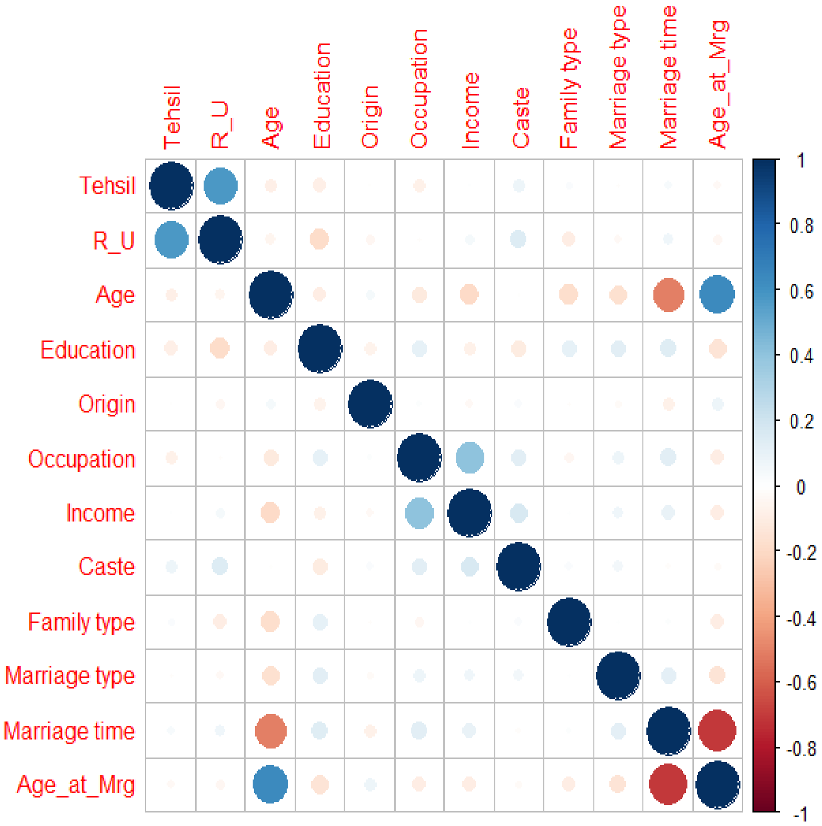


Figure 5.5: Plotting connections between variables using Correlation plot for Mardan district.

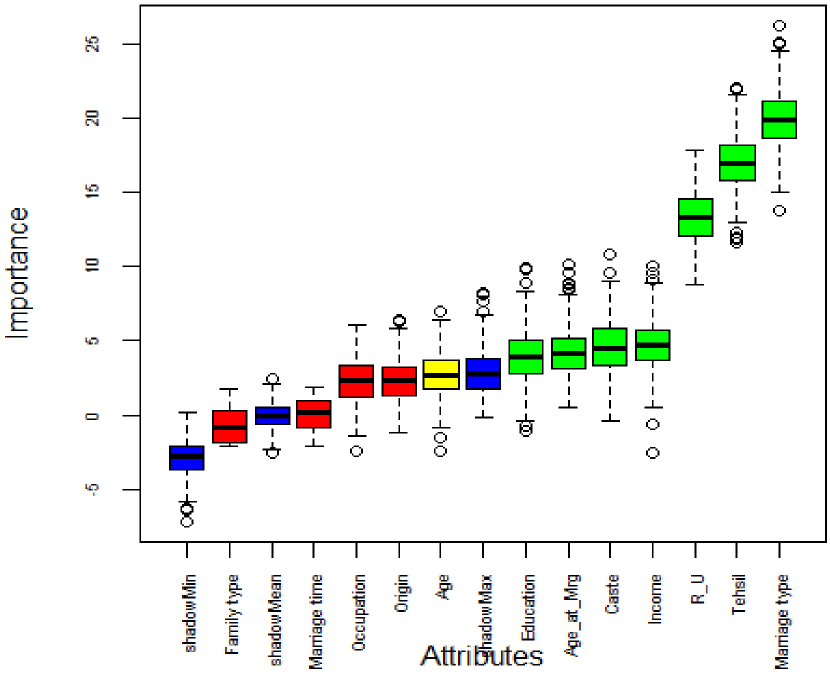


Figure 5.6: Using BORUTA technique to extract the important variables for the Mardan district.

the BORUTA feature selection strategy are shown in Fig 5.6 and can also be referenced in Table 5.1. with a focus on identifying significant variables for precisely predicting the target variable. The distribution of feature significance scores across all variables is depicted by each box plot in the picture. Features with box plots that are shifted to the right denote greater significance, whereas those with box plots that are overlapping or shifted to the left denote less significance. Relevant features are regarded as essential predictors, whilst less informative or noisy features may be disregarded to speed up the prediction process.

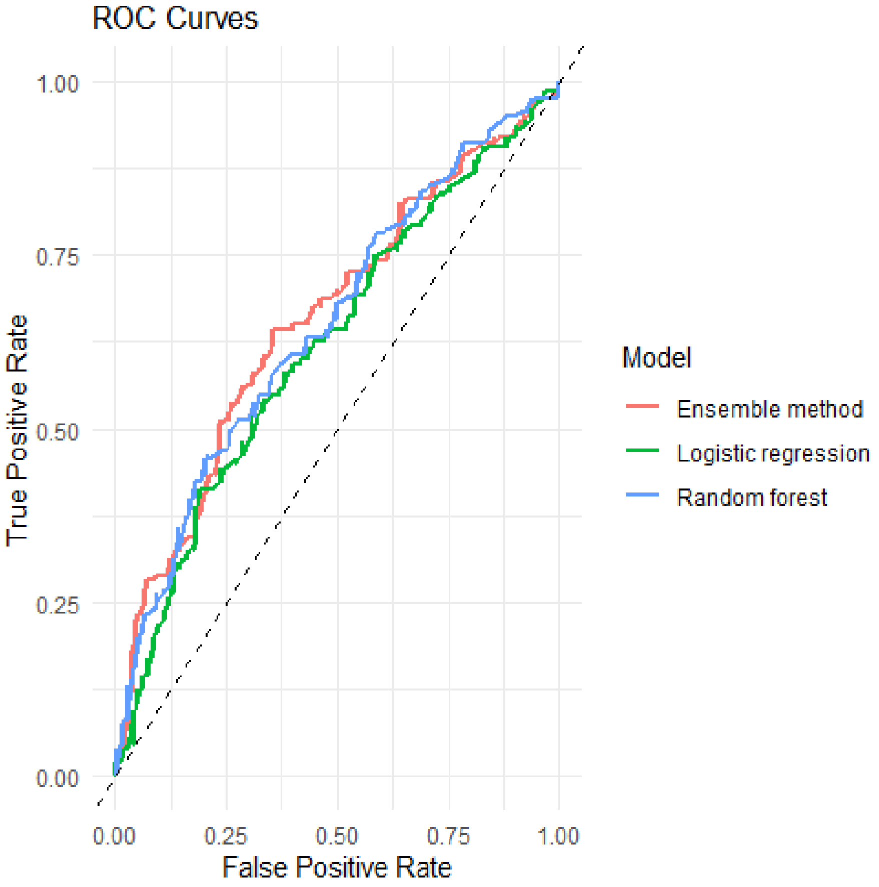


Figure 5.7: ROC Comparison of model performance for Mardan district.

Table 5.2: Comparison of model prediction using an independent set of Mardan district, Pakistan.

Models	**AUC**	**Sensitivity**	**Specificity**	**Prevalence rate**
Logistic regression	0.62	0.82	0.35	0.56
Random forest	0.64	0.74	0.47	0.56
Ensemble method	0.67	0.86	0.38	0.56

According to our second effort, the same machine-learning algorithm was applied to estimate consanguineous marriages in the Mardan district for external validation. The study compared the performance of various machine-learning techniques on a given dataset, with a particular focus on the (ROC) measure. Multiple comparison plots were utilized to evaluate the model performance with different metrics, as shown in 5.8. The AUC value, a useful metric for assessing a model's ability to differentiate across classes, was employed along with the ROC curve to check the model's performance. Higher AUC values indicate better prediction performance; Table 5.2 reveals that the ensemble method, stacked meta-model, has the highest AUC value of 0.67, indicating its superiority in capturing complicated relationships in the data.

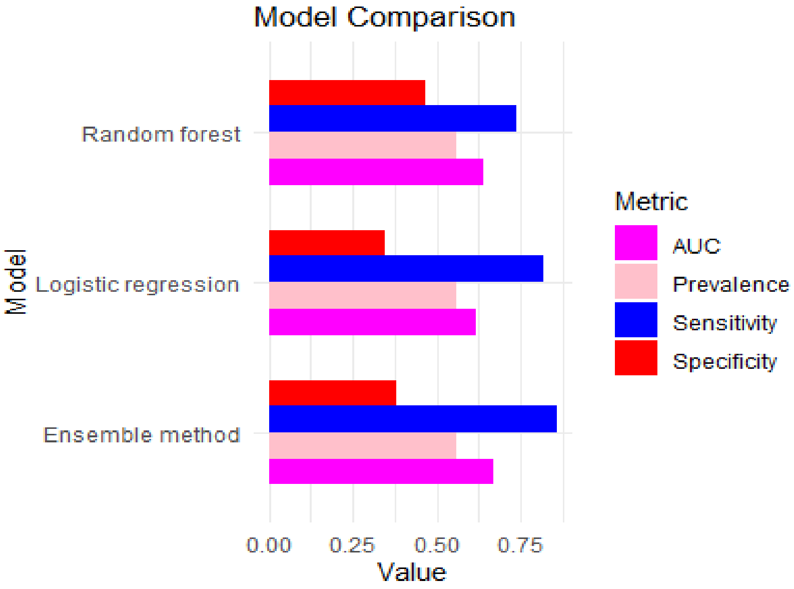


Figure 5.8: Comparison of model performance using various metrics for Mardan district.

Random forest follows with an AUC of 0.64, while logistic regression has the lowest AUC value at 0.62. The graphical representation of models using the ROC curve is displayed in Fig. 5.7, further supporting the AUC values of the models. Additional measures for evaluating model performance are presented in Table 5.2. Sensitivity assesses the models' capacity to correctly detect positive cases in the Mardan district, with logistic regression scoring 0.60, random forest scoring 0.80, and the ensemble technique scoring 0.86, demonstrating their usefulness in this regard. Similarly, specificity measures the models' ability to detect negative cases reliably, with logistic regression scoring 0.60, random forest scoring 0.43, and the ensemble technique scoring 0.38. The prevalence rate of 0.56 represents the total number of positive cases across all models.

Chapter 6

External validation of machine-learning algorithm predictive power across districts from Bhimber to Mardan.

6.1 Introduction

In this chapter, various machine-learning techniques for external validation were explored, focusing on the crucial process of testing models on new data. The primary objective was to enhance binary classification accuracy, a key consideration in real-world applications. A systematic procedure for evaluating the performance of machine-learning models on unfamiliar data was employed, ensuring that the models generalize effectively beyond their initial training data and accurately represent real-world efficacy. To achieve this, a diverse range of machine-learning algorithms, including decision trees, Support Vector Machines, random forests, ensemble methods (stacked meta-model), logistic regression, and bagging, were scrutinized. Emphasizing the improvement of binary classification accuracy, key performance metrics such as AUC value, sensitivity (true positive rate), specificity (true negative rate), and prevalence rate were selected and analyzed individually for each model. The ROC (Receiver Operating Characteristic) curve emerged as a dependable tool, visually representing a model's ability to distinguish between classes and facilitating an intuitive comparison for selecting the most effective model.

6.2 Methodology

6.3 Selection of study area and data source

The data for our investigation originated from two distinct locations in Pakistan. The initial dataset was acquired as part of a study conducted by (Jabeen and Malik, 2014a). This study focused on 1,584 married women from the Bhimber district, spanning three diverse localities and 24 sampling sites. This dataset served as our primary source of information on consanguineous marriages in that specific region, constituting the training set for the present investigation. To ensure the reliability and generalizability of our findings, we expanded our research to include a different dataset from the Mardan region. This new dataset, obtained from a 2017 study by (Tufail et al., 2017), is regarded as the validation set and comprises 1,204 participants. The incorporation of this additional dataset aimed to assess and validate the patterns and predictions established using the Bhimber district data. A Geographical Information System (GIS) was utilized to create a figure illustrating the geographical boundaries of this research, as depicted in Fig 6.1. The illustration below defines the study areas on a map, highlighting the specific regions where data collection took place.

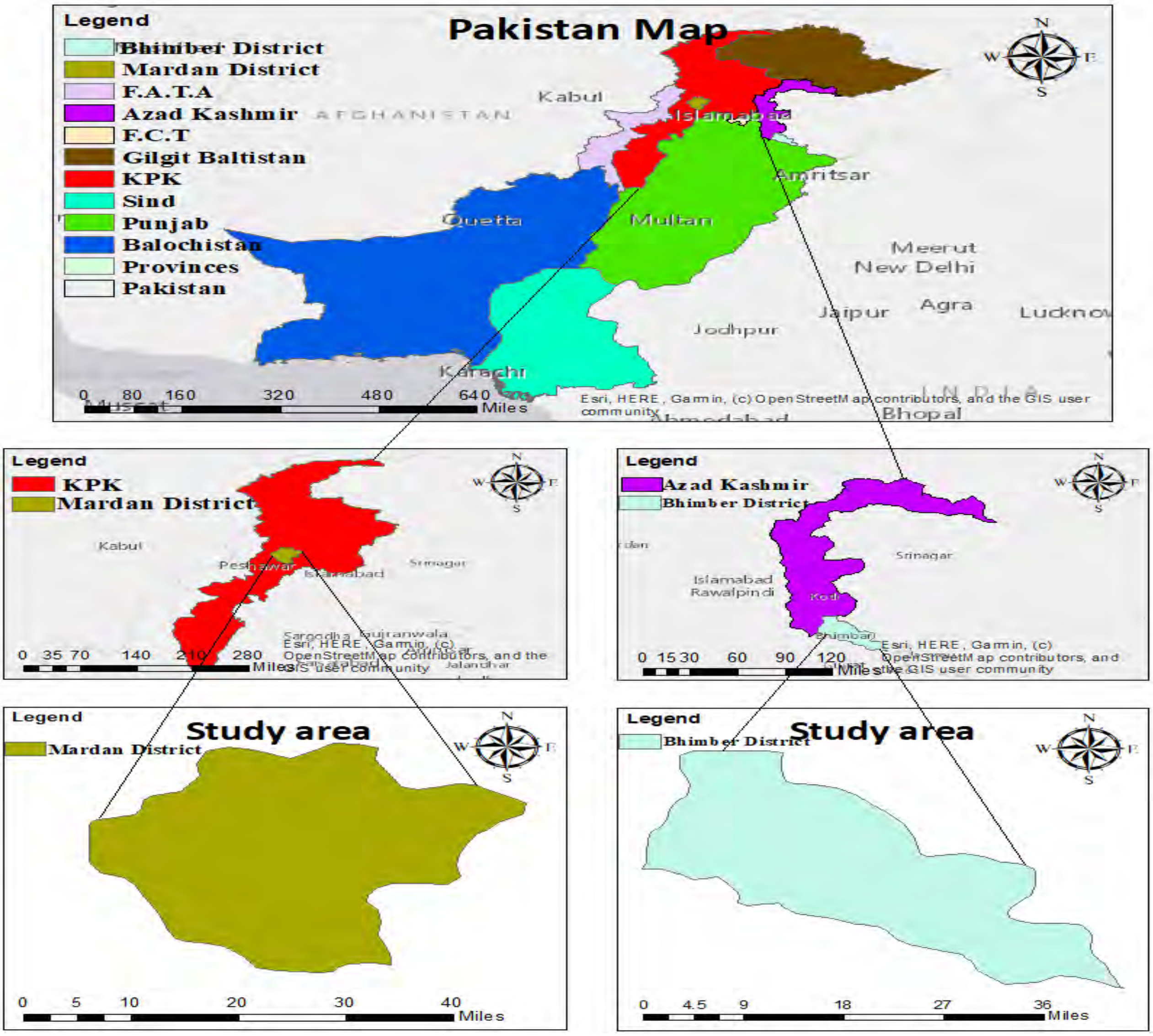


Figure 6.1: Study area map for Bhimber and Mardan district, Pakistan.

6.4 Mathematical model for consanguineous marriages

We are focusing on predicting consanguineous marriages using multiple machine-learning models in this next phase of our research. These models gain insight into data and predict consanguineous marriages. We are organizing data from two locations: one is the Bhimber district, serving as the training dataset, and the other is the Mardan district, serving as the test dataset with similar characteristics, to ensure that our predictions are robust and operate in diverse conditions. In this endeavor, diverse predictive algorithms take center stage. We carefully apply decision trees, support vector machines, random forests, logistic regression, and bagging. The essence of these models was previously discussed in chapters one and two, laying the groundwork for our current investigation. We incorporate five prediction approaches in the stacked model decision tree, random forest, bagging, SVM, and logistic regression. Each approach contributes its insights, and we combine their prediction to generate a more accurate prediction. This collaboration improves our capacity to predict consanguineous marriages by combining the qualities of each strategy.

6.5 Results and Discussion

Table 6.1: Evaluating the accuracy of Consanguineous marriage models tested for robustness on Mardan district and trained on Bhimber district, Pakistan.

Models	AUC	Sensitivity	Specificity	Prevalence rate
Random forest	0.53	0.28	0.78	0.56
Logistic regression	0.57	0.01	0.98	0.56
Decision tree	0.50	0	1	0.56
Support vector machine	0.49	0.01	0.98	0.56
Bagging	0.48	0.33	0.73	0.56
Ensemble method	0.58	0.83	0.29	0.56

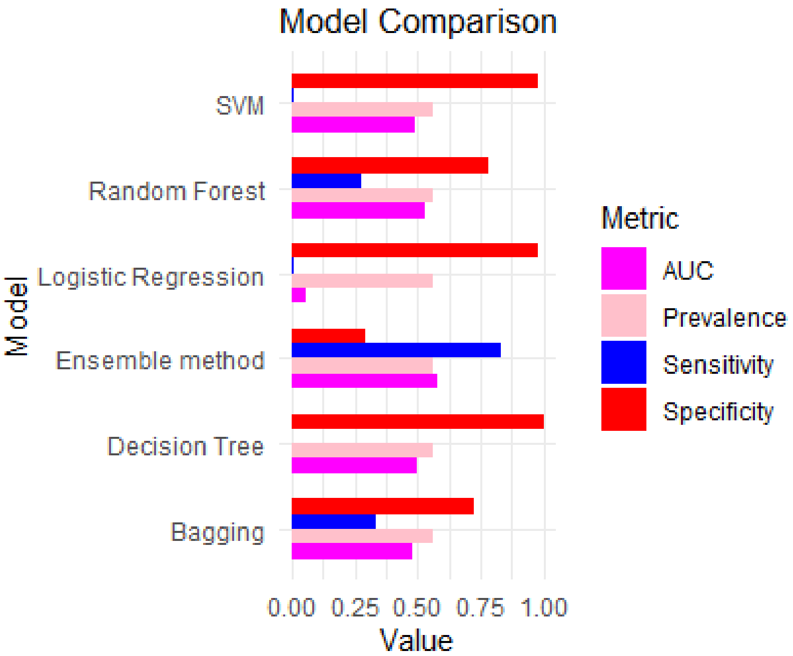


Figure 6.2: Comparison of model performance using various metrics for the external validation study.

To further assess the performance of the models and validate the results, we employed the same machine-learning models as in the previous studies for our third attempt to predict consanguineous marriage in Pakistan. For external validation, the dataset from the Bhimber

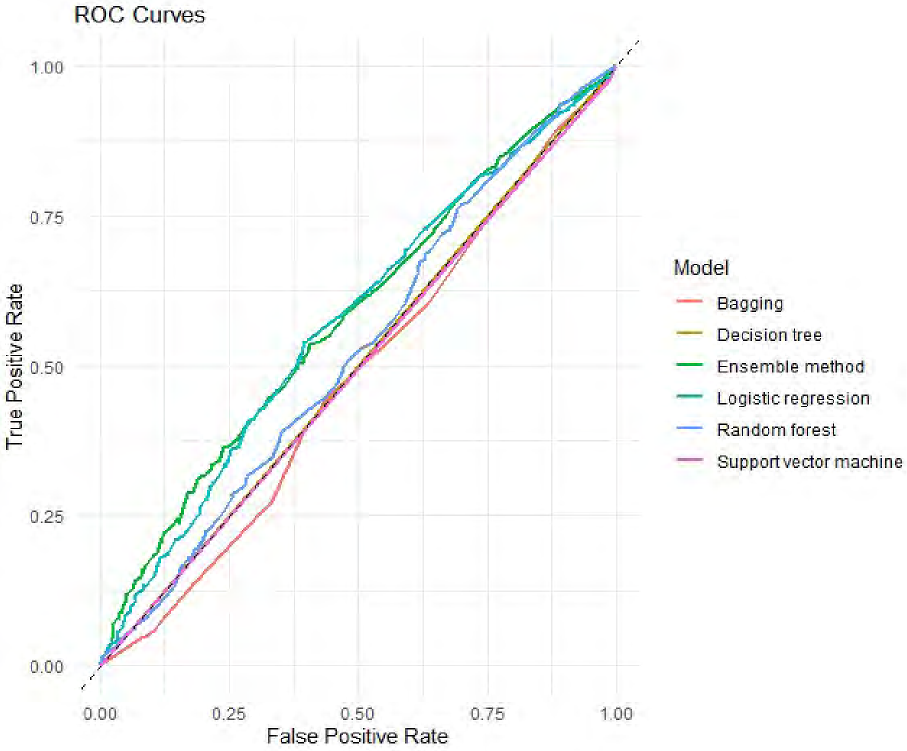


Figure 6.3: Comparing model discrimination using ROC curves on an external validation Set.

district served as the training set, and the dataset from the Mardan district served as the test set. The AUC values obtained for each model are presented in Table 6.1. The ensemble technique achieved the highest AUC value (0.58), closely followed by logistic regression (0.57), as shown in Fig 6.2. The decision tree model achieved an AUC of 0.50, while the random forest model exhibited moderate performance with an AUC of 0.53. The AUC values for the SVM and bagging models were lower, at 0.49 and 0.48, respectively. Sensitivity, specificity, and prevalence rate in Table 6.1 provide further insight into the models' ability to identify positive and negative cases, as well as the percentage of consanguineous cases in the dataset. ROC curves in Fig 6.3 visually confirm the effectiveness of the models, with the ensemble method's stacked meta-model ROC curve demonstrating a noteworthy area under the curve. Additionally, the logistic regression ROC curve exhibited strong classification capabilities, further enhancing its competitive performance. Sensitivity and specificity were evaluated to assess how well the models detect positive and negative instances, respectively. Among the models, the ensemble technique demonstrated the best sensitivity of 0.83, showcasing its strong capacity to detect positive cases. In contrast, the decision tree had the highest specificity of 1.0, indicating its superior ability to detect negative cases. The total prevalence rate of positive cases was 0.56 for all models.

Chapter 7

Discussion and conclusion

We conducted an extensive statistical investigation to ascertain the prevalence of consanguineous marriages in the Bhimber district of Kashmir, Pakistan. Employing statistical metrics, including odds ratio, chi-square, and confidence intervals, we identified significant confounders such as age, caste, mother's language, and years of marriage (refer to Table 4.1).

In our examination, age emerged as a critical determinant. Comparing women aged 18-25 to those aged 26-30, 31-40, and over 40 revealed unique odds ratios for consanguineous marriages of 0.96, 0.92, and 0.69, respectively. These findings indicate a decline in the likelihood of consanguineous marriages with increasing age, with the youngest age group having the highest odds.

Further, an unexpected revelation occurred when comparing the Choudary and Rajpot castes to the Jat reference group in terms of caste. Both Choudary and Rajpot had odds ratios of 0.67, signifying that individuals from these castes are 33% less likely to marry consanguineously than those from the Jat caste. Meanwhile, individuals from other castes exhibited an odds ratio of 0.96, indicating a slightly reduced incidence of consanguineous marriages. These findings shed light on the intricate dynamics of consanguineous marriages in the Bhimber district.

In addition, association studies on husbands (see Table 4.2) revealed statistically significant odds ratios concerning age and education. Husbands aged 18-30, compared to those aged 31-40, 40-50, and above 50, exhibited odds ratios of 0.72, 0.77, and 0.45, respectively, indicating a decrease in the frequency of consanguineous marriages with increasing age. Furthermore, husbands with metric education had an odds ratio of 1.29, while those with a graduate degree had an odds ratio of 1.05, both compared to the reference group of husbands with undergraduate education. This suggests a 29% greater risk for consanguineous marriages in individuals with less academic achievement.

The outlined approaches offer a comprehensive solution to the issue of consanguineous marriage in the Bhimber and Mardan districts. The steps included data exploration, feature importance assessment, model selection, performance evaluation, and visualization. To address the challenge of imbalanced data, we presented the distribution of the consanguineous class through bar charts (refer to Fig 4.5 and 5.4). The BORUTA technique was employed to assess the significance of factors related to consanguinity, enhancing the model's predictive capacity (see Fig 4.7 and Table 4.3).

To understand variable relationships and their importance in the prediction process, statistical techniques like chi-square, odds ratios, and confidence intervals were applied. Different machine-learning techniques, including support vector machine, decision tree, and bagging, were utilized in all chapters to predict consanguineous marriage. Notably, the bagging technique demonstrated superior performance for the Bhimber district, as shown in Table 4.4 and supported by the graphical representation of the ROC curve in Fig 4.8.

Emphasis on performance evaluation is evident in the multiple metrics used to assess the model's efficacy, encompassing AUC, sensitivity, specificity, and prevalence rate. Visualization of performance trade-offs through the ROC curve provides a comprehensive view of how sensitivity and specificity change for different classification thresholds.

Finally, to visually present the impact of significant variables on consanguineous marriage, bar charts are employed (see Fig 4.10). These graphical representations offer a clear understanding of how each variable influences the target class. Through these comprehensive steps, the proposed method aims to provide a robust predictive model for consanguineous marriage in the Bhimber district.

The Mardan dataset is segregated into training and testing sets to perform external validation of the models' generalization skills on independent sets. Various approaches, including logistic regression, random forest, and the ensemble technique, were employed. The ensemble stacked meta-model consistently outperformed other models, with logistic regression ranking second (refer to Table 5.2).

In our third attempt at external validation, both the Bhimber and Mardan districts in Pakistan were considered. The datasets were divided, with Bhimber serving as training data and Mardan as test data for external validation. Various approaches were employed to predict consanguineous marriages based on both district datasets, including decision trees, SVM, bagging, logistic regression, random forest, and the ensemble technique. The ensemble technique demonstrated superior performance among all other models (see Table 6.1), with logistic regression coming in second.

After conducting a comprehensive analysis using different machine-learning techniques for both districts in Pakistan, the outcomes indicate the ensemble method (stacked meta-model) outperforms all other models, demonstrating the best performance with greater AUC values and supported by ROC curves. Furthermore, the model maintains its superior performance in comparison to the Bhimber district when put to external validation using a Mardan dataset. A notable difference in the prevalence rate of consanguineous marriage between the Mardan district and Bhimber, Kashmir, Pakistan, with the Mardan district having a noticeably higher incidence, is revealed by the study. These results underscore the critical need to address and enhance the healthcare system, particularly in the Mardan district, where the occurrence of consanguineous marriage may pose a serious issue in the future. The ability of the bagging and ensemble approach, notably the stacked meta-model, to accurately predict the target variable also highlights its potential to inform new policies, and treatment guidelines, and improve the healthcare system as a whole to address the high occurrence of consanguineous marriage. This finding is particularly evident for the Mardan district when compared to the Bhimber district of Kashmir, Pakistan. The study's conclusion emphasizes the importance of the ensemble approach, specifically the stacked meta-model, for making precise predictions and offers insightful information about the incidence of consanguineous marriage in various areas of Pakistan. The findings demand that healthcare authorities and government officials take immediate steps to address the discrepancies found and develop focused interventions to enhance health outcomes in the impacted areas. The effective application of machine-learning techniques ultimately opens new paths for improving healthcare practices and policies, assuring better treatment for communities with a high prevalence of consanguineous marriage and improving well-being settlement.

References

Abdulrazzaq, Y., Bener, A., Al-Gazali, L. I., Al-Khayat, A., Micallef, R., and Gaber, T. (1997). A study of possible deleterious effects of consanguinity. *Clinical genetics*, 51(3):167–173.

Al Aqeel, A. I. (2007). Islamic ethical framework for research into and prevention of genetic diseases. *Nature genetics*, 39(11):1293–1298.

Alper, Ö., Erengin, H., Manguoğlu, A., Bilgen, T., Cetin, Z., Dedeoğlu, N., and Lüleci, G. (2004). Consanguineous marriages in the province of antalya, turkey. 47(2):129–138.

Anwar, W. A., Khyatti, M., and Hemminki, K. (2014). Consanguinity and genetic diseases in north africa and immigrants to europe. *The European Journal of Public Health*, 24(suppl_1):57–63.

Ayaz, A. and Saleem, S. (2010). Neonatal mortality and prevalence of practices for newborn care in a squatter settlement of karachi, pakistan: a cross-sectional study. *PLoS One*, 5(11):e13783.

Bagherzadeh-Khiabani, F., Ramezankhani, A., Azizi, F., Hadaegh, F., Steyerberg, E. W., and Khalili, D. (2016). A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results. *Journal of clinical epidemiology*, 71(3):76–85.

Barbouche, M.-R., Galal, N., Ben-Mustapha, I., Jeddane, L., Mellouli, F., Ailal, F., Bejaoui, M., Boutros, J., Marsafy, A., and Bousfiha, A. A. (2011). Primary immunodeficiencies in highly consanguineous north african populations. *Annals of the New York Academy of Sciences*, 1238(1):42–52.

Bener, A. and Mohammad, R. R. (2017). Global distribution of consanguinity and their impact on complex diseases: Genetic disorders from an endogamous population. *Egyptian Journal of Medical Human Genetics*, 18(4):315–320.

Bennett, R. L., Motulsky, A. G., Bittles, A., Hudgins, L., Uhrich, S., Doyle, D. L., Silvey, K., Scott, C. R., Cheng, E., McGillivray, B., et al. (2002). Genetic counseling and screening of consanguineous couples and their offspring: Recommendations of the national society of genetic counselors. *Journal of genetic counseling*, 11(2):97–119.

Berndt, D. J., Fisher, J. W., Rajendrababu, R. V., and Studnicki, J. (2003). Measuring healthcare inequities using the gini index. In *36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings*, pages 10–pp. IEEE.

Bibi, A., Uddin, S., Naeem, M., Syed, A., Qazi, W. U.-D., Rathore, F. A., and Malik, S. (2022). Prevalence pattern, phenotypic manifestation, and descriptive genetics of congenital limb deficiencies in pakistan. *Prosthetics and Orthotics International*, 47(5):479–485.

Bittles, A. H. (1994). The role and significance of consanguinity as a demographic variable. *Population and development review*, 20(3):561–584.

Ceriani, L. and Verme, P. (2012). The origins of the gini index: extracts from variabilità e mutabilità (1912) by corrado gini. *The Journal of Economic Inequality*, 10(3):421–443.

Charsley, K. (2005). Unhappy husbands: Masculinity and migration in transnational pakistani marriages. *Journal of the Royal Anthropological Institute*, 11(1):85–105.

De Andrade, M. R., Brandão, F., Carvalho, K. L., and Tibúrcio, J. D. (2022). Factors associated with early neonatal death. *International Journal of Development Research*, 11(12):52899–903.

Dronamraju, K. and Khan, P. M. (1963). The frequency and effects of consanguineous marriages in andhra pradesh. *Journal of Genetics*, 58(52):387–401.

Eertink, J. J., Heymans, M. W., Zwezerijnen, G. J., Zijlstra, J. M., de Vet, H. C., and Boellaard, R. (2022). External validation: a simulation study to compare cross-validation versus holdout or external testing to assess the performance of clinical prediction models using pet data from dlbcl patients. *EJNMMI research*, 12(1):1–8.

El Goundali, K., Chebabe, M., Laamiri, F. Z., and Hilali, A. (2022). The determinants of consanguineous marriages among the arab population: a systematic review. *Iranian Journal of Public Health*, 51(2):253.

Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*, 2(4):42–47.

Hamamy, H. (2012). Consanguineous marriages trends, impact on health and counseling. *Geneva (Online)*.

Haque, M. U., Waheed, M., Masud, T., Malick, W. S., Yunus, H., Rekhi, R., Oelrichs, R., and Kucheryavenko, O. (2016). The pakistan expanded program on immunization and the national immunization support project.

He, Y., Xie, R.-G., Lou, J.-W., Li, Y.-W., Wang, C.-L., Zhang, V. W., and Li, D.-Z. (2021). Exome-based preconception carrier testing for consanguineous couples in china. *Prenatal Diagnosis*, 41(11):1425–1429.

Hina, S. and Malik, S. (2015). Pattern of consanguinity and inbreeding coefficient in sargodha district, punjab, pakistan. *Journal of biosocial science*, 47(6):803–811.

Iqbal, S., Zakar, R., Fischer, F., and Zakar, M. Z. (2022). Consanguineous marriages and their association with women's reproductive health and fertility behavior in pakistan: Secondary data analysis from demographic and health surveys, 1990–2018. *BMC Women's Health*, 22(1):118.

Jabeen, N. and Malik, S. (2014a). Consanguinity and its sociodemographic differentials in bhimber district, azad jammu and kashmir, pakistan. *Journal of health, population, and nutrition*, 32(2):301.

Jabeen, N. and Malik, S. (2014b). Prevalence of congenital anomalies and non-communicable diseases in women of age 12-75 years in district bhimber, azad jammu and kashmir, pakistan. *Iranian Journal of Public Health*, 43(1):42.

Joseph, N., Pavan, K. K., Ganapathi, K., Apoorva, P., Sharma, P., and Jhamb, J. A. (2015). Health awareness and consequences of consanguineous marriages: A community-based study. *Journal of primary care & community health*, 6(2):121–127.

Khan, A., Zuhaid, M., Fayaz, M., Ali, F., Khan, A., Ullah, R., Zafar, J., Ullah, H., Baloch, S., and Gandapur, S. (2015). Frequency of congenital anomalies in newborns and its relation to maternal health in a tertiary care hospital in peshawar, pakistan. *International Journal of Medical Students*, 3(1):19–23.

Khan, M., Hussain, F., and Musharaf, S. (2014). Floristic composition and ecological characteristics of shahbaz garhi, district mardan, pakistan. *Global Journal of Science Frontier Research*, 14(1):7–17.

Khlat, M. (1988). Social correlates of consanguineous marriages in beirut: a population-based study. *Human biology*, 60 (4):541–548.

Kuntla, S., Goli, S., Sekher, T., and Doshi, R. (2013). Consanguineous marriages and their effects on pregnancy outcomes in india. *International Journal of Sociology and Social Policy*, 33(7/8):437–452.

Li, J., Zhou, Z., Dong, J., Fu, Y., Li, Y., Luan, Z., and Peng, X. (2021). Predicting breast cancer 5-year survival using machine learning: A systematic review. *PloS one*, 16(4):e0250370.

Little, J., Higgins, J. P., Ioannidis, J. P., Moher, D., Gagnon, F., Von Elm, E., Khoury, M. J., Cohen, B., Davey-Smith, G., Grimshaw, J., et al. (2009). Strengthening the reporting of genetic association studies (strega)—an extension of the strobe statement. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 33(7):581–598.

Mazharul Islam, M. (2017). Consanguineous marriage in oman: understanding the community awareness about congenital effects of and attitude towards consanguineous marriage. *Annals of human biology*, 44(3):273–286.

Metgud, C. S., Naik, V. A., and Mallapur, M. D. (2012). Consanguinity and pregnancy outcome among rural pregnant women of belgaum district. *National Journal of Community Medicine*, 3(04):681–684.

Modell, B. and Darr, A. (2002). Genetic counselling and customary consanguineous marriage. *Nature Reviews Genetics*, 3(3):225–229.

Napierala, K. and Stefanowski, J. (2016). Types of minority class examples and their influence on learning classifiers from imbalanced data. *Journal of Intelligent Information Systems*, 46(3):563–597.

Nath, A., Patil, C., and Naik, V. (2004). Prevalence of consanguineous marriages in a rural community and its effect on pregnancy outcome. *Indian Journal of Community Medicine*, 29(1):41.

Nawaz, A., Zaman, M., and Malik, S. (2021). Consanguinity, inbreeding coefficient, fertility and birth-outcome in population of okara district, pakistan. *Pakistan Journal of Medical Sciences*, 37(3):770.

Nouri, N., Nouri, N., Tirgar, S., Soleimani, E., Yazdani, V., Zahedi, F., and Larijani, B. (2017). Consanguineous marriages in the genetic counseling centers of isfahan and the ethical issues of clinical consultations. *Journal of Medical Ethics and History of Medicine*, 10(12):10–12.

Nybo Andersen, A.-M., Gundlund, A., and Villadsen, S. F. (2016). Stillbirth and congenital anomalies in migrants in europe. *Best Practice Research Clinical Obstetrics Gynaecology*, 32(4):50–59. Migration – Impact on Reproductive Health.

Obeidat, B. R., Khader, Y. S., Amarin, Z. O., Kassawneh, M., and Al Omari, M. (2010). Consanguinity and adverse pregnancy outcomes: the north of jordan experience. *Maternal and child health journal*, 14(2):283–289.

Ocko, J. K. (1991). Women, property, and law in the people's republic of china. *Marriage and inequality in Chinese society*, pages 313–46.

Oestergaard, M. Z., Inoue, M., Yoshida, S., Mahanani, W. R., Gore, F. M., Cousens, S., Lawn, J. E., Mathers, C. D., for Child Mortality Estimation, U. N. I.-A. G., and the Child

Health Epidemiology Reference Group (2011). Neonatal mortality levels for 193 countries in 2009 with trends since 1990: a systematic analysis of progress, projections, and priorities. *PLoS medicine*, 8(8):e1001080.

Rajadurai, H. and Gandhi, U. D. (2020). A stacked ensemble learning model for intrusion detection in wireless network. *Neural computing and applications*, 34(18):1–9.

Ramspek, C. L., Jager, K. J., Dekker, F. W., Zoccali, C., and van Diepen, M. (2021). External validation of prognostic models: what, why, how, when and where? *Clinical Kidney Journal*, 14(1):49–58.

Robinson, J. C. (1985). Of women and washing machines: Employment, housework, and the reproduction of motherhood in socialist china. *The China Quarterly*, 101:32–57.

Saeed, T. U., Aziz, A., Khan, T. A., and AttaUllah, H. (2012). Application of geographical information system (gis) to groundwater quality investigation: A case study of mardan district, pakistan. *International Journal of Physical Sciences*, 7(37):5421–5448.

Shafique, F., Ali, S., ul Hassan, M., and Andleeb, S. (2023). Knowledge, attitudes, and perceptions towards beta-thalassemia among residents of azad kashmir, pakistan. *Punjab University Journal of Zoology*, 38(1):65–73.

Shahzad, M. S., Akram, S. A., and Hashmi, S. B. H. (2016). Azad jammu and kashmir and gilgit baltistan: Historical, constitutional & administrative development. *Journal of Contemporary Studies*, 5(1):69–85.

Shawky, R. M., Elsayed, S. M., Zaki, M. E., El-Din, S. M. N., and Kamal, F. M. (2013). Consanguinity and its relevance to clinical genetics. *Egyptian Journal of Medical Human Genetics*, 14(2):157–164.

Tufail, M., Rehman, A. U., and Malik, S. (2017). Determinants of consanguinity and inbreeding coefficient in the multiethnic population of mardan, khyber pakhtunkhwa, pakistan. *Asian Biomedicine*, 11(6):451–460.

Urooj, Z. and Ali, A. (2016). Ladybird beetles (coleoptera: Coccinellidae) fauna of district swabi, nowshera and mardan. *Int. J. Agric. Environ. Res*, 2(1):86–92.

Warsy, A. S., Al-Jaser, M. H., Albdass, A., Al-Daihan, S., and Alanazi, M. (2014). Is consanguinity prevalence decreasing in saudis?: A study in two generations. *African health sciences*, 14(2):314–321.

Zayed, H. (2016). The arab genome: health and wealth. *Gene*, 592(2):239–243.

Zhu, Y., Gu, X., and Xu, C. (2016). A mitochondrial dna a8701g mutation associated with maternally inherited hypertension and dilated cardiomyopathy in a chinese pedigree of a consanguineous marriage. *Chinese Medical Journal*, 129(03):259–266.