Exploring the Prevalence and Triggering Factors of Migraine in University Students of Bangladesh Using Machine Learning.

by

Zawwad Bin Saif – 180021335

Nazmus Sakib – 180021313

Muhammad Adnan – 180021217

A Thesis Submitted to the Academic Faculty in Partial Fulfillment of the Requirements for the Degree of

BACHELOR OF SCIENCE IN ELECTRICAL AND ELECTRONIC ENGINEERING

Department of Electrical and Electronic Engineering

Islamic University of Technology (IUT)

The Organization of Islamic Cooperation (OIC)
Board Bazar, Gazipur-1704, Bangladesh

May 2023

CERTIFICATE OF APPROVAL

The thesis titled "Exploring the Prevalence and Triggering Factors of Migraine in University Students of Bangladesh Using Machine Learning." submitted by Zawwad Bin Saif (180021335), Nazmus Sakib (180021313), and Muhammad Adnan (180021217) has been found as satisfactory and accepted as partial fulfillment of the requirement for the degree of Bachelor of Science in Electrical and Electronic Engineering on 27th May 2023.

Approved by:

(Signature of the Supervisor)
Mirza Muntasir Nishat
Assistant Professor
Department of Electrical and Electronic Engineering
Islamic University of Technology

DECLARATION OF CANDIDATES

It is hereby declared that this thesis or any part of it has not been submitted elsewhere for the award of any degree or diploma.

(Signature of the Candidate)
Zawwad Bin Saif
Student ID: 180021335

(Signature of the Candidate)
Nazmus Sakib
Student ID: 180021313

(Signature of the Candidate)
Muhammad Adnan
Student ID: 180021217

DEDICATION

We would like to dedicate this thesis to our family members and everyone who has given us unwearied support throughout the entirety of our existence and every situation of our life. They have always been a source of motivation for us. They pushed us ahead and showed us how to make the correct decisions. They never fail to inspire us to work hard and move forward to overcome life's difficulties. They have provided us with the protection, wisdom, and fortitude we need to face difficult situations.

ACKNOWLEDGEMENTS

First, we would want to express our heartfelt gratitude to Almighty Allah, our Creator, for creating and instilling in us the intellect to educate ourselves with worldly knowledge and, therefore, complete our thesis research. ***Mr. Mirza Muntasir Nishat***, Assistant Professor, Department of EEE, IUT, is our respected supervisor. We owe him a debt of gratitude for his continuous advice, care, and support in our pursuit of a career in electrical and electronic engineering. We would not be exploring Machine Learning as well as the Biomedical Engineering area if it were not for his motivation. We are grateful for his constant mentoring and genuine efforts in our thesis research. We conducted a study and collected numerous informative analyses under his leadership to get positive results. Moreover, he gave us the most efficient technique to better understand our research. He has encouraged us to learn the fundamentals of specific fields and shown us how to proceed in the right direction

Mr. Fahim Faisal, Assistant Professor, Department of EEE, IUT, mentored us throughout the research process. He has always been encouraging and motivating us to complete our work correctly. In addition, he has motivated us to study the primary goal of our project, which has given us greater confidence in our ability to build skills while working on the thesis.

We would also like to express our gratitude to ***Prof. Dr. Mohammad Rakibul Islam***, Head, Department of EEE, and all the faculty members of the EEE Department, IUT, for their unwavering support, encouragement, and assistance.

Finally, we owe a debt of gratitude to our family for encouraging and assisting us in overcoming life's challenges, as well as enchanting us with their wonderful words. Last but not least, we would like to express our gratitude to our friends for their unconditional support and for keeping our spirits upbeat throughout this journey.

ABSTRACT

Migraine is a recurring neurovascular illness causing prolonged acute pain, nausea, vomiting, and autonomic nervous system dysfunction, resulting from disrupted blood vessels and nerve signals in the brain due to unbalanced activity, whose exact cause remains unknown influencing significantly the quality of life. The study aims to explore the prevalence of migraine among Bangladesh's university students, predict their occurrence based on triggering factors using machine learning, and raise awareness to facilitate the everyday activities of migraine patients. Around 303 students from various universities in Bangladesh participated in this cross-sectional survey. in an interval between August to October of 2022 via means of a voluntarily completed online platform-based questionnaire. For the survey structure, a total of twenty factors were sorted out after keen observation that triggers the migraine and subsequently, a dataset was structured based on the factors. The prevalence of migraine and these 20 triggering factors of migraine among university students were determined through this survey. To generate a probabilistic prediction of the occurrence of migraine, nine ML algorithms have been applied for male and female participants separately considering the headache-triggering factors. With some data preprocessing and feature engineering, GridSearchCV was used to optimize the hyperparameters for each of the nine classification models to achieve more efficient results. ML algorithms are compared by examining their several performance matrices like accuracy, train score, precision, recall, F1 score, and ROC-AUC value and after extensive simulation, the Logistic Regression algorithm emerged with the highest accuracy of 78.1% for the male participants. The stacking Classifier and Random Forest Classifier emerged with the highest accuracy of 85.3% in the case of the female participants. Making use of various machine learning algorithms and clinical data in this field has the potential to make it simpler for people with migraines to identify and avoid the triggers of their condition, allowing them to go about their daily lives more comfortably.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Migraine is a complex neurological disorder that affects a great number of people all over the world. In addition to severe and persistent headaches, patients may also experience other symptoms such as nausea, vomiting, and heightened sensitivity to light and sound. Attacks that incapacitate a person can have a significant and negative impact on their quality of life. The annual global prevalence of migraines is estimated to range anywhere from 2.6% to 21.7% [1]. It is estimated that women are two to three times more likely than men to suffer from migraine headaches [1]. Migraines can strike people of any age, and in fact, ten percent of youngsters suffer from them [2]. Although migraines have a multifaceted etiology that can be broken down into three main categories—genetics, physiology, and the environment—their origins remain a mystery. It is thought that abnormal activity in the brain is what causes the blood vessels and nerve signals to become disrupted, which in turn leads to the characteristic symptoms. Migraines have been connected to a wide number of reasons, some of which include hormonal fluctuations, emotional stress, physical exertion, dietary difficulties, environmental triggers, and even some medicines. A throbbing headache on one side, moderate to severe severity, aggravation by physical activity, and avoidance of routine physical exertion are all required for a diagnosis of migraine disease [4]. Migraine disease can be prevented by avoiding routine physical exertion. Migraine attacks can frequently last anywhere from four to seventy-two hours. People's lives are profoundly altered when they suffer from migraines. According to the Migraine Research Foundation [3,] more than four million people in the United States suffer from migraine headaches daily. In addition to the evident pain, additional symptoms may also appear, such as blurred vision, numbness, tingling, worry, melancholy, weariness, difficulty focusing, and dizziness. These symptoms may also occur in conjunction with the obvious discomfort. In light of the importance of migraine care, it should come as no surprise that people are interested in learning how to anticipate future headaches. If a person is aware of when an episode is likely to occur, they are better equipped to take precautions, steer clear of precipitating factors, and experience less discomfort. Additionally, it can assist medical professionals in customizing unique treatment plans and procedures. Machine learning has recently demonstrated a great deal of promise as a potential

tool for the creation of migraine-specific predictive models. The data on migraine triggers and patterns can be analyzed by machine learning algorithms, which can then detect patterns in the data and predict the likelihood of an attack. This approach, which has the potential to fundamentally alter the way migraines are treated, could be of assistance to millions of people. The absence of an accurate method for migraine forecasting is the impetus behind our work, which aims to fill this hole. Our objective is to employ machine learning to develop a trustworthy model that will enable us to identify the elements that play a role in the onset of migraine headaches. People will be able to take control of their health and take preventative steps as a result of this, rather than waiting for symptoms to become more severe. Our machine-learning approach will shed light on the specific characteristics of migraines that are unique to each individual by making use of feature selection, pattern recognition, and predictive modeling. It will take into account things like how frequently you experience headaches, how severe they typically are, where the pain is located, what triggers them, and how effectively pain relievers work for you specifically. Additionally, our approach will assist medical professionals in better individualizing treatment plans for migraine sufferers, which will result in improved patient outcomes. If doctors can properly predict the migraine episodes of their patients and have a stronger understanding of the migraine features of their patients, then they will be better positioned to help their patients effectively manage their condition. In a nutshell, migraines impact millions of individuals all over the world and can significantly restrict the daily lives of those who suffer from them. The development of a reliable prediction system that is based on machine learning could be of tremendous assistance in the prevention, diagnosis, and treatment of migraines. Our project makes use of data analysis and pattern recognition to provide patients and medical professionals with the skills necessary to minimize the negative effects of migraines and improve the quality of life for individuals who experience them.

CHAPTER 2

BACKGROUND

Migraine is a complicated neurovascular illness that often only affects one side of the patient's head and expresses itself as pounding headaches that come and go. Individuals frequently experience associated symptoms such as nausea and vision abnormalities [11], in addition to the incapacitating pain that the condition causes. Migraines affect more than just the person who suffers from them; they also place a significant strain on society as a whole. It results in a decline in productivity, restrictions in one's ability to participate in daily activities, and a substantial fall in one's overall quality of life [12]. Migraines are commonly misdiagnosed or treated incorrectly, which contributes to the continual difficulties that people who suffer from this ailment must deal with [13]. This is even though migraines are very common and have a significant impact. The distressing reality that headache problems are the leading cause of years lost to disability (YLDs) in young people between the ages of 15 and 49 was brought to light in research called The Global Burden of Disease. In young boys, migraines are the second leading cause of YLDs, whereas, in young girls, they are the major cause [14]. It is believed that anywhere between 10 and 18 percent of students attending universities around the world suffer from migraines [15]. Migraines can have significant repercussions for this young adult group, including higher disciplinary failures, absenteeism, comorbidity, disability, and a detrimental effect on academic performance [16-19]. Students are a useful resource for the collection of migraine data because they make up a sizable group that is also easy to reach. It has been discovered that university students who suffer from migraines have decreased academic performance and are limited in their ability to participate in daily activities [20]. The continuous demands of focus, tests, and other academic stresses can have a substantial influence on the quality of life of those individuals. People who suffer from migraines may miss more classes than their peers, which might make it more difficult for them to properly acquire new information [21]. Sensitization is an enhanced sensitivity of the brain that is known as one of the key characteristics of migraine patients. This heightened sensitivity can be caused by a variety of stimuli, both external and internal [22,23]. These precipitating events, which are also known as headache triggers, might vary from patient to patient and even from one headache attack

to the next [24]. Stress, the beginning of a woman's menstrual cycle, hunger, shifts in weather and climate, lack of sleep, intense scents, neck pain, bright lights, alcohol consumption, smoking, irregular sleep patterns, exposure to heat, particular foods, strenuous physical activity, and sexual activity are all common headache triggers that people report [25,26]. It is essential for people who suffer from migraines to gain an understanding of these triggers. Patients can lessen the frequency and intensity of their episodes by first determining what triggers them and then avoiding those triggers. In addition, medical practitioners have an important part to play in aiding patients in devising individualized plans to lessen the adverse effects of probable triggers and in teaching patients about these triggers. In light of the wide variety of factors that might set off migraines and the individual differences in how they manifest, machine learning techniques present some interesting possibilities for the diagnosis and prognosis of these headaches. Machine learning algorithms can uncover patterns and correlations between migraine triggers and subsequent episodes by utilizing techniques for pattern recognition and analysis of vast amounts of data. This information can provide people the ability to empower themselves to make informed decisions, such as adjusting their lifestyle or avoiding particular triggers, to prevent or reduce the onset of migraines. In conclusion, migraines are a condition that affects people all over the world and is characterized by a high prevalence rate as well as a significant burden. The impact is felt in academic contexts as well, with university students reporting particularly difficult circumstances as a result of their migraines. Efficient management and prevention of migraines must have a solid understanding of the triggers that lead to migraine attacks. There is a significant opportunity for improvement in the delivery of individualized insights and the empowerment of individuals to take proactive actions in the management of their disease if machine learning algorithms can be used to anticipate migraines. We can work towards improving the quality of life for people who live with migraines and alleviating the societal burden that is imposed by this neurological condition if we continue to investigate the relationships between migraine triggers and migraine episodes.

Numerous studies have been conducted to investigate the prevalence and impact of migraines among university students. These investigations have shed light on the clinical characteristics, trigger factors, and implications of this ailment. A study that was carried out in Bangladesh by A. Rafi et al. and M. A. Ra et al. utilized statistical analysis methods such as frequency distribution, chi-square tests, t-tests, and multiple logistic regression models to evaluate the prevalence of migraines among university students [27, 28]. The researchers utilized screening methods such as

ID MigraineTM and HIT-6, taking into consideration a variety of characteristics including physio-demographics, socio-demographics, lifestyle, behavioral factors, and migraine triggers. The findings of this study contribute to the growing body of knowledge regarding the epidemiology of migraines among university students and provide useful insights into the incidence of migraines in the setting of Bangladesh. In a similar vein, G. Hatem and colleagues carried out research in Lebanon to determine the prevalence of migraines among college students and to investigate the factors that were found to be connected with the condition [29]. In this particular study, the ID Migraine screening tool was utilized, and physio-demographics, socio-demographics, lifestyle factors, behavioral factors, and migraine triggers were all taken into consideration. This study contributes to our understanding of the global variance in migraine prevalence and the factors that are connected to it by investigating the specific characteristics and triggers of migraines that are experienced by the population of university students in Lebanon. Another noteworthy study was conducted by A. de Vitta and colleagues [30], who used a cross-sectional design to investigate primary headaches and the characteristics that are linked with them among university students. This study focused on demographic variables, socioeconomic features, and the usage of electronic devices, to offer insights into the probable association between these factors and the prevalence of migraines. The findings highlight how important it is to take into consideration a wider range of environmental and lifestyle factors when attempting to understand the prevalence of migraines and the triggers that cause them among university students. In China, X. Gu et al. conducted a study primarily targeting medical students to investigate the prevalence of migraines, the characteristics of migraine sufferers, and the typical triggers [31]. In this study, the ID Migraine™ screening tool was utilized, and thorough statistical analyses were conducted to provide insights into the distinctive characteristics of migraines among medical students. These findings make a significant contribution to the expanding corpus of research on the epidemiology of migraines and provide a platform upon which a better knowledge of the factors that precipitate migraines and the techniques used to treat them in this particular demographic can be achieved. A complete summary of the global prevalence of migraines and related trigger factors in university students was provided by O. Flynn et al. in the form of a systematic review and meta-analysis that was carried out by these researchers [32]. This meta-analysis provides a more comprehensive view of the global burden of migraines among university students by compiling data from a variety of research. It also assists in identifying common patterns and characteristics that are connected with the occurrence of

migraines. The research conducted by A. Rustom and colleagues [33] focused on the occurrence and implications of migraines among university students, with a particular emphasis on increasing awareness of the illness. This study quantified the impact that migraines have on day-to-day life by making use of the Migraine Disability Assessment Scale (MIDAS) score. The results of this study demonstrate the enormous burden that migraines place on students who are afflicted by them. The findings highlight the necessity for better knowledge, support, and appropriate management measures for kids who live with migraines. M. J. Marmura and colleagues reviewed various factors that can bring on migraines in patients who suffer from migraines, as well as management techniques for preventing and alleviating migraine symptoms [34]. This study offers significant insights into detecting and managing migraine triggers, which might be particularly relevant for university students seeking effective ways to prevent migraine attacks and minimize the impact they have on day-to-day life. In addition, M. K. Demirkirkan and his colleagues wanted to determine the prevalence of migraines among university students in Afyon, Turkey [35]. They also wanted to determine the extent of the handicap caused by migraines and investigate patients' treatment preferences. This study provides insights into the prevalence of migraines in a specific population of Turkish university students by using a standardized questionnaire from the International Headache Society (IHS) as well as the Migraine Disability Assessment (MIDAS) score. Additionally, this study highlights the disability and treatment preferences associated with migraines. Because of this, our study aims:

- Conducting an online survey and studying the prevalence of migraines among Bangladesh's university students;
- Analyzing migraine-triggering factors in terms of sexual dimorphism;
- Predicting migraine prevalence among university students using multiple machine learning algorithms.

CHAPTER 3

METHODOLOGY

3.1 Survey Objective:

The primary purpose of this survey is not only to increase knowledge and comprehension among college students who suffer from migraines but also to equip those students with useful tools for coping with their illnesses. The survey will be distributed to students enrolled in colleges and universities. This study intends to provide students with insights into the possibility of suffering a migraine attack on specific days by applying probabilistic predictions based on statistical analysis and machine learning techniques. These techniques are intended to provide students with information regarding the likelihood of having a migraine attack. Recognizing the difficulty of effectively forecasting migraine attacks, the survey focuses on identifying and analyzing a thorough set of 20 migraine triggers that have been related to the condition. This is done in recognition of the fact that reliable migraine prediction is difficult. These elements cover a wide range of characteristics of the lives of students, such as physiological, environmental, behavioral, and lifestyle influences. The predictive model can create personalized probabilistic predictions for the students by taking into account the performance of various triggers within a 24-48 hour timeframe before receiving input from the students. The purpose of the survey is not only to educate students about the risk of experiencing a migraine episode but also to provide them with the knowledge and tools necessary to take preventative actions in their day-to-day lives. Students may now make educated choices about their activities, such as avoiding recognized triggers, altering their schedules, regulating their stress levels, and prioritizing self-care practices when they are armed with this knowledge. Students can minimize the frequency and intensity of migraine attacks by adopting preventative techniques that are based on probabilistic predictions. This will lead to an improvement in the student's general well-being and quality of life. In addition, the results of the study highlight how essential it is for members of the university community to become more knowledgeable about migraines. The purpose of the project is to educate students as well as the larger academic environment about the influence that migraines have on students' day-to-day life. This will be accomplished by offering personalized probabilistic predictions. This

enhanced knowledge has the potential to establish a friendly and inclusive campus culture that makes accommodations for students who suffer from migraines, supports the development of appropriate support systems and resources, and promotes understanding. Although the findings and forecasts of the survey offer helpful insights, it is essential to recognize the unique characteristics of each person who suffers from migraines. As a result of the fact that migraine triggers and patterns can vary greatly between individuals, the forecasts must be interpreted as informative counsel rather than as absolute certainty. Students should be encouraged to keep a journal of their own experiences, collect data, and work together with medical specialists to hone and customize their tactics for managing migraines. In general, the results of this survey indicate a significant step towards personalized migraine care and illustrate the possibility of combining statistical analysis and machine learning to empower students on their journey toward efficiently managing migraines. The ongoing research and breakthroughs being made in this sector may have a positive impact on the lives of university students who suffer from migraines. This might enable these students to prosper academically and personally while also minimizing the disruption caused by migraine attacks.

3.2 Questionnaire Design:

For this investigation, an online questionnaire was designed to cover a complete collection of 23 characteristics. Participants were given access to the questionnaire via Google Forms for a predetermined amount of time, which ranged from the 29th of August to the 5th of October, 2022 [38]. The collection of information regarding migraine triggers and the incidence of those triggers among college students was the major purpose of the study. The poll received voluntary participation from a total of 303 students between the ages of 18 and 24, attending a variety of educational institutions located throughout the United States. The Department of Electrical and Electronic Engineering (EEE) of the Islamic University of Technology in Gazipur, Bangladesh, gave their stamp of approval to the study project from an ethical standpoint. Two questions inside the survey were explicitly designed to collect demographic information, namely the respondent's gender and age [28,31]. Another consideration was whether or not the individuals suffered from migraines as opposed to regular headaches; this was done to assist differentiate between the two groups [33]. The remaining twenty elements were devoted to the collection of information on trigger factors related to migraine headaches as well as normal headaches for all of the individuals.

To ensure an accurate and thorough investigation of the trigger variables, the researchers divided them into five primary categories: factors linked to diet, factors related to the environment and the weather, factors connected to the senses, factors related to stress, and other factors. [34] These categories were used to classify the trigger factors. Each of the primary categories was made up of several particular elements, each of which could bring on headaches [37]. In the category of dietary factors were foods such as chocolate or cocoa [28,31], sausage [37], oily foods, foods containing monosodium glutamate (e.g., fast foods, snacks, chips, seasoning blends), aspartame (e.g., yogurt, chewing gum, sugarless candy) and tyramine (e.g., smoked fish, pickled cucumber, hot dogs) [34,37]. Temperature variations [31,34], shifts in the seasons [31], and vulnerability to environmental pollution or dust allergies [34,37] were some examples of the weather and environmental conditions that were taken into consideration. High altitudes, which cause oxygen deprivation [37], exposure to sunlight or intense artificial light, and sensitivity to odors (osmophobia) [34,37] were some of the sensory elements that were identified as triggers for the condition. Anxiety, confrontations, or arguments [28,33], exhaustion or tiredness, and thoughts of despair or mental breakdown [37] were some of the stress-related elements that were identified as potential triggers. A variety of conditions, including disruptions in sleep cycles [28,31], travel or trips [34], noise pollution [28], fasting [31,34], and anemia owing to the menstrual cycle [28,34] were identified as potential causes. Based on the participant's responses to these 20 headache-triggering circumstances, statistical analysis, and data evaluation were carried out. The female participants took into consideration all 20 elements, while the male participants took into consideration 19 aspects. Every participant chose either "YES" or "NO" as their response to each of the factors [25]. The purpose of the study was to provide significant insights into the prevalence and impact of various triggers among university students by collecting data on triggering factors like the ones discussed above. This research has the potential to contribute to a better knowledge of the precise triggers associated with migraines and typical headaches, hence enabling the creation of personalized preventive interventions, lifestyle modifications, and awareness campaigns to support students in properly managing their headaches. The results of this study can provide information that is useful for healthcare professionals, educators, and people who have migraines or headaches in general. This means that the findings may have ramifications that extend beyond the realm of the institution. Individuals who are empowered to make informed decisions about their daily routines, regulate their environments, and implement lifestyle modifications that may

lessen the frequency of headache episodes and the severity of those episodes might benefit from an understanding of the prevalence of specific triggers and the impact those triggers have. Additionally, the study serves as a basis for future research endeavors, enabling additional inquiries into the complex nature of migraines and the creation of personalized interventions for individuals who are affected by this ailment.

3.3 Data Collection:

The study involved a total of 303 students (N) who came from a wide variety of educational institutions located throughout the United States. The ages of the people that took part ranged anywhere from 18 to 24 years old. Notably, the research found that male students had a greater participation rate than female students. There were a total of 215 participants, which accounted for 70.96% of the sample. On the other hand, 88 female students took part in the study, making up 29.04% of the total sample. The inclusion of a participant group that is diverse concerning gender and age contributes to the robustness of the study and enables a more in-depth understanding of the occurrence of migraine triggers among university students as well as the impact that they have. The research intends to capture any potential gender-based differences in the triggering variables and their connections with migraines and normal headaches. To do this, a considerable number of participants from both sexes will be included in the study. The gender breakdown of the sample demonstrates how crucial it is, when analyzing the findings, to take into account any relevant gender-related issues. It's possible that hormonal, physiological, or behavioral differences are to blame for the differences in migraine prevalence and triggers that men and women experience. As a result, the researchers conducting this study understand the significance of doing independent analyses and interpretations of the data for the male and female participants to identify any potential gender-related trends or differences. A gender-based analysis of the data can provide useful insights into the distinctive characteristics of migraines that are experienced by both male and female college students. Using this knowledge to influence future research, treatment procedures, and awareness campaigns that target specific gender-related triggers and management measures is a possibility. The purpose of this research is to contribute to a more nuanced knowledge of migraines and to enhance the efficacy of migraine management options for university students of both genders. This will be accomplished by taking into account gender-specific issues such as those listed above.

TABLE I. Prevalence of triggering factors

Triggering Factors	Normal Headache				Migraine Headache			
	Female		Male		Female		Male	
	N	%	N	%	N	%	N	%
Anemia	3	10.714	Null	Null	23	38.333	Null	Null
Fasting	10	35.714	19	17.273	42	70	48	45.714
Chocolate/cocoa	0	0	1	0.909	8	13.333	16	15.238
Sausage	0	0	0	0	3	5	8	7.619
Oily Food	0	0	5	4.545	16	26.667	38	36.190
MSG	1	3.571	0	0	8	13.333	18	17.143
Aspartame	1	3.517	5	4.545	10	16.667	15	14.286
Tyramine	0	0	5	4.545	6	10	17	16.190
Temperature Fluctuations	10	35.714	48	43.636	55	91.667	93	88.571
Season Change	8	28.571	42	38.182	53	88.333	86	81.905
Environmental Pollution	6	21.429	38	34.525	47	78.333	80	76.190
High Altitude	4	14.286	19	17.273	32	53.333	48	45.714
Sunlight/Artificial Bright Light	20	71.429	55	50	56	93.333	88	83.810
Osmophobia	4	14.286	23	20.909	34	56.667	44	41.905
Anxiety	15	53.571	53	48.182	56	93.333	87	82.857
Fatigue	13	46.429	52	47.273	53	88.333	80	76.190
Depression	18	64.286	40	36.364	56	93.333	77	73.333
Sleep	19	67.857	81	73.636	56	93.333	93	88.571
Travels/Trips	7	25	22	20	49	81.667	61	58.095
Noise pollution	16	57.143	64	58.182	56	93.333	89	84.762

To gain insight into the prevalence of migraines and normal headaches among 303 students, a statistical analysis was performed on a full collection of individual samples. The results of this study are shown in TABLE I. According to the findings of the study, 165 students had a diagnosis of migraines (54.455%), whereas 138 students had a diagnosis of normal headaches (45.544%). It was discovered that females were more likely to suffer from migraines than males were, in contrast to typical headaches, which were found to be more common in males. Sixty of the 165 students who were diagnosed with migraines were female (68.182%), while 105 of the students were male (48.837%). On the other hand, out of the 138 students who reported normal headaches, 28 (31.818%) were female whereas 110 (51.163%) had normal headaches. According to these data, there is a greater propensity for women to suffer from migraines than there is for men. Further investigation of TABLE I enables us to investigate the factors that are related to migraines and normal headaches among the student populations of both male and female students. The vast majority of female students who have been diagnosed with migraines have stated that more than one cause is to blame for their condition. To be more specific, 56 students (93.33%) named sunlight as a trigger, which suggests that being exposed to strong light makes their migraines worse. Additionally, the same number of female students (56, or 93.33%) related their migraines to anxiety and depression, which suggests a correlation between emotional well-being and the occurrence of migraines. Sleep disturbances and exposure to excessive noise were indicated as factors that lead to the migraines of 56 female students, which indicates that disruptions in sleep patterns and exposure to excessive noise play a substantial role in playing role in triggering migraines among females. Temperature swings and disruptions to the normal sleep-wake cycle were shown to be the most significant risk factors for migraines in male students. The fact that 93 out of 105 male students with migraines (88.571%) ascribed their headaches to variations in temperature demonstrates that migraine attacks can be brought on by shifts in temperature in males. In addition, 88.571% of the male students surveyed said that disruptions in their normal sleep cycle were a factor in the development of their migraines. These findings underline the necessity of keeping a regular sleep cycle as a means of treating migraines in males and highlight the influence of sensory stimuli, such as variations in temperature. A gender-specific pattern appears when looking at the kids who have been diagnosed with normal headaches. Twenty of the female students (71.429%) reported that sunshine was a trigger for their normal headaches. This

was the most common cause given. This provides support for the hypothesis that females are more susceptible to headaches when exposed to intense light. On the other hand, among male students who have normal headaches, 81 (73.636%) relate their condition to disruptions in the regular sleep cycle. This highlights the necessity of maintaining consistent sleep patterns when it comes to the management of normal headaches among males. According to the findings, females have a higher migraine prevalence than males. Migraine headaches are more prevalent in women than men, particularly when hormonal fluctuations are involved. Noise pollution, stress, depression, a lack of natural light, and trouble sleeping are all factors in this category. However, temperature swings and shifts in sleep patterns have a greater negative impact on men. Similarly, women are more prone to blame too much sun exposure for their headaches than men, while men are more likely to blame disruptions in their sleep. These findings improve our understanding of the frequency and characteristics shared by migraines and everyday headaches. Based on a person's gender and individual stressors, this data can be used to create individualized treatment plans.

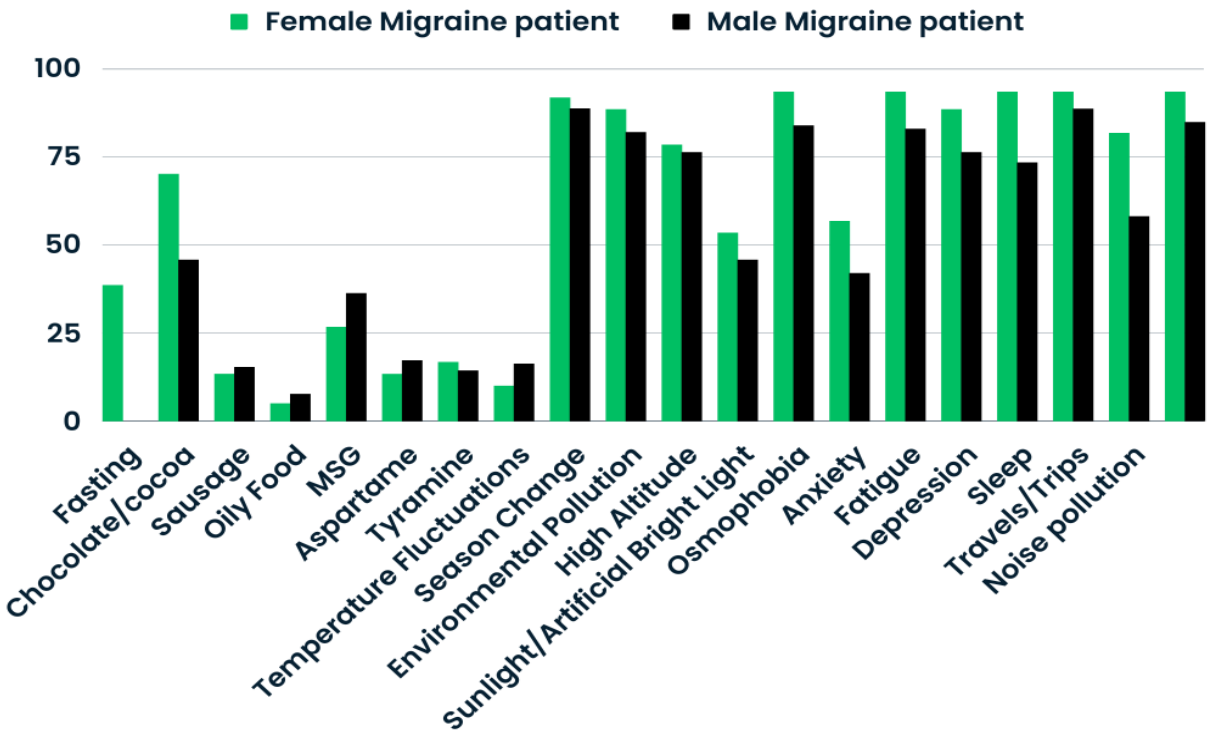

Figure 3.1: Prevalence of triggering factors among migraine patients.

By analyzing the data presented in Figure 1, we can acquire a deeper understanding of the unique causes of headaches that are experienced by women who have been diagnosed with migraines as opposed to those who have typical headaches. The majority of headaches experienced by women who have typical headaches are brought on by exposure to strong light, whether natural or artificial. This shows that exposure to bright light sources, whether natural or artificial, can lead to the beginning of headaches in this population. It's possible that eye strain or other physiological responses are to blame for the light sensitivity experienced by women who otherwise have normal headaches. But there are also other possible explanations. In contrast, the migraine triggers that affect females who have been diagnosed with the condition go beyond sensitivity to light. Migraines in females are frequently brought on by feelings of worry and depression, in addition to exposure to bright light, whether natural or artificial. In this particular demographic, the onset of migraines can be significantly influenced by a variety of emotional factors, such as stress, worry, or mood changes. In addition, problems sleeping and exposure to loud noises have been identified as additional migraine causes for females. Both disruptions in sleep patterns, such as not getting enough sleep or having poor quality sleep, and exposure to excessive noise have been linked to the development of migraines. When we turn our attention to men, we find that the majority of headaches, even mild ones, are brought on by disruptions in their normal sleep patterns. Headaches can be brought on by shifts or disruptions in a person's normal sleep pattern, such as not getting enough sleep or keeping an unpredictable sleep schedule. A man's risk of developing headaches can be directly impacted by disruptions in his normal sleep cycle, given that sleep is an essential component in the process of preserving his general health and well-being. Alterations in temperature are one of the migraines triggers that men who have been diagnosed with the condition may be susceptible to, in addition to disruptions in their sleep cycles. Alterations in temperature, such as being subjected to surroundings of severe heat or cold, might act as additional triggers for migraines in males. The onset of migraines in this population may be precipitated by sensitivity to temperature or by disruptions in the processes that control thermoregulation. These gender-specific distinctions in headache triggers underline how important it is to personalize treatment and management techniques based on the individual's gender as well as the unique triggers that they experience. Both common headaches and migraines can be alleviated to some degree in women by utilizing techniques that lessen their sensitivity to light, learn to better control their stress, and develop more restful sleeping patterns. In addition, it is important to address emotional well-being

and to manage environmental issues like noise pollution when attempting to treat migraines in females. On the other hand, for men, focusing on maintaining regular sleep patterns, treating sleep difficulties, and managing temperature swings can all contribute to the effective management of both typical headaches and migraines. It is crucial to highlight that the data reported here are based on the study of Figure 1, which gives visual representations of the unique headache triggers that are found within each gender and headache type. These discoveries broaden our comprehension of the intricate dynamic at play between the sexes, the precipitating factors, and the occurrence of headaches. As a result, we can design headache treatments that are more individualized and attentive to the needs of both sexes.

3.4 Data Preprocessing and Feature Engineering:

Using the Jupyter Notebook provided by the Python Navigator platform, several preprocessing processes were carried out to prepare the dataset for the training of various algorithms on it. These steps were taken to prepare the data for optimal analysis and training of the model. The first phase involved distinguishing between the data that was input and the data that was output. Input data consisted of the history of performance on 20 trigger variables for migraine patients, as well as demographic information and their experience concerning migraines. Additionally, the data included their experience. The output data showed the likelihood of experiencing a migraine attack at some point in the future. The dataset was organized into different input and output variables thanks to the separation of these components, which made future analysis much simpler. Some columns were eliminated so that the dataset may be improved in terms of both its quality and its usefulness. This stage was carried out to improve the values of various metrics and to refine the dataset as a whole. Eliminating columns that are superfluous or duplicative is an effective way to streamline the data and direct attention to the most vital aspects. Using label encoding, the alphabetical data, more especially the responses of "YES" and "NO," were then turned into numerical values. The method of label encoding involves giving categorical variables their distinct numeric labels to render them understandable by computers. The data can then be used by machine learning algorithms, which only accept quantitative inputs after the categorical replies have been converted into their numerical equivalents. Following this stage, you can rest assured that the classifiers will be able to read and analyze the data accurately. Following the completion of the data preprocessing step, feature engineering strategies were implemented to further improve the

dataset before the training of the classification machine learning models. In the beginning, the demographic information, particularly the age element, was left out of the equation. Excluding this information helped to streamline the dataset and minimize potential biases because the age range was generally comparable, and the input type was unsuitable for the classification models. Additionally, the gender section was left out of the Jupyter Notebook. Instead, the index selection method was used in the data table to offer gender-related inputs. Because only females experience menstruation, it was important to handle null values in the "Anaemia" component while dealing with male inputs. This approach was required to accomplish this task. The dataset was modified to exclude null values in the "Anaemia" component and assure its consistency. This was accomplished by deleting the gender section and integrating gender-related inputs via the index selection approach. In addition, the "experience" component was changed into the likelihood of experiencing migraines, which was the output variable for this study. Because of this change, the machine learning models were able to discern between the triggers that cause migraines and the triggers that cause typical headaches based on the patients' individual experiences. The output variable was produced by taking into account the 20 different kinds of migraine triggers and the effect that each of those triggers has on the occurrence of migraines. The dataset was divided into a training set and a test set with a 7:3 ratio to avoid the problem of overfitting. By dividing the data in this way, we were able to ensure that the models were trained on a specific section of the data and evaluated using the remaining data. The effectiveness of the models and their ability to generalize may be properly evaluated by testing them on data that they had not before encountered. The dataset was suitably prepared for the training of the classification machine learning models by conducting these preprocessing processes and feature engineering techniques. These methods ensured the quality of the data, made it compatible with algorithms, and made it suitable for accurate predictions of the likelihood of a migraine occurring based on the identified trigger factors.

3.5 Hyperparameter Optimization:

Instead of utilizing RandomizedSearchCV as the method for hyperparameter optimization in all nine classification models, the GridSearchCV strategy was chosen because it yields superior results. In contrast to the parameters that are automatically learned by the models, the values for the hyperparameters need to be manually modified for each dataset individually. GridSearchCV is a methodical strategy that methodically investigates a wide variety of parameter tunings and then cross-validates the results to determine which settings produce the best results. The goal of this exercise is to identify the hyperparameters that will lead to the highest level of performance for the model. GridSearchCV performs a comprehensive search over a collection of hyperparameters that have been provided, taking into account every conceivable combination inside a grid-like framework. By using such an exhaustive search strategy, one may be certain that the optimal mix of hyperparameters will be located. GridSearchCV can determine the hyperparameters that result in the best possible model performance by doing a cross-validation test on each possible combination of inputs. On the other hand, RandomizedSearchCV takes a distinctively different method. It does a random search over the distributions that have already been established for each hyperparameter. The RandomizedSearchCV algorithm makes arbitrary choices for the combinations of hyperparameters to use, and then trains the model using those specific values. The search procedure is carried out several times, and after each iteration, the performance of the model is evaluated to determine which collection of hyperparameters produces the best results. GridSearchCV investigates every conceivable combination of random hyperparameters, in contrast to RandomizedSearchCV, which only takes into account a predetermined set of these possibilities. Because of this crucial difference, GridSearchCV is certain to locate the hyperparameter combination that produces the best results for the given dataset, whereas RandomizedSearchCV might not always be able to achieve the same level of optimization. GridSearchCV was selected as the best available choice after taking into consideration the classification dataset that was discussed before. The dataset had a sufficient amount of computer resources and time available for the comprehensive search. Due to the nature of GridSearchCV, which is thorough, it guarantees that no potentially ideal hyperparameter combination will be overlooked. GridSearchCV provides confidence in discovering the greatest potential hyperparameters for maximizing model performance since it searches the whole hyperparameter space thoroughly and completely. On the other hand, it is essential to point out that utilizing

RandomizedSearchCV would have been a more appropriate choice if the amount of time, as well as the available computational resources, were restricted. The random sampling of hyperparameter combinations that RandomizedSearchCV performs is less computationally intensive and has the potential to deliver good results in situations when a full search is not practicable. In conclusion, GridSearchCV is the method of choice for datasets that have plenty of computational resources and time, as this method ensures the discovery of the optimal hyperparameter combination. On the other hand, RandomizedSearchCV is a more effective alternative when computer resources and time are limited. This is because it examines a subset of hyperparameter combinations to approximate the optimal solution, rather than exploring all of the possible combinations.

3.6 Classification Models:

The classification models subcategory of machine learning algorithms is an essential part of the field and is used to make predictions about categorized outcomes. They examine the incoming data and assign each instance to a certain class according to the characteristics and patterns of the data. In the field of machine learning, one of the key goals of classification models is to accurately place previously unseen data into the right category or class based on the properties of that data. There have been many different classification algorithms developed, and each one has its own set of advantages and disadvantages. The nature of the problem and the qualities of the data should both be taken into consideration when selecting a categorization model to use. The following is a list of well-known classification algorithms that are frequently employed, along with the suitable features of each:

3.6.1 Random Forest Classifier (RF):

Random forest classifier is treated also as supervised learning. It is architected in such a way that it produces a predicted outcome class that depends on the majority votes of decision trees and conversely, mean value is obtained during random forest regression. A particular decision tree has three sections- the leaf node, the decision node, and the root node. The whole dataset is driven through the root node and subsequently, the decision node acting as a linking bridge takes the required decision and the dataset is chopped into several subsets. The leaf node indicates the ending point of passed datasets and prediction is done at this stage. The accuracy of the ultimate predicted outcome can be elevated based on the dataset's number. Here are some key points about the model:

Supervised Learning and Ensemble Method: The Random Forest classifier is a supervised learning algorithm and is considered an ensemble method. It combines the predictions of multiple decision trees to make the final prediction. Each decision tree is trained on a subset of the data and contributes to the overall prediction through a voting mechanism.

Decision Tree Structure: A decision tree consists of three main components: the root node, decision nodes (also known as internal nodes), and leaf nodes (also known as terminal nodes). The root node represents the starting point of the decision tree, while the decision nodes make decisions based on the input features. The leaf nodes represent the outcomes or predictions.

Root Node and Decision Nodes: The entire dataset is initially passed through the root node, which makes a decision based on a specific feature or attribute. The decision node acts as a linking bridge, branching the data into different subsets based on the decision made at each node. This process continues until the data reaches the leaf nodes.

Leaf Nodes and Predictions: The leaf nodes indicate the ending points of the data subsets and are responsible for making the final predictions. Each leaf node represents a specific class label or a numerical value, depending on the task (classification or regression). In a classification problem, the majority vote of the decision trees determines the predicted outcome class. In a regression problem, the mean value of the predictions from different decision trees is taken as the final prediction.

Accuracy and Number of Decision Trees: The accuracy of the predicted outcome in a Random Forest classifier can be enhanced by increasing the number of decision trees in the ensemble. The diversity and number of trees help in capturing different aspects of the data, reducing overfitting, and improving overall prediction performance.

Random Forest classifiers are known for their robustness, scalability, and ability to handle high-dimensional data. They are widely used in various applications, including classification, regression, feature selection, and anomaly detection.

3.6.2 Decision Tree Classifier (DT):

The decision nodes are responsible for making decisions based on specific features or attributes of the input data. These nodes have multiple branches leading to different child nodes, each representing a possible outcome or value of the decision. The decision node poses a question or tests a condition on a particular attribute, and the outcome of that test determines the subsequent path to follow in the tree. On the other hand, leaf nodes represent the outcomes or predictions of the decision tree. They do not have any branches and signify the end of a particular path in the tree. The leaf nodes provide the predicted class labels or values for the input data based on the decisions made at the decision nodes. The main purpose of a decision tree is to partition the data into subsets or subtrees based on the responses to the questions posed at the decision nodes. By recursively partitioning the data, the decision tree can capture complex decision rules and relationships between the input features and the target variable. During the prediction phase, when a new record or instance is given to the decision tree, it starts at the root node and compares the attribute value of the record with the attribute condition of the root node. Based on the comparison, it follows the corresponding branch to traverse the tree until it reaches a leaf node. The class label or value associated with that leaf node is then assigned as the prediction for the record. Decision trees are useful for generating interpretable models and making accurate predictions based on learned decision rules from the training data. In addition, here are some further points about decision trees: Splitting Criteria: When constructing a decision tree, the choice of attribute and the splitting criteria at each decision node is crucial. Common splitting criteria include Gini impurity and information gain. These criteria help determine the attribute that best separates the data and maximizes the homogeneity within each subset.

Recursive Process: The construction of a decision tree is a recursive process. Once a decision node splits the data into subsets, the same splitting process is applied to each subset, creating further decision nodes and leaf nodes. This recursive process continues until a stopping criterion is met, such as reaching a maximum depth or a minimum number of samples required to create a leaf node.

Overfitting and Pruning: Decision trees tend to overfit the training data, meaning they can memorize the training examples and perform poorly on new, unseen data. To mitigate overfitting, techniques such as pruning can be employed. Pruning involves removing decision nodes or

collapsing branches that do not contribute significantly to improving the tree's predictive accuracy on unseen data.

Handling Categorical and Numerical Features: Decision trees can handle both categorical and numerical features. For categorical features, the decision node compares if the attribute value matches a specific category. For numerical features, different strategies can be used, such as binary splitting (e.g., is the value greater than a threshold?) or multiway splitting (e.g., is the value within a specific range?).

Ensemble Methods: Decision trees can be combined using ensemble methods to improve their predictive performance. Random Forests and Gradient Boosting are popular ensemble methods that utilize multiple decision trees to make predictions. These techniques help reduce overfitting and enhance the overall accuracy and robustness of the model.

3.6.3 Support Vector Machine (SVM):

SVM is a supervised learning technique that arranges data into categories and uses machine learning to analyze data for classification. SVM attempts to develop a computationally efficient learning method by dividing hyperplanes in high-dimensional attribute space. This algorithm, built on the statistical learning framework, is one of the most powerful ones available. it may be used to solve both regression and classification issues. Because of its ability to use the kernel method, SVM is capable of classifying both linear and non-linear datasets. The kernel function changes the form of the input as per the requirement. Regarding the SVM algorithm, the kernel trick addresses nonlinearity and greater dimensions. The SVM performs well because there is a clear division called margin between the different classes, and it is also more efficient in high-dimensional spaces. An input space with low dimensions is stretched by the kernel to produce an output space with a higher dimension. In plainer language, when the kernel is applied to a problem, it increases the number of dimensions, turning it from a non-separable to a separable one. It increases the efficiency, adaptability, and precision of SVM. Here are some further details:

Supervised Learning and Classification: SVM is a supervised learning algorithm primarily used for classification tasks. It analyzes labeled data and aims to separate the data into different categories or classes based on their attributes or features.

Hyperplane and High-dimensional Space: SVM works by finding an optimal hyperplane that divides the data into different classes. The hyperplane is a decision boundary that maximizes the margin, which is the distance between the hyperplane and the closest data points from each class. SVM seeks to find the best hyperplane that achieves the largest margin, which leads to better generalization and robustness.

Statistical Learning and Efficiency: SVM is built upon the statistical learning framework and is considered one of the most powerful algorithms available. It aims to find a computationally efficient learning method by utilizing optimization techniques and linear algebra operations.

Kernel Trick and Non-linearity: One of the key strengths of SVM is its ability to handle non-linear datasets. This is achieved through the kernel trick. The kernel function allows SVM to transform the input space into a higher-dimensional feature space, where the data becomes separable. The kernel function effectively computes the dot product between two points in the higher-dimensional space without explicitly calculating the transformation.

Regression and Classification: While SVM is primarily known for classification tasks, it can also be used for regression. In regression, SVM estimates a continuous target variable instead of discrete class labels by finding a hyperplane that best fits the data while minimizing the errors.

Margin and Efficiency in High-dimensional Spaces: SVM performs well in high-dimensional spaces, especially when there is a clear separation or margin between different classes. It is particularly efficient when the number of features or attributes is large, as the kernel trick allows SVM to operate in the transformed feature space without explicitly mapping each data point.

Kernel Flexibility and Performance: By applying different kernel functions, such as linear, polynomial, radial basis function (RBF), or sigmoid kernels, SVM can handle a wide range of data distributions and capture complex relationships between attributes. The choice of the kernel depends on the specific problem and the nature of the data.

SVM is a widely used algorithm in machine learning due to its ability to handle non-linear data, its effectiveness in high-dimensional spaces, and its strong theoretical foundation. It offers versatility, adaptability, and precision in solving classification and regression problems.

3.6.4 Logistic Regression (LR):

Regression is categorized under supervised learning. Regression is a process to predict continuous outcomes, normally a real number, based on the correlation between the independent variable and the dependent one. Logistic regression follows the tack to auspicate a dependent variable or outcome, given a set of independent variables or inputs. The outcome has two polarities, either binary 1 or 0. It finds the probable output depending on the threshold value. For example, above the edge point, it'll give binary 1(P=1) or yes, and for the reverse situation, the output is no or 0(P=0). The output(Y) is a function of the input variable(X). The domain for this method is bounded to [0,1]. The curve line indicates the probability region for several variable inputs. This algorithm goes along with the sigmoid function(z) which results in between 0 and 1. This algorithm does well because it is simple to implement and it trains quickly and effectively, even when faced with nonlinearity. The optimal curve for a dataset can be demonstrated by combining a linear regression line and a sigmoid function. The equation regarding the linear regression algorithm is,

$$log(\frac{Y}{1-Y}) = B + C1X1 + C2X2$$

Where,

Y is the predicted outcome.

X1, X2 is the set of inputs.

B is the constant term and output when no factors have an effect.

The following points contain the key characteristics of the model:

Supervised Learning and Continuous Outcomes: Regression is a supervised learning technique that aims to predict continuous outcomes. It analyzes the relationship between the independent variables (also known as predictors or inputs) and the dependent variable (the outcome) to make predictions. An outcome variable is typically a real number, and regression models estimate its value based on the correlation with the independent variables.

Logistic Regression for Binary Classification: Logistic regression is a specific form of regression used for binary classification tasks. It predicts the probability of an event or the likelihood of an outcome falling into one of two categories (e.g., 1 or 0, yes or no). Logistic regression is suitable when the dependent variable is binary, and the goal is to determine the probability of a specific class label.

Threshold and Binary Output: Logistic regression uses a threshold value (usually 0.5) to classify the predicted probabilities into binary outcomes. If the predicted probability is above the threshold, the predicted class is assigned as 1 (or yes), and if it's below the threshold, the predicted class is assigned as 0 (or no).

Sigmoid Function and Probability: Logistic regression models the relationship between the independent variables and the probability of the binary outcome using the sigmoid function (also known as the logistic function). The sigmoid function maps any real-valued number to a value between 0 and 1, creating an S-shaped curve. It transforms the linear combination of the input variables (known as the log-odds or logits) into a probability value.

Range and Interpretation: The predicted probability from logistic regression lies within the range of [0,1], representing the probability of belonging to a certain class. This probability can be interpreted as the likelihood or confidence of the predicted outcome.

Advantages of Logistic Regression: Logistic regression has several advantages. It is relatively simple to implement, computationally efficient and can handle large datasets. It also performs well even when faced with non-linear relationships between the independent variables and the outcome.

Combining Linear Regression and Sigmoid Function: The logistic regression model can be visualized as a combination of a linear regression line and a sigmoid function. The linear regression line captures the linear relationship between the independent variables and the log-odds, while the sigmoid function maps the log-odds to a probability value between 0 and 1.

Logistic regression is widely used in various fields, such as healthcare, finance, and social sciences, for binary classification tasks. Its simplicity, interpretability, and ability to handle non-linear relationships make it a popular choice for predictive modeling.

3.6.5 K- Nearest Neighbors (KNN):

K-nearest-neighbor classification, also known as KNN classification, is one of the most basic and simplest classification methods. It is a non-parametric classifier that lies under supervised learning. KNN predicts the class of new data points through the feature similarity method. The new data points are classified on how closely they are connected to the training dataset. The point is then put into the category of the class that has the most characteristics in common with its K closest neighbors. The Euclidean distance method is the most common procedure to sort out the class. The distance between the new point and the recorded data points needs to be measured. The distance matrices form the region to decide the point's class. The value of k in this algorithm indicates the number of neighboring points that will be checked to determine the class of the unknown point. It's better to choose an odd integer for k which results in untied voting. The euclidean distance is written as,

$$Euclidean\ Distance = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Here are some further details:

Supervised Learning and Non-parametric Classifier: KNN classification is a supervised learning algorithm, meaning it requires labeled training data to make predictions. It is a non-parametric classifier because it does not make any assumptions about the underlying data distribution or learn explicit parameters during training.

Feature Similarity and Neighbor Voting: KNN predicts the class of a new data point by measuring the similarity or distance between the features of the new point and the training data. The new point is classified based on the classes of its K closest neighbors in the feature space. The class that occurs most frequently among the K neighbors is assigned as the predicted class for the new point.

Euclidean Distance and Distance Metrics: The most common distance metric used in KNN is the Euclidean distance. It calculates the straight-line distance between two points in the feature space. However, other distance metrics like Manhattan distance, Minkowski distance, or cosine similarity

can also be used depending on the nature of the data and the problem at hand. The choice of distance metric affects the notion of proximity between data points.

Determining K and Tied Voting: The value of K in KNN determines the number of neighboring points considered when making a prediction. It is important to choose an appropriate value for K. If K is too small, the model may be sensitive to noise and outliers. If K is too large, the model may oversimplify the decision boundary and lose local patterns. It is often recommended to choose an odd value for K to avoid tied voting, where multiple classes have an equal number of votes among the neighbors.

KNN is known for its simplicity and ease of implementation. It can be effective for both binary and multi-class classification problems. However, KNN can be computationally expensive when dealing with large datasets, as it requires calculating distances between the new point and all training points.

3.6.6 XGBoost Classifier (XGB):

Extreme Gradient Boosting (XGBoost) is an extremely effective decision tree-based ensemble machine learning classifier. Because of its high processing speed, scalability, and improved performance, this approach employs a gradient-boosting framework. Gradient-boosted decision trees build a powerful learner by using a series of decision trees, each of which learns from the one before it and influences the one after it. In XGBoost, weak classifiers are combined to produce a powerful one. The feedback from previously approved decision trees is incorporated into XGBoost. Gradient boosting optimizes the provided loss function with each iteration. The objective is to reduce the residual from the preceding phase, which may be understood as the difference between the anticipated estimation and the true estimation. The final model is declared ready for usage once the residual value reaches a certain level. However, training is halted and the final model is chosen if several decision trees fall below a threshold value before the residual may do so. The XGB model has several notable characteristics, including the adoption of regularization, and the use of parallel execution. The objective function of XGBoost for rating the efficiency of the model may be expressed by the equation $P(q) = t(q) + r(q)$, where q stands for the parameters, r for the regularization term, and t for the training loss. Here are some key points:

XGBoost and Gradient Boosting: XGBoost is an optimized implementation of the gradient boosting framework, which is a machine learning technique that combines weak learners (typically decision trees) to create a powerful ensemble model. Gradient boosting sequentially adds decision trees to the ensemble, with each tree learning from the mistakes of the previous trees.

Combination of Weak Classifiers: XGBoost combines multiple weak classifiers to create a strong learner. Weak classifiers are individual decision trees that may have limited predictive power. By combining the predictions of these weak classifiers, XGBoost builds a more accurate and robust model.

Feedback and Gradient Optimization: XGBoost incorporates feedback from previously trained decision trees into the model. It optimizes a provided loss function by reducing the residuals, which represent the differences between the predicted values and the true values. Each new decision tree focuses on reducing the residual errors of the ensemble.

Training Stopping Criteria: XGBoost iteratively adds decision trees until a certain stopping criterion is met. This criterion can be defined based on the number of trees or the level of the residual error. If the number of decision trees falls below a threshold value before the residual error reaches a desired level, training is halted, and the final model is selected.

Regularization: XGBoost incorporates regularization techniques to prevent overfitting. Regularization helps control the complexity of the model by adding penalties or constraints to the objective function. It prevents the model from becoming too complex and improves its generalization performance on unseen data.

Parallel Execution: XGBoost takes advantage of parallel processing to speed up training and prediction. It leverages parallelism at various levels, such as column blockings, parallel tree construction, and distributed computing, to efficiently handle large datasets.

Objective Function: The objective function in XGBoost evaluates the performance of the model based on the parameters (q), the regularization term (r), and the training loss (t). The objective function is optimized during the training process to find the best model parameters that minimize the overall loss.

Boosting and Iterative Model Building: XGBoost follows the boosting principle, where weak learners are sequentially added to the ensemble to correct the mistakes made by the previous models. Each weak learner is trained on the residuals (the differences between the predicted and actual values) of the previous models, allowing the ensemble to focus on the challenging samples and improve overall performance.

Tree Pruning and Regularization: XGBoost employs regularization techniques to control the complexity of the individual decision trees. It uses both L1 and L2 regularization terms to add penalties to the loss function based on the complexity of the trees. This helps to prevent overfitting and improve generalization by discouraging the creation of overly complex trees.

Feature Importance: XGBoost provides a measure of feature importance, indicating the relevance of each input feature in the model. By tracking how much each feature contributes to reducing the loss function during the training process, XGBoost can rank features based on their importance. This information can be valuable for feature selection, feature engineering, and understanding the underlying patterns in the data.

Handling Missing Values: XGBoost has built-in capabilities to handle missing values in the dataset. During the training process, XGBoost automatically learns the best direction to assign missing values, optimizing the decision tree splits and improving overall performance. This feature is particularly useful when dealing with real-world datasets that often contain missing values.

Parallel Execution and Scalability: XGBoost is designed to take advantage of parallel processing capabilities, making it highly scalable and efficient. It leverages multi-threading on a single machine and distributed computing across multiple machines to speed up the training and prediction processes. This allows XGBoost to handle large-scale datasets and significantly reduce training times.

Early Stopping: XGBoost incorporates an early stopping mechanism during training. It monitors the performance on a validation set at each iteration and stops the training process if the performance does not improve for a certain number of consecutive iterations. Early stopping helps prevent overfitting and saves computation time by avoiding unnecessary iterations.

Handling Imbalanced Classes: XGBoost provides options to address class imbalance problems in classification tasks. By adjusting the weights or sampling strategy during training, XGBoost can

give more emphasis to the minority class, helping to improve the model's ability to correctly classify samples from the underrepresented class.

External Memory Support: XGBoost supports external memory, allowing it to handle datasets that cannot fit into memory. By loading data in chunks from disk, XGBoost enables training on large-scale datasets that exceed the memory capacity of the system.

XGBoost is widely used in various machine-learning competitions and real-world applications due to its excellent performance, scalability, and regularization capabilities. It is effective for both classification and regression tasks and has become a popular choice among data scientists and practitioners.

3.6.7 Gaussian Naïve Bayes Classifier (GNB):

Naive Bayes is an easy-to-understand probabilistic classifier that works on Bayes' theorem principle. The theorem elaborates on the likelihood of an event occurring in any condition. The formula looks like this,

$$\mathrm{P(C|D)} = \frac{P(D|C) * P(C)}{P(D)}$$

Where P(C|D) represents the probability that event D will take place given that event C has already occurred. If event D has already happened, P(D|C) is the likelihood that event C will also occur. P(C) is the likelihood that event C will occur. P(D) is the likelihood that event D will occur. According to NB, each feature variable is treated as an independent variable. The main benefit of Naive Bayes is that it only needs a relatively small amount of training data, which is required for classification and characterization. But Naive Bayes' conditional independence assumption between characteristics is weak and rarely valid in most cases of actual problems, except for when the attributes are collected from independent processes. There have been some attempts to enhance naive Bayes by relaxing the conditional independence assumption. This Gaussian naive Bayes classification is a special example of the naive Bayes approach, in which the attribute values are assumed to follow a Gaussian distribution in light of the class label. Gaussian distribution is used to figure out how the continuous values for each feature are spread out. The training data are separated by class, and the mean and standard deviation of each class are calculated. So, the following equation can be used to estimate the odds of a set of continuous data,

$$p(x_i|c) = \frac{1}{\sqrt{2\pi\sigma_{c,i}^2}} e^{\left(-\frac{(x_i-\mu_{c,i})^2}{2\sigma_{c,i}^2}\right)}$$

Where, for the given class label c, it's assumed that the ith attribute is continuous and that its mean and variance are represented by $\mu_{c,i}$ and $\sigma_{c,i}^2$ respectively. x_i is the likelihood of finding a value in ith attribute given a class label c.

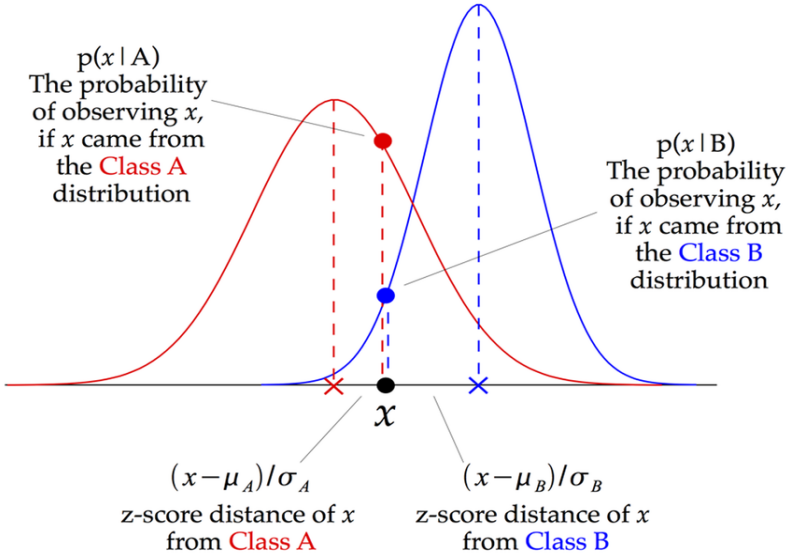

Figure 3.2: Gaussian Naïve Bayes.

This diagram demonstrates the operation of the Gaussian Naive Bayes classifier. Every data point's z-score length from each class means is calculated by dividing the length from the class mean by the class standard deviation. It is evident that the GNB model uses a fractionally different methodology and is effective in its application.

TABLE II. Classification Models Parameters

Classification Models	Features	Value
RF	n_estimators	20,40,60,80,100,120,150
	Min_samples_split	2,4,8,10,15,20
	Max_depth	5,10,15,20,40
DT	Min_samples_split	2,4,6,8,10,12,14,16,20
SVM	Tol	1e-3,1e-4,1e-5
	C	1.0,2.0,3.0
	Degree	3,4,5,6
LR	Tol	1e-3,1e-4,1e-5
	C	1.0,2.0,3.0
KNN	n_neighbors	5,6,7,8,9,10,11,12
	Leaf_size	30,35,40,45
	p	2,3,4,5,6
GNB	var_smoothing	1e-9,1e-10,1e-11
XGB	Eta	0.3,0.4,0.5,0.6,0.7
	Gamma	0,10,20,160,50,100
	Max_depth	6,7,8,9,10,20,50

3.6.8 Stacking Classifier (SC):

Stacking is usually a two-tiered ensemble learning algorithm that uses the result from the base classifier level as input data to a meta-learning algorithm for fusion. Stacking merges classifiers that were made using various methods of learning on the same data set. This makes it possible to make more accurate predictions. First, a base classifier level is developed. Then, the results of the base-level classifiers are used to learn a meta-level classifier. To make a training dataset for the meta-level classifier, the cross-validation method is used. A meta-classifier is a classifier that uses all of the predictions as features to generate a final prediction out of them all. It chooses the final

class as the desired outcome using the classes projected by many other classifiers. The stacking process involves the following steps:

Base Classifier Level: Several different base classifiers, each potentially trained using different learning methods, are trained on the same dataset. These base classifiers can be diverse, such as decision trees, support vector machines, or neural networks. Each base classifier makes predictions on the training data.

Meta-learning: The predictions made by the base classifiers become the input data for a meta-learning algorithm, often referred to as the meta-classifier. This meta-learning algorithm is trained on the predictions of the base classifiers along with the actual target values of the training data. It learns how to combine or fuse the predictions from the base classifiers to generate a final prediction.

Training Dataset: To create a training dataset for the meta-classifier, a cross-validation technique is commonly used. The training data is divided into multiple subsets or folds, and each fold is used as a validation set while training the base classifiers. The predictions of the base classifiers on the validation sets are then collected and combined to form the training dataset for the meta-classifier.

Meta-classifier: The meta-classifier takes the combined predictions from the base classifiers as input features and learns to predict the outcome or class label. It can be any classifier, such as logistic regression, random forest, or another decision tree. The meta-classifier is trained on the training dataset created in the previous step.

Prediction: Once the stacking ensemble is trained, it can be used to make predictions on new, unseen data. The base classifiers generate predictions on the new data, and these predictions are fed into the meta-classifier. The meta-classifier then combines the base classifier predictions and generates the final prediction or class label.

Stacking allows for the combination of multiple classifiers and their predictions, leveraging the strengths of different algorithms and learning methods. By using a meta-classifier to learn from the base classifier predictions, stacking aims to improve prediction accuracy compared to using individual classifiers alone.

It's worth noting that there are variations and extensions to the basic stacking algorithm, such as using multiple layers of base classifiers and meta-classifiers or incorporating weighted averaging of the base classifier predictions in the meta-classifier. These variations aim to further enhance the performance of the stacking ensemble.

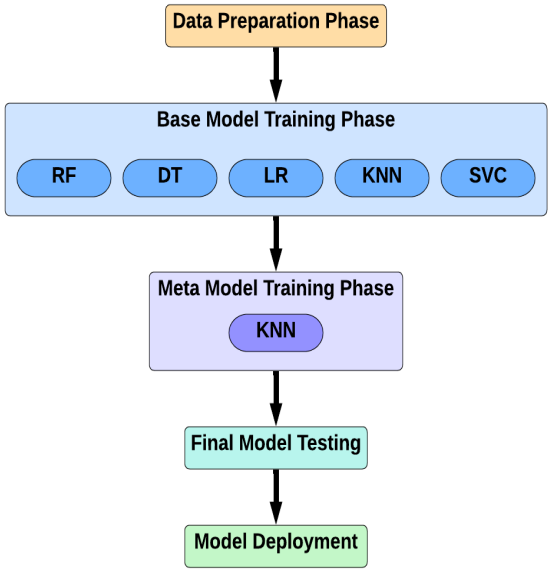

Figure 3.3: Flowchart for SC

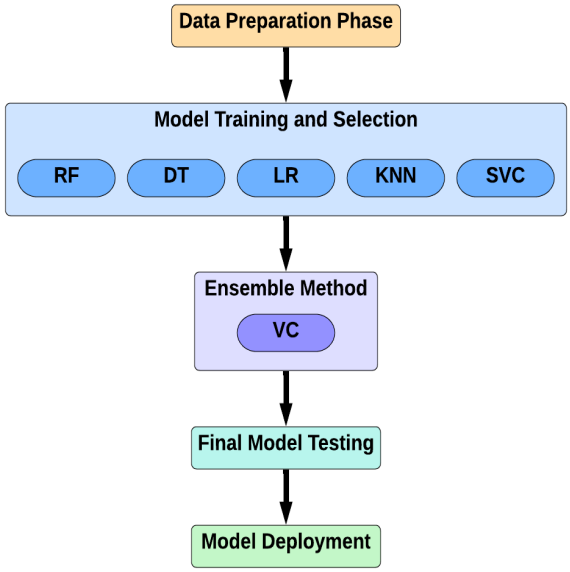

Figure 3.4: Flowchart for VC

3.6.9 Voting Classifier (VC):

Ensemble learning combines the predictions of multiple classifiers to make more accurate predictions compared to using individual classifiers alone. When building an ensemble, it is important to use diverse classifiers. Diversity can be achieved by using different learning algorithms, varying hyperparameters, or employing different subsets of the training data. Diversity helps to capture different aspects of the data and reduces the risk of overfitting. By using a diverse set of classifiers or learning algorithms, ensemble methods can capture different patterns and improve overall prediction performance.

A voting classifier is a popular ensemble learning technique where multiple base models, each trained on the same training data, make predictions individually. The predictions of these base models are then combined to generate a final prediction. There are two types of voting classifiers: hard voting and soft voting.

Hard-Voting Classifier: In a hard-voting classifier, each base model votes for a specific class label. The final prediction is determined by the majority vote of the base models. The class label that receives the most votes is selected as the predicted class label. This approach is also known as the majority-vote classifier.

Soft-Voting Classifier: In contrast, a soft-voting classifier takes into account the probabilities or confidence scores assigned by each base model for each class label. Instead of considering only the majority vote, the soft-voting classifier calculates the average probabilities of each class label across all the base models. The class label with the highest average probability is chosen as the final prediction.

The advantage of the soft-voting classifier is that it takes into account the confidence or likelihood information provided by each base model, which can result in more nuanced predictions. However, soft-voting classifiers typically require models that can provide probability estimates for each class, while hard-voting classifiers can work with models that provide only class labels.

The performance of an ensemble typically improves with an increase in the number of diverse base classifiers. However, there is a point of diminishing returns, and adding too many classifiers may result in longer training and prediction times without substantial improvement.

Ensemble methods, including voting classifiers, can improve prediction accuracy, increase robustness to noise, and handle complex decision boundaries. They are commonly used in machine learning to solve a wide range of problems.

3.7 Work Flow Diagram:

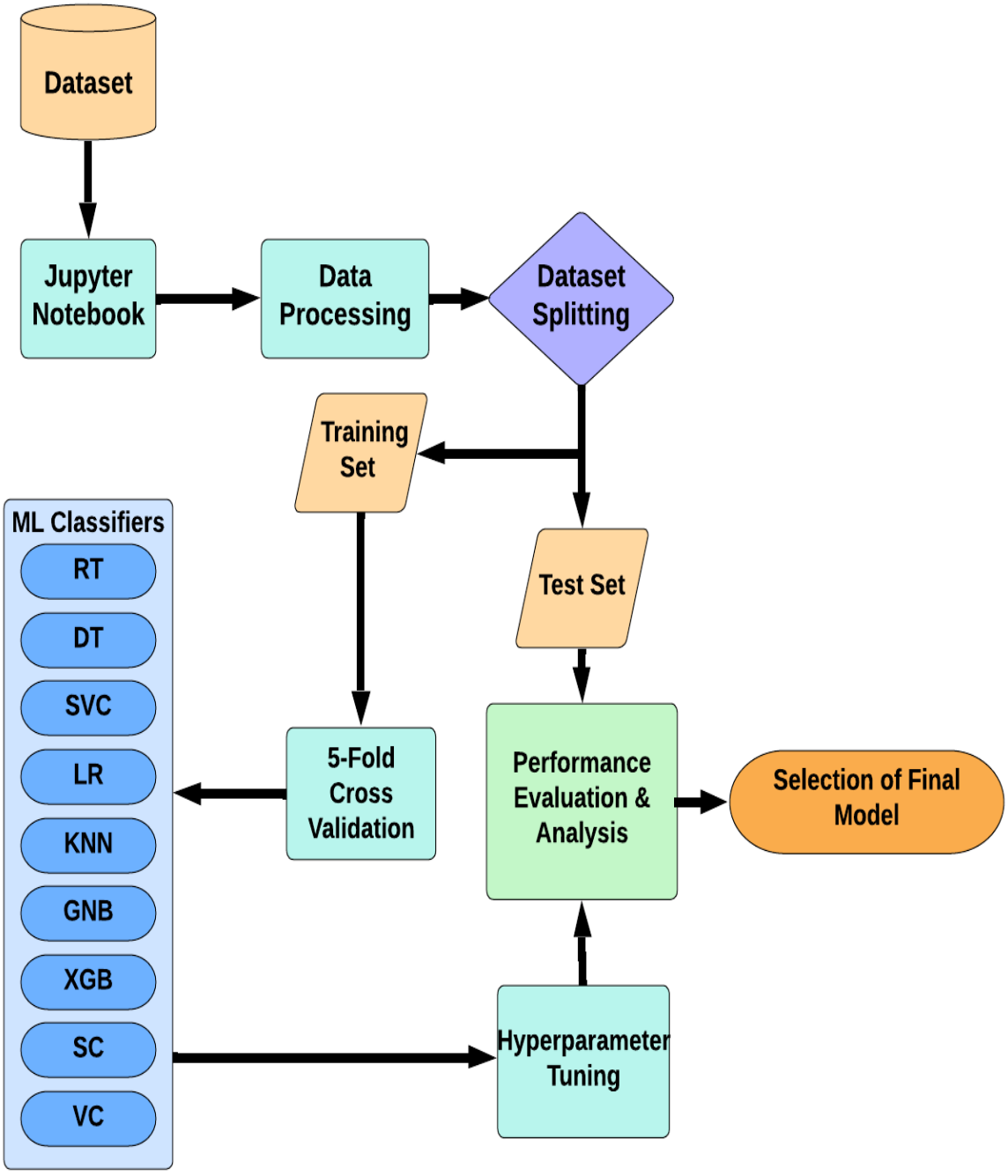

Figure 3.5: Algorithm of the work

CHAPTER 4

RESULT

4.1 Performance parameters:

When it comes to determining how successful machine learning models are, performance parameters are necessary tools. They make it possible to conduct a quantitative analysis of the model's performance by contrasting the outcomes that the model predicted with the outcomes that were seen. The selection of performance parameters is contingent on the particular activity at hand as well as the goals of the model, and this selection can vary from one application to another. When choosing performance parameters, it is essential to take into account the nature of the problem that is being tackled as well as the goals that the model is attempting to achieve. The terms "accuracy," "precision," "recall," "F1 score," and "area under the receiver operating characteristic curve" (ROC AUC) are examples of performance parameters that are frequently utilized. Note that these are just a few examples of performance parameters; the selection you make should be based on the particular requirements and goals of the machine learning task you are working on. It is crucial to keep in mind that these are only a few of the available performance parameters. Choosing proper performance criteria is essential for accurately evaluating and comparing models, guiding changes, and eventually getting the intended results. In this study, the used performance parameters along with their corresponding mathematical equations are given below:

4.1.1 Accuracy:

Accuracy is a regularly used performance metric that measures the proportion of right predictions made by a machine learning classification model concerning the total number of predictions. Accuracy is a commonly used performance metric that measures the proportion of correct predictions made by a machine learning classification model. It indicates how successfully the model can categorize or classify the data. In its most basic form, the accuracy of a machine learning model is denoted by the proportion of examples, relative to the total number of occurrences, for which it is possible to be certain that the model has correctly predicted the outcomes. By calculating the accuracy, we can have a better picture of the overall performance of the model as

well as its capacity to correctly assign labels to new data points that have not been seen before. A greater accuracy score indicates a higher proportion of true predictions, which suggests that the model is working effectively and can be trusted to make accurate forecasts. On the other hand, a lower accuracy means that the model's predictions may be less dependable or less aligned with the actual events. Although accuracy is a performance characteristic that sees widespread application, it is essential to keep in mind the limitations of this metric. It is possible that accuracy alone does not provide a whole view of the effectiveness of a model, particularly in circumstances in which there are class imbalances present within the dataset. In situations in which the distribution of the classes is unequal, a high accuracy score may give an incorrect impression. For instance, in a binary classification task with 95% negative instances and 5% positive examples, a naive model that always predicts negative might achieve 95% accuracy, but it would fail to capture the genuine positive cases effectively. This is because the naive model would always predict negative. Accuracy is a valuable performance parameter for reviewing a machine learning classification model, but it should be understood in conjunction with other metrics to acquire a thorough knowledge of the model's performance. To summarise, while accuracy is a valuable performance measure for assessing a machine learning classification model, it should be evaluated in conjunction with other metrics. We can make more educated judgments about the dependability and efficiency of the model when it is applied to situations that take place in the real world if we evaluate and understand a range of performance parameters.

4.1.2 Train score:

There is a crucial juncture in the process for machine learning at which point the model is said to have reached its final state. This indicates that we have reached a point in the process where we have not previously dealt with or interacted with the dataset that is being presented to us. In situations that take place in the real world, it is possible to come across data that is foreign to you. This instance depicts that possibility. The achievement of completeness in model development is a significant milestone since it enables us to analyze the performance of the model and determine whether or not it is valid. Examining a model's performance score is one method for determining whether or not the model is complete. A higher score indicates a higher level of validity, which suggests that the model is more dependable in making correct predictions since it has a higher level of reliability. This score is a measurement of how well the model generalizes to data that it

has not before seen. When the model performs well on the validation or test datasets, it instills confidence that it will likely perform similarly on new instances that it has not seen before. This is because it has learned from its previous successes. However, it is essential to keep in mind that merely having a high score does not in and of itself guarantee perfect or unfailing forecasts. It is important to evaluate the performance of the model using a thorough set of performance indicators that have been adapted to the particular undertaking and dataset. A more complete comprehension of the model's advantages and disadvantages can be attained by evaluating it from a variety of perspectives. In addition, having all of the pieces of a model does not indicate that the model is fixed or unalterable in any way. It is possible that the model will need to be reevaluated and may be revised when new data becomes available or as the context of the problem evolves. This will ensure that the model continues to maintain its performance and validity. To guarantee the model's dependability in dynamic contexts, ongoing monitoring and development are very necessary. We can acquire insights into the model's validity and its potential for deployment in the actual world by using completeness as a checkpoint in the process of developing the model and by employing performance scores. Attempting to achieve higher scores not only proves that the model can accurately classify or predict events but also lays the groundwork for making well-informed decisions based on the information it generates.

4.1.3 Precision:

The precision score is a performance metric that compares the projected positive labels to the actual labels to determine how accurately they match up. In some circumstances, this concept is also referred to as the positive predictive value. When performing a classification assignment, precision is an extremely important factor to consider when trying to achieve a healthy balance between the number of false positives and false negatives. The distribution of classes contained within the dataset has the potential to have an impact on the precision score. The precision score offers a reliable measure of the accuracy of the prediction in circumstances in which there is a considerable difference between the classes. while the goal is to reduce the number of cases in which the model mistakenly predicts a positive label while the actual label is negative, it becomes very useful to focus on minimizing the number of false positives. This means reducing the number of instances in which the model incorrectly predicts a positive label. When calculating the accuracy score, the number of true positives (instances that were successfully predicted as positive) is

divided by the total number of true positives and false positives (instances that were wrongly projected as positive). This gives the percentage of instances that were correctly predicted as positive. This ratio provides a quantitative representation of the proportion of the model's positive predictions that turn out to be accurate. The accuracy score is especially helpful in circumstances in which the cost or repercussions associated with false positives are significant. For example, in the field of medical diagnostics, it is essential to accurately identify individuals as having a particular ailment to avoid making incorrect diagnoses and wasting time and resources on unneeded treatments. In situations like these, accuracy is a key metric for determining whether or not a model can correctly identify positive instances while also minimizing the chances of producing false positives. It is essential to be aware of the fact that the distribution of classes contained within the dataset affects the precision score. The precision score can be impacted when there is a considerable class imbalance, which indicates that one class is significantly more prevalent than the other. In situations like this, it is possible to earn a high precision score by merely guessing the majority class the majority of the time. However, this may not be an accurate reflection of the model's actual performance. It is necessary to take into account precision in addition to other performance indicators such as recall, accuracy, and F1 score to produce a thorough evaluation. These metrics provide a more holistic assessment of the performance of the model, taking into consideration both false positives and false negatives, and they can assist in making educated decisions based on the particular demands and requirements of the task at hand. In conclusion, the precision score is an important metric that may be used to evaluate the correctness of positive predictions concerning the actual labels. It helps strike a balance between false positives and false negatives, which is especially helpful in situations in which it is essential to minimize the number of false positives. It is essential, however, to evaluate precision in the context of class distributions and in conjunction with other performance indicators to have a thorough view of the effectiveness of the model.

4.1.4 F1 score:

The F1 score is a popular performance metric that provides a full evaluation of the performance of a model by combining the scores for precision and recall. This score also takes into account how well the model performs overall. The F1 score, in contrast to more conventional measures of accuracy, accords equal weight to a question's precision as well as its recall. As a result, this score

is useful in circumstances in which achieving a healthy balance between the two metrics is essential. It is possible to think of the F1 score as an alternative to accuracy metrics because it does not require prior knowledge of the entire number of observations and delivers a single value that provides insights into the overall quality of the model's output. In other words, the F1 score is a more straightforward metric than accuracy measures. The F1 score is especially helpful in situations in which there is an uneven distribution of classes or where the impacts or costs of false positives and false negatives are not the same. The F1 score provides a fair assessment of the model's capacity to properly categorize positive cases while simultaneously minimizing the number of false positives and false negatives. This is accomplished by taking into account both precision and recall. The F1 score is determined mathematically by finding the harmonic mean of the respondent's precision and recall responses. Because the harmonic mean lays a greater emphasis on lower values, the F1 score will be lower if either precision or recall is low. This is because the harmonic mean places more emphasis on lower values. This makes certain that the F1 score accurately reflects the performance of the model in terms of both precision and recall, as opposed to being excessively biased by only one parameter on its own. The F1 score has the benefit of providing a single figure that is condensed while yet providing a comprehensive summary of the overall performance of the model. This facilitates the comparison of various models or variants of the same model, as well as the communication of the accuracy of the model's predictions to relevant stakeholders and decision-makers. However, it is essential to keep in mind that the F1 score might not always be the most applicable statistic for all circumstances. This is something that should be taken into consideration. It is possible that alternative metrics, such as accuracy, precision, or recall, are more pertinent to the task at hand. This is something that will depend on the unique requirements and priorities of the work to be done. When analyzing and interpreting the performance of the model, it is necessary to take into account several performance measures and the trade-offs associated with each of them. In conclusion, the F1 score is a helpful performance metric that provides a balanced evaluation of the performance of a model by combining precision and recall to deliver accurate results. It delivers a single value that provides insights into the overall quality of the model's output and provides an alternative to accuracy measurements. Even though the F1 score is rather popular, it is essential to take into account the particulars of the situation and the prerequisites of the activity to choose the performance metric that is going to be the most accurate reflection of how well something was done.

4.1.5 Recall:

A performance metric that evaluates a model's capacity to reliably predict positive instances based on a training set that contains actual positives is called the recall score. This score is also referred to as the sensitivity rate or the rate of true positives. It places a particular emphasis on the model's capacity to differentiate between true positives and false positives. If the model has a high recall score, it shows that it can successfully recognize both positive and negative examples of the target characteristic. It is essential to have a high recall rate because this is directly correlated to having a high sensitivity and true positive rate. To put it another way, a model that has a high recall score is very good at identifying occurrences of the researched phenomenon that are in a positive light. This indicates that there is a reduced risk of missing or failing to recognize positive cases as a result of using it. On the other hand, a low recall score may indicate that the model is not very good at accurately detecting positive situations. It suggests that the model may have a greater rate of false negatives, which occurs when the model fails to identify true positive cases. A low recall score indicates a reduced sensitivity and true positive rate, which can lead to missed opportunities or misclassifications in applications where the accurate identification of positive cases is critical. Having a low recall score also denotes a lower rate of recall accuracy. Achieving a high recall score is frequently prioritized across a variety of areas, including medical diagnostics and anomaly detection, to reduce the likelihood of missing significant positive cases. A high recall score does not, however, necessarily guarantee that there will not be any false positives. This is a crucial point to keep in mind. To achieve an evaluation that is more complete of the model's overall performance, it is important to take into consideration other performance measures in addition to recall. One such metric is precision. To summarize, the recall score examines the capability of the model to accurately recognize positive cases. A high recall score implies a great skill to recognize positive examples and reflects a high sensitivity and true positive rate. In other words, it suggests that the rate of real positives is also high. On the other hand, a low recall score suggests a decreased capacity to correctly identify positive cases. In applications in which the correct detection of affirmative cases is of the utmost importance, it is especially crucial to work towards achieving a high recall score. To get a deeper and more comprehensive comprehension of the performance of the model, it is necessary to take into account other performance indicators in addition to recall.

TABLE III. Result Parameters

Parameters	***Formulae***
Accuracy	$\frac{(TP + TN)}{(TP + FN + TN + FP)}$ (Unseen data) [39]
Train score	$\frac{(TP + TN)}{(TP + FN + TN + FP)}$ (Seen/Trained data) [39]
Precision	$\frac{TP}{(FP + TP)}$ [39]
F1 Score	$\frac{2 \times Precision \times Recall}{(Precision + Recall)}$ [39]
Recall	$\frac{TP}{(FN + TP)}$ [39]

4.2 Analysis of the Result:

Several different machine learning models, such as Random Forest (RT), Decision Tree (DT), Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbours (KNN), Gaussian Naive Bayes (GNB), XGBoost (XGB), Stacking Classifier (SC), and Voting Classifier (VC), were subjected to a comprehensive analysis. The purpose of this study was to evaluate the effectiveness of these models in terms of their ability to predict outcomes based on a given dataset. The results of the calculation of several important performance measures provided insights into the efficiency of each model.

TABLE IV. Analysis of result parameters for ML Algorithms

Gender	Labels	RF	DT	SVC	LR	KNN	GNB	XGB	SC	VC
Female	**Accuracy**	0.853	0.736	0.818	0.84	0.839	0.542	0.851	0.853	0.829
	Train Score	0.989	0.937	0.952	0.947	0.881	0.549	0.995	0.913	0.958
	Precision	0.859	0.81	0.79	0.808	0.707	0.616	0.805	0.835	0.796
	F1 Score	0.799	0.714	0.728	0.758	0.716	0.505	0.776	0.812	0.746
	Recall	0.817	0.74	0.75	0.775	0.75	0.633	0.792	0.833	0.767
Male	**Labels**	**RF**	**DT**	**SVC**	**LR**	**KNN**	**GNB**	**XGB**	**SC**	**VC**
	Accuracy	0.772	0.736	0.768	0.781	0.754	0.695	0.749	0.676	0.777
	Train Score	0.923	0.937	0.904	0.882	0.87	0.683	0.995	0.816	0.909
	Precision	0.805	0.81	0.755	0.815	0.796	0.768	0.785	0.672	0.816
	F1 Score	0.74	0.714	0.725	0.748	0.718	0.639	0.715	0.643	0.738
	Recall	0.777	0.74	0.773	0.786	0.759	0.69	0.754	0.679	0.782

Upon conducting an in-depth study of TABLE II, the performance measures, such as accuracy, train score, precision, F1 score, and recall, display differences between males and females. These differences were found to be statistically significant. These inequalities can be attributable to several different causes, such as the presence of an extra characteristic (anemia) that is unique to

females and variances in the number of samples gathered from each gender. Let's investigate the findings even further and offer some additional perspectives: According to the findings of the study, both the SC model and the RF model performed extremely well in terms of accuracy for females, attaining a high score of 0.853. This was the case for both models. This suggests that these models were able to generate correct predictions on a sizeable percentage of the female samples that were tested. In addition, the XGB model accomplished a remarkable value of 0.995 for the training score, which represents the maximum possible value. A high train score indicates that the model did a good job of capturing the underlying relationships and patterns that were present in the training data for females. The RF model stood out from the other models that were tested because it achieved the highest accuracy score of 0.859 for females. accuracy is a measurement that determines a model's ability to minimize the number of false positives that it produces. This suggests that the RF model had a rather low rate of incorrectly categorizing situations as positive when they were, in fact, negative. In addition to this, the SC model had the greatest F1 score of 0.812 and the highest recall score of 0.833 out of all of the models that were tested for females. The F1 score offers a well-rounded evaluation because it takes into account both precision and recall, and the SC model performed exceptionally well when it came to striking a healthy balance between reducing the number of false positives and the number of false negatives. A high recall score suggests that the SC model had a higher rate of properly recognizing positive cases for females. This may be inferred from the fact that the model had a higher rate of overall accuracy. The investigation showed that when the focus was shifted to men, the LR model achieved the greatest accuracy score of 0.781, suggesting its competency in making accurate predictions on the male samples. This was discovered when the focus was shifted from females to males. In a manner analogous, the LR model achieved the greatest F1 score of 0.748 and the highest recall score of 0.786 among the models that were put to the test for males. This provides support for the hypothesis that the LR model was successful in accurately detecting positive events within the male subset. The XGB model also achieved the highest train score for males, which was 0.995, which is comparable to the results that were found for females. This suggests that the XGB model successfully learned from the training data for males and successfully caught the underlying patterns. It is interesting to note that the VC model emerged as the top performance for guys. It achieved the maximum precision score of 0.816, making it the top performer in this category. This suggests that the VC model had a reasonably low rate of incorrectly categorizing cases as positive

for males overall. In a nutshell, the investigation of the data included in TABLE II yields insightful information regarding the capabilities of various machine learning models concerning the prediction of outcomes in both males and females. The findings demonstrate the various models' superiorities in terms of accuracy, train score, precision, F1 score, and recall for each gender. These findings can be used as a guide for the selection of appropriate models depending on the specific criteria and objectives of the prediction task, taking into mind gender-specific characteristics and considerations.

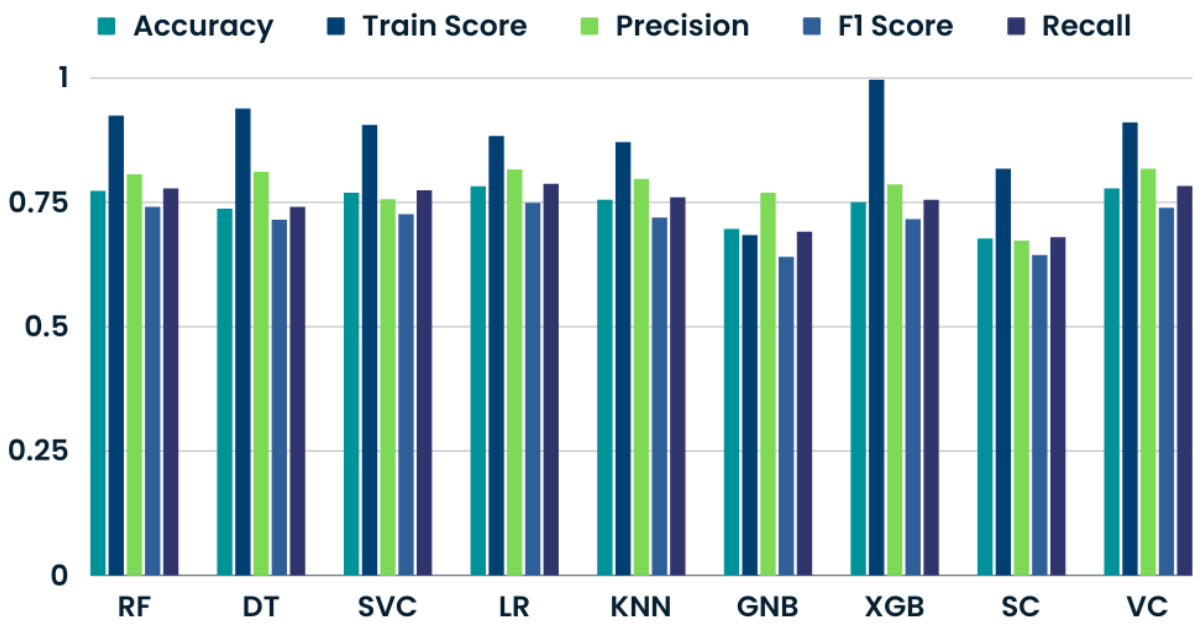

Figure 4.1: Comparison of result parameters for Males

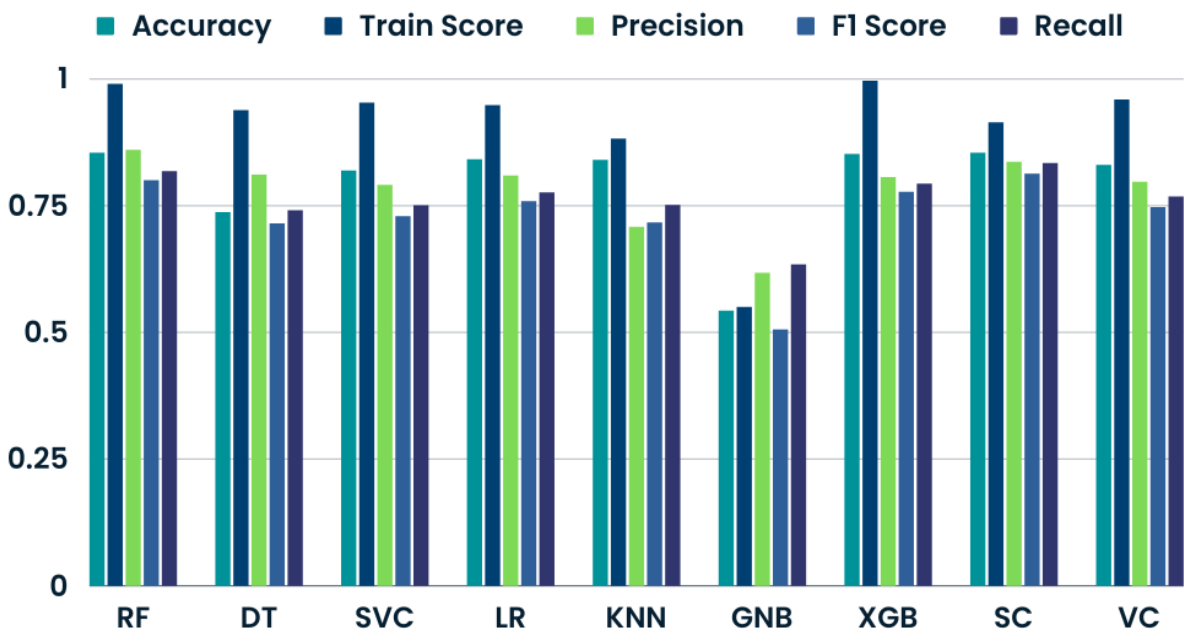

Figure 4.2: Comparison of result parameters for Females

In addition to analyzing performance parameters such as accuracy, train score, precision, F1 score, and recall, the generation of ROC curves and the calculation of ROC-AUC values offer additional valuable insights into the predictive capabilities of the machine learning models that were applied to the provided dataset. These insights can be used to improve the predictive capabilities of machine learning models. The ROC curve and the ROC-AUC values also exhibit changes between males and females, just like the performance metrics that were discussed previously. These variations are caused by the inclusion of an extra feature (Anaemia) that is unique to females, as well as differences in the number of samples obtained from each gender. The Receiver Operating Characteristic (ROC) curve is a graphical depiction that depicts the trade-off between the true positive rate (TPR) and the false positive rate (FPR) at various categorization thresholds. The curve can be found by typing "Receiver Operating Characteristic" into a computer's search engine. It

enables us to visualize the performance of a model across a range of thresholds and evaluate the model's ability to differentiate between positive and negative examples. The ROC curve compares the TPR with the FPR, and the best-case scenario is one in which the curve hugs the upper-left corner of the graph, which indicates that the TPR is high while the FPR is low. In addition, the ROC-AUC (Area Under the ROC Curve) gives a single numerical value that summarises the performance of the model over all possible classification thresholds. This value may be found in the area under the ROC curve. It is the likelihood that the model will rank a randomly selected positive instance higher than a randomly selected negative instance when the two are compared head-to-head. If the ROC-AUC value is higher, this implies that the model performs better in terms of its capacity to discriminate and in general. When taking into consideration the gender-specific analysis, the ROC curves particular to females and males were constructed for each of the nine machine-learning algorithms that were used on the dataset. These ROC curves offer a graphical depiction of how well the models can tell the difference between positive and negative examples for each gender. By analyzing the ROC curves, it is feasible to determine which models have greater capacities for discrimination and to arrive at well-informed choices on the selection of models. In addition to this, a quantitative comparison of the performances of the models is made possible by the generation of ROC-AUC values. We can discover the models that regularly give superior predicting capabilities and better differentiate positive and negative examples by comparing the ROC-AUC values of the various models and seeing which ones have the advantage. The construction of ROC curves and the calculation of ROC-AUC values together offer a thorough evaluation of the predictive abilities of the models in terms of distinguishing between positive and negative examples. It is feasible to determine whether models are superior in terms of their capacity to discriminate between males and females by conducting separate analyses of these curves and values for males and females, taking into consideration the factors that are special to each gender. This information helps pick the models that are most fit for the given dataset and prediction job, taking into consideration the particular needs and goals that are associated with each gender.

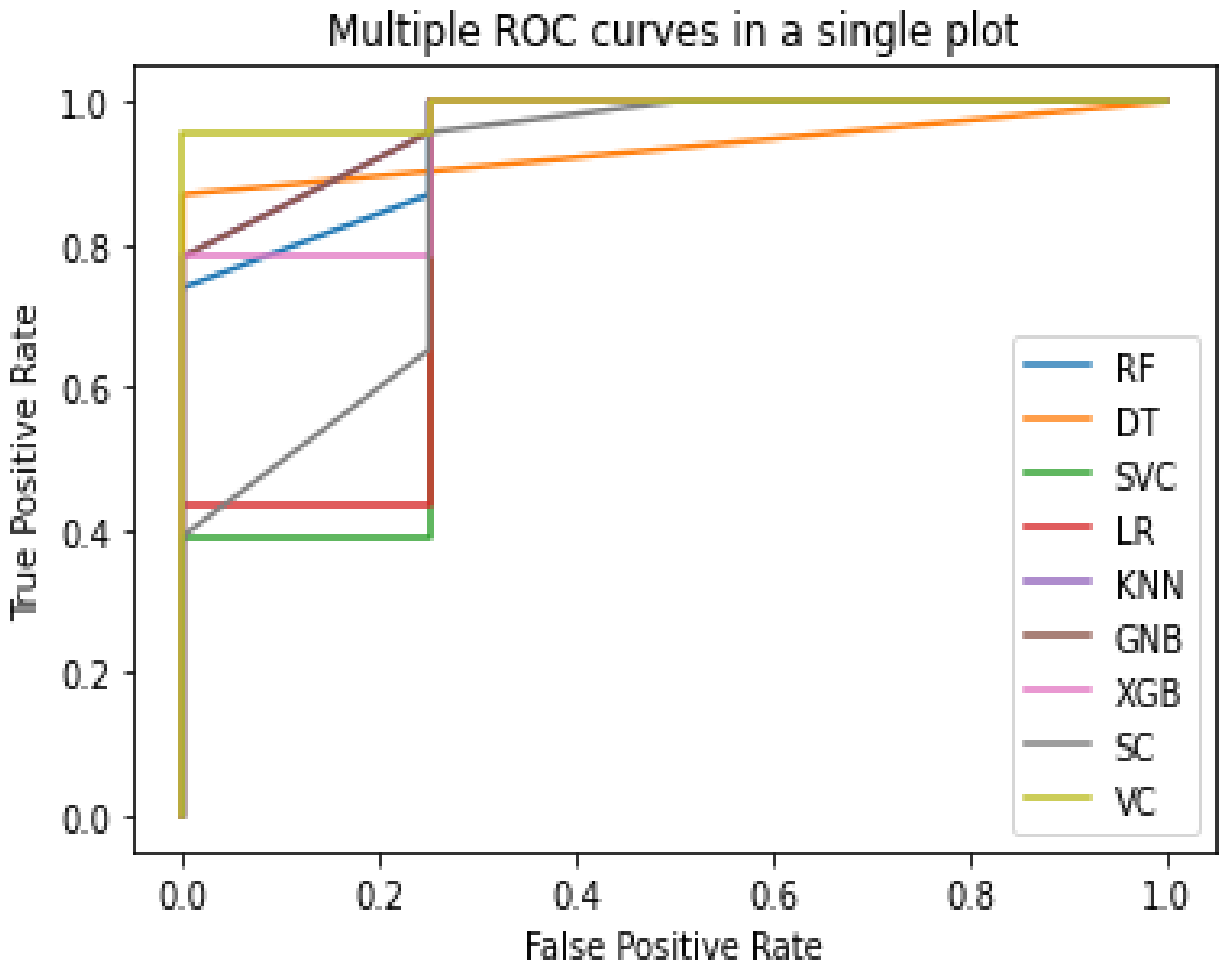

Figure 4.3: ROC curves for females

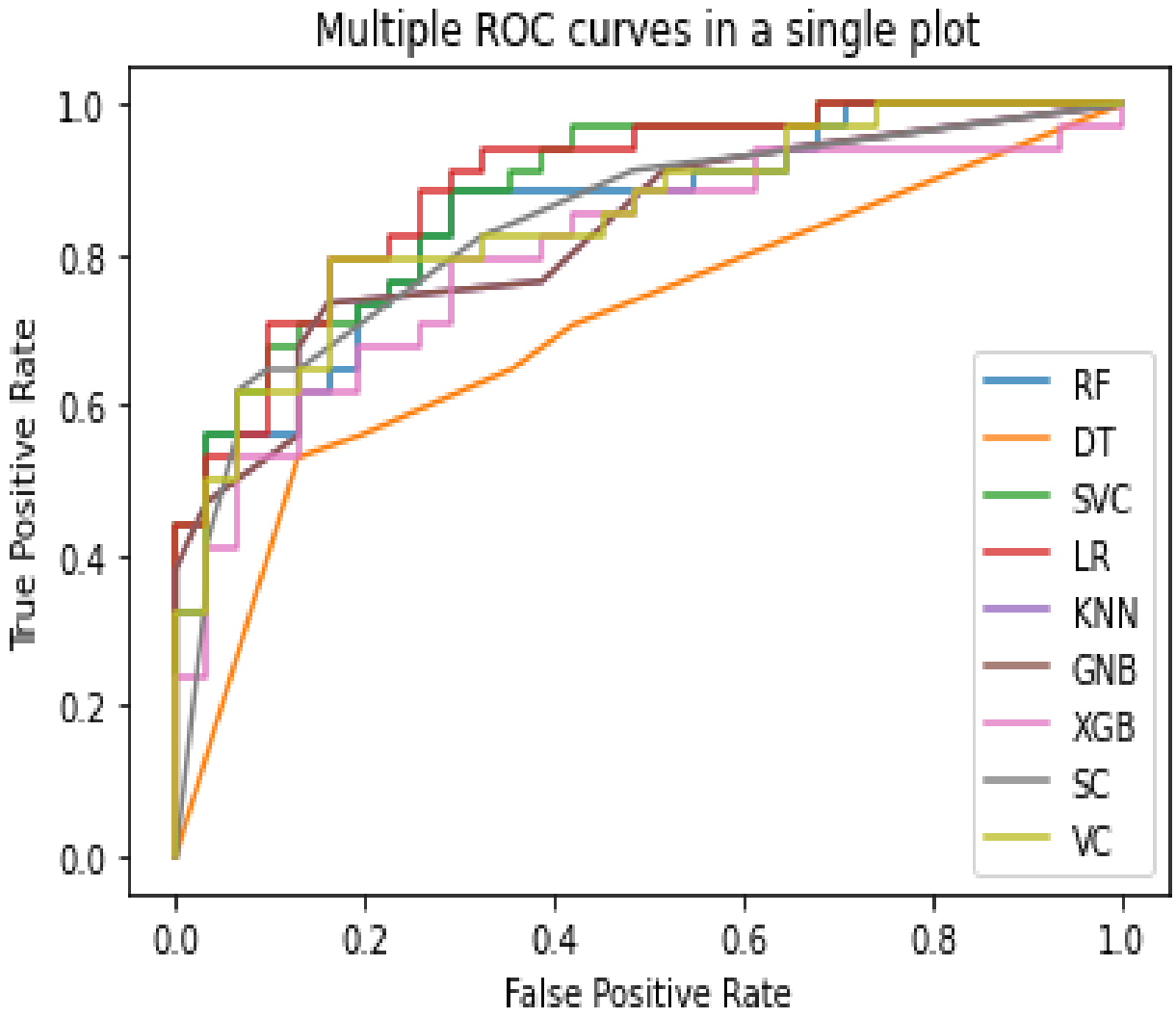

Figure 4.4: ROC curves for males

Upon analyzing Fig. 7 and Fig. 8, it is evident that the ROC curves generated by the female samples generally exhibit a larger area under the curve (AUC) compared to the ROC curves generated by the male samples. This discovery lends credence to the hypothesis that the machine learning models trained on female samples tend to have higher overall performance when comparing them to the models trained on male samples to discriminate between positive and negative instances. The existence of an additional characteristic (Anaemia) that is unique to females as well as probable changes in the dataset makeup between genders can both be attributed to the difference in the AUC values that have been observed. The ROC-AUC value is used as a quantitative measurement of the performance of the models in terms of their ability to discriminate between different types of data. It is the probability that the model will rank a randomly selected positive instance higher than a randomly selected negative instance. This probability is expressed as a percentage. When the ROC-AUC value is larger, it suggests that the overall performance is better and that the capacity to correctly categorize positive and negative examples is stronger. To conduct a more in-depth analysis of the performance of the models, the ROC-AUC values were computed for each algorithm and then applied to the samples collected from both male and female subjects. These results provide a numerical representation of the performance of the models in differentiating between positive and negative instances for each gender. The performance of the various machine learning algorithms can be compared across genders by inspecting the ROC-AUC values. This makes it conceivable. Higher ROC-AUC values suggest that the models have greater discriminative abilities and can produce more accurate predictions. ROC-AUC values range from 0 to 1. In conclusion, the examination of Figs. 7 and 8 reveal that the ROC curves created by the female samples generally exhibit bigger areas under the curve in comparison to those generated by the male samples. This shows that the models applied to the female samples have a better overall performance in differentiating between positive and negative cases. The resulting ROC-AUC values further quantify the performance of the models and enable a comparative evaluation of the algorithms for both males and females. These new insights can help select the most effective models for forecasting outcomes based on gender-specific considerations and can contribute to the development of more accurate models.

TABLE V. ROC-AUC values for males

Algorithms	ROC-AUC value
RF	0.856
DT	0.686
SVC	0.884
LR	0.893
KNN	0.828
GNB	0.828
XGB	0.799
SC	0.829
VC	0.853

TABLE VI. ROC-AUC values for females

Algorithms	ROC-AUC value
RF	0.967
DT	0.935
SVC	0.848
LR	0.859
KNN	0.967
GNB	0.967
XGB	0.946
SC	0.897
VC	0.989

The fact that the LR algorithm got the greatest ROC-AUC value of 0.893 for men, which indicates its greater discriminative power in distinguishing between positive and negative occurrences, is a noteworthy discovery that can be gleaned from the data presented in Tables III and IV. On the other hand, the VC algorithm had the greatest ROC-AUC value of 0.989 for females, indicating that it had a strong predictive performance in distinguishing between positive and negative instances. This was the case when it came to males. Based on these findings, it appears that the LR algorithm performs particularly well when analyzing the dataset for males, whilst the VC method appears to produce more promising results when applied to females. The inputs will be offered in the context of the final output of the research, and they will be based on a list of 20 triggering elements related to the activities carried out by the patient in the preceding twenty-four hours. These inputs can be analyzed by making use of appropriate machine learning models that have been selected based on their performance parameters (accuracy, train score, precision, F1 score, and recall). This allows for the probability of a migraine headache occurring for the patients to be determined while taking into consideration the particular characteristics of the dataset that is being used. The selected models, which have shown promising performance in terms of their capacity to predict outcomes, can handle the input data well and produce a probability assessment of the likelihood of an individual experiencing a migraine headache. The output has the potential to

provide insightful information regarding the likelihood of experiencing migraine headaches based on the inputs that have been provided by capitalizing on the models' experience in identifying patterns and relationships within the dataset. This strategy makes it possible to conduct a customized risk assessment for the occurrence of migraine headaches by taking into account the individual's unique behaviors and migraine-inducing events that occurred within the preceding twenty-four hours. The output can give a reasonable probability assessment that is personalized to the patient's particular circumstances if it incorporates machine learning models that are reliable and accurate based on the set performance parameters. In a nutshell, the purpose of this research is to employ suitable machine learning models, chosen based on the performance characteristics they offer, to analyze inputs that are associated with actions that have taken place in the previous twenty-four hours. The research outcome can provide a probability assessment for the occurrence of migraine headaches by utilizing the models that are most suitable for each gender based on their performance in the given dataset. This can facilitate personalized insights and potential preventive strategies for individuals.

CHAPTER 5

CONCLUSION

A migraine typically causes a severe, excruciating headache surrounding the area of the forehead, side of the head, or eye area having pounding or pulsating pain that can continue for several hours or even days. Any kind of movement, activity, bright light, or loud noise contributes to the aggravation of pain. Having many symptoms besides frequent nausea and vomiting, migraines are thought to originate from fleeting changes in neurotransmitters, neurons, and blood vessels in the brain, albeit this is only speculation. Through this study, it is highlighted that various triggers including specific foods and drinks, stress, changes in the weather and surroundings, and sensory aspects are connected to migraine attacks. Concerning that, there arises a desire for migraine prediction using machine learning taking into account the triggering elements to assist physicians in advising migraine patients on the best course of action to avoid future migraine attacks. In this study, multiple performance evaluation methodologies, including accuracy, precision, train score, recall, F1 score, ROC curve, and ROC-AUC value, are used to evaluate and analyze various machine learning algorithms. The most effective algorithm among all other algorithms in our research is determined to be the random forest algorithm having an accuracy rate of 85.9%, and the voting classifier algorithm, which had an accuracy rate of 81.6% for male and female participants, respectively. Future, this study will guide future research in developing a more precise predictive analysis and treatment strategies for migraine patients. Besides this study will pave the way for the creation of an electronic healthcare system and maintain the quality of life by creating awareness among migraine patients because of its high accuracy and quick processing time.

REFERENCES:

[1] Z. Farhadi *et al.*, "The Prevalence of Migraine in Iran: A Systematic Review and Meta-Analysis," *Iran Red Crescent Med J*, vol. 18, no. 10, Sep. 2016, doi: 10.5812/ircmj.40061.

[2] "The Facts About Migraine | American Migraine Foundation." https://americanmigrainefoundation.org/resource-library/migraine-facts/ (accessed Mar. 06, 2023).

[3] "Prevalence Of Migraines - How Common Are Migraines? | MI." https://www.themigraineinstitute.com/migraine-overview/prevalence-of-migraines/ (accessed Mar. 06, 2023).

[4] "1.1 Migraine without aura - ICHD-3." https://ichd-3.org/1-migraine/1-1-migraine-without-aura/ (accessed Mar. 06, 2023).

[5] D. E. Desouky, H. A. Zaid, and A. A. Taha, "Migraine, tension-type headache, and depression among Saudi female students in Taif University," *Journal of the Egyptian Public Health Association*, vol. 94, no. 1, p. 7, Dec. 2019, doi: 10.1186/s42506-019-0008-7.

[6] R. B. Domingues, C. C. H. Aquino, J. G. Santos, A. L. P. da Silva, and G. W. Kuster, "Prevalence and impact of headache and migraine among Pomeranians in Espirito Santo, Brazil," *Arq Neuropsiquiatr*, vol. 64, no. 4, pp. 954–957, Dec. 2006, doi: 10.1590/S0004-282X2006000600013.

[7] M. I. Oraby, R. H. Soliman, M. A. Mahmoud, E. Elfar, and N. A. Abd ElMonem, "Migraine prevalence, clinical characteristics, and health care-seeking practice in a sample of medical students in Egypt," *Egypt J Neurol Psychiatr Neurosurg*, vol. 57, no. 1, p. 26, Dec. 2021, doi: 10.1186/s41983-021-00282-8.

[8] M. E. Bigal, J. M. Bigal, M. Betti, C. A. Bordini, and J. G. Speciali, "Evaluation of the Impact of Migraine and Episodic Tension-type Headache on the Quality of Life and Performance of a University Student Population," *Headache: The Journal of Head and Face Pain*, vol. 41, no. 7, pp. 710–719, Jul. 2001, doi: 10.1046/j.1526-4610.2001.041007710.x.

[9] S. van Hemert *et al.*, "Migraine Associated with Gastrointestinal Disorders: Review of the Literature and Clinical Implications," *Front Neurol*, vol. 5, Nov. 2014, doi: 10.3389/fneur.2014.00241.

[10] L. S. Pahim, A. M. B. Menezes, and R. Lima, "Prevalência e fatores associados à enxaqueca na população adulta de Pelotas, RS," *Rev Saude Publica*, vol. 40, no. 4, pp. 692–698, Aug. 2006, doi: 10.1590/S0034-89102006000500020.

[11] N. Magnavita, "Headache in the Workplace: Analysis of Factors Influencing Headaches in Terms of Productivity and Health," *Int J Environ Res Public Health*, vol. 19, no. 6, p. 3712, Mar. 2022, doi: 10.3390/ijerph19063712.

[12] M. I. Oraby, R. H. Soliman, M. A. Mahmoud, E. Elfar, and N. A. Abd ElMonem, "Migraine prevalence, clinical characteristics, and health care-seeking practice in a sample of medical students in Egypt," Egyptian Journal of Neurology, Psychiatry and Neurosurgery, vol. 57, no. 1, Dec. 2021, doi: 10.1186/s41983-021-00282-8.

[13] X. Wang, H. B. Zhou, J. M. Sun, Y. H. Xing, Y. L. Zhu, and Y. S. Zhao, "The prevalence of migraine in university students: a systematic review and meta-analysis," *Eur J Neurol*, vol. 23, no. 3, pp. 464–475, Mar. 2016, doi: 10.1111/ene.12784.

[14] N. Magnavita, "Headache in the Workplace: Analysis of Factors Influencing Headaches in Terms of Productivity and Health," Int J Environ Res Public Health, vol. 19, no. 6, Mar. 2022, doi: 10.3390/ijerph19063712.

[15] B. Menon and N. Kinnera, "Prevalence and characteristics of migraine in medical students and its impact on their daily activities," *Ann Indian Acad Neurol*, vol. 16, no. 2, p. 221, 2013, doi: 10.4103/0972-2327.112472.

[16] C. M. O. de Almeida, P. A. M. da S. Lima, R. Stabenow, R. S. de S. Mota, A. L. Boechat, and M. Takatani, "Headache-related disability among medical students in Amazon: a cross-sectional study," *Arq Neuropsiquiatr*, vol. 73, no. 12, pp. 1009–1013, Dec. 2015, doi: 10.1590/0004-282X20150172.

[17] S. Chahine, S. Wanna, and P. Salameh, "Migraine attacks among Lebanese university medical students: A cross sectional study on prevalence and correlations," *Journal of Clinical Neuroscience*, vol. 100, pp. 1–6, Jun. 2022, doi: 10.1016/j.jocn.2022.03.039.

[18] V. Caponnetto *et al.*, "Comorbidities of primary headache disorders: a literature review with meta-analysis," *J Headache Pain*, vol. 22, no. 1, p. 71, Dec. 2021, doi: 10.1186/s10194-021-01281-z.

[19] G. Sarıcam, "Relationship between migraine headache and hematological parameters," *Acta Neurol Belg*, vol. 121, no. 4, pp. 899–905, Aug. 2021, doi: 10.1007/s13760-020-01362-x.

[20] A. Klein and C. J. Schankin, "Visual snow syndrome, the spectrum of perceptual disorders, and migraine as a common risk factor: A narrative review," *Headache: The Journal of Head and Face Pain*, vol. 61, no. 9, pp. 1306–1313, Oct. 2021, doi: 10.1111/head.14213.

[21] A. Falavigna *et al.*, "Prevalence and impact of headache in undergraduate students in Southern Brazil," *Arq Neuropsiquiatr*, vol. 68, no. 6, pp. 873–877, Dec. 2010, doi: 10.1590/S0004-282X2010000600008.

[22] G. F. Carvalho, J. Mehnert, H. Basedau, K. Luedtke, and A. May, "Brain Processing of Visual Self-Motion Stimuli in Patients With Migraine," *Neurology*, vol. 97, no. 10, pp. e996–e1006, Sep. 2021, doi: 10.1212/WNL.0000000000012443.

[23] L. Papetti *et al.*, "Truths and Myths in Pediatric Migraine and Nutrition," *Nutrients*, vol. 13, no. 8, p. 2714, Aug. 2021, doi: 10.3390/nu13082714.

[24] K. R. Merikangas, "Contributions of Epidemiology to Our Understanding of Migraine," *Headache: The Journal of Head and Face Pain*, vol. 53, no. 2, pp. 230–246, Feb. 2013, doi: 10.1111/head.12038.

[25] S. Kurt and Y. Kaplan, "Epidemiological and clinical characteristics of headache in university students," *Clin Neurol Neurosurg*, vol. 110, no. 1, pp. 46–50, Jan. 2008, doi: 10.1016/j.clineuro.2007.09.001.

[26] S. Kurt and Y. Kaplan, "Epidemiological and clinical characteristics of headache in university students," *Clin Neurol Neurosurg*, vol. 110, no. 1, pp. 46–50, Jan. 2008, doi: 10.1016/j.clineuro.2007.09.001.

[27] A. Rafi, S. Islam, M. T. Hasan, and G. Hossain, "Prevalence and impact of migraine among university students in Bangladesh: findings from a cross-sectional survey," *BMC Neurol*, vol. 22, no. 1, Dec. 2022, doi: 10.1186/s12883-022-02594-5.

[28] M. A. Ra, M. T. Hasan, and M. G. Hossain, "Prevalence and Impact of Migraine among University Students in Bangladesh: Findings from a Cross-sectional Survey", doi: 10.21203/rs.3.rs-120597/v1.

[29] G. Hatem, R. Mosleh, M. Goossens, D. Khachman, A. Al-Hajje, and S. Awada, "Prevalence and risk factors of migraine headache among university students: A cross-sectional study in Lebanon," *Headache Medicine*, vol. 13, no. 3, pp. 213–221, Oct. 2022, doi: 10.48208/headachemed.2022.23.

[30] A. de Vitta, R. dal B. Biancon, G. P. Cornélio, T. P. F. Bento, N. M. Maciel, and P. de O. Perrucini, "Primary headache and factors associated in university students: a cross sectional study," *ABCS Health Sciences*, vol. 46, Mar. 2021, doi: 10.7322/abcshs.2020005.1793.

[31] X. Gu and Y. J. Xie, "Migraine attacks among medical students in Soochow university, southeast China: A cross-sectional study," *J Pain Res*, vol. 11, pp. 771–781, Apr. 2018, doi: 10.2147/JPR.S156227.

[32] O. Flynn, B. M. Fullen, and C. Blake, "Migraine in University Students: A Systematic Review and Meta-Analysis," *European Journal of Pain*, Jan. 2022, doi: 10.1002/ejp.2047.

[33] A. Rustom, F. Audi, H. al Samsam, R. Nour, A. M. Mursi, and I. Mahmoud, "Migraine awareness, prevalence, triggers, and impact on university students: a cross-sectional study," *Egyptian Journal of Neurology, Psychiatry and Neurosurgery*, vol. 58, no. 1, Dec. 2022, doi: 10.1186/s41983-022-00555-w.

[34] M. J. Marmura, "Triggers, Protectors, and Predictors in Episodic Migraine," *Current Pain and Headache Reports*, vol. 22, no. 12. Current Medicine Group LLC 1, Dec. 01, 2018. doi: 10.1007/s11916-018-0734-0.

[35] M. K. Demirkirkan, H. Ellidokuz, and A. Boluk, "Prevalence and clinical characteristics of migraine in university students in Turkey," *Tohoku Journal of Experimental Medicine*, vol. 208, no. 1, pp. 87–92, Dec. 2005, doi: 10.1620/tjem.208.87

[36] N. Karsan and P. J. Goadsby, "Biological insights from the premonitory symptoms of migraine," *Nature Reviews Neurology*, vol. 14, no. 12. Nature Publishing Group, pp. 699–710, Dec. 01, 2018. doi: 10.1038/s41582-018-0098-4.

[37] Timy Fukui, P., Rachel Tranquillini Gonçalves, T., Giunchetti Strabelli, C., Maria Fernandes Lucchino, N., Cunha Matos, F., Pinto Moreira dos Santos, J., Zukerman, E., Zukerman-Guendler, V., Prieto Mercante, J., Rodrigues Masruha, M., Sávio Vieira, D., Fernando Prieto Peres, M., & Prieto Peres -Rua, M. F. (2008). Trigger facTors in Migraine PaTienTs. In *Arq Neuropsiquiatr* (Vol. 66, Issue A) .

[38] “Triggering-Factors-of-Migraine-dataset-/Prevalence and Triggering Factors of Migraine in University Students of Bangladesh .csv at main · Zawwad26/Triggering-Factors-of-Migraine-dataset-.”https://github.com/Zawwad26/Triggering-Factors-of-Migraine-dataset-/blob/main/Prevalence%20and%20Triggering%20Factors%20of%20Migraine%20in%20University%20Students%20of%20Bangladesh%20.csv (accessed Mar. 13, 2023).

[39] “Accuracy, Precision, Recall & F1-Score - Python Examples - Data & Digital.” https://vitalflux.com/accuracy-precision-recall-f1-score-python-example/ (accessed Feb. 27, 2023).