JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 07 Issue 01
2024
Research Article
https://doi.org/10.56979/701/2024

Predicting Pediatric Diarrhea in Pakistan: Impact of Water Quality

Laraib Noor[1*], Muhammad Rashad[1], Umar Rauf[1], and Muhammad Azam[1]

[1]Superior University, Lahore, 54000, Pakistan.
[*]Corresponding Author: Laraib Noor. Email: Laraibnoor361@gmail.com

Received: January 22, 2024 **Accepted:** April 04, 2024 **Published:** June 01, 2024

Abstract: According to 2023 study, 45.54% of Pakistan's population is under 18 years old, with 13.34% being under 5 years old. A 2023 report indicated that there are 6.4 million cases of pediatric diarrhea. Infectious disease account for 70% of mortality in Pakistan, with 60% of these deaths related to diarrhea. This study aims to analyze the impact of poor water quality on the incidence of diarrhea. In environments with inadequate sanitation, hygiene, and access to clean, safe water for drinking, cooking and cleaning, infections are more likely to occur. The hypothesis posists that investing in water infrastructure, strictly enforcing water quality regulations, and expanding access to safe drinking water through purification and treatment techniques will reduce diarrheal cases. Using a Chi-square test to assess associations at a 95% confidence interval, we found a statistically significant association between the use of water in cooking food (CF) and the incidence of diarrhea, with 49.9% of deaths attributed to diarrhea and children aged 0-1 (male) being the most effected. Employing various machine learning techniques, the support vector machine (SVM) emerged as the most accurate, with 88% accuracy, 93% sensitivity, and 83% specificity. These results can inform the development of appropriate policies to ensure the provision of clean water and mitigate the incidence of diarrhea.

Keywords: Diarrhea; Water Quality; Chi-Square; Naïve Bayes; Logistic Regression; Decision Tree; Random Forest; Support Vector Machine.

1. Introduction

The materials or substances that are found in the environment and used by living things are referred to as natural resources. Natural resources(fossil fuels, minerals, metals, water, air, and solar energy are abundant in Pakistan and have a big impact on the country's economy and way of life. Water resources support the maintenance of a healthy environment, including the preservation of natural ecosystems. Access to reliable water is imperative for facilitating essential medical and sanitary procedures within healthcare environments, thereby guaranteeing the provision of safe and efficient healthcare services. Many other health issues are also increasing in Pakistan. The primary source of contamination is the widespread release of sewage, or feces, into drinking water systems. The discharge of toxic chemicals from industrial effluents and the accidental entrance of fertilizers and pesticides from agricultural sources into water bodies are two other sources of contamination. A number of diseases can spread as a result of poor water quality. Chemical Contamination: Drinking water contaminated with pesticides, industrial chemicals, or heavy metals can lead to various health issues, including neurological abnormalities, heavy metal toxicity, and cancer. Exposure to dirty water can lead to the development of skin disorders such dermatitis, rashes, and infections. Human activities lead to waterborne illnesses, accounting for roughly 80% of all diseases and contributing to 33% of fatalities. The main cause of microbial contamination is the intertwining of drinking water supply and sewer lines. Surface water is used for drinking in many rural regions of Pakistan, where chlorination and slow sand filtering are not frequently carried out at filtration stations. As per the World Health Organization (WHO), diarrhea is a major contributor to illness and mortality within the nation, particularly among children below the age of five. During September 2021, diarrhea emerged as a notable public health issue in Pakistan, specifically affecting the younger population.

The occurrence of diarrhea has been linked to multiple factors, such as restricted availability of safe drinking water, substandard sanitation infrastructure, insufficient healthcare provisions, and limited understanding of appropriate hygiene practices in certain areas. More than 66% of Pakistani homes use water that is tainted with germs, leading to widespread diarrheal illness as well as other illnesses including hepatitis. Cholera, Amoebiasis, gastroenteritis, giardiasis, and Campylobacteriosis are just a few of the organisms that can cause diarrheal disease that are spread by contaminated water. A major public health concern is diarrhea, particularly in areas with limited access to sanitary facilities and clean drinking water. Definition of diarrhea is having three or more instances of daily loose, watery feces. Waterborne Diarrhea is the term used to describe Diarrhea that is brought on by drinking tainted or unclean water. Many factors can cause the diarrhea among children all over the world. But the main factor is dirty water which is the main issue in Pakistan. The objectives of my study are in terms of drinking water quality, Pakistan is ranked 80th number out of 122 countries. To overcome the mortality rate and to aware the people specially mothers to take serious steps and awareness of the importance of cleanliness. To make strict rules for factories which are the main cause of this disease? To aware the government about the severe condition of poor quality of water, that they take special steps to control diarrhea by improving the sewage and water cleaning system.

The main concern in this research is the children age, their water resources and have they ever effected by Diarrhea.

In this research, my target population is the rural and urban area of Pakistan which has less clean water, poor sanitation and sewage system and health facilities because in that areas diarrhea may be highly occurred. Significance In the previous study different authors used mostly statistical tools for the analysis or forecasting diarrhea in children under five and take almost same factors/variables like age, mother education, age, children vaccination and breast feeding children. But in this study we have discussed the main factor of diarrhea which is quality of water. We have predicted this disease by using statistical tools and specially Machine Learning tools. By these tools we have checked the sensibility and specificity after checking its accuracy. According to my observation nobody has yet use these tools for the analysis of diarrhea in Pakistan. There are many issues because of contaminated water some of them are controllable and some are severe. The death of children rate is increasing day by day with diarrhea due to weak quality of water. In Pakistan, concerning authorities like water and sanitation Agency (WASA) probably not paying attention to it. Government should take serious action to reduce the number of children infected by diarrhea and solve the given problems of sewage system and drinking water. Different researcher has used different machine learning and statistical tools for the prediction. According to my study for last 3 to 4 years nobody has used AI tools to find the accuracy of this study in Pakistan so and after this work many other authors can also work by using these tools and taking different factors to suggest the government to overcome the death ratio from this disease.

2. Literature Review

Many researchers have been predicting the diarrhea by using different statistical and machine learning tools. Md. Maniruzzaman et al. (2020) investigate the prediction of children diarrhea in Bangladesh by using machine learning algorithm. Age, maternal education, and wealth were found to be significant variables influencing diarrhea in children. Numerous techniques were used, including Logistic Regression, Support Vector Machine (SVM), Quadratic Discriminant Analysis (QDA), and Linear Discriminant Analysis (LDA). The Support Vector Machine with a radial basis kernel outperformed other machine learning methods, achieving the highest accuracy (65.61%), sensitivity (66.27%), and specificity (52.28%), according to the results [1].

Metadel Adane et al. (2017) investigate the facilities of sanitation in slums of Addis Ababa and aim to find the key factor link to accurate Diarrhea among children 0-5 and analysis the data by using binary logistics regression model at 95% C.I. and found that 94.6% facilities of sanitation were unimproved by using multivariable Logistic Regression and finally found that acute diarrhea has been associated with several factors, such as the presence of shared restrooms, the cleanliness of these spaces, their proximity to residential areas, and the negligence with which household waste is disposed of [2].

Adam C Levine et al. (2015) determine whether children with severe Diarrhea are dehydrated in Dhaka by using machine learning algorithm Decision Tree. They led to the development of the first diagnostic tool that was both internally validated and backed by empirical evidence for assessing dehydration in

children with severe diarrhea. In times of resource scarcity, general practice nurses may employ this strategy. The sensitivity was 81%, specificity was 67%, LR+ of the DHAKA Dehydration Tree was 2.5, and severe dehydration's result has an LR-of 0.28 [3].

Metadel Adane1 et al. (2017) analyzed the data using Strata and Multivariable Logistics Regression with 95% C.I. of the people of Ethiopia slums of Addis Ababa wash hands with soap and water to avoid the onset of acute diarrhea in children younger than five years old. The 4.4% households with access to soap and water for hand washing. During the five suggested situations, the average number of times that individuals washed their hands with soap was 19.8 %. Before feeding a child, the caretaker used soap to wash their hands in 24.8% of cases. Before preparing food, 23.8% of people wash their hands.17.1% following manure [4].

Muhammad Ali et al. (2022) Identify children under five who are experiencing diarrhea by utilizing Bivariate and Multivariate Logistic Regression with a 95% Confidence Interval. For estimation with different factors. This study shows that in Pakistan, children who are stunted are more likely to have diarrhea in Multivariate models that look at the entire population, rural areas, and the bottom 40%.According to the Bivariate model, girls in Pakistan are 11% less likely than boys to have diarrhea [5].

Ni Komang Ayu Santikaa et al. (2020) identify the variables associated with diarrhea in children under the age of two by using Chi-Square and Binary Logistic Regression with a p-value of 0.05 and a 95% confidence interval. 17.16% of children in Indonesia under the age of two have diarrhea. Regarding the child's gender, males experienced Diarrhea at a higher rate 51.14% than females 48.86%. About 20.57% of the participants were from lower-class families, and 51.81% of the participants lived in rural areas [6].

Helen Powell, et al. (2023) have used Conditional Logistic Regression for vaccine impact on Diarrhea in Africa (VIDA).The statistical techniques employed in VIDA aim to optimize the utilization of accessible data, enabling the generation of more resilient estimates of the disease burden specific to pathogens. These estimates help underscore the potential impact of effective interventions. By this rotavirus vaccine impact of Diarrhea can be reduce by this [7].

Robinah Nantege et al. (2022) used Bivariate and Multivariate Regression analyses with 95% C.I. The frequency of diarrheal sickness in kids younger than five in Entebbe municipality's shanty settlements was found to be significant, with 62.4% of the 378 children in the study having diarrhea at the outset [8].

Heather K. Amato et.al (2022) researcher used statistical tools for analysis the biogas cook stove intervention and children diarrhea by daily observations and compared to children who were not exposed to biogas stoves, the impact of utilizing a biogas stove on the incidence of diarrhea was stronger in breastfed children than in non-breastfed children, according to research employing univariate log-linear regression models with 95% confidence interval and alpha = 0.10, 0.05. Among the children exposed to the usage of biogas cook stoves, the children who had recently experienced Acute Lower Respiratory Infections (ALRI) had the highest mean risk of diarrhea, measured in person-days [9].

Xiang Yang et.al (2021) researcher used Random Forest Algorithm to analysis the Metrological and social conditions contribute infectious diarrhea in china and discovered that seasonal changes, particularly those brought on by weather patterns, have long been linked to the [11] Na Zhang et.al (2019) Used interrupted time series with quasi-poisoned data; researchers in the Chinese province of Anhui determined that flooded locations had a higher incidence of diarrhea (rate ratio 1.147, 95%CI, Random effect). According to meta-analyses and meta-regression, living close to the Yangtze River has been linked to an increased risk; kids between the ages of 5 and 14 are the most susceptible to diarrhea in flooded areas [10]. [12] Sokhna Thiam et.al (2017) researcher found that the frequency of diarrhea and associated risk factors in Mbour's among less than 5 years old population, Senegal, by using univariate and multivariable logistic regression. According to the survey's results, 26% of youngsters said they had diarrhea during the two weeks it was conducted. With the highest prevalence occurring in the urban central zone (3.3%), and 44.8% in peri-central areas.

3. Materials and Methods

3.1. Dataset

In this research article we have used the qualitative data by survey which is collected from the different areas of rural and urban areas of Punjab in Johar Town, Sheikhupura, Nankana Sahab (2024). We collected total 1500 dataset from these areas by questionnaire.

3.2. Dependent Variables

As a dependent variable in this study, survival from diarrhea was defined as:

0 → (Yes): Survival from diarrhea

1 → (No): Death from diarrhea

3.3. Independent variables

In this study, we have used total 33 independent variables in which age, gender, and variables related to water quality are included some variables are binary and others are Likert scale responses. The main predictors are water quality, source of water and nearby industry and many other variables. Survival from diarrhea is used as dependent variable.

3.4. Statistical Analysis

Utilizing the Chi Square test, one may examine the relationship between categorical variables.(age, water in cooking food, mother HWP, nearby industry, WQ Test, Availability of SW, Water source, Drinkable WS, Water color, Waste SDS, IAU observe SPM) with p-value = 0.05 for testing and 95% confidence level. Age (in years) with a χ^2 = 70.7 and a very low p-value (3.028×10^{-15}), this result shows that $p < 0.05$ so the null hypothesis is rejected.

This indicates that expanding access to safe drinking water, enforcing water quality rules, and investing in water infrastructure do help lower the number of diarrheal cases is supported by a lot of research, and showing a strong relationship between age and other outcomes. Gender with χ^2 = 0.68 and p-value (4.095×10^{-1}) this result shows that $p > 0.05$ so null hypothesis will be accepted and indicating that there is no association between the outcomes and gender. Caregiver education with χ^2 = 3.83 and p-value (2.806×10^{-1}) this result shows that $p > 0.05$ so the findings shows that caregiver education and the efficiency of the interventions to lower the number of diarrheal cases are not significantly associated ,the null hypothesis is accepted since the impact of caregiver education on the outcome is not strongly supported by statistical evidence.

Water in CF with a χ^2 = 252 and a very low p-value (2.434×10^{-53}), this result shows that $p < 0.05$. So the null hypothesis is rejected. This indicates that the alternative hypothesis—which holds that enhancing access to safe drinking water, enforcing water quality standards, and investing in water infrastructure—is strongly supported by the data. Mother HWP with a χ^2 = 43.2 and a very low p-value (2.250×10^{-9}), this result shows that $p < 0.05$. The null hypothesis is rejected. This means that, especially when taking into account the variable of mothers' hand washing practices, there is very strong evidence to support the hypothesis that investing in water infrastructure, enforcing water quality regulations, and expanding access to safe drinking water do help lower the number of diarrheal cases. Nearby industry with a χ^2 = 4.93 and a very low p-value (2.648×10^{-2}), this result shows that $p < 0.05$.

The null hypothesis is rejected. This indicates that, even after accounting for the influence of surrounding industries, there is evidence to support the premise that enhancing access to safe drinking water, enforcing water quality rules, and investing in water infrastructure help reduce the number of diarrheal cases. Water quality Test with a χ^2 = 7.88 and a very low P-value (4.999×10^{-3}), this result shows that $p < 0.05$.it shows association between this intervention. This indicates that there is substantial evidence to support the hypothesis that, as shown by improvements in water quality test results, investing in water infrastructure, upholding water quality regulations, and increasing access to safe drinking water through purification and treatment techniques help lower the number of diarrheal cases.

Reliable WS Test with a χ^2 = 2.61 and a very low p-value (1.060×10^{-13}), this result shows that $p > 0.05$. The data examined do not provide sufficient evidence to reject the null hypothesis. Association is not statistically significant. Water treatment Test with a χ^2 = 0.40 and a very low P-value (5.254×10^{-1}), this result shows that $p > 0.05$.it shows a weak association between water tr and diarrheal cases and accepted the null hypothesis. Availability of safe drinking water with a χ^2 = 207 and a very low P-value (1.334×10^{-44}), this result shows that $p < 0.05$; it shows strong association between variables. The data provides strong evidence to reject the null hypothesis [33] [34].

Water color with a χ^2 = 161 and a very low p-value (1.122×10^{-34}), this result shows that $p < 0.05$.It shows a strong association between water color and diarrheal cases and rejects the null hypothesis. More specifically, the correlation between diarrheal incidents and water color, an indicator of water quality, highlights how critical it is to enhance water quality through these treatments. Water sewage disposal system with a χ^2 = 96.2 and a very low p-value (8.375×10^{-23}), this result shows that $p < 0.05$ and shows a

strong association between water SDS and occurrence of diarrheal cases. Reject the null hypothesis and is statistically significant. The reduction of diarrheal illnesses is significantly influenced by the quality of water, sewage, and disposal systems, underscoring the need of appropriate infrastructure and management in public health campaigns [32] [35].

DWI projects Test with a χ^2 = 0.002 and a very low P-value (9.620×10^{-1}), this finding indicates that P > 0.05, suggesting that there may not be much of a relationship between the occurrence of diarrheal illnesses and the development of water infrastructure projects.as indicated by a very low chi-square value of 0.002. There is no statistical significance in the results. Put another way, we are unable to reject the null hypothesis based on the data that we have examined. IAU observes SPM Test with a χ^2 = 4.036 and a very low p-value (4.453×10^{-2}), this result shows that p < 0.05. A moderate correlation between the incidence of diarrheal cases and the adoption of stringent pollution control measures in these industries is shown by a χ^2 = 4.036.

This indicates that there is statistical significance in the findings. Stated differently, there is enough evidence to reject the null hypothesis based on the data that have been analyzed.

4. Results

Results shows that among all age groups total (439) and Non survival are (61.0%) 0-1 year children most affected by this among them and are male. Total 598 people use Tap water to cooking food and death ratio is (70.9%). 52.8% children did not survive in which industries are near to their residential areas. Total 1205 mother did not have WQ test and death ratio is more in this (51.7%).Total 1109 have no reliable WS and (50.2%) did not survive.

Table 1. Dependent and independent variables association

Variables	χ^2	P-value	significant
Age(in years)	70.7	3.028×10^{-15}	True
Gender	0.68	4.095×10^{-1}	False
caregiver education	3.83	2.806×10^{-1}	False
water in CF	252	2.434×10^{-53}	True
mother HWP	43.2	2.250×10^{-9}	True
nearby industry	4.93	2.645×10^{-2}	True
WQ Test	7.88	4.999×10^{-3}	True
Reliable WS	2.61	1.060×10^{-1}	False
Water tr	0.40	5.254×10^{-1}	False
Availability of SW	207	1.334×10^{-44}	True
Water source	49.0	5.726×10^{-10}	True
Drinkable ws	120	8.033×10^{-26}	True
Water color	161	1.122×10^{-34}	True
Waste SDS	96.2	8.375×10^{-23}	True
DWI projects	0.002	9.620×10^{-1}	False
IAU observe SPM	4.036	4.453×10^{-2}	True

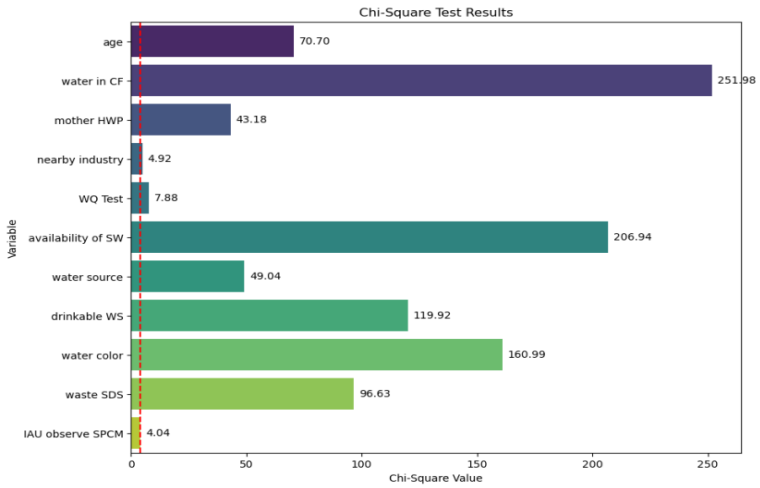

Figure 1. Chi Square Test

Table 2. Odd Ratios (OR) for the effect of Non-survival from diarrhea

Variables	Coefficient	Odd Ratio	P value	95% C.I for OR	
Age (in years)				**Upper**	**Lower**
0-1	0.555	1.743	$2.668x10^{-5}$	0.803	0.284
2-3	0.174	1.190	$1.497x10^{-1}$	0.410	-0.058
4-5	-0.276	0.759	$3.057x10^{-2}$	-0.041	-0.527
Above 6	-0.453	0.636	$5.602x10^{-3}$	-0.146	-0.755
Gender					
Female	-0.019	0.981	$7.966x10^{-1}$	0.124	-0.168
Male	0.019	1.020	$7.967x10^{-1}$	0.168	-0.124
Caregiver education					
Higher	0.027	1.028	$8.713x10^{-1}$	0.362	-0.307
No formal education	-0.230	0.795	$7.915x10^{-2}$	0.026	-0.471
Primary	0.175	1.191	$1.247x10^{-1}$	0.404	-0.031
Secondary	0.027	1.028	$8.245x10^{-1}$	0.265	-0.218
Water in CF					
Boiled	-0.541	0.582	$1.495x10^{-3}$	-0.208	-0.884
Canal	0.424	1.527	$1.890x10^{-2}$	0.784	0.080
Purified water	-0.612	0.542	$8.727x10^{-3}$	-0.154	-1.074
Tap	0.923	2.518	$1.222x10^{-11}$	1.178	0.660
Tube well	-0.194	0.824	$1.994x10^{-1}$	0.111	-0.489
Mother HWP					
Always	-0.230	0.795	$1.205x10^{-1}$	0.053	-0.510
Frequently	0.143	1.153	$3.555x10^{-1}$	0.450	-0.154
Never	0.188	1.206	$5.460x10^{-1}$	0.795	-0.411
Rare	-0.101	0.904	$6.206x10^{-1}$	0.321	-0.501
Nearby Industry					
No	-0.127	0.881	$8.902x10^{-2}$	0.022	-0.270
Yes	0.127	1.135	$8.905x10^{-2}$	0.269	-0.228
WQ Test					
No	0.056	1.057	$5.361x10^{-1}$	0.228	-0.125
Yes	-0.056	0.946	$5.360x10^{-1}$	0.126	-0.228
Reliable WS					
No	-0.064	0.938	$4.605x10^{-1}$	0.109	-0.230
Yes	0.064	1.066	$4.606x10^{-1}$	0.230	-0.109
Water Treatment					
No	0.050	1.051	$5.414x10^{-1}$	0.207	-0.111
Yes	-0.050	0.951	$5.413x10^{-1}$	0.111	-0.208
Availability of SW					
Excellent	-0.580	0.560	$5.934x10^{-2}$	0.003	-1.219
Good	-0.155	0.856	$4.337x10^{-1}$	0.197	-0.541
Normal	-0.132	0.876	$4.091x10^{-1}$	0.212	-0.429
Poor	0.868	2.382	$3.373x10^{-8}$	1.179	0.574
Water source					
Canal	-0.104	0.901	$5.177x10^{-1}$	0.223	-0.407
Filter	0.253	1.288	$1.636x10^{-1}$	0.601	-0.100
Mineral Water	0.052	1.053	$8.013x10^{-1}$	0.473	-0.321
Tap	0.186	1.204	$1.762x10^{-1}$	0.450	-0.083
Tube Well	-0.386	0.680	$6.045x10^{-3}$	-0.123	-0.668
Drinkable WS					
Bottle	-0.082	0.921	$5.239x10^{-1}$	0.168	-0.328
Clay Pot	0.115	1.122	$3.406x10^{-1}$	0.356	-0.116
Other	-0.251	0.778	$9.318x10^{-2}$	0.047	-0.540

Tap	0.218	1.243	$1.655x10^{-1}$	0.510	-0.082
Water Color					
Clear &transparent	-0.577	0.562	$3.310x10^{-3}$	-0.191	-0.976
Contain Particles	0.265	1.304	$5.802x10^{-2}$	0.548	-0.003
Discolored	0.340	1.405	$5.193x10^{-3}$	0.590	0.105
Unclear	-0.029	0.972	$8.138x10^{-1}$	0.210	-0.268
Waste SDS					
No	0.313	1.368	$1.555x10^{-4}$	0.479	0.159
Yes	-0.313	0.731	$1.548x10^{-4}$	-0.160	-0.479
DWI Projects					
No	-0.245	0.783	$1.177x10^{-2}$	-0.062	-0.440
Yes	0.245	1.278	$1.177x10^{-2}$	0.439	0.062
IAU Observe SPCM					
No	0.103	1.108	$5.103x10^{-1}$	0.398	-0.208
Yes	-0.103	0.902	$5.103x10^{-1}$	0.212	-0.398
Diarrhea Category					
Acute bloody diarrhea	0.397	1.488	$4.311x10^{-3}$	0.669	0.127
Acute watery diarrhea	-0.382	-0.185	$2.601x10^{-4}$	-0.185	-0.587
Persistent diarrhea	-0.016	0.209	$8.896x10^{-1}$	0.209	-0.223

4.1 Explanation

A statistical tool for assessing the degree of correlation between an exposure and an outcome is the odds ratio (OR). When analyzing case-control studies and logistic regression models in particular, odds ratios can be used to better understand the direction and strength of the relationship between variables. In the above table odd ratios are finding by using 95% C.I. The results of the analysis show that children under the age of one year have a significantly higher odds ratio (OR = 1.743, 95% CI: 0.284-0.803, $p = 2.668 \times 10^{-5}$) than children over the age of six years (OR = 0.636, 95% CI: -0.755 to -0.146, $p = 5.602 \times 10^{-3}$), suggesting a significant age-related risk difference.so children of younger age has more related to non-survival from diarrhea. Compared to acute watery diarrhea, which has a negative association with an odds ratio of -0.185 (p-value = $2.601x10^{-4}$), acute bloody diarrhea has a positive odds ratio of 1.488 (p-value = $4.311x10^{-3}$), indicating a significant increase in the likelihood. For both situations, the confidence intervals (C.I.) do not include 1, demonstrating their statistical significance. This implies that, in the context of this study, acute bloody diarrhea is more strongly associated with the outcome than acute watery diarrhea. From above table the overall Using tap water considerably enhances the likelihood of the outcome compared to other factors, as indicated by the greatest odds ratio (OR) of 2.518 (p-value = $1.222x10^{-11}$) for "Tap water" as a water source in the presented data. "Acute watery diarrhea" had the lowest odds ratio (OR of -0.185; p-value = $2.601x10^{-4}$), indicating a substantial reduction in the chance of the outcome when compared to other variables.

5. Classifiers

5.1. Logistic Regression

For binary classification tasks, a logistic regression classifier is a statistical model that estimates the likelihood that a given input will fall into one of two categories. It uses the logistic (sigmoid) function, which produces values between 0 and 1, to represent the link between the input features and the probability of a particular result. It divides inputs into one of the two groups by applying a decision threshold, usually set at 0.5.

5.2. Decision Tree

A non-parametric supervised learning method called decision trees is used for both regression and classification. Its internal nodes, leaf nodes, branches, and root node make up its hierarchical tree structure.

5.3. Gaussian NB

An algorithm for supervised machine learning is the Naïve Bayes classifier. They accomplish categorization jobs by applying probability principles. It is also called normal distribution. It makes the assumption that each feature is independent, which may not always be the case.

5.4. Support Vector Machine

A machine learning algorithm called the support vector machine (SVM) draws boundaries between data points using preset outputs, labels, or classes. To tackle difficult problems in regression, outlier identification, and classification, it makes use of supervised learning models. SVMs excel at solving binary classification problems, which divide the elements of a data set into two groups.

5.5. Random Forest

Random forest is a machine learning technique used for classification and regression issues. It achieves extraordinary prediction accuracy by utilizing the power of decision tree aggregation.

Table 3. Results of the Classifiers

Classifier Types	Measures		
	Accuracy (%)	**Sensitivity (%)**	**Specificity (%)**
Logistic Regression	87	93	81
Decision Tree	77	76	80
Gaussian NB	84	90	78
Support Vector Machine	88	93	83
Random Forest	87	88	86

The performance metrics of several classifiers, such as accuracy, sensitivity, and specificity, are shown in the table. With 88% accuracy, 93% sensitivity, and 83% specificity, the Support Vector Machine performs best overall. Strong results are also shown by Random Forest and Logistic Regression, which yielded 87% Accuracy, 88% Sensitivity, and 81% Specificity and 87% Accuracy, 88% Sensitivity, and 86% Specificity, respectively. With 77% Accuracy, 76% Sensitivity, and 80% Specificity, the Decision Tree classifier performs the worst.

Table 4. The Area of the Receiver Operating Characteristic (ROC) curve

Classifier	AUC Score
Logistic Regression	0.922
Decision Tree	0.763
Gaussian NB	0.890
Support Vector Machine	0.934
Random Forest	0.926

Based on their Area under the Curve (AUC) values, the classifiers' performances are displayed in the table. Out of all the classifiers, the Support Vector Machine performed the best, earning the highest AUC score of 0.934. Then, with scores of 0.926 and 0.922, respectively, came the Random Forest and Logistic Regression. With an AUC score of 0.763, the Decision Tree performed comparatively worse in this investigation.

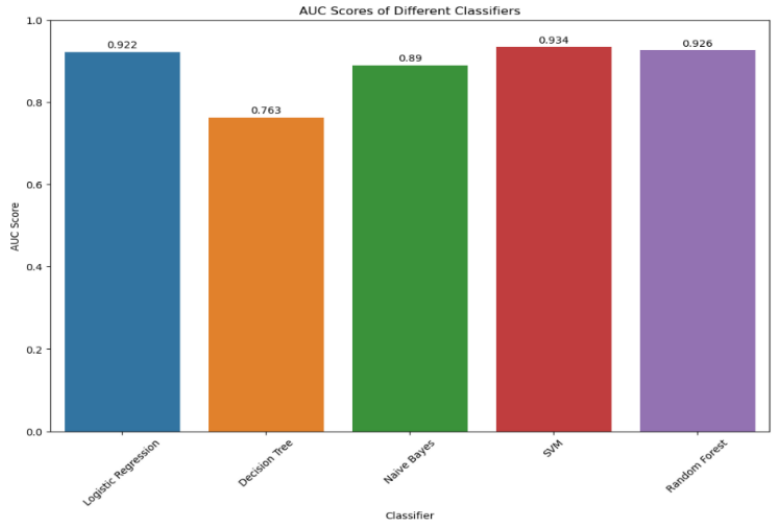

Figure 2. Comparison of Classifiers

6. Comparatively analyzing the classifiers

Table displays a comparison of these classifiers' output.2. I have employed five machine learning classifiers: Random Forest, Naïve Bayes, Decision Tree, Support Vector Machine, and Logistic Regression.

The decision tree has the lowest AUC Score, accuracy, specificity, and sensitivity. But Support vector machine gives the highest accuracy, sensitivity, specificity and AUC score among all other algorithm.so SVM is the best classifier to predict the water is the main element which causes the diarrhea among. Mostly the children age 0-1 years most affected by this and non-survival. By support vector machine we have found that availability of SW (Poor) most areas are that which has poor water quality children.

7. Limitations and Scope

In this research work we have taken water quality as the main variable which is the main cause of diarrhea and predict by using Machine Learning algorithms. While many other variables like toilet condition, mother education breastfeeding these can be taken as variables and predict using ML algorithm.

8. Conclusions

In this study we have investigate that survival from diarrhea and also suggest a prediction model for the prediction of survival or non-survival from diarrhea. This study shows that some main factors like age water quality, nearby industry, gender, mother HWP, availability of SW and many other factors are associated with the survival from diarrhea and SVM gives the better result to show the high specificity and accuracy.

References

1. Maniruzzaman, I. S., & Abedin, M. (2020). Prediction of childhood diarrhea in Bangladesh using machine learning approach. *Insights Biomed Res*, 4(1), 111-116.
2. Adane, M., Mengistie, B., Kloos, H., Medhin, G., & Mulat, W. (2017). Sanitation facilities, hygienic conditions, and prevalence of acute diarrhea among under-five children in slums of Addis Ababa, Ethiopia: baseline survey of a longitudinal study. PloS one, 12(8), e0182783.
3. Levine, A. C., Glavis-Bloom, J., Modi, P., Nasrin, S., Rege, S., Chu, C., ... & Alam, N. H. (2015). Empirically derived dehydration scoring and decision tree models for children with diarrhea: assessment and internal validation in a prospective cohort study in Dhaka, Bangladesh. Global Health: Science and Practice, 3(3), 405-418.
4. Adane, M., Mengistie, B., Mulat, W., Medhin, G., & Kloos, H. (2018). The most important recommended times of hand washing with soap and water in preventing the occurrence of acute diarrhea among children under five years of age in slums of Addis Ababa, Ethiopia. Journal of community health, 43, 400-405.
5. Ali, M., Abbas, F., & Shah, A. A. (2022). Factors associated with prevalence of diarrhea among children under five years of age in Pakistan. Children and Youth Services Review, 132, 106303.
6. Komang, N., Santika, A., Efendi, F., Rachmawati, P. D., Mishbahatul, E., & Kusnanto, K. (2020). Children and Youth Services Review Determinants of diarrhea among children under two years old in Indonesia. Child Youth Serv Rev, 111(104838), 10-1016.
7. Walker, C. L. F., Rudan, I., Liu, L., Nair, H., Theodoratou, E., Bhutta, Z. A., ... & Black, R. E. (2013). Global burden of childhood pneumonia and diarrhoea. The Lancet, 381(9875), 1405-1416.
8. Powell, H., Liang, Y., Neuzil, K. M., Jamka, L. P., Nasrin, D., Sow, S. O., ... & Kotloff, K. L. (2023). A description of the statistical methods for the Vaccine Impact on Diarrhea in Africa (VIDA) study. Clinical Infectious Diseases, 76(Supplement_1), S5-S11.
9. Nanan, D., White, F., Azam, I., Afsar, H., & Hozhabri, S. (2003). Evaluation of a water, sanitation, and hygiene education intervention on diarrhoea in northern Pakistan. Bulletin of the World Health Organization, 81, 160-165.
10. Khatun, U. H. F., Malek, M. A., Black, R. E., Sarkar, N. R., Wahed, M. A., Fuchs, G., & Roy, S. K. (2001). A randomized controlled clinical trial of zinc, vitamin A or both in undernourished children with persistent diarrhea in Bangladesh. Acta paediatrica, 90(4), 376-380.
11. Woldu, W., Bitew, B. D., & Gizaw, Z. (2016). Socioeconomic factors associated with diarrheal diseases among under-five children of the nomadic population in northeast Ethiopia. Tropical medicine and health, 44, 1 – 8.
12. Nantege, R., Kajoba, D., Ddamulira, C., Ndoboli, F., & Ndungutse, D. (2022). Prevalence and factors associated with diarrheal diseases among children below five years in selected slum settlements in Entebbe municipality, Wakiso district, Uganda. BMC pediatrics, 22(1), 394.
13. Shah, S. M., Yousafzai, M., Lakhani, N. B., Chotani, R. A., & Nowshad, G. (2003). Prevalence and correlates of diarrhea. The Indian Journal of Pediatrics, 70, 207-211.
14. Ogwel, B., Mzazi, V., Nyawanda, B. O., Otieno, G., & Omore, R. (2024). Predictive modeling for infectious diarrheal disease in pediatric populations: A systematic review. Learning Health Systems, 8(1), e10382.
15. Sinmegn Mihrete, T., Asres Alemie, G., & Shimeka Teferra, A. (2014). Determinants of childhood diarrhea among underfive children in Benishangul Gumuz regional state, north West Ethiopia. BMC pediatrics, 14, 1-9.
16. Muluken Dessalegn, M. D., Abera Kumie, A. K., & Worku Tefera, W. T. (2011). Predictors of under-five childhood diarrhea: Mecha District, West Gojam, Ethiopia.
17. Amato, H. K., Hemlock, C., Andrejko, K. L., Smith, A. R., Hejazi, N. S., Hubbard, A. E., ... & Pokhrel, A. (2022). Biodigester cookstove interventions and child diarrhea in semirural Nepal: a causal analysis of daily observations. Environmental Health Perspectives, 130(1), 017002.
18. Umair Ahmed, U. A., Rafia Mumtaz, R. M., Hirra Anwar, H. A., Shah, A. A., Rabia Irfan, R. I., & García-Nieto, J. (2019). Efficient water quality prediction using supervised machine learning.
19. Yang, X., Xiong, W., Huang, T., & He, J. (2021). Meteorological and social conditions contribute to infectious diarrhea in China. Scientific Reports, 11(1), 23374.
20. Troeger, C., Forouzanfar, M., Rao, P. C., Khalil, I., Brown, A., Reiner, R. C., ... & Mokdad, A. H. (2017). Estimates of global, regional, and national morbidity, mortality, and aetiologies of diarrhoeal diseases: a systematic analysis for the Global Burden of Disease Study 2015. The Lancet infectious diseases, 17(9), 909-948.
21. Anderson, J. D., Bagamian, K. H., Muhib, F., Amaya, M. P., Laytner, L. A., Wierzba, T., & Rheingans, R. (2019). Burden of enterotoxigenic Escherichia coli and shigella non-fatal diarrhoeal infections in 79 low-income and lower middle-income countries: a modelling analysis. The Lancet Global Health, 7(3), e321-e330.

22. Zhang, N., Song, D., Zhang, J., Liao, W., Miao, K., Zhong, S., ... & Huang, C. (2019). The impact of the 2016 flood event in Anhui Province, China on infectious diarrhea disease: An interrupted time-series study. Environment international, 127, 801-809.
23. Platts-Mills, J. A., Babji, S., Bodhidatta, L., Gratz, J., Haque, R., Havt, A., ... & Houpt, E. R. (2015). Pathogen-specific burdens of community diarrhoea in developing countries: a multisite birth cohort study (MAL-ED). The Lancet Global Health, 3(9), e564-e575.
24. Fischer Walker, C. L., Perin, J., Aryee, M. J., Boschi-Pinto, C., & Black, R. E. (2012). Diarrhea incidence in low-and middle-income countries in 1990 and 2010: a systematic review. BMC public health, 12, 1-7.
25. Thiam, S., Diène, A. N., Fuhrimann, S., Winkler, M. S., Sy, I., Ndione, J. A., ... & Cissé, G. (2017). Prevalence of diarrhoea and risk factors among children under five years old in Mbour, Senegal: a cross-sectional study. Infectious diseases of poverty, 6(04), 43-54.
26. Alebel, A., Tesema, C., Temesgen, B., Gebrie, A., Petrucka, P., & Kibret, G. D. (2018). Prevalence and determinants of diarrhea among under-five children in Ethiopia: a systematic review and meta-analysis. PloS one, 13(6), e0199684.
27. Arif, A., & Arif, G. M. (2012). Socio-economic determinants of child health in Pakistan. Academic Research International, 2(1), 398.
28. Stanly, A. M., Sathiyasekaran, B. W., & Palani, G. (2009). A population based study of acute diarrhoea among children under 5 years in a rural community in South India. Sri Ramachandra Journal of Medicine, 1(1), 17.
29. Solomon, E. T., Robele, S., Kloos, H., & Mengistie, B. (2020). Effect of household water treatment with chlorine on diarrhea among children under the age of five years in rural areas of Dire Dawa, eastern Ethiopia: a cluster randomized controlled trial. Infectious Diseases of Poverty, 9, 1-13.
30. Ma, T., Villot, C., Renaud, D., Skidmore, A., Chevaux, E., Steele, M., & Guan, L. L. (2020). Linking perturbations to temporal changes in diversity, stability, and compositions of neonatal calf gut microbiota: prediction of diarrhea. The ISME journal, 14(9), 2223-2235.
31. Grampurohit, S., & Sagarnal, C. (2020, June). Disease prediction using Machine Learning algorithms. In 2020 international conference for emerging technology (INCET) (pp. 1-7). IEEE.
32. Shaker, B., Ullah, K., Ullah, Z., Ahsan, M., Ibrar, M., & Javed, M. A. (2023, November). Enhancing grid resilience: Leveraging power from flexible load in modern power systems. In 2023 18th International Conference on Emerging Technologies (ICET) (pp. 246-251). IEEE.
33. Munir, A., Sumra, I. A., Naveed, R., & Javed, M. A. (2024). Techniques for Authentication and Defense Strategies to Mitigate IoT Security Risks. Journal of Computing & Biomedical Informatics, 7(01).
34. Ali, H., Iqbal, M., Javed, M. A., Naqvi, S. F. M., Aziz, M. M., & Ahmad, M. (2023, October). Poker Face Defense: Countering Passive Circuit Fingerprinting Adversaries in Tor Hidden Services. In 2023 International Conference on IT and Industrial Technologies (ICIT) (pp. 1-7). IEEE.
35. Khan, M. F., Iftikhar, A., Anwar, H., & Ramay, S. A. (2024). Brain Tumor Segmentation and Classification using Optimized Deep Learning. Journal of Computing & Biomedical Informatics, 7(01), 632-640.