Arch Biol Sci. 2023;75(4):459-474 https://doi.org/10.2298/ABS230909037G

Use of data mining to establish associations between Indian marine fish catch and environmental data

Joseph Gladju[1], Ayyasamy Kanagaraj[2,*] and Biju Sam Kamalam[3]

[1]*Department of Computer Science, Nallamuthu Gounder Mahalingam College*, Pollachi, Tamil Nadu - 642001, India
[2]*Department of Computer Science, Kristu Jayanti College*, Bengaluru, Karnataka - 560077, India
[3]*ICAR-Directorate of Coldwater Fisheries Research*, Bhimtal, Uttarakhand - 263136, India

***Corresponding author**: kanagaraj.a@kristujayanti.com; a.kanagaraj@gmail.com

Received: September 9, 2023; **Revised**: October 11, 2023; **Accepted**: October, 20 2023; **Published online**: November 16, 2023

Abstract: For decades, changes in fish catch composition and the marine environment have been monitored worldwide and recorded in databases like FAO FishStatJ and the European Union Copernicus Marine Service. However, the complexity and high variability in the dataset makes it challenging to find meaningful information through conventional data analytical methods. Therefore, in this pilot data mining study, we employed association rule mining algorithms (Apriori, ECLAT, and FP-Growth) to find frequently occurring itemsets in the fish-catch composition and marine environment data of the west and east coasts of India during the past decade (2011-2020). Firstly, the inherent spatial and temporal variations in fish-catch composition and marine environment (sea surface temperature and chlorophyll) on the west and east coasts of India were statistically analyzed and described. Then, the data were preprocessed, selected, and transformed into categorical attributes. By applying the association rule mining algorithms written in the Python language in the Google Colab workspace, we obtained frequent itemsets of fish catch and marine environment with different levels of minimum support and confidence. The preliminary results showed linear and inverse associations between changes in the sea surface temperature, chlorophyll concentration, and major catch groups, such as anchovies, Indian oil sardine, Indian mackerel, hairtails, butterfish-pomfrets, Bombay duck, flatfish, tunas, giant tiger prawn, crabs, lobsters, and cephalopods. Among the tested data mining algorithms, FP-Growth was found to be more efficient and reliable in finding associations between the spatiotemporal dynamics of the marine environment and fish distribution and abundance. Therefore, it can be potentially used to support marine fisheries' resource assessment and management strategies after refinement.

Keywords: association rule mining (ARM), marine fish production, Apriori, ECLAT, FP-Growth

Abbreviations: ARM – association rule mining; ECLAT – equivalence class clustering and bottom-up lattice traversal; FP-Growth – frequent pattern growth; SST – sea surface temperature

INTRODUCTION

Oceans and seas comprise the largest ecosystem on earth. According to the World Register of Marine Species (WoRMS), they harbor more than 240,000 known marine organisms and provide food and livelihood to more than three billion people [1,2]. Globally, 78.8 million tons of marine fish and other aquatic organisms were caught by the fishing industry in 2020, 4.7% of which was contributed by India [1]. In terms of marine resources, India is naturally endowed with 8,118 km of coastline, 0.53 km^2 of continental shelf area, and 2.02 million km^2 of exclusive economic zone. A wide variety of fish species, including pelagic and demersal fishes, crustaceans, mollusks, and other aquatic creatures are found in India's marine regions. From this pool of multispecies resources, annually about four million tons of living marine organisms are caught using an array of fishing gears and crafts. Sardines, mackerel, anchovies, tuna, and shrimp are some of the marine species that are most often captured. As India has an extensive tradition of fishing, marine capture fisheries provide livelihood to nearly five million people in more than 3,200 fishing villages [3]. The sustainability of marine fisheries is thus critical from both ecological and socioeconomic standpoints.

Data related to fish catch and the environment form the backbone of fisheries' management programs that are drafted to ensure the sustainable exploitation of marine resources. Therefore, extensive efforts are being made to systematically collect and analyze marine fisheries' information on global and national scales [4]. In India, much of the fish-catch information has conventionally come from long-running fisheries' observer programs (from 1950 onwards), with representative coverage in each maritime state and union territory. This data is collected by a national research institution (ICAR-Central Marine Fisheries Research Institute), consolidated in the national repository (National Marine Fisheries Data Center) and a summary of it is reported in an international open-source database like FAO FishStatJ [5]. Worldwide, the uptake of electronic monitoring of fisheries has been slow despite the availability of advanced global positioning system (GPS) technology, digital cameras, and network communication [6]. On the other hand, marine observation programs and monitoring networks across the globe such as the Copernicus Marine Service (European Union) are routinely collecting gargantuan volumes of complex multidimensional ocean environment data through satellites and hydrological sensor networks [7]. Obviously, the information stored in this massive heterogeneous and complex spatiotemporal dataset on fish-catch composition and the marine environment cannot be manually interpreted to identify ecological relationships and causal associations and forecast potential fish-catch dynamics in marine fisheries. As spatiotemporal marine fish-catch and environmental data, such as sea surface temperature and chlorophyll, are dynamic, widespread, and multi-sourced, they need to be organized, processed and analyzed to make them useful for decision support applications [8,9].

Artificial intelligence has already crept into various aspects of marine fishing for the prediction of fish biomass and species composition, real-time tracking of fishing efforts, identification of profitable and safe fishing routes and spots, and decision support related to fisheries' management and regulatory strategies [6,10]. Integration of artificial intelligence in fisheries databases might enable the further discovery of meaningful information and hidden knowledge related to cause-and-effect associations, trends, and anomalies, and strategic solutions for fisheries' management [7]. Data mining, a branch of artificial intelligence, has been previously employed in capture fisheries to study the dynamic relationship between environmental changes and spatiotemporal abundance and distribution of aquatic organisms. Examples of data mining applications in environment-resource association studies is the identification of ecosystem patterns (El Niño and cold-warm regimes) associated with fluctuations in sardine (*Sardinops sagax*) and anchovy (*Engraulis ringens*) landings in Chile, using *k*-means combined with multivariate analysis and time-series decomposition [11]. Likewise, a spatiotemporal assignment mining model that incorporated fuzzy knowledge, neighbor rules and a decision table was used to find the link between temperature variation and the occurrence of an active fishing ground with more fish assemblages in the Yellow Sea, China [12]. Genetic programming was used to identify relevant hydrological indicators (water flow) that reflected the abundance and diversity of the fish community in an Illinois river [13], and a self-organizing feature map was used to find relationships between water quality and fish community composition in the Dahan River in Taiwan [14]. Overall, it can be noted that different data mining applications have shown the potential ability to decipher the complex relationship between environmental changes and the life cycle processes of fish species.

Over the years, data mining has evolved from simple statistical analysis to sophisticated methods that involve artificial intelligence, big data science, and data warehousing. Data mining methods are multidisciplinary, automated, scalable, and meant to handle massive volumes of data. These tools are being developed on discrete scales to help in pattern recognition, predictive analytics, and decision support [15]. Association rule mining is a data mining method used to identify correlations and associations between variables in a dataset, based on co-occurring frequent itemsets in a substantial proportion of interactions. From these connected variables and co-occurrence patterns, association rules may be constructed with an antecedent and a consequent. The intensity and validity of an association rule is determined by its support and confidence scores. The greater the connection between the components of the association rule, the higher the support and confidence levels [16]. Among the association rule mining techniques, Apriori is a widely used algorithm that is simple and efficient. It works by scanning the dataset multiple times to find frequent itemsets, starting with individual items, and

generating candidate itemsets iteratively. The Apriori algorithm was found to be useful in extracting meaningful patterns of biological associations in trawl fishery data from Chennai fishing harbor, India [17]. ECLAT (equivalence class clustering and bottom-up lattice traversal) is another notable approach in association rule mining that traverses the itemset lattice and finds frequent itemsets using a depth-first search technique. FP-Growth is a slightly more advanced strategy that operates by scanning the dataset and constructing a tree structure known as the FP-tree in the first phase and in the second phase it counts and locates frequent itemsets from the FP-tree using a recursive algorithm [16,18]. Each of these ARM algorithms has certain advantages and limitations, depending on the datasets analyzed. They can be implemented using Python in a cloud-based notebook environment such as Google Colab because of the large set of available tools and libraries.

Generally, marine fisheries data are heterogeneous, intricately spatiotemporal, and gathered over a longer period of time. Although this data has been available, improved data analysis methods are required to make sense of the data and derive insights helpful to fisheries' management [7]. Therefore, in this study, we examined the performance efficiency of three common ARM algorithms (Apriori, ECLAT, and FP-Growth) in identifying the associations and hidden patterns in Indian marine fish-catch composition and environmental data. For this purpose, chlorophyll content and sea surface temperature (SST) data for the west and east coasts of India were collected from the European Union Copernicus Marine Service, and species-wise information of marine fish catch was taken from FAO FishStatJ [19,20]. We found certain important associations between environmental (temperature and chlorophyll) changes and spatiotemporal marine fish-catch composition dynamics in India.

MATERIALS AND METHODS

Data source, description, and selected attributes

The primary information and dataset used in this study were marine fish-catch composition and two key environmental variables – sea surface temperature (SST) and chlorophyll content, of the west and east coasts of India from 2011 to 2020. The data on marine fish-catch composition were sourced from the open-access database of the UN Food and Agricultural Organization FishStatJ [19]. This dataset contained time-series statistics on the production of marine organisms by fish species, biomass, and maritime coast, from 2011 to 2020. On the other hand, the geospatial oceanic environmental data were obtained from the EU Copernicus Marine Service [20]. It comprised grid-wise satellite data of SST and chlorophyll levels (as shown in Supplementary Fig. S1) extending out up to 12 nautical miles from India's entire coastline (189 grids - one degree GPS) and spanning 2011 to 2020. SST data is an essential indicator of marine health and changes in the climate, whereas sea chlorophyll data is an essential indication of the overall condition and productivity of the marine environment.

Feature selection plays a critical role in data mining analysis, as it can enhance the accuracy of the algorithm while minimizing the computing cost and complexity of the analysis. By picking the most useful and relevant characteristics, and eliminating redundant variables, feature selection prevents overfitting, increases generalization, and enables greater effectiveness and efficiency in data analysis [21]. In this study, two datasets with 147 data attributes of selected marine environmental factors and fish-catch composition data from the west and east coasts of India were collected for ARM analysis. During preprocessing and cleaning, the dataset was corrected for missing, duplicate, and inconsistent values, and the data were transformed into categorical attributes. Then by performing dimensionality reduction with principal component analysis, the two datasets were reduced to 33 and 31 attributes on the west and east coasts, respectively (Supplementary Table S1). Following the basic statistical description of the selected dataset (minimum, maximum, mean, coefficient of variation, and correlation), the association rule mining algorithms were applied to extract meaningful frequent itemsets.

Execution environment and software description

For the execution of data mining analysis using the ARM algorithms, we used Google Colab, a free-to-use online platform for executing Python programs in virtual machines [22]. Google Colab provides a handy environment for executing Python code with the flexibility to simply scale up or down as required. When the program was executed as a code cell in Colab, the code

was sent to a distant virtual machine operating in one of Google's data centers. The code was then executed by the virtual machine and the outputs were sent to the web browser being used. Python was our programming language choice for ARM analysis due to its simplicity, versatility, strong libraries, and data manipulation tools [23]. Python libraries and modules for ARM algorithms such as NumPy and Pandas were installed and used in the Colab's virtual machine. Python also provided a high level of scalability, allowing users to customize. For ARM application in this study, we used the Python Apriori package, which is a module within the MLxtend library. Likewise, the pyECLAT and pyfpgrowth modules were used for mining frequent itemsets with minimum support and confidence threshold.

Apriori algorithm

The Apriori algorithm works by finding the most common items in a dataset and then using those items to generate a set of candidate itemsets. It then checks the frequency of each candidate itemset in the dataset and removes those that do not meet a minimum support threshold [18]. The Apriori algorithm uses a breadth-first search approach to explore the space of possible itemsets. It starts by finding all frequent 1-itemsets, then uses these frequent itemsets to generate candidate 2-itemsets, and so on [24]. The result of the Apriori algorithm is a set of frequent itemsets and their associated support values. These frequent itemsets can then be used to identify interesting associations between items in the dataset. An association rule is a statement of the form "if X, then Y", where X and Y are sets of things. The strength of an association rule is assessed by its support and confidence. The support of an association rule is the proportion of transactions that contain both X and Y, whereas the confidence of an association rule is the fraction of transactions that contain Y providing they contain X [16,25]. Association rules with high support and confidence are regarded as strong. The flow chart of the Apriori algorithm implemented in this study is presented in Supplementary Fig. S2.

ECLAT algorithm

The ECLAT algorithm works by traversing a lattice of itemsets in depth-first order, where each node denotes an itemset and its child nodes reflect its subsets. The lattice was built by first producing all conceivable itemsets of size 1, followed by recursively combining frequent itemsets of size k-1 to produce itemsets of size k [25]. The algorithm subsequently traversed the lattice in order to discover frequently occurring itemsets. The ECLAT technique employed vertical data representation, which means that the dataset was expressed as a collection of itemsets and their associated transaction IDs. This structure enabled the method to establish the support associated with each itemset efficiently by counting the number of transactions that contain the itemset [16]. The transactional database is represented as a matrix in the vertical data format, with each row representing a transaction and each column representing an item, making it easy to spot common item groupings. These pairings were then used to construct a new data structure called the tidset, which is a tree-like structure that displays the frequency of itemsets in the database (Supplementary Fig. S2). Every itemset's support is calculated by intersecting the associated equivalence classes. Because of the vertical data format, the ECLAT technique uses less memory for storing the dataset and is capable of processing high-dimensionality datasets.

FP-Growth algorithm

The FP-Growth algorithm functions by efficiently encoding a collection of frequent itemsets with a data structure known as an FP-tree, and it uses the divide-and-conquer technique for frequent-pattern mining [24]. The FP-tree comprises a compressed representation of the underlying dataset, with each node representing an item and associated count representing the frequency with which that item appears in the dataset. The FP-tree is constructed from the bottom up, with infrequent items pruned off the tree. The remaining items were then placed into the FP-tree in such a way that their frequency order was preserved. By iteratively mining conditional patterns, the FP-tree was subsequently utilized to build frequent itemsets. In a nutshell, the key processing steps of the FP-Growth algorithm were to construct the frequent tree pattern, mine frequent itemsets, prune infrequent itemset, and generate association rules [16,26]. With a low support threshold, the technique is efficient, scalable, and capable of handling enormous datasets. The FP-tree structure saves time by reducing the amount of database searches necessary for frequent itemset creation.

For comparing multiple ARM algorithms, it is critical to establish the proper assessment metric based on the expectations of the results and limitations. The evaluation metrics such as minimum support and confidence are useful for determining the most effective algorithm to deal with marine fish production-related datasets, and to fine-tune the algorithm's specifications [16]. The most often used statistic, i.e., support, quantifies the frequency of recurrence of a certain itemset. Confidence assesses the strength of the link between elements and indicates the likelihood of witnessing the subsequent item given the antecedent item.

RESULTS

Spatiotemporal dynamics in Indian marine fisheries

Globally, oceanographic changes and fish-catch composition are highly complex, heterogenous, dynamic, and variable. Therefore, it is challenging to normalize data attributes collected from a larger spatiotemporal scale. To highlight the spatial and temporal dynamics of marine fish production in India, the state and union territory-wise apportionment of marine resources and capture fisheries' production is shown in Supplementary Table S2 and described hereunder. In terms of length of the coastline, Gujarat, Tamil Nadu, Andhra Pradesh, Maharashtra, and Kerala are the states with the maximum resources, and among the union territories, the Andaman and Nicobar Islands have the maximum coastline. With respect to the continental shelf area, Gujarat, Maharashtra, Tamil Nadu, Kerala, and Andhra Pradesh have the maximum coverage in the listed order. Though the length of the eastern coastline is longer than that of the west coast, the continental shelf area available for fishing activities is substantially larger on the west coast. Correspondingly, the total catch of marine organisms on the west coast is higher than that from the east coast, with maximum contribution from Gujarat, Tamil Nadu, Andhra Pradesh, Kerala, and Maharashtra. When unit marine capture fish production is calculated per km of coastline, Karnataka, West Bengal, Goa, Kerala, and Maharashtra are the most resource-efficient states. Likewise, in terms of unit fish production per km^2 of the continental shelf area, Andhra Pradesh, Karnataka, Tamil Nadu, Kerala, and Goa are the most productive states. Due to the absence of an open-source database on state-wise fish-catch composition, it was not possible to find associations between marine environmental data and fish-catch composition on a complete spatial scale. However, in this pilot study, using the available open-source data (FAO FishStatJ) and a data-mining approach, we made a coast-wise analysis of the associations between fish-catch composition and environmental changes.

Statistical interpretation of variations in marine fish production data

Concerning the catch composition attributes in both the west and east coasts of India, fishes from the taxonomical order Perciformes were the most abundantly caught species group, followed by fishes of the order Clupeiformes and total crustaceans (Fig. 1). With respect to individual species, Indian oil sardine (*Sardinella longiceps*) was the predominantly caught fish species on the west coast, followed by Bombay duck (*Harpadon nehereus*), Indian mackerel (*Rastrelliger kanagurta*) and giant tiger prawn (*Penaeus monodon*). Similarly in the east coast, giant tiger prawn, Indian mackerel, and Indian oil sardines are the main species caught`. The catch volume attributes of most fish groups were higher on the west coast as compared

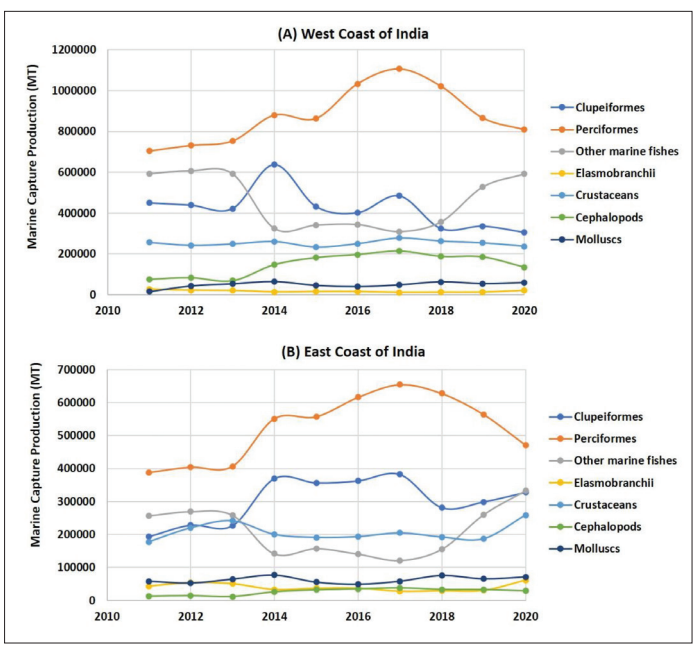


Fig. 1. Temporal changes in marine fish catch composition of west coast (**A**) and east coast (**B**) of India during 2011-2020.

Table 1. Basic statistics of the temporal fish-catch composition (2011-2020) attributes from west and east coasts of India

Fish catch in metric tons	Minimum		Maximum		Mean		Coeff. Variation (%)		Correlation (R^2)	
	West	East	West	East	West	East	West	East	West	East
Total fish and invertebrates	2110504	1129416	2451264	1552000	2223096	1372523	4.8	9.2	0.086	0.804
Total Clupeiformes	305000	193093	637204	382196	422796	302722	22.7	22.5	0.331	0.298
Indian oil sardine	77000	29842	435000	109684	236679	62268	47.6	38.4	0.675	0.328
Anchovies	67615	42219	114964	80479	87506	57225	17.5	20.9	0.827	0.324
Other clupeoids	49389	91240	134887	254000	98611	183230	34.8	35.6	0.469	0.456
Total Perciformes	704172	387252	1106138	654447	876914	523737	15.6	19.0	0.298	0.379
Indian mackerel	48395	39637	158936	129580	114378	86576	34.4	40.1	0.408	0.290
Tunas	63024	20423	102772	51335	79134	31032	19.2	30.5	0.201	0.125
Hairtails and Scabbard fishes	83007	26610	179599	59756	135265	42628	23.7	30.0	0.336	0.188
Croakers and drums	108073	27378	221290	63046	150805	39621	29.7	34.0	0.523	0.551
Carangids	94465	40277	224264	103418	151577	72439	30.2	31.6	0.444	0.551
Butterfish and pomfrets	16304	20259	47000	53000	22227	30112	40.8	30.1	0.134	0.239
Seerfish and ponyfish	41568	55684	60831	114072	49179	82007	13.2	24.1	0.016	0.025
Other percoids	76916	97980	318406	194826	174349	139323	44.8	26.5	0.225	0.272
Bombay duck	103000	-	183014	-	129202	-	22.3	-	0.208	-
Flatfishes	28698	-	46489	-	38937	-	15.5	-	0.313	-
Marine catfishes	28308	25860	46808	55000	40540	38518	14.8	20.1	0.279	0.004
Other marine finfishes	90097	79361	396846	277900	249818	171116	48.1	40.9	0.027	0.000
Total crustaceans	232367	177664	278840	259000	252171	206630	5.4	12.5	0.008	0.037
Giant tiger prawn	99905	76353	122000	160000	112343	90918	6.0	27.2	0.250	0.369
Crab and lobster	7660	17000	22000	47780	11700	37845	32.8	29.3	0.457	0.035
Other marine crustaceans	93000	56994	148272	121108	128128	77868	12.7	27.5	0.086	0.320
Total cephalopods	68306	11919	213889	37790	147412	26467	37.0	36.8	0.481	0.614
Total marine mollusks	15700	48650	63014	77000	48053	62725	28.9	15.7	0.369	0.145
Total elasmobranchs	11662	28509	27682	61000	17253	40608	30.9	27.4	0.363	0.037

Table 2. Basic statistics of the temporal marine environmental (2011-2020) attributes from west and east coasts of India

Environmental index*	Minimum		Maximum		Mean		Coeff. Variation (%)		Correlation (R^2)	
	West	East	West	East	West	East	West	East	West	East
Minimum temperature	18.0	21.2	22.3	23.6	20.7	22.5	6.4	3.0	0.029	0.025
Maximum temperature	30.8	31.2	32.1	32.0	31.5	31.6	1.5	0.7	0.606	0.135
Average temperature	28.3	28.2	28.9	28.9	28.6	28.5	0.7	0.8	0.591	0.516
Minimum chlorophyll	0.08	0.08	0.10	0.13	0.08	0.09	7.9	15.6	0.410	0.346
Maximum chlorophyll	1.77	3.08	4.24	5.60	3.04	4.07	27.2	21.7	0.039	0.657
Average chlorophyll	0.16	0.25	0.32	0.42	0.20	0.29	21.6	18.0	0.317	0.434

* Sea surface temperature (°C) and chlorophyll concentration

to the east coast, except for seerfish (Scombridae), ponyfish (Leiognathidae), pomfrets and butterfish (Bramidae), crabs (Brachyura), lobsters (Nephropidae) and elasmobranchs (Elasmobranchii). As shown in Table 1, during the study period (2011-2020) the temporal total fish-catch volume coefficient of variation was higher on the east coast (9.2%) than on the west coast (4.8%). The coefficient of variation of the different fish-catch composition attributes ranged from 5.4 to 48.1% on the west coast and 12.5 to 40.9% on the east coast. This indicates the fluctuations in the fish-catch composition, including the predominant species groups such as Indian oil sardines and Indian mackerel, in the last decade. On the other hand, the correlation estimates did not indicate linear trends in most of the fish-catch attributes from the west and east coasts of India. On the west coast, anchovies (0.827) and Indian oil sardine (0.675) showed a high

correlation; on the east coast, total fish-invertebrate catch (0.804) and total cephalopod catch (0.614) showed a high correlation. These statistical estimates show the complexity and variability in the temporal fish-catch composition of India.

Statistical interpretation of variations in marine environment data

Concerning the marine environmental data attributes that were collected from 189 geospatial grids monthly, substantial geospatial and seasonal variation was observed in the studied temporal scale. On both the west and east coasts of India, the seasonal variation in sea surface temperature in the northern grids was more pronounced (~6°C) than on the southern coastline (1-2°C). The minimum SST was usually recorded in January, February, and December, and the maximum was recorded in April, May, and June. On a year-on-year basis, the average SST (28.2-28.9°C) did not show a greater degree of fluctuation, but the maximum (30.8-32.1°C) and minimum (18-22.3°C) temperature records were found to vary considerably, more on the west coast than on the east coast (Table 2 and Supplementary Fig. S3).

With respect to chlorophyll and primary productivity, the average and maximum concentration was generally found to peak in July, August, and September, with rare anomalous changes on both the west and east coasts of India. Chlorophyll concentrations (minimum, maximum, and average) were higher on the east coast as compared to the west coast. On a year-on-year basis, the coefficient of variation was high for maximum chlorophyll records, followed by average estimates and minimum records. The decadal trends did not show strong linear changes in the key marine environmental factors. Nevertheless, average SSTs showed a correlation of 0.59 and 0.52 on the west and east coasts, respectively, indicating the potential impact of climate change. Likewise, maximum chlorophyll on the east coast (0.66) and maximum temperature on the west coast (0.61) showed some correlation, as compared to the other environmental variables. The dynamic changes in fish-catch and marine environmental data attributes of the west and east coasts of India are clearly depicted in the heat maps, and reflect the direction and magnitude of the temporal changes during the study period.

Heatmap of fish-catch composition and environment attributes

To depict the temporal trends in marine fish, catch composition, sea surface temperature, and chlorophyll, heatmaps (Figs. 2 and 3) were generated and used as a base for interpretation. Through the heat maps, preliminary trends were derived in the temporal changes of the selected data attributes in marine fish-catch composition of the west and east coasts of India, and marine environmental changes with respect to sea surface temperature and chlorophyll concentration. Some of the fish-catch and marine environment attributes showed a progressive increase over the time scale, whereas others showed decreasing or fluctuating trends.

Performance efficiency of ARM algorithms

Implementing the extensively used Apriori algorithm written in Python and operated in a Google Colab virtual machine, frequent itemsets and associations between the marine fish-catch composition on the west and east coasts of India and marine environmental changes (temperature and chlorophyll) were identified. The bottom-up strategy yielded feasible itemsets with their minimum support and confidence. The pair of frequent itemsets that involve a fish-catch composition and marine environment attributes, along with their ARM performance indicators (support and confidence), are presented in Table 3 (west coast) and Table 4 (east coast). The species that are likely to be more abundant or less in the catch composition on the west and east coasts due to changes in environmental conditions (sea surface temperature and chlorophyll concentration) are listed in these tables.

Similarly, ECLAT algorithm implementation in Google Colab using the Python code program picked up frequently occurring itemsets from the marine fish-catch and environment transactional database (Tables 3 and 4). ECLAT traversed the itemset tree efficiently and found frequent itemsets by using a depth-first search approach and vertical data format. The marine fisheries' datasets were searched by the ECLAT algorithm vertically for all pairs of items that occur within the transaction. By using these pairings, another data structure called the tidset with the transactions from

(A)

Code	[2011]	[2012]	[2013]	[2014]	[2015]	[2016]	[2017]	[2018]	[2019]	[2020]
FS1	2120674	2166440	2157017	2326728	2110504	2280455	2451264	2226308	2234575	2156999
FS2	449211	438608	421838	637204	431887	401393	483954	323427	335436	305000
FS3	322079	321372	299028	435000	211000	194600	267993	123476	115241	77000
FS4	71697	67615	73421	86700	86000	83150	85377	102137	114964	104000
FS5	55435	49621	49389	115504	134887	123643	130584	97813	105231	124000
FS6	704172	731501	753534	880207	863511	1033669	1106138	1021074	866099	809234
FS7	83876	48395	64172	130500	131000	137300	158936	157581	91019	141000
FS8	74361	83842	102772	95636	63750	89680	63595	90429	64254	63024
FS9	102662	83007	103921	159000	133000	162900	179599	145434	164128	119000
FS10	200349	221290	215750	129500	124000	125900	119874	108073	108312	155000
FS11	101094	94465	96012	148500	169500	171600	196737	224264	187600	126000
FS12	22231	24507	18665	20192	16304	17663	18749	16470	20489	47000
FS13	42683	50850	41568	60831	55803	54931	50242	46343	45983	42554
FS14	76916	125145	110674	136047	170154	273695	318406	232481	184315	115656
FS15	119885	175121	183014	105000	103000	135200	135353	105123	109328	121000
FS16	28698	30531	46489	44000	36700	36400	37649	42364	41537	45000
FS17	46808	44997	42832	35700	43600	42100	46081	28308	33972	41000
FS18	396846	356431	319504	139840	157521	130120	90097	180749	342811	384265
FS19	256178	241117	248253	260779	232367	249875	278840	263206	254097	237000
FS20	113105	104040	99905	117200	113500	114000	119353	109464	110868	122000
FS21	10684	7660	10743	9579	9867	11875	11215	11859	11518	22000
FS22	132389	129417	137605	134000	109000	124000	148272	141883	131711	93000
FS23	75494	83481	68306	147200	181500	196550	213889	187689	185012	135000
FS24	15700	42570	52430	63014	45184	39808	47601	62142	53584	58500
FS25	27682	22083	20817	13784	15234	15340	11662	12227	12698	21000

(B)

Code	[2011]	[2012]	[2013]	[2014]	[2015]	[2016]	[2017]	[2018]	[2019]	[2020]
FS1	1129416	1242028	1260055	1398870	1386775	1433846	1487116	1395712	1439410	1552000
FS2	193093	228222	226523	369429	356754	362676	382196	281402	298928	328000
FS3	59634	82926	77161	109684	54667	50392	69397	31974	29842	57000
FS4	42219	46834	50855	56845	60287	58284	59836	71607	80479	45000
FS5	91240	98462	98507	202900	241800	254000	252963	177822	188607	226000
FS6	387252	403732	406150	551036	556937	615984	654447	627496	563732	470600
FS7	40003	39637	52558	106556	106801	111941	129580	128475	74207	76000
FS8	24672	33716	29956	31787	22108	26492	20423	42047	51335	27788
FS9	27734	26610	33315	50405	44259	54200	59756	48388	54608	27000
FS10	63046	56374	54963	32364	31383	31893	30367	27378	27438	41000
FS11	50560	40592	40277	68590	80091	80699	103300	103418	83858	73000
FS12	20259	35925	27360	30000	24500	26500	28128	24709	30738	53000
FS13	62998	66344	63198	114072	97448	93594	88067	76980	101689	55684
FS14	97980	104534	104523	117262	150347	190665	194826	176102	139858	117128
FS17	39802	41089	39112	32975	39754	38459	42096	25860	31035	55000
FS18	217158	228582	219764	108794	117863	102536	79361	129790	229409	277900
FS19	177664	219807	241492	200123	191090	193315	204545	192174	187095	259000
FS20	76502	79513	76353	88402	85695	86116	90160	82690	83750	160000
FS21	21398	31394	44031	38050	39600	47780	45125	47719	46351	17000
FS22	79764	108900	121108	73671	65795	59419	69260	61764	56994	82000
FS23	12835	14567	11919	26013	31977	34726	37790	33155	32687	29000
FS24	58300	52430	64570	77000	55200	48650	58173	75944	65485	71500
FS25	43312	53598	50525	33500	37200	37500	28509	29891	31040	61000

Fig. 2. Heat map of the year-on-year changes of fish-catch attributes of west coast (**A**) and east coast (**B**) of India during 2011-2020.

(A)

Code	[2011]	[2012]	[2013]	[2014]	[2015]	[2016]	[2017]	[2018]	[2019]	[2020]
WQ1	20.9	20.6	21.0	19.0	20.5	22.3	22.0	21.7	20.5	18.0
WQ2	30.8	31.0	31.1	31.5	31.6	32.1	31.1	31.6	31.9	32.1
WQ3	28.4	28.3	28.3	28.5	28.9	28.7	28.6	28.6	28.7	28.9
WQ4	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.09	0.10
WQ5	2.32	2.62	3.32	3.01	2.18	3.97	4.24	3.87	1.77	3.05
WQ6	0.18	0.18	0.20	0.18	0.16	0.21	0.20	0.20	0.18	0.32

(B)

Code	[2011]	[2012]	[2013]	[2014]	[2015]	[2016]	[2017]	[2018]	[2019]	[2020]
WQ1	22.0	22.6	22.3	22.6	21.2	23.6	23.3	22.3	22.7	22.1
WQ2	31.4	31.2	31.7	31.7	31.4	32.0	31.6	31.5	31.5	31.7
WQ3	28.4	28.3	28.2	28.4	28.6	28.8	28.6	28.4	28.7	28.9
WQ4	0.08	0.08	0.08	0.08	0.08	0.09	0.09	0.09	0.09	0.13
WQ5	3.61	3.08	3.69	3.41	3.43	4.58	3.67	4.15	5.60	5.51
WQ6	0.25	0.26	0.28	0.29	0.25	0.27	0.26	0.28	0.31	0.42

Fig. 3. Heat map of the year-on-year changes of marine environmental attributes of west coast (**A**) and east coast (**B**) of India during 2011-2020.

the marine fish-catch composition database and marine temperature and chlorophyll database, was generated. The transaction numbers in the tidset reflected the support count of the itemsets. The speed of the ECLAT algorithm was high.

As compared to Apriori and ECLAT algorithms, the FP-growth algorithm was found to yield more instances from the marine fish-catch and environment datasets with a higher degree of minimum support and confidence. Through its efficient divide-and-conquer mining approach for frequent itemsets, it was found to be suited to handle the marine fish-catch datasets with a lower support threshold. The FP-growth algorithm was able to construct a Frequent Pattern tree that could represent and describe the marine fish-catch composition and temperature-chlorophyll dataset in a compact manner. The output from this study is consolidated and presented in Tables 3 and 4, for the west and east coasts, respectively.

Interpretation of marine fish-catch-environment associations

With respect to practical interpretations (Table 5), changes in sea surface temperature on the west coast of India were found to be associated with linear changes in the catch abundance of fish belonging to the order Clupeiformes and, in particular, anchovies and other clupeoids, and Perciformes fish like hairtails, scabbard fishes (Trichiuridae) and Indian mackerel, and cephalopods. Concurrently, an increase in the average SST was found to be correlated with a decrease in the catch of Bombay duck and Indian oil

Table 3. Frequent itemsets in the west coast marine fish-catch-environment data along with ARM performance measures

Frequent Itemset	Apriori		ECLAT		FP-Growth	
	Support	Confidence	Support	Confidence	Support	Confidence
{WQ1, FS2}	0.033	0.95	0.09	1.00	0.13	1.00
{WQ1, FS3}	0.033	0.95	0.09	1.00	0.13	1.00
{WQ1, FS9}	0.015	0.90	0.05	0.95	0.09	0.98
{WQ1, FS23}	0.015	0.85	0.05	0.95	0.11	0.95
{WQ2, FS4}	0.033	0.85	0.09	1.00	0.13	0.98
{WQ2, FS5}	0.033	0.70	0.09	0.95	0.13	0.95
{WQ2, FS7}	0.015	0.75	0.05	1.00	0.11	1.00
{WQ3, FS4}	0.033	0.95	0.08	1.00	0.15	1.00
{WQ3, FS5}	0.033	0.90	0.08	0.95	0.12	1.00
{WQ3, FS7}	0.015	0.80	0.05	0.90	0.08	0.90
{WQ4, FS21}	0.033	0.95	0.09	1.00	0.14	1.00
{WQ4, FS4}	0.033	0.95	0.08	1.00	0.13	1.00
{WQ4, FS12}	0.015	0.80	0.05	0.95	0.09	0.95
{WQ4, FS16}	0.015	0.80	0.05	0.95	0.08	0.95
{WQ5, FS7}	0.033	0.90	0.09	1.00	0.12	0.98
{WQ6, FS7}	0.015	0.85	0.07	0.90	0.07	0.95
{WQ6, FS20}	0.005	0.65	0.025	0.85	0.03	0.78
{WQ6, FS21}	0.005	0.70	0.025	0.90	0.02	0.86
{WQ3, FS15}	0.005	0.65	0.025	0.80	0.03	0.75
{WQ3, FS3}	0.005	0.65	0.025	0.80	0.01	0.80

Table 4. Frequent itemsets in the east coast marine fish-catch-environment data along with ARM performance measures

Frequent Itemset	Apriori		ECLAT		FP-Growth	
	Support	Confidence	Support	Confidence	Support	Confidence
{WQ1, FS9}	0.025	0.90	0.07	0.95	0.11	1.00
{WQ2, FS2}	0.025	0.85	0.05	0.95	0.09	0.95
{WQ3, FS5}	0.033	0.90	0.09	0.90	0.13	0.98
{WQ3, FS20}	0.025	0.85	0.06	0.90	0.11	0.90
{WQ4, FS1}	0.033	0.95	0.09	0.95	0.14	1.00
{WQ4, FS5}	0.015	0.80	0.04	0.90	0.07	0.90
{WQ4, FS20}	0.010	0.75	0.03	0.90	0.05	0.85
{WQ5, FS1}	0.033	0.95	0.08	1.00	0.13	1.00
{WQ6, FS12}	0.015	0.85	0.04	0.90	0.07	0.95
{WQ6, FS24}	0.015	0.80	0.05	0.85	0.08	0.90
{WQ1, FS8}	0.005	0.75	0.01	0.80	0.02	0.85
{WQ3, FS1}	0.005	0.70	0.01	0.75	0.03	0.80

sardine. On the other hand, regarding primary productivity, chlorophyll changes on the west coast were found to be associated with the catch abundance of Indian mackerel, anchovies, butterfish, pomfrets, flatfish, crabs, lobsters, and giant tiger prawn. On the east coast, an increase in mean SST and maximum chlorophyll was found to be associated with a higher overall catch of fish and invertebrates. Like the west coast, an increase in mean and maximum SST was found to be correlated with a higher catch of Clupeiformes fish. The low minimum temperature was linked to a low catch of hairtails, scabbard fish, and tunas, whereas the high mean temperature was associated with a higher abundance of giant tiger prawn. With respect to primary productivity, low chlorophyll was linked to a smaller catch of giant tiger prawn and other clupeoids, while a high mean chlorophyll concentration was associated with higher catch abundance of butterfish, pomfrets, and total marine mollusks.

DISCUSSION

India is one of the world's major producers of fish and seafood, with a rich diversity of fish and invertebrate species found in its vast marine and inland aquatic resources. Marine fish catch in India comes from diverse environments, such as estuaries, lagoons, bays, the continental shelf, and open sea. Coastal fisheries are typically small-scale, traditional, and operate within a few km of the shore, and data are poor. Fishing is carried out with a variety of fishing equipment, such as gillnets, hooks and lines, traps, and seines. Small, mechanized boats, or traditional crafts are used for fishing activities. Common marine fish species captured in coastal fisheries include sardines, mackerel, anchovies, and numerous crustaceans, such as prawns, and crabs. Offshore or deep-sea fisheries function beyond the shoreline of a continental shelf area and pursue pelagic and demersal species, such as tuna, shark, and squid; they are relatively better documented and monitored. Longlines, purse seines, and trawls are commonly utilized by these fisheries, which use large, mechanized vessels. In the last decade, the volume of total annual Indian marine catches ranged between 3.3 to 3.9 million tons, with marked variations in fish-catch composition [3,5]. Climate change and associated environmental fluctuations with respect to temperature, pH, salinity, chlorophyll, and primary productivity

Table 5. Summary of the common association rules extracted from marine fish-catch and environment data through data mining

Environmental parameter	Fish catch composition
West coast of India	
Low minimum sea surface temperature	Low catch of Clupeiformes, Indian oil sardines, hairtails, scabbard fish and cephalopods
High maximum sea surface temperature	High catch of anchovies, other clupeoids, and Indian mackerel
High average sea surface temperature	High catch of anchovies, other clupeoids, and Indian mackerel; low catch of Bombay duck and Indian oil sardine
Low minimum chlorophyll concentration	Low catch of anchovies, butterfish, pomfrets, flatfishes, crabs, and lobsters
High maximum chlorophyll concentration	High catch of Indian mackerel
High average chlorophyll concentration	High catch of Indian mackerel, giant tiger prawn, crabs, and lobsters
East coast of India	
Low minimum sea surface temperature	Low catch of hairtails, scabbard fish, and tunas
High maximum sea surface temperature	High catch of total Clupeiformes
High average sea surface temperature	High catch of other clupeoids, giant tiger prawn, and overall fish-invertebrate catch
Low minimum chlorophyll concentration	Low catch of other clupeoids, giant tiger prawn, and overall fish-invertebrate catch
High maximum chlorophyll concentration	High total fish and invertebrate catch
High average chlorophyll concentration	High catch of butterfish, pomfrets, and total marine mollusks

can be some of the major reasons for variations in the recruitment, abundance, distribution, and catch of pelagic, demersal, and oceanic fisheries' resources.

In this milieu, based on the available open-source information, we undertook this study and elucidated hidden associations between the marine environmental changes (SST and chlorophyll levels) and corresponding fish-catch composition in the two maritime coastlines of India. For this purpose, we employed three prominent association rule mining algorithms (Apriori, ECLAT, and FP-Growth) written in Python and executed in the open-source Google Colab. The data mining strategies employed by Apriori, ECLAT, and FP-Growth differed in the depth and breadth of the search, execution speed, accuracy, minimum support, and confidence [16]. As this was a pilot investigation, we were able to analyze only a smaller number of instances and frequent itemsets. Initially, two datasets with 147 data attributes from the west and east coasts of India were collected for ARM analysis (from FAO FishStatJ and EU Copernicus Marine Service). After preprocessing, the two datasets were reduced to 33 and 31 attributes on the west and east coasts, respectively. Following the data preprocessing and dimensionality reduction, the ARM algorithms Apriori, ECLAT, and FP growth were implemented using the combination of selected marine environmental factors and fish-catch composition data attributes from the west and east coasts of India, to derive meaningful frequent itemsets and associations. Except for one earlier study that used the Apriori algorithm to extract meaningful patterns of biological associations in trawl fishery data from Chennai fishing harbor [17], to date there is no other study that explored the potential of data mining approaches to extract meaningful patterns and information from Indian fisheries' databases. On the other hand, globally, the environment-resource approach has been the focus area for the implementation of data mining approach in capture fisheries [7]. Some of the data mining techniques that have been shown for application in fisheries and environment datasets include *k*-means with time series decomposition and multivariate analysis, spatiotemporal assignment mining model, genetic programming, and self-organizing feature map and structuring index [11-14]. Interestingly, nearest-neighbor clustering of species-wise catch per unit effort data based on local fisheries' statistics were able to identify a correlation between stock depletion and the market price of the fish [27]. Marine scientists have also developed data mining techniques to predict and identify the locations of fish aggregation or potential fishing zones from the dynamic multi-dimensional marine environment (e.g., sea surface temperature and chlorophyll concentration data), and fisheries' resource (catch statistics) datasets [28-30]. This application in accurately predicting potential fishing zones has significant economic benefits for fishers, as it reduces the time, effort, and resources spent in searching for fish shoals.

The comparative performance and results of the conventional algorithms, namely Apriori, ECLAT, and FP-Growth, were analyzed for robustness, performance,

and efficiency in identifying frequent itemsets. Based on the comparative performance evaluation of Apriori, ECLAT, and FP-Growth algorithms, it was found that each algorithm exhibited distinct advantages in data processing. Nevertheless, FP-Growth showed higher efficiency in identifying the associations between species-wise fish-catch composition, sea surface temperature, and chlorophyll concentration data, as indicated by improved processing indicators such as support and confidence. Comparatively, in the west coast data, the higher number of instances and frequent itemsets were extracted with minimum support and confidence, as compared to the east coast data. Previous comparative studies of ARM algorithms have reported that the performances of these algorithms are related to the characteristics of the analyzed dataset and threshold values. FP-Growth overcomes the limitation of the Apriori algorithm that requires multiple passes over the source data during the candidate generation phase [31]. Numerical attribute management during the preprocessing step is also a determinant of the performance efficiency of the ARM algorithm [32].

The experimental results show that fish-catch volume on the west coast was stagnant and fluctuating, while it increased moderately on the east coast throughout the time period studied. The second set of marine fish-capture statistics included the major fish species groupings taken on the west and east coasts over the same time period. Out of the 75 fish species groups in the dataset, 25-27 dominating groups were chosen for association mining based on production over the established confidence level. Perciformes fishes dominated the catch composition on both the west and east coasts, followed by temporal changes in the catch of Clupeiformes, and other finfishes. Cephalopods were abundant on the west coast, while other mollusks and elasmobranchs were abundant on the east coast. Crustacean production data revealed temporal changes, with the west coast plateauing and the east coast catching up [1,3]. The lowest, maximum, and averages of both the environmental factors were evaluated using marine environment data, namely monthly average temperature and chlorophyll concentration data from 113 geographical grids on the west coast and 76 geographical grids on the east coast. The temporal and regional patterns in environment-fish productivity were investigated [20]. In comparison to Apriori, the ECLAT method was found to have higher computational speed and efficacy in discovering common fish-environment itemsets. FP-Growth provided a significant advantage in terms of association mining efficiency as well as processing performance to extract information from complex and heterogeneous spatiotemporal fisheries' data.

Results showed linear and inverse associations between changes in the sea surface temperature, chlorophyll concentration, and major catch groups, such as anchovies, Indian oil sardine, other clupeoid fish (Clupeiformes), Indian mackerel, hairtails, butterfish, and pomfrets (Perciformes), other species groups such as Bombay duck, flatfish and tunas, and invertebrates like giant tiger prawn, crabs, lobsters, and cephalopods. Considering the prominent associations, increasing chlorophyll content (primary productivity) and temperature (to a certain extent) were observed to result in enhanced fish output, particularly for low trophic level species. Increases in maximum SST have been related to stagnation or a decrease in the catch of some fish species such as Indian oil sardines. Similar observations of associations in spatiotemporal patterns of temperature and other hydrological indicators, and fish assemblage and catch per unit effort have been previously deciphered using specific data mining algorithms [12-14]. Overall, the ARM algorithms demonstrated capabilities of defining associations in fish-catch composition, sea surface temperature, and chlorophyll concentration data. FP-Growth demonstrated better performance and processing metrics, such as support, confidence, and lift, as it used conditional FP-trees and link nodes. Therefore, the implementation of the FP-Growth algorithm might substantially support fisheries' management actions, as it provides future insights and potential forecasts.

With respect to the limitations of the study, the diversity and complexity of the spatiotemporal marine fish-catch composition and environment data were a challenge for effective and error-free prediction of fish capture and environmental associations by using the ARM techniques. Similarly, the accurate spatial segmentation and stratification of marine environmental data throughout the two coastal areas could not be accounted for in the present research. These challenges highlight the need for a digitized and unified data-collecting infrastructure for fish-catch and marine environmental parameters to construct

successful models for predicting marine fish-stock status and abundance. For this study, India's marine environmental data were collected from 189 grids of 1° latitude and longitude, up to 12 nautical miles only, but this could be extended geographically in future investigations. Future prediction models could be developed more efficiently for other areas in fisheries, such as fish trade and fish nutritional composition datasets, as the inherent complexities in data collection is minimal. Also, comparisons between ARM and other data mining techniques can be made to identify the best technique to extract information and predict marine fish production.

CONCLUSIONS

Despite the challenges related to data complexity, quality, and volume in marine fisheries, the development and adoption of intelligent data analytical systems and tools are inevitable in fisheries' management. As a prelude to this, in this pilot study, two datasets with 147 data attributes related to fish-catch composition and marine environment from the west and east coasts of India were collected for ARM analysis. After preprocessing, the two datasets were reduced to 33 and 31 attributes on the west and east coasts, respectively. Following the data preprocessing and dimensionality reduction, the ARM algorithms Apriori, ECLAT, and FP-Growth were applied to a combination of selected marine environmental factors and fish-catch composition datasets from the west and east coasts of India, and meaningful frequent itemsets and associations were revealed. Based on the comparative performance evaluation of the ARM algorithms, it was found that each algorithm exhibits distinct advantages in data processing. Based on the ARM support and confidence indicators, FP-Growth showed comparatively higher efficiency in mining associations between fish-catch composition, sea surface temperature, and chlorophyll concentration data.

Funding: The author(s) received no specific funding for this work.

Acknowledgments: The academic guidance and support of Dr. Antony Selvadoss Thanamani, Head, Department of Computer Science, NGM College, in the PhD research of the first author is gratefully acknowledged.

Author contributions: Conceptualization, GJ, KA and BSK; methodology and software, GJ and KA; formal analysis and data curation, GJ, KA and BSK; writing - original draft preparation, GJ and BSK; writing - review and editing, KA; supervision, KA. All authors have read and agreed to the published version of the manuscript.

Conflict of interest disclosure: There is no conflict of interest to be declared.

Data availability: The raw data used in this study were sourced from the publicly accessible repository FAO FishStatJ [https://www.fao.org/fishery/en/topic/166235/en] and the European Union Copernicus Marine Service [https://marine.copernicus.eu/]. Data underlying the reported findings have been provided as a raw dataset which is available here: [https://www.serbiosoc.org.rs/NewUploads/Uploads/Gladju%20et%20al_Dataset.xlsx]

REFERENCES

1. FAO. The state of world fisheries and aquaculture 2022 - towards blue transformation. Rome: Food and Agriculture Organization of the United Nations; 2022. 236 p. https://doi.org/10.4060/cc0461en
2. Costello MJ, Chaudhary C. Marine biodiversity, biogeography, deep-sea gradients, and conservation. Current Biology. 2017;27(11): R511-R527. https://doi.org/10.1016/j.cub.2017.04.060
3. Fisheries Statistics Division. Handbook on fisheries statistics 2022. New Delhi: Department of Fisheries, Ministry of Fisheries, Animal Husbandry and Dairying, Government of India; 2022. 198 p. https://dof.gov.in/sites/default/files/2023-01/HandbookFisheriesStatistics19012023.pdf
4. Malde K, Handegard NO, Eikvil L, Salberg AB. Machine intelligence and the data-driven future of marine science. ICES Journal of Marine Science. 2020;77(4):1274-85. https://doi:10.1093/icesjms/fsz057
5. Mohamed KS, Sathianandan TV, Padua S. Integrated spatial management of marine fisheries of India for more robust stock assessments and moving towards a quota system. Marine Fisheries Information Service Technical and Extension Series. 2018;236:7-15.
6. van Helmond AT, Mortensen LO, Plet-Hansen KS, Ulrich C, Needle CL, Oesterwind D, Kindt-Larsen L, Catchpole T, Mangi S, Zimmermann C, Olesen HJ, Bailey N, Bergsson H, Dalskov J, Elson J, Hosken M, Peterson L, McElderry H, Ruiz J, Pierre JP, Dykstra C, Poos JJ. Electronic monitoring in fisheries: Lessons from global experiences and future opportunities. Fish and Fisheries. 2020;21(1):162-89. https://doi:10.1111/faf.12425
7. Gladju J, Kamalam BS, Kanagaraj A. Applications of data mining and machine learning framework in aquaculture and fisheries: A review. Smart Agricultural Technology. 2022;2:100061. https://doi.org/10.1016/j.atech.2022.100061
8. He Y, Su F, Du Y, Xiao R. Web-based spatiotemporal visualization of marine environment data. Chinese Journal of Oceanology and Limnology. 2010;28(5):1086-1094. https://doi:10.1007/s00343-010-0029-8
9. Su T, Cao Z, Lv Z, Liu C, Li X. Multi-dimensional visualization of large-scale marine hydrological environmental data. Advances in Engineering Software. 2016;95:7-15. https://doi.org/10.1016/j.advengsoft.2016.01.009

10. Bradley D, Merrifield M, Miller KM, Lomonico S, Wilson JR, Gleason MG. Opportunities to improve fisheries management through innovative technology and advanced data systems. Fish and Fisheries. 2019;20(3):564-83. https://doi.org/10.1111/faf.12361
11. Plaza F, Salas R, Yáñez E. Identifying ecosystem patterns from time series of anchovy (*Engraulis ringens*) and sardine (*Sardinops sagax*) landings in northern Chile. Journal of Statistical Computation and Simulation. 2018;88(10):1863-81. https://doi.org/10.1080/00949655.2017.1410150
12. Su F, Zhou C, Lyne V, Du Y, Shi W. A data-mining approach to determine the spatiotemporal relationship between environmental factors and fish distribution. Ecological Modelling. 2004;174(4):421-31. https://doi.org/10.1016/j.ecolmodel.2003.10.006
13. Yang YCE, Cai X, Herricks EE. Identification of hydrologic indicators related to fish diversity and abundance: A data mining approach for fish community analysis. Water Resources Research. 2018;44(4):W04412. https://doi.org/10.1029/2006WR005764
14. Tsai WP, Huang SP, Cheng ST, Shao KT, Chang FJ. A data-mining framework for exploring the multi-relation between fish species and water quality through self-organizing map. Science of the Total Environment. 2017;579:474-83. https://doi.org/10.1016/j.scitotenv.2016.11.071
15. Han J, Kamber M, Pei J. Data mining: Concepts and techniques. 3rd ed. Morgan Kaufmann Publishers. 2011. https://doi.org/10.1016/C2009-0-61819-5
16. Kotsiantis S, Kanellopoulos D. Association rules mining: A recent overview. GESTS International Transactions on Computer Science and Engineering. 2006;32(1):71-82.
17. Pugazhendi D. Apriori algorithm on marine fisheries biological data. International Journal of Computer Science & Engineering Technology. 2013;4(12):1409-11
18. Jiang N, Gruenwald L. Research issues in data stream association rule mining. ACM Sigmod Record. 2006;35(1):14-9. https://doi.org/10.1145/1121995.1121998
19. FAO FishStatJ Fisheries and Aquaculture Statistical Time Series [Internet]. Rome: Food and Agriculture Organization of the United Nations. 2022 - [Cited 2023 September 8]. Available from: https://www.fao.org/fishery/en/topic/166235/en
20. European Union Copernicus Marine Service [Internet]; European Union. 2022 - [Cited 2023 September 8]. Available from: https://marine.copernicus.eu/
21. Mukhlash I, Sitohang B. Spatial data preprocessing for mining spatial association rule with conventional association mining algorithms. In: Proceedings of the International Conference on Electrical Engineering and Informatics; 2007 June 17-19; Bandung, Indonesia. Bandung: Institute Teknologi Bandung, Indonesia; 2007. p. 531-34.
22. Bisong E. Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners. Berkeley CA: Apress; 2019. p. 709. https://doi.org/10.1007/978-1-4842-4470-8
23. McKinney W. Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. 2nd ed. Sebastopol: O'Reilly Media; 2017. p. 522.
24. Abdullah Z, Adam O, Herawan T, Deris MM. Lecture notes in Electrical Engineering: A review on sequential pattern mining algorithms based on apriori and patterns growth. Singapore: Springer; 2019. p. 646. https://doi.org/10.1007/978-981-13-1799-6
25. Borgelt C. Efficient implementations of apriori and eclat. In: Zaki MJ, Goethals B, editors. Proceedings of FIMI'03 Workshop on Frequent Itemset Mining Implementations; 2003 November 19; Melbourne. RPI CS Department Technical Report TR 03-14; 2003. p. 154.
26. Borgelt C. An Implementation of the FP-growth Algorithm. In: Goethals B, Nijssen S, Zaki MJ, editors. Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations; 2005 August 21; Chicago. New York: Association for Computing Machinery; 2005. p. 83.
27. Enomoto K, Ishikawa S, Hori M, Sitha H, Song SL, Thuok N, Kurokura H. Data mining and stock assessment of fisheries resources in Tonle Sap Lake, Cambodia. Fisheries Science. 2011;77:713-22. https://doi.org/10.1007/s12562-011-0378-z
28. Fitrianah D, Hidayanto AN, Gaol JL, Fahmi H, Arymurthy AM. A spatiotemporal data-mining approach for identification of potential fishing zones based on oceanographic characteristics in the Eastern Indian Ocean. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2015;9(8):3720-8.
29. Hidayanto AN, Fahmi H, Fitrianah D, Arymurthy AM. Oceanographic features selection to predict the tuna potential fishing zones using SFFS method. In: International Mathematical Forum. 2016;11(24):1157-66.
30. Fitrianah D, Fahmi H, Hidayanto AN, Arymurthy AM. A data mining based approach for determining the potential fishing zones. International Journal of Information and Education Technology. 2016;6(3):187-91.
31. Yıldız B, Ergenç B. Comparison of two association rule mining algorithms without candidate generation. In: Hamza MH, editor. Proceedings of the 10th IASTED International Conference on Artificial Intelligence and Applications; 2010 February 15-17; Innsbruck, Austria. Innsbruck: ACTA Press; 2010. 450-457 p.
32. Moreno MN, Segrera S, López VF, Polo MJ. Improving the quality of association rules by preprocessing numerical data. In: Proceedings of the II Congreso Español de Informática; 2007 September 11-14; Zaragoza, Spain. Asociación de Técnicos de Informática; 2007. 223-30 p.

SUPPLEMENTARY MATERIAL

Supplementary Table S1. List of selected data attributes for association mining

Attribute category	Attributes / Code	Data Source
Fish catch composition West coast of India East coast of India (2011-2020)	FS1 - Total fish and invertebrate catch	FAO FishStat J (Fisheries and Aquaculture Statistics Database)
	FS2 - Total Clupeiformes	
	FS3 - Indian oil sardine	
	FS4 - Anchovies	
	FS5 - Other clupeoids	
	FS6 - Total Perciformes	
	FS7 - Indian mackerel	
	FS8 - Tunas	
	FS9 - Hairtails and Scabbard fishes	
	FS10 - Croakers and drums	
	FS11 - Carangids	
	FS12 - Butterfish and pomfrets	
	FS13 - Seerfish and ponyfish	
	FS14 - Other percoids	
	FS15 - Bombay duck	
	FS16 - Flatfishes	
	FS17 - Marine catfishes	
	FS18 - Other marine finfishes	
	FS19 - Total crustaceans	
	FS20 - Giant tiger prawn	
	FS21 - Crab and lobster	
	FS22 - Other marine crustaceans	
	FS23 - Total cephalopods	
	FS24 - Total marine molluscs	
	FS25 - Total elasmobranchs	
Marine environment data West coast of India East coast of India (2011-2020)	WQ1 - Minimum temperature - annual	Copernicus Marine Service (European Union - Sentinel)
	WQ2 - Maximum temperature - annual	
	WQ3 - Average temperature - annual	
	WQ4 - Minimum chlorophyll - annual	
	WQ5 - Maximum chlorophyll - annual	
	WQ6 - Average chlorophyll - annual	

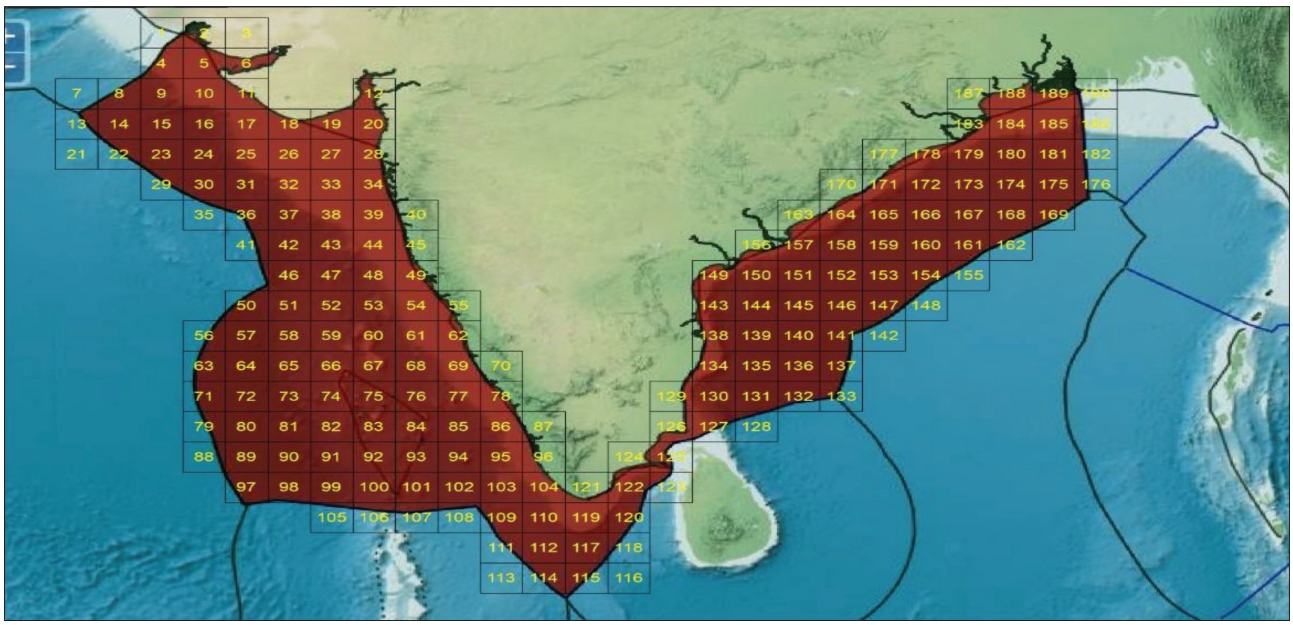
Supplementary Table S2. State-wise and coast-wise annual marine fish production of India during 2010-2020

Coastal State / Union Territory	Coastline**		Continental shelf		Annual marine fish production (expressed in '000 tonnes) #									
	km	MT/km	km²	MT/km²	2010-11	2011-12	2012-13	2013-14	2014-15	2015-16	2016-17	2017-18	2018-19	2019-20
West Coast of India														
Gujarat	1600	438	184000	3.8	689	692	694	696	698	697	699	701	699	701
Maharashtra	720	615	112000	4.0	447	434	449	467	464	434	463	475	468	443
Daman & Diu[1]	27	1185	-	-	17	17	19	19	32	23	23	24	28	32
Goa	104	971	10000	10.1	90	86	74	110	115	107	114	118	115	101
Karnataka	300	1343	27000	14.9	341	347	357	357	400	412	399	414	390	403
Kerala	590	805	40000	11.9	560	553	531	522	524	517	431	414	609	475
Lakshadweep[1*]	132	152	4000	5.0	12	12	12	19	13	16	30	21	22	20
West Coast TOTAL	3473	626	377000	5.8	2156	2143	2135	2190	2246	2206	2159	2167	2331	2175
East Coast of India														
West Bengal	158	1032	17000	9.6	197	182	152	188	179	178	177	185	163	163
Odisha	480	329	26000	6.1	133	114	118	120	133	145	153	151	159	158

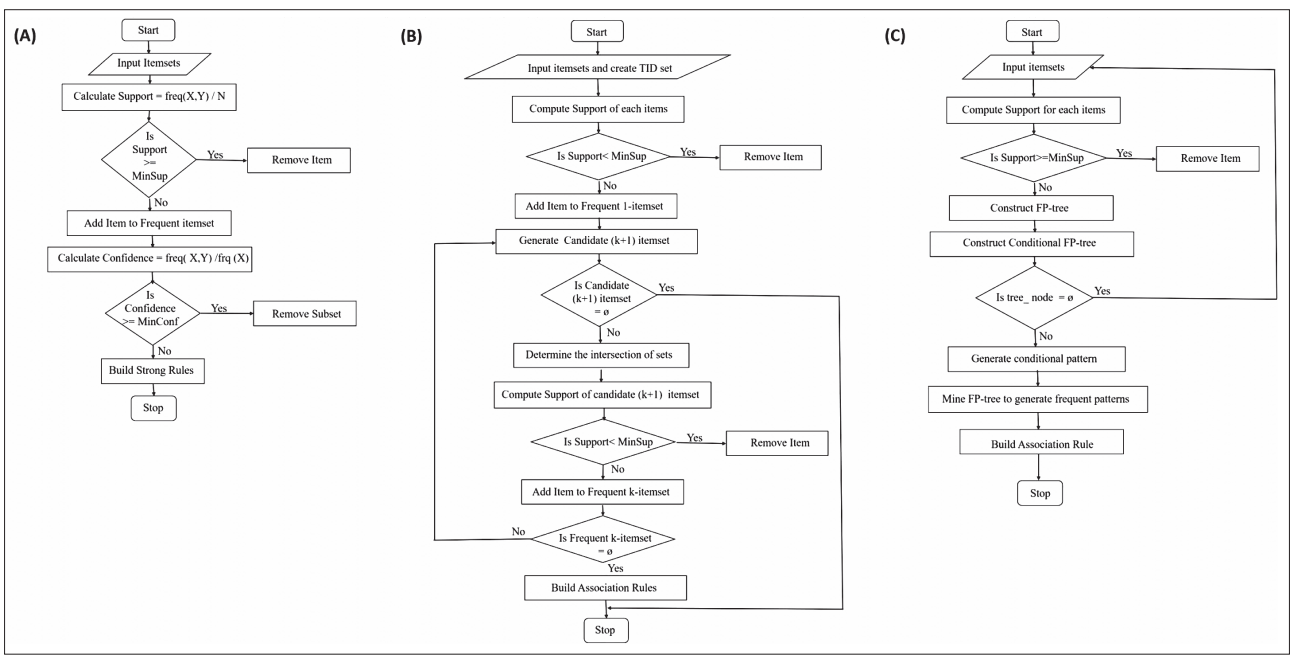
Table S2 continued

Andhra Pradesh	974	579	33000	17.1	289	433	414	438	475	520	580	605	600	564
Puducherry[1]	45	978	1000	44.0	36	38	36	38	42	47	46	42	40	44
Tamil Nadu	1076	542	41000	14.2	405	427	428	432	457	467	472	497	520	583
Andaman & Nicobar[1*]	1912	21	35000	1.1	34	35	36	37	37	37	39	39	40	40
East Coast TOTAL	4645	334	153000	10.1	1094	1229	1185	1253	1323	1394	1467	1519	1522	1552

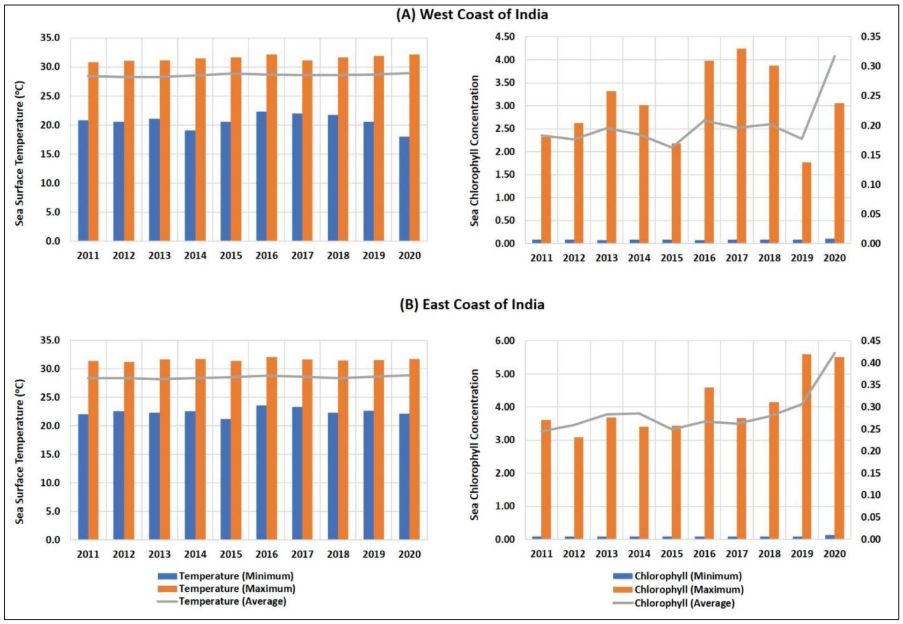
[1]Union territory; * Islands; # Data sourced from the Ministry of Fisheries, Animal Husbandry and Dairying, Government of India
** Fish production per km of coastline and per km^2 of continental shelf area was calculated based on 2019-2020 marine fisheries data.



Supplementary Fig. S1. Representation of the geospatial (1° labelled grid-wise) map from which multi-temporal marine environment data were collected from the Copernicus Marine Service (Sentinel satellite data) after atmospheric and terrain correction, editing and calibration.



Supplementary Fig. S2. Methodological flow charts of the association rule mining algorithms (A) Apriori, (B) ECLAT and (C) FP-Growth.



Supplementary Fig. S3. Temporal changes in the marine environment of west coast (A) and east coast (B) of India during 2011-2020.