Enhancing Riverine Flood Forecasting capability of Gaussian Process Classifier using Hyper-Parameter Optimization

Mirza Imran[1]*, Yatoo Nuzhat Ahmad[2] and Sheikh Firdos Alam[1]
1. Department of Computer Science Engineering, Poornima University, Jaipur, INDIA
2. B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, INDIA
*imranmirza100@gmail.com

Abstract
Climate change has increased the rate of melting glaciers and snow cover. The precipitation patterns have also changed along with the increased rate of melting snow and glaciers which contribute heavily to the floods in rivers. The floods are most damaging, whether we talk about the economy or human lives. This makes the forecasting or prediction of floods important. Several methods and techniques were tried for decades to predict floods. Machine Learning (ML) is the latest and most advanced technique used for forecasting and prediction in almost every field. So, in this study, we developed a machine learning model, Gaussian Process Classifier (GPC) with hyperparameter tuning by Random SearchCV to predict the flood risk in the Himalayan river Jhelum.

We pre-processed our dataset using techniques like feature scaling and data balancing. Feature importance evaluation was also done using an ensemble machine learning algorithm, extra trees classifier (ETC). Our model showed a ROC/AUC score of 0.803, average precision of 0.79 and the most important metric, the recall value of 0.85.

Keywords: Support vector machine, Floods, Himalayan rivers, Climate change, Flood prediction, Machine learning.

Introduction

Climate change has increased the rate of natural disasters and the prominent ones are hydrological disasters. The increase of carbon in the atmosphere and increase in temperature are the main contributors to flooding when talking about the glacier and snow-fed rivers[17]. Extreme weather and climate events are natural features of the climate system; nevertheless, as the climate changes, so do the frequency, severity, spatial extent, length and timing of these occurrences[5]. Rising greenhouse gas levels induce anthropogenic climate change, sometimes known as global warming.

By prohibiting radiation from escaping into space, these gases retain the heat into the atmosphere[1]. With the increase in temperature, the glaciers and snow cover are melting quickly[25] and with the heavy precipitation, the magnitude of floods has increased drastically[7]. This phenomenon has led to several glacial lake outbursts of floods (GLoF)[24]. The river floods occur more frequently than other types of hydrological disasters and leave a devastating mark on the pages of history whenever they occur. The frequency and intensity of flooding have increased a lot in Asia as shown in figure1. This is because the Himalayan rivers flow through various countries of Asia.

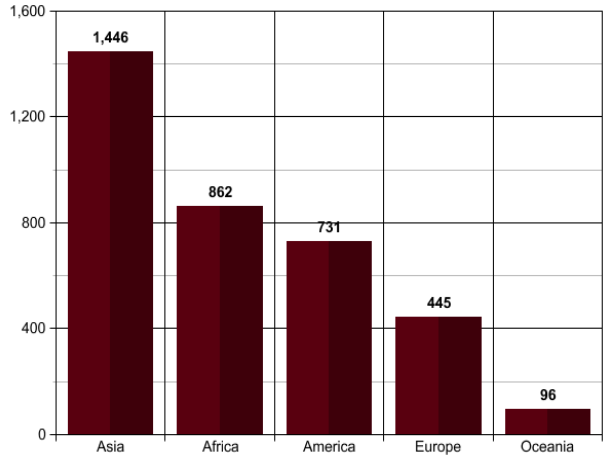

Fig. 1: Occurrences of Flooding continent wise 1999 to 2020

India, Pakistan and Bangladesh have also experienced an increase in the frequency and magnitude of floods[21]. The damages associated with these floods are also worth billions spanning different areas like farming, infrastructure, livestock, etc. The floods are not only devastating in developing countries but also in developed countries like the United States. According to a study, the United States incurred a loss of 3986 million USD per year since 1996-2016[36]. However, the major difference is in the type of damages the economically weak and developed nations incur. The developed countries incur more economic and less loss of human lives while on the other hand, developing nations where we have more homeless people or people living in mud houses, experience more loss of human lives.

Currently, around 58 million people are affected by river floods each year around the world, with Asia accounting for more than half of those affected. Global flood mortality increases quite dramatically with increasing temperature, rising from an average of almost 5,700 deaths each year in the reference climate to 9,700 (+70%), 11,500 (+103%) and 15,900 (+180%) with 1.5, 2°C and 3°C global temperatures respectively[11].

Keeping these consequences of floods in view, a common man can have an opinion that we should have a system that could predict the floods and warn the community. By delivering information that allows people and communities to secure their lives and livelihoods, early warning systems assist to decrease financial losses and the number of accidents and fatalities caused by disasters[14]. The need to predict any event or value gave rise to a State-of-the-Art method, Machine learning (ML). ML is used extensively in forecasting or predicting events. Machine learning is the science of computer algorithms that may learn to enhance their task performance based on their own prior experience or data[23]. Machine learning is a technique for teaching computers how to interpret data more effectively. Often, after seeing the data, we are unable to evaluate the extracted information.

Various ML models like Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), K-Nearest Neighbour (KNN), Decision Trees (DT), Bagging, Boosting etc. have been used to carry out several tasks in different fields. An algorithm is chosen according to the nature of the problem and dataset[19]. One of the algorithms which are used in prediction is Gaussian Process Classifier (GPC) and its another variant is Gaussian Process Regressor (GPR). A Gaussian process is a stochastic process that may be found in many disciplines of research, including data communication, networking and computer science. It is commonly used to represent the properties of the target system as a non-parametric and probabilistic technique. For noisy, distorted, or erroneous data, a Gaussian process is a good option. These qualities made the Gaussian process be utilized in different fields of research. Bui et al[9] used Gaussian process to predict the workload of each core of CPU in a computer system and showed its application in energy efficiency. However, they used GPR and described it as slow when paired with huge datasets.

Zhu and Chen[37] employed the Gaussian process in combination with partial least squares analysis to predict and to evaluate the energy efficiency of large-scale chemical plants. The application of the algorithm is vast as long as it suits well to the posed problem. In healthcare, several researchers used GPC to predict different diseases and mortality rates[28]. Rinta-Koski et al[28] used GPC to predict the death rate in hospitalized infants. The authors used the Receiver Operating Characteristic curve (ROC) which is an excellent metric to measure the efficiency of a classification model.

In 2014, Mehdipour et al used GPR to predict the mortality rates of children under age five[20]. For mortality rate prediction, Ludkovski et al[18] used GPR and they also worked on the mortality rate improvement factors. Gaussian processes are also used in forecasting several environment variables like temperature, humidity, rainfall[22], wind speed[16] etc. Grbic et al[15] developed a model using GPR to predict the daily mean water temperature of Drava river Croatia. GPR is also used in material science to measure different parameters.

Zhang and Xu[35] used it to predict the critical temperature of doped magnesium boride. The authors have also compared and found that GPR shows better results than SVM in their case. But, this is not the case in every situation, rather it depends on the type of problem and type of data. Another important application of the Gaussian process is a prediction of water level whether it is groundwater level[4], reservoir water level, or river water level. In this study our objective is to predict the flood risk for that we propose a hyperparameter tuned Gaussian Process Classifier model to predict the water level of river Jhelum Jammu and Kashmir, India at Sangam gauge.

Dataset

The Jhelum River is a perennial river that flows throughout the year and gets most of its water from 24 tributaries sourced from the snow-capped mountains of Pir Panjal and Himalayas but here in this research work we considered four watersheds which converge at Sangam gauge station. These four watersheds have three India Meteorological Department (IMD) stations to measure the precipitation and temperature. We obtained rainfall and temperature data from the IMD and water level data of the Jhelum river from the Irrigation and Flood Control Department of Jammu and Kashmir.

The rainfall and temperature data of three meteorological stations viz. Pahalgam, Kokernag and Qazigund were taken into consideration as shown in figure 2. The map has been developed using ArcMap 10.3. The water level data from the Sangam gauge station was used as an output variable. We combined the datasets and created a new dataset with six input variables of temperature and rainfall from three meteorological stations and output variable, water level from Sangam gauge station. The dataset is composed of daily data for these variables for ten years.

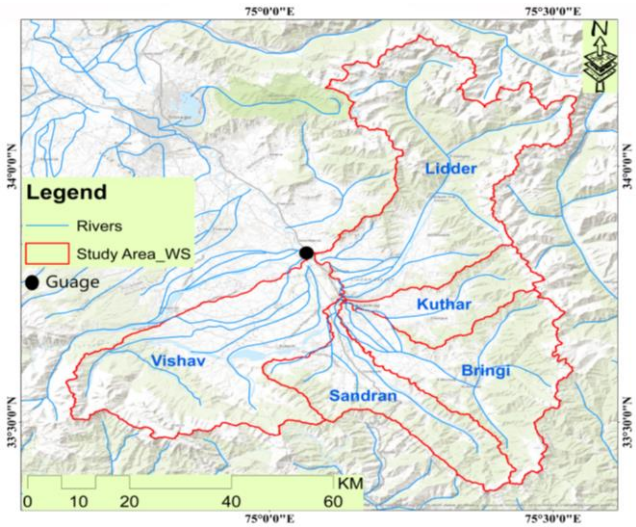

Fig. 2: Watershed map of study area with three IMD stations

Material and Methods

To develop the model we used Python 3.8.5 which is powerful and easy to code programming language. It includes high-level data structures like list and associative arrays (also known as dictionaries), dynamic typing and dynamic binding, modules, classes, exceptions and automated memory management, among other things. It features a syntax that is both basic and beautiful[30]. The tool we used to code python and develop our model is the Jupyter notebook. The Jupyter notebook is an open-source, browser-based application that serves as a virtual lab notebook for processes, code, data and visualizations related to the research process[10].

The dataset was free from any missing values, so it was not necessary to use any technique to address the issue. In pre-processing the data, we used a boxplot to detect outliers. The boxplot approach is a graphically-based way of finding outliers that are appealing not just for their simplicity, but also for the fact that it does not employ extreme possible outliers when computing a measure of dispersion[31]. To remove the outliers we used Inter-Quartile Range and then replaced these extreme values with median imputation. In the boxplot, the (IQR) refers to the distance between the lower quartile (Q1) and higher quartile (Q3). Inner "fences" are 1.5 IQR below Q1 and above Q3, while outside "fences" are 3 IQR below Q1 and above Q3.

A potential outlier is a value that falls within the inner and outer fences, but a number that falls outside the outer fences is a probable outlier. A value that falls outside the outer fence is most certainly an outlier[29]. In circumstances where the data distribution is skewed and there are outliers, median imputation is preferred. The mean mode and median imputation methods are simple yet powerful to deal with the missing values[32]. Our dataset is a bit imbalanced needed to be balanced and for that purpose, we used Synthetic Minority Over-sampling Technique (SMOTE) which has been proven to improve the results[2]. The standard scaler feature scaling method was utilized to normalize the range of features as it is compulsory to scale the features having a difference in magnitude, units and range so that the machine learning algorithm can interpret all of the features on the same scale[6].

The Standard Scaler (SS) is a feature scaling approach that removes the mean of each feature and scales the variation to one. Because the normalized value is determined only by the mean and variance, it has several advantages including being linear, reversible, rapid and extremely scalable[12]. To determine the feature importance, an ensemble machine learning technique Extra Tree Classifier (ETC) was used. It is important to evaluate the importance of features so that we can decide which feature contributes to the results and which one does not.

According to the traditional top-down process, the Extra-Trees technique generates an ensemble of unpruned decision or regression trees[13]. The data was divided into training and testing datasets. Two performance evaluation metrics ROC/AUC, average precision and recall value were chosen to evaluate the developed model. These metrics were chosen in view of the nature of the research problem. In our case we cannot afford to have huge number of false negatives. The flood risk prediction can tolerate the false positive rate to some acceptable level but false negative rate should be as low as possible.

For example, if in a situation, the developed model predicts flood but that turns out to be a false alarm, the only thing that can happen is that people and administration will take precautionary measures and be ready for the catastrophe. But if our system predicts that there is no risk of flood but the disaster strikes, then the economic and life loss will be huge. Recall is the number of recovered relevant items as a percentage of all relevant things for any given retrieved collection. As a result, recall is a measurement of efficacy in recovering (or choosing) performance and it can also be interpreted as a measure of inclusion of relevant things in the recovered set[8]. When compared to overall accuracy, the area under the ROC curve (AUC) has been found to display a variety of admirable characteristics as a classification performance indicator[3].

The ML model we developed in this study is based on Gaussian Process, abbreviated as GP, which is a generalisation of the Gaussian probability distribution or in other words we can say that a Gaussian process is a group of random variables with combined Gaussian distributions in any limited number of them[27]. GP's are also known as bell function because of the bell curve in distribution graph as shown in figure 3.

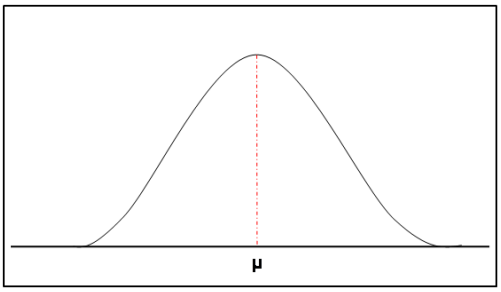

Fig. 3: Bell curve of a Gaussian Process

The distribution of random variables is summarised by Gaussian probability distribution functions, while the features of the functions are summarised by Gaussian processes. The continuous probability distribution is the connection between the occurrences for a continuous random variable and their probabilities, which is expressed by a probability density function expressed by the following equation:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}}\, e^{\frac{-1}{\sqrt{2\pi\sigma^2}}(x-\mu)^2} \quad (1)$$

where **σ** is standard deviation, **σ**^{2} is variance and μ is mean.

The cumulative distribution function, or CDF, defines the likelihood of an occurrence being equal to or less than a certain number. The CDF for discreet variable and continuous variable can be defined as shown in equation 2 and 3 respectively:

$$FX(x) = P(X \leq x) \quad (2)$$

$$FX(x) = \int_{-\infty}^{x} FX(x)dx \quad (3)$$

where X is a random variable which takes values from the sample space, x is the value and P is the probability.

The values for the parameters of the GPC model were set to default and given in table 1.

Random Search hyperparameter tuning technique was used to tune the hyperparameters of the model. The value for cross-validation was varied and desired results were achieved at 5. To improve the results further, we varied the values of "max_iter_predict, n_restarts_optimizer and random_state" that are given in table 2.

Results and Discussion

We used the boxplot to detect outliers in the dataset of the three meteorological stations as shown in figure 4. The results show that only precipitation data has outliers and temperature data is free from outliers. The detection of outliers is very important in the development of classification models and that also conforms with the research work done by Shieh and Hung[33]. As we experimented and executed the model without outlier removal, the results were unacceptable, so they were removed by the Inter Quartiles and then those missing values were replaced with median imputation. The boxplot of the data after median imputation is shown in figure 5. The developed GPC model before hyperparameter tuning showed a very good recall value of 0.85 but average results for the average precision and ROC/AUC score.

The model was then subjected to hyperparameter tuning technique and the method chosen was random search instead of grid search because upon experimenting we experienced that grid search was extremely slow in operation, same is the finding of Swamy et al[34].

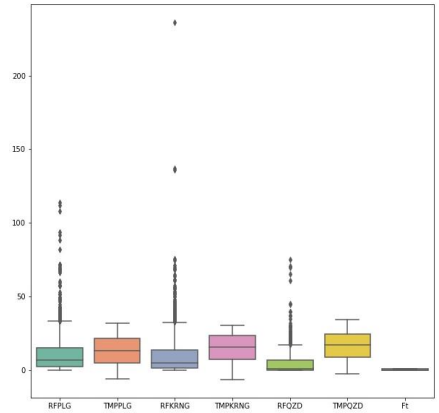

Fig. 4: Boxplot before outlier removal

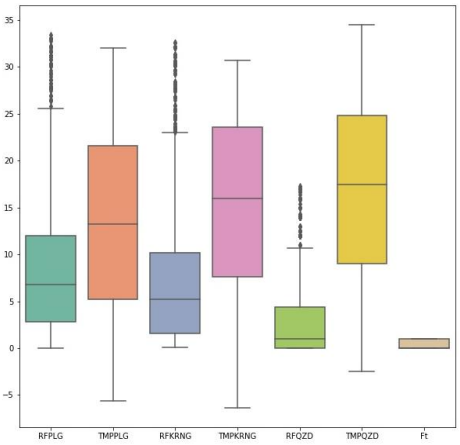

Fig. 5: Boxplot after outlier removal

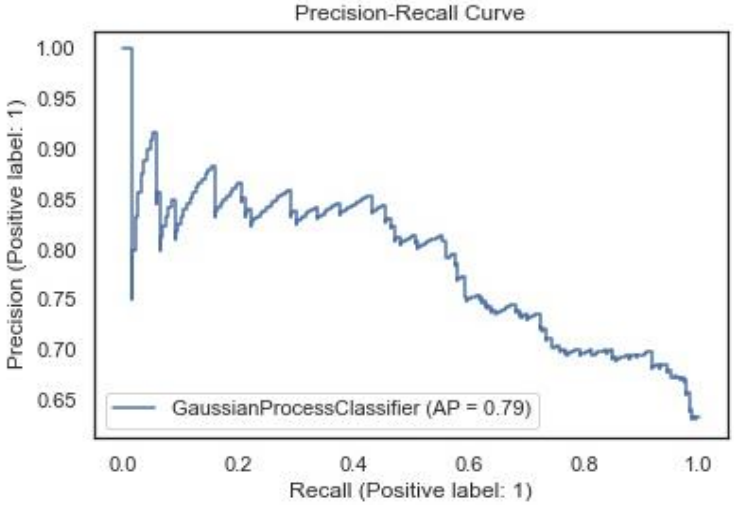

Fig. 6: Precision-Recall curve of developed GPC model

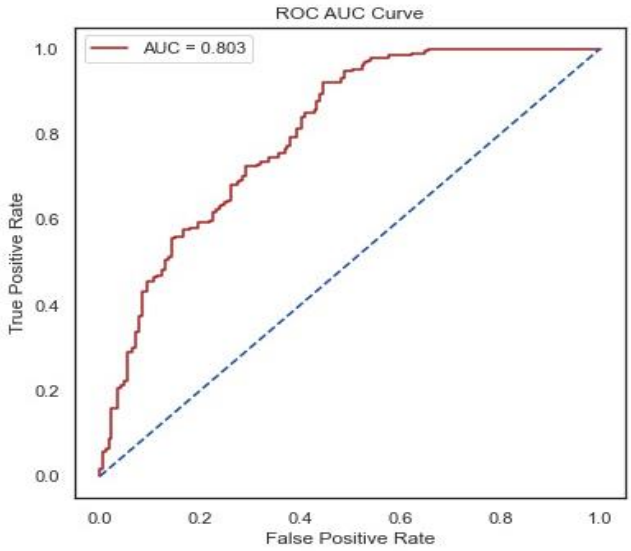

Fig. 7: ROC/AUC of developed GPC model

Table 1
Default parameters of GPC model

Kernel	max_iter_predict	n_restarts_optimizer	random_state	optimizer
RBF	100	0	0	fmin_1_bfgs_b

Table 2
Changed values of the parameters of GPC model

Kernel	max_iter_predict	n_restarts_optimizer	random_state	optimizer
RBF	800	15	1	fmin_1_bfgs_b

Table 3
Results of the developed model before and after hyperparameter tuning

Metric	GPC without hyperparameter tuning	GPC after hyperparameter tuning
ROC/AUC	0.7153	0.8032
Recall	0.8537	0.8537
Average Precision	0.69	0.79

The cross-validation score of 5 provided the best results for which the precision/recall curve and ROC/AUC score are shown in figures 6 and 7 respectively. The results for the developed model before hyperparameter tuning and after hyperparameter tuning are shown in table 3.

Conclusion

The rainfall and temperature data of the snow-capped mountainous region like the Himalayas can be used to predict the risk of flood in the rivers that get most of the water from melting snow and glaciers. By looking at the data, we cannot determine which features contribute to the output and which ones do not. To solve that issue, evaluating feature importance is a must to do practice. So to accomplish that task, an ensemble machine learning algorithm like ETC proved to be very useful. The outlier detection and removal of the data are necessary as neglecting this part can degrade the results. IQR technique to remove the outliers and median Imputation to replace the missing values play a significant role in improving the results.

To improve the results, further hyperparameter tuning technique is important and after experimentation it implied that between two major hypertuning techniques, grid search and random search, the later one is fast. The grid search technique took a lot of time to execute. By using this technique, the recall value remained the same but ROC/AUC and average precision score increased.

Acknowledgement

We are thankful to Irrigation and Flood control Department of Jammu and Kashmir and India Meteorological Department for providing data to do this research work.

References

1. Aalst M.K. Van, The impacts of climate change on the risk of natural disasters, *Disasters*, **30**, 5–18 **(2006)**

2. Ahsan M., Gomes R. and Denton A., SMOTE implementation on phishing data to enhance cybersecurity, IEEE International Conference on Electro/Information Technology (EIT) IEEE, 531–536 **(2018)**

3. Andrew P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern Recognition*, **30**, 1145–1159, https:doi:https://doi.org/10.1016/S0031-3203(96)00142-2 **(1997)**

4. Band S.S. et al, Mechanics Groundwater level prediction in arid areas using wavelet analysis and Gaussian process regression, *Engineering Applications of Computational Fluid Mechanics*, **15**, 1147–1158, https: doi:10.1080/19942060.2021.1944913 **(2021)**

5. Banholzer S., Kossin J. and Donner S., The impact of climate change on natural disasters, In Reducing disaster: Early warning systems for climate change, Springer, Dordrecht, 21–49, doi:10.1007/978-94-017-8598-3 **(2014)**

6. Bollegala D., Dynamic Feature Scaling for Online Learning of Binary Classifiers, *Knowledge-Based Systems*, **129**, 97–105, doi: https://doi.org/10.1016/j.knosys.2017.05.010 **(2017)**

7. Bronstert A., Floods and Climate Change: Interactions and Impacts, *Risk Analysis*, **23**, 545–557, doi: https://doi.org/10.1111/1539-6924.00335 **(2003)**

8. Buckland M. and Gey F., The Relationship between Recall and Precision, *Journal of the American Society for Information Science*, **45**, 12–19 **(1994)**

9. Bui D.M. et al, Gaussian process for predicting CPU utilization and its application to energy efficiency, *Applied Intelligence* **43**, 874–891, doi:10.1007/s10489-015-0688-4 **(2015)**

10. Cardoso A., Leitão J. and Teixeira C., Using the Jupyter Notebook as a Tool to Support the Teaching and Learning Processes in Engineering Courses, International Conference on Interactive Collaborative Learning, 227–236, doi:10.1007/978-3-030-11935-5 **(2020)**

11. Dottori F. et al, Increased human and economic losses from river flooding with anthropogenic warming, *Nature Climate Change*, **8**, 781–786, doi:10.1038/s41558-018-0257-z **(2018)**

12. Ferreira P., Le D.C. and Zincir-Heywood N., Exploring Feature Normalization and Temporal Information for Machine Learning Based Insider Threat Detection, *Pattern Recognition Letters*, **128**, 544–550 **(2019)**

13. Geurts P., Ernst D. and Wehenkel L., Extremely randomized trees, *Machine Learning*, **63**, 3–42, doi:10.1007/s10994-006-6226-1 **(2006)**

14. Grasso V.F. and Singh A., Early warning systems: State-of-art analysis and future directions, Draft report, UNEP, 1 **(2011)**

15. Grbic R., Kurtagic D. and Sliškovic D., Expert Systems with Applications Stream water temperature prediction based on Gaussian process regression, *Expert Systems with Applications*, **40**, 7407–7414, doi:10.1016/j.eswa.2013.06.077 **(2013)**

16. Hoolohan V., Tomlin A.S. and Cockerill T., Improved near surface wind speed predictions using Gaussian process regression combined with numerical weather predictions and observed meteorological data, *Renewable Energy*, **126**, 1043–1054, doi: https://doi.org/10.1016/j.renene.2018.04.019 **(2018)**

17. Loukas A. and Quick M.C., The effect of climate change on floods in British Columbia, *Hydrology Research*, **30**, 231-256 **(1999)**

18. Ludkovski M., Risk J. and Zail H., Gaussian process models for mortality rates and improvement factors, *ASTIN Bulletin: The Journal of the IAA*, **48**, 1307-1347 **(2018)**

19. Mahesh B., Machine learning algorithms-a review, *International Journal of Science and Research (IJSR)*, **9**, 381-386, doi:10.21275/ART20203995 **(2020)**

20. Mehdipour P. et al, Application of Gaussian Process Regression (GPR) in estimating under-five mortality levels and trends in Iran 1990-2013 study protocol, *Archives of Iranian Medicine*, **17**, 189–192 **(2014)**

21. Mirza M., Climate change, flooding in South Asia and implications, *Regional Environmental Change*, **11**, 95-107, doi:10.1007/s10113-010-0184-7 **(2011)**

22. Mishra N. and Kushwaha A., Rainfall Prediction using Gaussian Process Regression Classifier, *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, **8**, 392–397 **(2019)**

23. Mjolsness E. and Decoste D., Machine Learning for Science: State of the Art and Future Prospects, *Science*, **293**, 2051–2055, doi:10.1126/science.293.5537.2051 **(2009)**

24. Rafiq M. et al, Modelling Chorabari Lake outburst flood, Kedarnath, India, *Journal of Mountain Science*, **16**, 64–76 **(2019)**

25. Rafiq M. and Mishra A., Investigating changes in Himalayan glacier in warming environment: a case study of Kolahoi glacier, *Environmental Earth Sciences*, **75**, 1–9, doi:10.1007/s12665-016-6282-1 **(2016)**

26. Randles B.M. et al, Using the Jupyter Notebook as a Tool for Open Science: An Empirical Study, ACM/IEEE Joint Conference on Digital Libraries (JCDL), IEEE, 1–2 **(2017)**

27. Rasmussen C.E., Gaussian Processes in Machine Learning, Summer school on machine learning, Springer, Berlin, Heidelberg, **2**, 63–71, https:doi:https://doi.org/10.1007/978-3-540-28650-9_4 **(2003)**

28. Rinta-koski O. et al, Neurocomputing Gaussian process classification for prediction of in-hospital mortality among preterm infants, *Neurocomputing*, **298**, 134–141, https:doi:10.1016/j.neucom.2017.12.064 **(2018)**

29. Salgado C.M. et al, Noise Versus Outliers, Secondary Analysis of Electronic Health Records, 163–183, https:doi:10.1007/978-3-319-43742-2 **(2016)**

30. Sanner M.F., Python: a programming language for software integration and development, *J Mol Graph Model*, **17**, 57–61 **(1999)**

31. Schwertman N.C., Owens M.A. and Adnan R., A simple more general boxplot method for identifying outliers, *Computational Statistics & Data Analysis*, **47**, 165–174, https:doi:10.1016/j.csda.2003.10.012 **(2004)**

32. Sessa J. and Syed D., Techniques to Deal with Missing Data, 5th International Conference on Electronic Devices, Systems and Ppplications (ICEDSA), IEEE, 1–4 **(2016)**

33. Shieh A.D. and Hung Y.S., Data Detecting Outlier Samples in Microarray Data, *Statistical Applications in Genetics and Molecular Biology*, **8**, 1-24 **(2009)**

34. Swamy S., Kundale J. and Jadhav D., Sentiment Analysis of Multilingual Mixed-Code, Twitter Data Using Machine Learning Approach, International Conference on Innovative Computing and Communications, **1388**, 683–697, doi:doi.org/10.1007/978-981-16-2597-8_58 **(2022)**

35. Zhang Y. and Xu X., Predicting doped MgB 2 superconductor critical temperature from lattice parameters using Gaussian process regression, *Physica C: Superconductivity and its Applications*, **573**, 1353633, doi:10.1016/j.physc.2020.1353633 **(2020)**

36. Zhou Q., Leng G. and Peng J., Recent changes in the occurrences and damages of floods and droughts in the United States, *Water*, **10**, 1109, doi:10.3390/w10091109 **(2018)**

37. Zhu L. and Chen J., Energy efficiency evaluation and prediction of large-scale chemical plants using partial least squares analysis integrated with Gaussian process models, *Energy Conversion and Management*, **195**, 690–700 https:doi:10.1016/j.enconman.2019.05.023 **(2019)**.

(Received 24th March 2024, accepted 12th September 2024)