PROJECT REPORT

CLIMATE MODEL SIMULATION CRASHES

SUBMITTED BY

KENCHE KOUSHIK

12217949

SECTION: KM067

COURSE CODE: INT 354

UNDER THE GUIDENCE OF Dr. Ankita Wadhavan

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

ABSTRACT

In this paper, we have used the dataset “Climate Model Simulation Crashes” which has 540 instances and 21 features. The preprocessing technique used for this model is standard scaler and the methods used are logistic regression, SVM and KNN.

Keywords: KNN, SVM, accuracy, MSE, Standard Scaler

INTRODCTION

Modern global three-dimensional climate models are extraordinarily complex pieces of science (e.g., Randall et al., 2007; Gent et al., 2011; The HadGEM2 Development Team, 2011) and software engineering (Easterbrook et al., 2011; Rugaber et al., 2011; Easterbrook, 2010). They contain over a million lines of code (Easterbrook and Johns, 2009; Easterbrook, 2012) and use hundreds to thousands of files, functions, and subroutines to solve equations of state and conservation laws for the flows of matter, energy, and momentum within and between the atmosphere, oceans, land, and other reservoirs of the Earth system (Washington and Parkinson, 2005). They also use numerous algorithms of biological, chemical, geologic, and anthropogenic processes to simulate the cycles of carbon, nitrogen, sulphur, aerosols, ozone, greenhouse gases, and other climate-relevant quantities of interest. To compound this complexity, these algorithms operate across many orders of magnitude in space and time, and contain constituents that exist in gas, liquid, solid and mixed phases.

Given this enormous range of scientific complexity, climate models are vulnerable to many types of software design and implementation issues. Software issues aside, many potential problems still arise with scientific representations in complex models. As code verification can be used to find software bugs, emerging tools being developed in the field of uncertainty quantification (UQ) can help pinpoint scientific discrepancies in simulation models, the knowledge of which can be used to guide and improve model development. Primary UQ targets for climate models are schemes containing parameters with adjustable values. Small perturbations to the values of the adjustable parameters can amplify and lead to large changes in simulation outputs. In some cases, the simulations may fail altogether. So, our goal is to use classification to predict simulation outcomes (fail or succeed) from input parameter values.

The rest of the paper is structured as follows: Literature Review where the work of previous researchers is mentioned; Proposed Methodology where the methods used for implementation of model are mentioned; Experimental Analysis where the result of the model is discussed; Conclusion and Future Scope.

LITERATURE REVIEW

Peter D. Dueben and Peter Bauer (2018) have worked on the paper "Challenges and design choices for global weather and climate models based on machine learning" whose objective is to identify challenges and fundamental design choices for a forecast system. The dataset which they have used is ERA5. The results of their work show that it may indeed be possible to generate global weather predictions based on NNs for short-range prediction but whether a NN prediction system will ever be competitive with state-of-the-art weather prediction models remains an open question. It would certainly require a serious level of complexity with as many (or more) degrees of freedom as conventional models. [1]

D. D. Lucas et al. (2013) have worked on the paper "Failure analysis of parameter-induced simulation crashes in climate models" whose objective to predict the success of a climate simulation model. The dataset used is Climate Crashes Simulation Model. A highly predictive SVM classification system was trained from a dataset containing only 32 failure instances out of 360 simulations and validated using an independent set of 180 simulations. The resulting classification system had a prediction AUC score exceeding 0.96 and achieved discrimination accuracies above 97%. Global sensitivity analysis was then used to identify eight model parameters from four different modules that drive high probabilities of failing, the results of which can be used to increase the robustness of CCSM4 to parameter perturbations. These methods can be used to characterize simulation failures in other complex scientific computer models.[2]

Table 1: Summarised review of literature

AUTHORS	DATASET	DESCRIPTION	MODEL USED
Peter D. Dueben and Peter Bauer[1]	ERA5	To identify challenges and fundamental design choices for a forecast system	Neural Networks (NN)
D. D. Lucas et al.[2]	Climate Crashes Simulation Model	To predict the success of a climate simulation model	Support Vector Machines (SVM)

PROPOSED METHODOLOGY

The libraries used to implement the model are:

1. Pandas library
2. Scikit-Sklearn library
3. Matplotlib library

The steps followed to build the prediction model are

1. Dataset

The first step to implement this model is to import dataset. The dataset is of 540 instances and 21 features including the target feature. We import the dataset using pandas library.

2. Data Preprocessing

The first thing we have done after importing the dataset is to check whether the dataset has any null values in it. On ensuring that no null values are present, we split the data into two groups. One group is input group and other is target. We then use "train_test_split" method to split the both input data and target data into train set and test set. We then preprocess the dataset. For this, we use "StandardScaler" method in which we have standardised the data.

3. Model Training

The next step is to train the model. For this we have used three different algorithms to implement our prediction model.

a. Logistic Regression: Logistic regression is a supervised machine learning algorithm used for classification tasks where the goal is to predict the probability that an instance belongs to a given class or not. Logistic regression is a statistical algorithm which analyse the relationship between two data factors.
b. SVM: Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. SVM algorithms are very effective as we try to find the maximum separating hyperplane between the different classes available in the target feature.
c. KNN: K-Nearest Neighbors (KNN) is a supervised machine learning algorithm employed to tackle classification and regression problems. It is widely disposable in real-life scenarios since it is non-parametric, meaning it does not make any underlying assumptions about the distribution of data.

4. Prediction

After training the model, predict the output of the model using “accuracy_score” method for both training set and testing set. Also, calculate the mean squared error for both training set and testing set and finally compare all the three models.

EXPERIMENTAL ANALYSIS

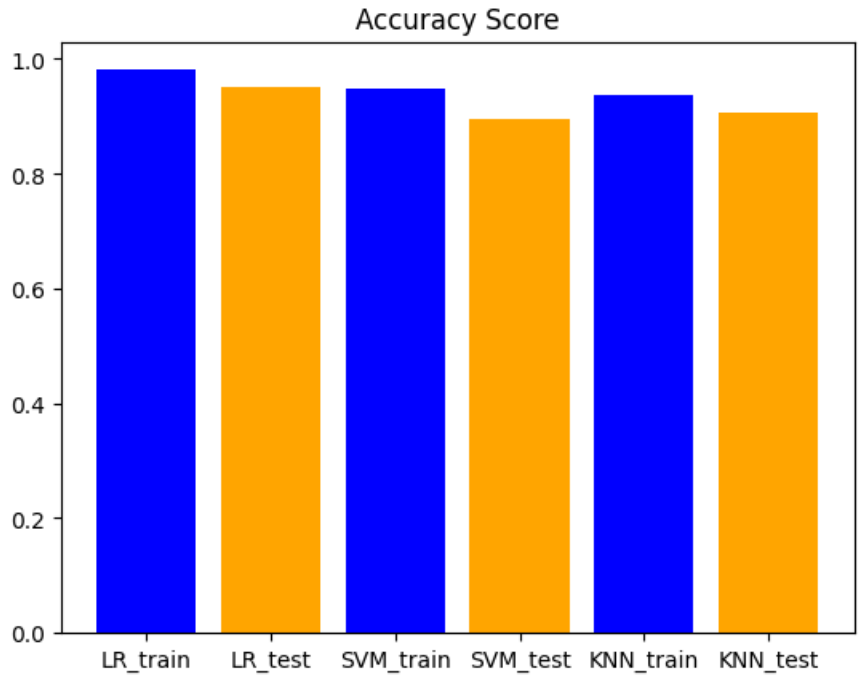
After training the model we predict the accuracy of the model using “accuracy_score” model.

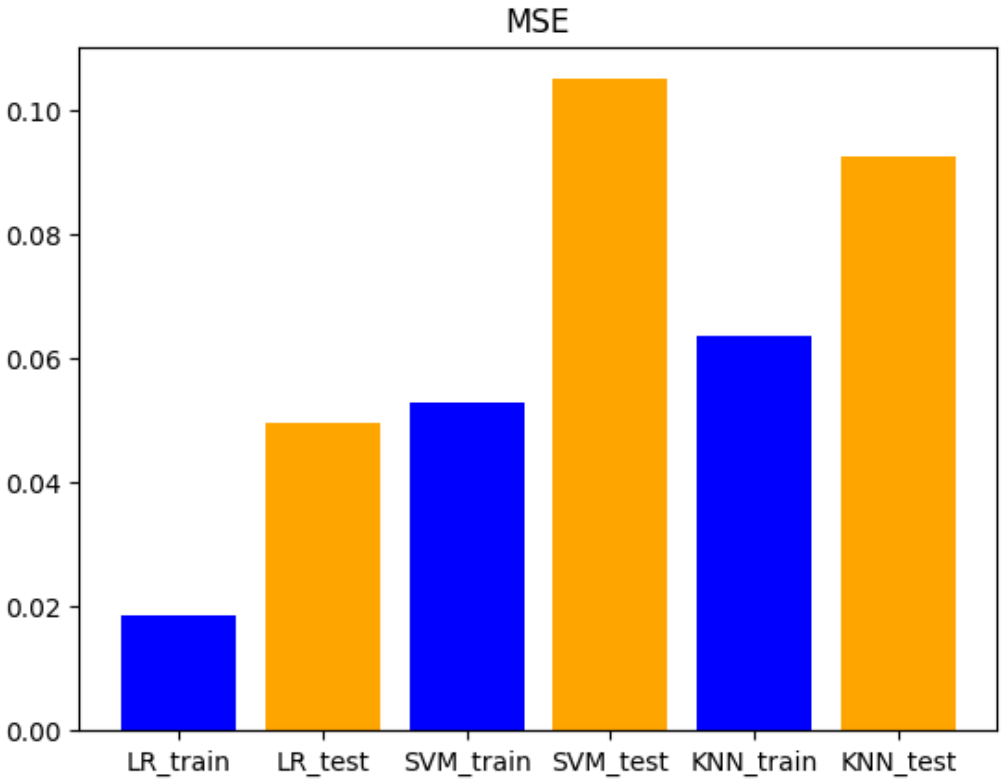
Table 2: Accuracy and mean squared error of Training dataset

MODEL	ACCURACY (accuracy_score)	MSE
Logistic Regression	0.9814	0.0185
Support Vector Machine (SVM)	0.9471	0.0529
K-Nearest Neighbors (KNN)	0.9365	0.0634

Table 3: Accuracy and mean squared error of Testing dataset

MODEL	ACCURACY (accuracy_score)	MSE
Logistic Regression	0.9506	0.0493
Support Vector Machine (SVM)	0.8951	0.1049
K-Nearest Neighbors (KNN)	0.9074	0.0925





From the above tables (Table 2 and Table 3), it is proved that the accuracy of the model is highest for logistic regression algorithm and also, the mean squared error for logistic regression is least compared to the other two algorithms (SVM and KNN).

CONCLUSION AND FUTURE SCOPE

The goal of our project is to develop a model to predict the outcomes of the simulation and from the above analysis, we found that the model with logistic regression algorithm has highest accuracy. Thus, our model is able to predict the outcomes of the simulation with an accuracy of 95%.

FUTURE SCOPE

The future scope of this project is, if the dataset has more records, then the model can be trained to predict more accurately than right now.

REFERENCES

[1] Dueben, P. D. and Bauer, P.: Challenges and design choices for global weather and climate models based on machine learning, Geosci. Model Dev., 11, 3999–4009, https://doi.org/10.5194/gmd-11-3999-2018, 2018.

[2] Lucas, D. D., Klein, R., Tannahill, J., Ivanova, D., Brandon, S., Domyancic, D., and Zhang, Y.: Failure analysis of parameter-induced simulation crashes in climate models, Geosci. Model Dev., 6, 1157–1171, https://doi.org/10.5194/gmd-6-1157-2013, 2013.

Code:

https://drive.google.com/drive/folders/1Spf2CviQin84fsOV2yBumjsxdOzzgkmP?usp=sharing