Pilot-Study Proposal

Assignment: *Design and Application of a Machine Learning System for a Practical Problem*

CE802-Machine Learning And Data Mining

Name: Thara Jeni Samala Anthony Raj

Reg. No: 2200940

MSC Data Science with Professional Placement

Word Count: 673

Area of research and Domain of research:

Machine Learning along with its techniques, mainly on Classifiers and Regressions.

Objective of the proposal:

The aim of this proposal is to investigate the performance of a number of machine learning procedures on the given training set to predict the best procedure to perform on a given test data set.

Short description:

The given training set contains the information about several features of the customers and whether they are currently struggling with paying their energy bills or not. The task is to predict whether a customer is going to suffer from the increasing energy prices or not.

Type of Predictive task:

The given data set comes under supervised machine learning, since machines are using labelled data (given training set) to predict results. The finding results are discrete values (to predict Yes or No) not a numeric values. Therefore the identified type of predictive task is "Classification type".

Examples of possibly informative features:

I would like to be provided with the following possibly informative features to proceed further with machine learning techniques:

- Customer's income – will class them based on annual income
- Customer's Expenditure – to know his annual expenses is necessary
- Dependencies – total members depending on customer's income
- Loan details – to know whether he has additional cost to repay or not
- House details –customer staying in rental/own house
- Vehicle details, if any – to know the extra additional maintenance cost

Learning Procedures:

There are many machine learning procedures for predictive analysis. For this problem, we have to investigate classifiers as well as regressions, so I would suggest the following learning procedures:

- Decision Tree & Random Forest
- Support Vector Machine
- Linear Regression
- K-NN

The reason for my choice :

1)Decision Tree:
It classifies unknown records very fast. In the presence of redundant attributes decision tree work very good. Decision trees are somewhat strong in the presence of noise if the methods likes over fitting are provided. It has many applications in medicine, image processing and intrusion detection[2].

2)Random Forest:
It is fast & scalable and robust against overfitting. It gives better results with the increasing number of examples[3].

3)Support Vector Machine:
It works relatively more effectively for high-dimensional datasets, because the complexity of the training dataset does not depend upon the dimensionality of the dataset[1].

4)Linear Regression:
Works with almost any kind of dataset &gives information about the relevance of features.

Evaluation Method:

After making each technique to gain knowledge from the training set and test set with the help of labels to predict events. Will check the system is capable of providing results to an input data with adequate training process. Before deploying it, I would evaluate the performance of machine by the accuracy scores of the machine learning procedures. I would compare the machine learning procedure's results with actual and expected results to identify errors to change the model based on results. After that I would compare the accuracy score of each investigating procedure with before and after GridSearchCV. The accuracy score indicates the total number of correct predictions, predicted by linear procedures. Which learning procedure's score is high, I would suggest it as best model.

References:

1) R. Saravanan and P. Sujatha, "A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification," *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, 2018, pp. 945-949, doi: 10.1109/ICCONS.2018.8663155.
2) M. Somvanshi, P. Chavan, S. Tambade and S. V. Shinde, "A review of machine learning techniques using decision tree and support vector machine," *2016 International Conference on Computing Communication Control and automation (ICCUBEA)*, Pune, India, 2016, pp. 1-7, doi: 10.1109/ICCUBEA.2016.7860040.
3) Liang Zhang, Jin Wen, Yanfei Li, Jianli Chen, Yunyang Ye, Yangyang Fu, William Livingood, "A review of machine learning in building load predictions", Applied Energy,Volume285,2022,116452,ISSN03062619,*doi:10.1016/j.apenergy.2021.1164.*