CS 534 Group Project Assignment 3 (100 Points)

Due: 11:59 p.m. on 04/09/2023

Project Objective: The goals of this assignment are to help you understand the concepts and the algorithms that we will discuss/have discussed in Week 7 ~ Week 11.

Note:

*(1) This project is to be done by **EACH GROUP**. No help besides the textbook, materials, and the instructor/TA should be taken. Copying any answers or part of answers from other sources, including your classmate groups, will earn you a grade of zero.*
*(2) Your program must be developed and implemented in the PyCharm-like IDE, or 10% of the graded score is deducted. Please check and choose the one from here: https://realpython.com/python-ides-code-editors-guide/, as the suggestion. Note that we **DO NOT** use the Jupyter as the IDE.*
(3) Assignments are accepted in their assigned Canvas drop box without penalty if they are received by 11:59PM EST on the due date, or 10% of the graded score is deducted for the late submission per day. Work submitted after one week of its original due date will not be accepted.

Project Deliverables: Submit a **zip** file that includes your answers for the below questions in the **pdf** file and the **MachineFailurePredictorPipeline.py with the original raw data** to complete your group project assignment to Canvas.

A. Project Questions:

Question 1 (10 Points). Convert the following set of sentences into the conjunctive normal form (CNF).

S1: $A \Leftrightarrow (B \vee E)$
S2: $E \Rightarrow D$
S3: $C \wedge F \Rightarrow \neg B$

Note: Each answer needs to be shown in the step-by-step process. Each step must be provided with the reason, i.e., the rule of the standard logical equivalences.

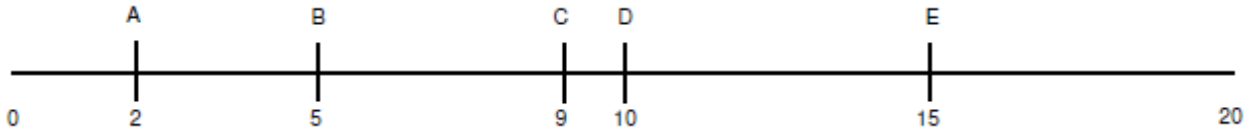
Question 2 (10 Points). Assuming predicates ***Parent(p, q)*** and ***Female(p)*** and constants ***Joan*** and ***Kevin***, with the obvious meanings, express each of the following sentences in first-order logic. (You may use the abbreviation $\exists^1$ to mean "there exists exactly one.")

1. Joan has a daughter (possibly more than one, and possibly sons as well).
2. Joan has exactly one daughter (but may have sons as well).
3. Joan has exactly one child, a daughter.
4. Joan and Kevin have exactly one child together.
5. Joan has at least one child with Kevin, and no children with anyone else.

Note: Each answer needs to be shown clearly.

Question 3 (10 Points): *K*-means Clustering

Five Customers' Rating on a New Car on a 20-point Scale.



(a) Assume $K = 2$ and the two initial centroids are 3 and 4.
   1. Use the *K*-means algorithm and show all the computational steps with the numerical answers and solution to determine the two-cluster solutions. **No point** is given if only the answer is provided.

   2. Show all the computational steps with the numerical answers and solution to calculate Silhouette Coefficient Index, Davies–Bouldin Index, and Calinski-Harabasz Index. **No point** is given if only the answer is provided.

(b) Assume $K = 2$ and the two initial centroids are 11 and 12.
   1. Use the *K*-means algorithm and show all the computational steps with the numerical answers and solution to determine the two-cluster solutions. **No point** is given if only the answer is provided.

   2. Show all the computational steps with the numerical answers and solution to calculate Silhouette Coefficient Index, Davies–Bouldin Index, and Calinski-Harabasz Index. **No point** is given if only the answer is provided.

(c) Use the results from (a) and (b) to determine which two-cluster solution should be chosen. Please describe and explain your answer in detail.

B. Project Development

Project (70 Points). In this project, we develop and implement a pipeline that will perform the data pre-preprocessing, train five machine-learning (ML) binary classifiers, select the best-trained models, and then test those best-trained models from which we can select the best one among five of them to develop an application to predict the machine failure in the future.

(1). Study and investigate the raw dataset provided by http://archive.ics.uci.edu/ml/datasets/AI4I+2020+Predictive+Maintenance+Dataset#.

(2). Read the paper (https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9253083) that will help you understand more about the meaning of each attribute in the raw dataset.

(3). As the raw dataset is severely imbalanced having only 339 datapoints labeled as machine failure ("1"), the rest of the datapoints labeled as machine normal ("0"). For the learning and practice purposes, we will perform the under-sampling by using **RandomUnderSampler**. That is, we would like to use 339 datapoints labeled as machine failure ("1") and 339 datapoints labeled as machine failure ("0") to train and test the ML binary classifiers. In total, the size of our datasets used for this project is 678 datapoints. You can review these two resources, as below, for your reference that will help you complete this step:

   (a) The "3. Random under-sampling with imblearn" section from this website (https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/) and
   (b) The "RandomUnderSampler" from this website (https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.RandomUnderSampler.html?highlight=randomundersampler).

   **** At the end of the above step, your code will display all the 678 data instances on the console output.**

(4). After (3), divide the above pre-processed dataset into training (70%) and testing (30%) sets (https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html). Using the training portion to do the following step and do not touch the testing dataset at this step.

   More specifically, your team needs to perform the 5-fold cross validation (https://scikit-learn.org/stable/modules/cross_validation.html#computing-cross-validated-metrics) on the training data to develop and implement **EACH** ML model by fine-tuning their hyperparameters until one set of their parameter values of each model delivers the best performance in terms of its F1-score (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html). The ML models with the indicated hyperparameters that your team needs to develop include:

(a) Artificial Neural Networks (*hidden_layer_sizes*, *activation*): https://scikit-learn.org/stable/modules/neural_networks_supervised.html
(b) Support Vector Machine (*C*, *kernel*): https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#
(c) BaggingClassifier (*n_estimators*, *max_samples*), where the default *DecisionTreeClassifier* is used: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingClassifier.html#
(d) AdaBoost (*n_estimators*, *learning_rate*), where the default *DecisionTreeClassifier* is used: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html
(e) Random Forest (*n_estimators*, *criterion*, *max_features*, *max_depth*, *max_samples*), where the default *DecisionTreeClassifier* is used: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

You can decide to use either GridSearch or RandomizedSearch or both to learn those hyperparameter values of each ML model.

(a) Hyperparameter Tuning - https://scikit-learn.org/stable/modules/grid_search.html#
(b) GridSearch - https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
(c) RandomSearch - https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html

**** At the end of this step, your code will display the below table on the console output.**

ML Trained Model	Its Best Set of Parameter Values	Its F1-score on the 5-fold Cross Validation on Training Data (70%)
Artificial Neural Networks	(hidden_layer_sizes, activation)	
Support Vector Machine	(C, kernel)	
BaggingClassifier	(n_estimators, max_samples, max_features)	
AdaBoost	(n_estimators, learning_rate)	
Random Forest	(n_estimators, criterion, max_depth, max_samples)	

(5). After (4), use the testing dataset to evaluate the performance among all the best-trained models obtained from above and then select the one that beats all the other four models in terms of their F1-scores.

**** At the end of this step, your code will display the below table on the console output, as well as indicate which the ML model among five should be the chosen one that will be used to predict the machine failure in the future.**

ML Trained Model	Its Best Set of Parameter Values	Its F1-score on Testing Data (30%)
Artificial Neural Networks	(hidden_layer_sizes, activation)	
Support Vector Machine	(C, kernel)	
BaggingClassifier	(n_estimators, max_samples, max_features)	
AdaBoost	(n_estimators, learning_rate)	
Random Forest	(n_estimators, criterion, max_depth, max_samples)	

(6). Submit the **original raw dataset** and your **pipeline** (**MachineFailurePredictorPipeline.py**), including (3), (4), and (5), using the given construct to Canvas. That is, when your TA executes your pipeline with your provided raw data, all the above outputs will be displayed on your TA's computer console output.

```
def main():
    pass

if __name__ == "__main__":
    main()
```

Grading Criteria: Your answers must be complete and clear.

Checkpoints	Points Possible
Project Question 1	10 Points
Project Question 2	10 Points
Project Question 3	10 Points
Project Development	70 Points
(1). Proper Naming Conventions and Program Documentation on Your Codes	5 Points
(2). Compliable Code: MachineFailurePredictorPipeline.py	5 Points
(3). Random under-sampling	5 Points
(4). Pre-processed Dataset into Training (70%) and Testing (30%)	5 Points
(5). 5-fold Cross validation on Training Data by Fine-tuning their Hyperparameters in Terms of its F1-score: (a) Artificial Neural Networks, (b) Support Vector Machine, (c) Bagging, (d) AdaBoost, and (e) Random Forest (6). Your code will display the performance table on the console output (See Page 3 of the assignment)	35 Points
(7). Testing Dataset to Evaluate the Performance among All the Best-Trained Models Obtained from Above and Then Select the One That Beats All the Other Four Models in Terms of Their F1-Scores. (8). Your code will display the performance table on the console output and indicate which the ML model among five should be the chosen one that will be used to predict the machine failure in the future. (See Page 3 of the assignment)	15 Points
Total	100 Points