****

**Work Integrated Learning Programmes Division**

**M.Tech (AIML/ DSE)**

**Machine Learning**

**Assignment**

**Problem Statement 2**

Cervical cancer is one of the most preventable and treatable forms of cancer when detected early through proper screening and risk assessment. In many developing regions, however, lack of access to healthcare and diagnostic services contributes to late-stage detection, leading to higher mortality rates.

Here provided with a dataset containing anonymized information about 858 women. Primary objective is to build machine learning models that can **predict the likelihood of a positive biopsy diagnosis** for cervical cancer based on these risk factors.

.

**Dataset description**

Dataset: risk\_factors\_cervical\_cancer.csv

Target Variable: Biopsy (0 = No cervical cancer, 1 = Cancer detected)

The dataset contains the following types of features:

* Demographic: Age, number of sexual partners, pregnancies
* Behavioral: Smoking, contraceptive use
* Medical history: STDs, HPV infection
* Diagnosis results: Hinselmann, Schiller, Citology, and Biopsy (targets)

1. **Import Libraries/Dataset** 
   1. Download the dataset.
   2. Import the required libraries.
2. **Data Visualization and Exploration [1M]** 
   1. Print 2 rows for sanity check to identify all the features present in the dataset and if the target matches with them.
   2. Provide appropriate data visualizations to get an insight about the dataset.
   3. Do the correlational analysis on the dataset. Provide a visualization for the same. Will this correlational analysis have effect on feature selection that you will perform in the next step? Justify your answer. **Answer without justification will not be awarded marks.**
3. **Data Pre-processing and cleaning [2M]** 
   1. Do the appropriate pre-processing of the data like identifying NULL or Missing Values if any, handling of outliers if present in the dataset, skewed data etc. Mention the pre-processing steps performed in the markdown cell.
   2. Apply appropriate feature engineering techniques. Apply the feature transformation techniques like Standardization, Normalization, etc. You are free to apply the appropriate transformations depending upon the structure and the complexity of your dataset. Provide proper justification. **Techniques used without justification will not be awarded marks**. Explore a few techniques for identifying feature importance for your feature engineering task.
4. **Model Building [11M]**
   1. Split the dataset into training and test sets. **Answer without justification will not be awarded marks.** [1]
      1. Train = 80 % Test = 20%
      2. Also, try to split the dataset with different ratios of your choice.
   2. 1) Implement predictive models/classifiers using the following classification approaches: [8]
      1. Logistic Regression
      2. Decision tree
      3. K-Nearest Neighbour
      4. Ensemble Methods (any one of your choice)
5. Tune hyperparameters (e.g., number of trees, maximum depth) using cross-validation. For each of the above models. Justify your answer. [2]
6. **Performance Evaluation [6M]**
   1. Compare the performances of each model/classifier considering the given dataset using different evaluation measures such as Precision, Recall, F1-Score, AUC-ROC. Show the comparison chart in Python notebook [4]
   2. Identify the model, which you think is the best amongst all the models that have been trained. Also, explain why you think this is the best model. Answer this question in the notebook itself. [2]

**Instructions for Assignment Evaluation**

1. Organise your code in separate sections for each task. Add comments to make the code readable.
2. Deep Learning Models are strictly not allowed. You are encouraged to learn classical Machine learning techniques and experience their behaviour.
3. Notebooks without output shall not be considered for evaluation.

**For clarifications, contact**

**Dr. P. Rajesh (rajesh.p@wilp.bits-pilani.ac.in)**