**Birla Institute of Technology & Science, Pilani**

**Work Integrated Learning Programmes Division**

# Applied Machine Learning (Merged - SEZG568/SSZG568)(S1-23)

**Programming Assignment-II**

|  |
| --- |
| **Dataset: Fire\_alarm.csv** |
| Dataset consists of 62630 rows and 16 attributes . Build any 2 classification models to predict smoke. The last column, “Fire alarm” is the label.  The final deliverables of the Programming Assignment-II are:Python code in ipynb format with documenting all the findings of every stage  Save in a folder, zip and upload. |

# Tools and Techniques

Python libraries for data analysis. (**NumPy**,**SciPy**,**Matplotlib**,**Pandas**,**ScikitLearn**,**Statsmodels**,**Seaborn**,**Bokeh**,**Blaze**,**Scrapy**,**Req uests**,**BeautifulSoup**)

# Sample Exploratory Data Analysis Case Studies

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques> <http://ucanalytics.com/blogs/exploratory-data-analysis-retail-case-study-example-part-3/>

**Programming Assignment -II Guidelines**

These are the guidelines and questions that you are expected to answer. The student will have to analyze the data that he/she has been given and come up with meaningful insights for the given dataset. The steps that have to be taken are explained below.

1. **Descriptive Statistics**

Exploratory Data Analysis (EDA) is used to tackle specific tasks such as:

* + 1. Spotting mistakes and missing data;
    2. Mapping out the underlying structure of the data;
    3. Identifying the most important variables;
    4. Listing anomalies and outliers;

1. **Data Visualization**

Data should be visualized using the various types of charts and graphs that the student has learnt. Utilizing these visualizations, there should be insights from every visualization that is submitted and they should help frame the problem statement that is intended to be solved.

1. **Balance the dataset if required**
2. **Classification Model**

Build any two classification models

(Models to choose: Logistic regression, Naïve bayes, KNN, Random forest, Adaboost, Artificial Neural network)

**5)Evaluating the model**

The classification model should be evaluated using confusion matrix.

6) **Performance comparison**

Compare the performance of the two chosen models. Write your inferences.

# Presentation

* Presentation is key. Ensure that your notebook is capable of explaining your insights and visualizations by itself. Section your questions and emphasize your results. **Do not** hide your final result in a sea of code or debugging cells.

# Examples:

* + If your question is on data cleaning, highlight the rows which need to be cleaned and show the results of your data cleaning before and after it has been applied on those rows.
  + If your question asks you to prove a statement using visualizations, ensure that you actually have a concluding statement after your graphs. **Do not** leave the conclusion unstated after visualizing the data in your notebook.
* It is recommended to have short bullet points explaining what you have done before each task, especially for non-visualization tasks. This will help us understand your approach to the problem and can help with partial marks even if you are unable to solve the entire question.
* Prioritize interpretability over design. While it is encouraged to have visually appealing graphs, make sure that you do not lose interpretability of the data in the pursuit of aesthetic visualizations.