**Empirical Software Engineering (SE-404)**

**LAB A1-G2**

**Laboratory Manual**



**Department of Software Engineering**

**DELHI TECHNOLOGICAL UNIVERSITY(DTU)**

Shahbad Daulatpur, Bawana Road, Delhi-110042

**Submitted to: - Submitted by:-**

Ms. Priya Singh Name: ASHISH KUMAR

Roll number: 2K18/SE/041

**INDEX**

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No.** | **EXPERIMENT** | **DATE** | **REMARKS** |
| **10.** | Perform a comparison of the following data analysis tools. WEKA, KEEL, SPSS, MATLAB, R. | 04-01-2022 |  |
| **1.** | Consider any empirical study of your choice (Experiments, Survey Research, Systematic Review, Postmortem analysis and case study). Identify the following components for an empirical study:a. Identify parametric and nonparametric testsb. Identify Independent, dependent and confounding variablesc. Is it Within-company and cross-company analysis?d. What type of dataset is used? Proprietary and open-source software | 18-01-2022 |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Empirical Software Engineering LAB – A1 G2  
EXPERIMENT 1**

**-** ASHISH KUMAR  
 - 2K18/SE/041

# Experiment Objective:- Consider any empirical study of your choice (Experiments, Survey Research, Systematic Review, Postmortem analysis and case study). Identify the following components for an empirical study:

# a. Identify parametric and nonparametric tests.

# b. Identify Independent, dependent and confounding variables.

# c. Is it Within-company and cross-company analysis?

# d. What type of dataset is used? Proprietary and open-source software.

**Introduction:-**

* **Parametric and non-parametric tests:** Parametric tests are used for data samples having normal distribution (bell-shaped curve), whereas non-parametric tests are used when the distribution of data samples is highly skewed.
* **Independent variables:** Independent variables (or predictor variables) are input variables that are manipulated or controlled by the researcher to measure the response of the dependent variable.
* **Dependent variables:** The dependent variable (or response variable) is the output produced by analyzing the effect of the independent variables. The dependent variables are presumed to be influenced by the independent variables.
* **Confounding variables:** A confounding variable is a third variable that influences both the independent and dependent variables. Failing to account for confounding variables can cause you to wrongly estimate the relationship between your independent and dependent variables.
* **Within-company analysis:** In within-company analysis, the empirical study collects the data from the old versions/ releases of the same software, predicts models, and applies the predicted models to the future versions of the same project.
* **Cross-company analysis:** The process of validating the predicted model using data collected from different projects from which the model has been derived is known as cross-company analysis.
* **Proprietary software:** Proprietary software is licensed software owned by a company. For example, Microsoft.
* **Open source software:** Open source software is usually a freely available software, developed by many developers from different places in a collaborative manner. For example, Google Chrome, Android operating system, and Linux operating system.

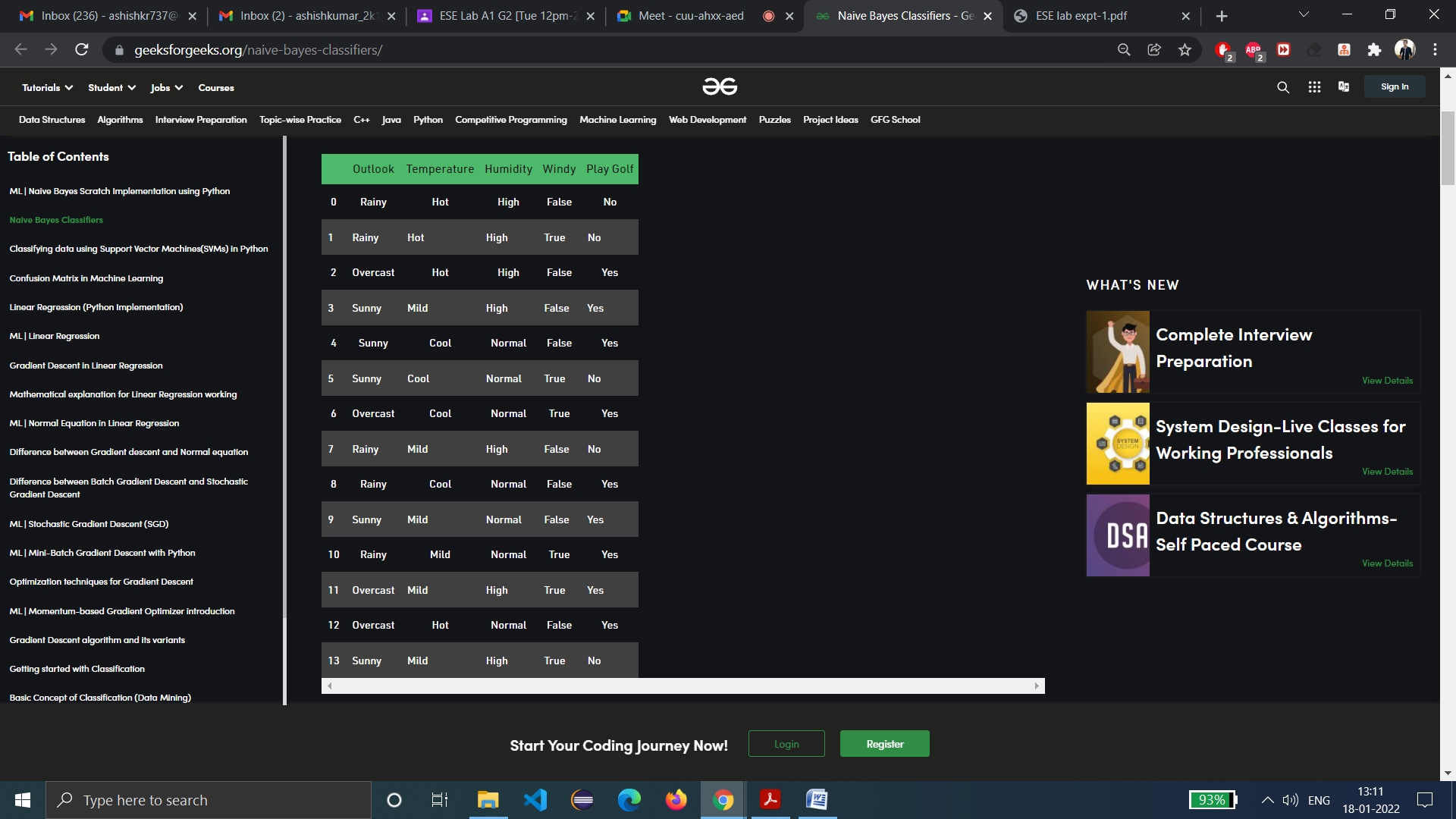
I have chosen [**Naïve Bayes Classifier**](https://www.geeksforgeeks.org/naive-bayes-classifiers/) as a case study for this experiment.

Naive Bayes classifiers are a collection of classification algorithms based on Bayes’ Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

To start with, let us consider a dataset.

Consider a fictional dataset that describes the weather conditions for playing a game of golf. Given the weather conditions, each tuple classifies the conditions as fit (“Yes”) or unfit (“No”) for playing golf.

Here is a tabular representation of chosen dataset:



[Source: Geeksforgeeks]

**Result:-**

In the given case study of Naive Bayes’ Classifier, following are the identified attributes:

1. **Parametric Test:** Since the attributes mentioned in the dataset have normal distribution. So, parametric test can be used is t-test since dataset is small and have normal distribution of data.
2. **Non-Parametric Test:** None
3. **Independent Variables: ‘**Outlook’, ‘Temperature’, ‘Humidity’ and ‘Windy’.
4. **Dependent Variables:** ‘Play Golf’
5. **Cofounding variables:** None as no variable is there that influences both the independent and dependent variables.
6. **Within Company and cross-company analysis:** Since the data is taken from single source, hence it is within company.
7. **Dataset:** The dataset is an open-source dataset, publicly available on the Geeksforgeeks website. This dataset describes the weather conditions whether for playing golf is fit (“Yes”) or unfit (“No”).

**Learning from experiment:**- We have successfully learned about parametric and non-parametric test. I was able to identify dependent and independent variables in chosen case study. There is no cofounding variable. It was open-source software and has conducted i.e. parametric and non-parametric tests as well. We have also learned about the Difference between within-company and cross-company analysis.