EXPERIMENT: 1

(2K17/SE/79 PARV GUPTA)

AIM: Identify the following components for an empirical study for the case study: Naive Bayes Classifier

- Identify parametric and non-parametric tests.
- Identify Independent, dependent and confounding variables.
- Is it Within-company and cross-company analysis?
- What type of dataset is used? Proprietary and open source software.

THEORY:

- **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 a 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.

OUTPUT:

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

Independent Variables: GND1, GND2, GND3, and so on that describe the generalizations among the voters. One top-level generalization, GND2, describes congressmen from agricultural states with high levels of school expenditures who voted for an education bill, parks in Alaska, and so forth. The 24th Texas Congressional District is stored under this generalization, along with two sub-generalizations. Someone familiar with U.S. polities would describe this voting pattern as 'liberal.' Similarly, the second top-level node in this example, GND3, would be considered 'conservative.'

Dependent Variables: describes whether a voter is "for" or "against" voting for a party.

Within Company and cross-company analysis: The analysis is a standalone experiment and has thus been performed by researchers within a single company. Although, the data used has been collected from multiple sources.

Dataset: The datasets used in this study are unique and proprietary which have been used to assess patterns in voting records.

Parametric Test: Friedmach and t-test.

Non-Parametric Test: No significant one.

CONCLUSION:

In the given case study I learned about parametric and non-parametric test. I was able to identify dependent and independent variables. The case study use confounding variables. It was made using both i.e. open and proprietary software and have both test conducted i.e. parametric and non-parametric. The dataset used in the case study was cross company analysis as the dataset is publicly available provided by the authors of the paper.