**BIG DATA – UE19CS322**

**PROJECT REPORT**

PES UNIVERSITY

Electronic City, Bengaluru

Team ID: BD2\_091\_106\_170\_363

Team members:

Bhoomika Bhavimath - PES2UG19CS091

Deepali Attavar - PES2UG19CS106

K Keerthana - PES2UG19CS170

Sanjana S - PES2UG19CS363

Github Repository URL: <https://github.com/sanjana87/Big-data-Spark>

I. PROJECT TITLE

**Machine Learning with Spark ML Lib – Spam Detection**

II. DESIGN DETAILS

Apache Spark is a data processing framework that can quickly perform processing tasks on very large data sets and can also distribute data processing tasks across multiple computers, either on its own or in tandem with other distributed computing tools.

Spark MLlib is used to perform machine learning in Apache Spark.The MLlib algorithm we are going to use will be Naïve Bayes Classifier.

Naïve Bayes algorithm is a supervised learning algorithm, which is based on the Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset.

Naive Bayes Classifier Formula

A, B = events

P(A|B) = probability of A given B is true

P(B|A) = probability of B given A is true

P(A), P(B) = the independent probabilities of A and B

III. SURFACE LEVEL IMPLEMENTATION DETAILS ABOUT EACH UNIT

Unit 1- Introduction – implementing pipelines; A pipeline allows us to maintain the data flow of all the relevant transformations that are required to reach the result.

Unit 3 – In memory computation - Spark introduction and architecture, working details, data-frames

Unit 4 – Streaming Analysis – streaming the model, real-time processing, spark streaming of the spam dataset

IV. REASON BEHIND DESIGN DECISIONS

Naïve Bayes is one of the fast and easy ML algorithms to predict a class of datasets.

It can be used for Binary as well as Multi-class Classifications. It can be used in real-time predictions because Naïve Bayes Classifier is an eager learner. It is used in Text classification such as Spam filtering and Sentiment analysis.

Naive Bayes spam filtering is a baseline technique for dealing with spam that can tailor itself to the email needs of individual users and give low false positive spam detection rates that are generally acceptable to users. It is one of the oldest ways of doing spam filtering.

With Bayes' Rule, finding the probability an email is spam, given it contains certain words. We do this by finding the probability that each word in the email is spam and then multiplying these probabilities together to get the overall email spam metric to be used in classification.

V. TAKEAWAY FROM THE PROJECT

From this project, we learnt how to stream the given dataset. Using spark MLlib, we learnt how we can do machine learning using the spark python library. Preprocessing the data, feature transformation, training the model and testing the model for accurate predictions.

Researched online for the best method possible to implement the project. Practically implementing what we have learnt in this course is been helpful.

Using GitHub for pushing the code and commit has given us learn about version management and branch management.