**inNutshell**

**Project Proposal**



inNutshell Enterprise aims at benefiting people who work in the news industry who wish to categorize their content and publish it. Categorization helps users navigate or browse through collections, Web sites or search results. By grouping too many discrete items into understandable categories, users can quickly eliminate what is irrelevant or not interesting, and just pay attention to what matters most.

Our enterprise offers an additional feature for our client to summarize huge articles to provide the short glimpse of it to accomplish readers to dive deep into the article. For Summarization there are 2 approaches Extractive Summarization and Abstractive Summarization. Out of which we have used Extractive Summarization. This will be beneficial for both the users as well as the editors working for a News Industry.

We have implemented 6 models in our project for news classification and calculated their accuracy and performance metrics. For news summarization we have use 1 algorithm that is text extraction.

**Project Goals**

**The Goal of our project is:**

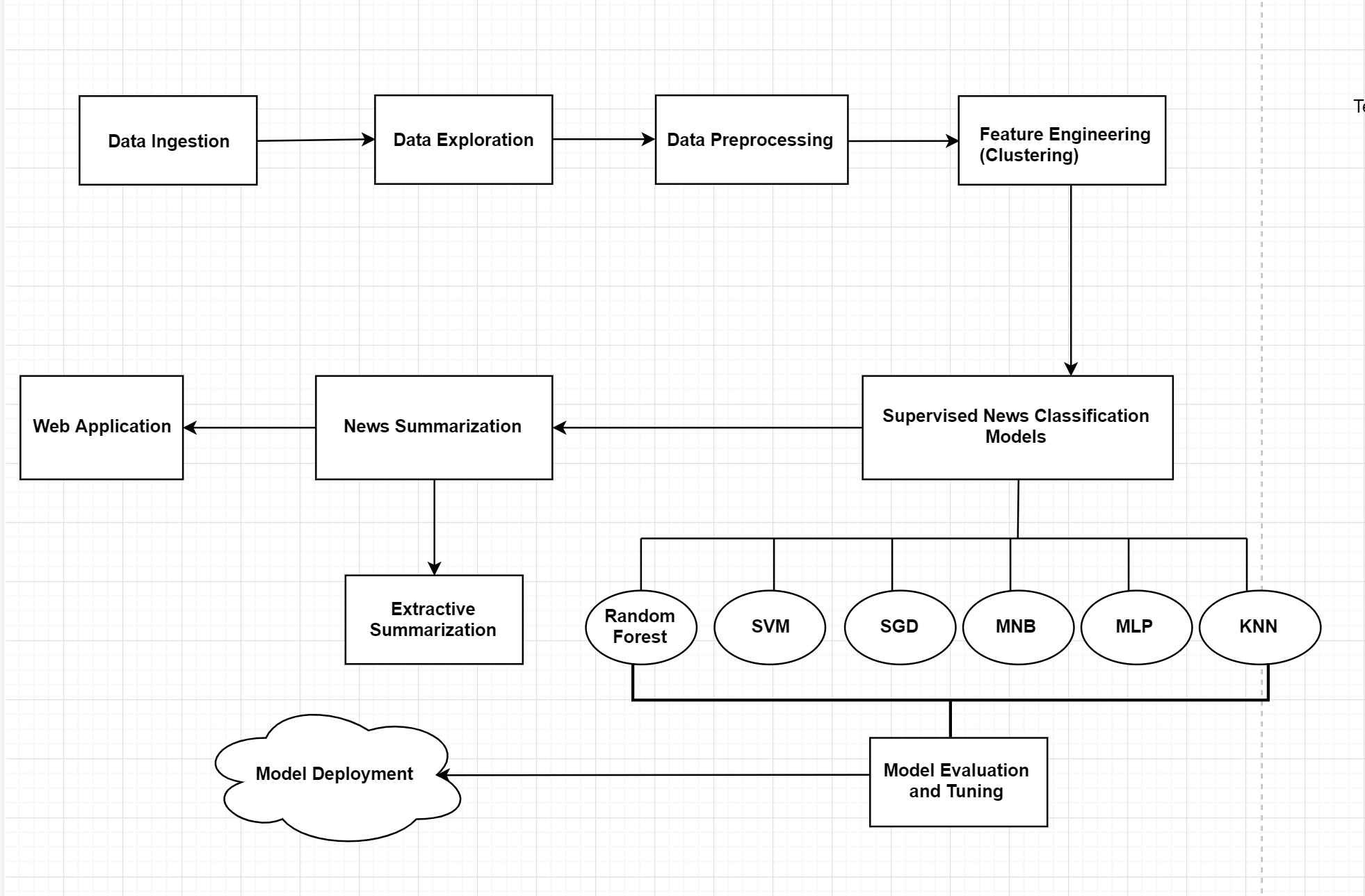
1. **News Classification:**

For news classification we have used multiclass classification concept. Multiclass classification is a popular problem in supervised machine learning. In our dataset the classes used are different categories of news. The goal is to categorize the user’s input news into category.

1. **News Summarization:**

For news summarization we have used NLP technique i.e. Text Extraction. The main aim of this module is

**Pipeline Design**



**Implementation details**

1. **Data Ingestion and Data Exploration**

Data ingestion is the process of obtaining and importing data for immediate use or storage in a database. To ingest something is to "take something in or absorb something." We had dataset of around 23k which consisted columns of title and category.

1. **Data Preprocessing**

In data Preprocessing we removed the rows containing missing values in title, categories and description. Remove the html tags from the description. Cleaned the data further by removing the Unicode error.

1. **Feature Engineering**

For feature engineering we created clusters for categories. Initially there were around 170 categories which cannot be considered as core categories. They were sub categories of some main categories. So we manually created 5 clusters and added all the related categories in those 5 categories.

1. **Supervised News Classification Model**

We have implemented total 6 models for Multiclass News Classification. The 6 models are:

1. Random Forest
2. Support Vector Machine (SVM)
3. Stochastic Gradient Descent (SGD)
4. Multinomial Naïve Bayes (MNB)
5. multilayer perceptron (MLP)
6. K-Nearest Neighbours (KNN)
7. **Model Evaluation and tuning**

We performed hyperparameter tuning on all the models to increase the accuracy of the model. Also calculated the performance metrics for all the models to check which model is considered as best.

1. **News Summarization**

For summarizing the news we have used Extractive Summarization Algorithm.

1. **Web Application**
2. **Model Deployment**

**Analysis of models**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **MODELS** | **ACCURACY** | **PRECISION** | **RECALL** | **F1-SCORE** | **SUPPORT** |
| **RANDOM**  **FOREST** | **81.95** | **0.82** | **0.82** | **0.82** | **3360** |
| **SVM** | **80.88** | **0.81** | **0.81** | **0.81** | **5040** |
| **SGD** | **82.34** | **0.82** | **0.82** | **0.82** | **5040** |
| **MNB** | **80.55** | **0.81** | **0.81** | **0.81** | **5040** |
| **MLP** | **79.00** | **0.79** | **0.79** | **0.79** | **5040** |
| **KNN** | **78.27** | **0.8** | **0.79** | **0.79** | **3360** |

**Details on how to run the model**