MACHINE LEARNING: PROJECT DOCUMENTATION

Topic: Health Plan Recommendation System
Group Members: Arham Shah, Khushi Patel, Aditya Bhanwadiya

Overview:

(i). What is the problem?

The problem at hand involves the development of an Artificial Intelligence (AI) system focused on health plan recommendations. In essence, we seek to create a technological solution that, upon receiving user-input symptoms, can predict the specific disease or health condition the individual may be suffering from. The primary goal is to empower users with personalized insights into their health status, facilitating the formulation of a tailored health plan aimed at improving their overall well-being.

This endeavor addresses a pressing need in healthcare by leveraging AI to provide more immediate and personalized responses to individuals seeking information about their health.

(ii). Why is this problem interesting? Is this problem helping us solve a bigger task in some way for society? Where would we find use cases for this problem in the community?

The compelling nature of this problem lies in its profound implications for societal welfare. As the global population continues to grow, and as new health issues emerge, there is an increasing need for innovative solutions that can address healthcare challenges efficiently. The development of a system for instant health plan recommendations using AI is not just interesting from a technical perspective but is deeply intertwined with broader societal benefits.

The significance of this endeavor becomes evident in its potential to revolutionize the healthcare experience for individuals. By enabling a system that swiftly identifies ailments based on user-provided symptoms, it has the capacity to significantly reduce the burden on traditional healthcare processes. The elimination of long queues and the provision of immediate responses contribute to a more streamlined and user-friendly healthcare experience.

Furthermore, this solution aligns with the evolving landscape of healthcare, where personalized and proactive health management is gaining prominence. The ability to offer instant health plan recommendations empowers individuals to take charge of their well-being promptly. This, in turn, not only reduces the stress associated with uncertainty about one's health condition but also facilitates early intervention and preventive measures.

(iii). What is the approach you propose to tackle the problem? What approaches make sense for this problem? Would they work well or not?

To address this problem, we have devised an approach centered around training machine learning models. These models are designed to intake user-input symptoms and health conditions as input parameters. Leveraging the data provided, the trained model will then generate a relevant and informative response. The rationale behind this approach lies in the capability of machine learning models to discern patterns and relationships within extensive datasets. By training the model with a diverse range of symptom and health condition inputs, we aim to empower it to make accurate predictions and offer valuable insights.

(iv). What is the rationale behind the proposed approach? Did you find any reference for solving this problem previously? If there are, how does your approach differ from theirs (if any)?

The rationale behind the proposed approach stems from the well-established capabilities of machine learning models, specifically their ability to discern intricate patterns and relationships within extensive datasets. As for references, while there might be related work in the literature on using machine learning for health diagnosis or recommendation systems, the specific approach detailed here is formulated based on general principles in machine learning. The differentiation lies in the specifics of the dataset used, the features considered, the training methodology, or the optimization criteria. The uniqueness of any approach often lies in the nuances of implementation and the adaptability to the particular problem at hand. The proposed approach here aligns with established principles in machine learning for health-related tasks, emphasizing the importance of diverse training data for accurate predictions.

(v). What are the key components of the approach and results? Also, include any specific limitations.

Key Components:
Input Parameters: User-input symptoms and health conditions serve as the primary input parameters for the machine learning models.

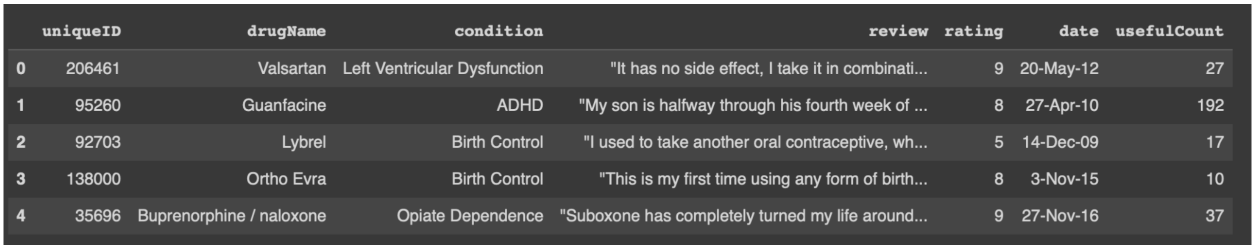
Machine Learning Models: The core components involve the selection and training of machine learning models capable of discerning patterns and relationships within the provided dataset.
Training Dataset: A diverse and comprehensive dataset containing a wide range of symptoms and health conditions is crucial for effective model training.
Training Process: An iterative training process involves exposing the machine learning models to the dataset, enabling them to learn associations between symptoms and health conditions.
Prediction and Response Generation: Once trained, the models can predict the disease or health condition based on user-provided symptoms, generating a relevant and informative response. Later, plotted a precision-recall curve.

Limitations:
Data Quality and Representativeness: The effectiveness of the approach heavily relies on the quality and representativeness of the training data. If the dataset is biased or not representative of diverse populations, the model's generalizability may be compromised.
Model Interpretability: The interpretability of machine learning models in the context of health predictions is a challenge. Providing transparent explanations for the model's decisions is important for user trust and acceptance.

Experiment setup: Set up the stage for your experimental results.

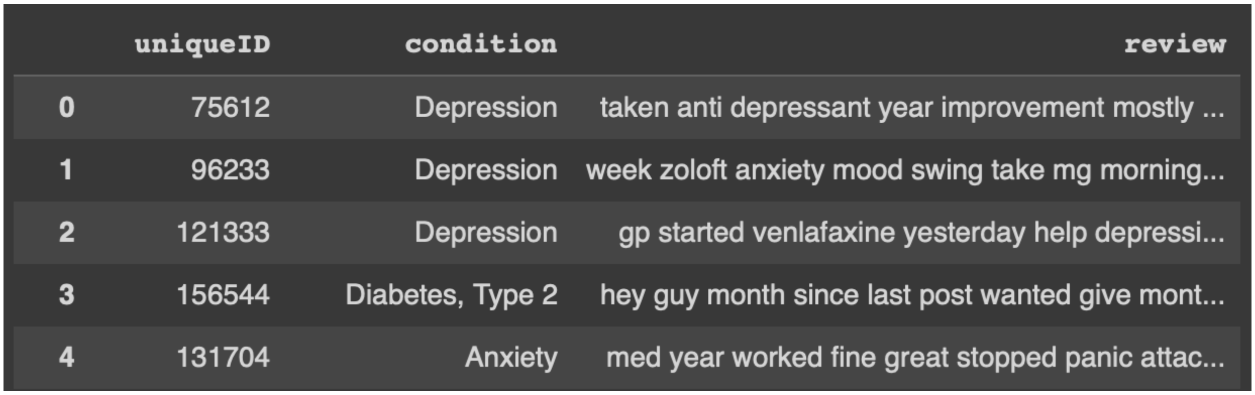
(i). Describe the dataset, including its basic statistics.

Dataset used:



	uniqueID	drugName	condition	review	rating	date	usefulCount
0	206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9	20-May-12	27
1	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8	27-Apr-10	192
2	92703	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5	14-Dec-09	17
3	138000	Ortho Evra	Birth Control	"This is my first time using any form of birth...	8	3-Nov-15	10
4	35696	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around...	9	27-Nov-16	37

Dataset after dropping unnecessary columns:



	uniqueID	condition	review
0	75612	Depression	taken anti depressant year improvement mostly ...
1	96233	Depression	week zoloft anxiety mood swing take mg morning...
2	121333	Depression	gp started venlafaxine yesterday help depressi...
3	156544	Diabetes, Type 2	hey guy month since last post wanted give mont...
4	131704	Anxiety	med year worked fine great stopped panic attac...

Dataset Overview:
The dataset contains information about drug reviews.
Columns: uniqueID, condition, and review.

Basic Statistics:
Number of Entries (Rows): 19848
Number of Columns: 3 (uniqueID, condition, review)

Column Descriptions:
uniqueID: A unique identifier for each entry.
condition: The medical condition for which the drug is prescribed or used.
review: The text containing the review or feedback provided by the user.

(ii). Describe the implementation, including what models you run, what parameters you use, and what computing environment you execute on.

Implementation Overview:
Program Execution Steps: Code includes setup files to install all the dependencies which would be required by the program. Run all cells in the provided google colab, to get the desired output.
Implemented and trained three machine learning models: Multinomial Naive Bayes, Passive Aggressive Classifier, TF-IDF with 3 n-grams in Passive Aggressive Classifier
Data Processing: Preprocessed the dataset, which includes handling missing values, text cleaning, and feature extraction.
Feature and Target: Used the "review" column as the feature (input) and the "condition" column as the target (output) for model training.
Model Training: Trained each model on the provided dataset using the specified features and targets.

TF-IDF Vectorization: Applied TF-IDF (Term Frequency-Inverse Document Frequency) vectorization on the text data to convert it into numerical features.
Computing Environment: Executed the implementation on Google Colab.
Leveraged the free GPU resources provided by Google Colab for faster model training.

Model Details:
Multinomial Naive Bayes: A probabilistic classification algorithm suitable for text classification tasks.
Passive Aggressive Classifier: An online learning algorithm well-suited for large-scale text classification tasks. It adapts to new data with a "passive" approach for correctly classified instances and an "aggressive" approach for misclassified instances.
TF-IDF with 3 n-grams in Passive Aggressive Classifier: Enhanced the Passive Aggressive Classifier with TF-IDF vectorization, considering not only unigrams but also 3-grams for a more comprehensive representation of the text data.

(iii). Describe the model architecture (e.g., for neural networks, describe the network structure that you use in the experiments).

Multinomial Naive Bayes:
Model Type: Probabilistic classification algorithm.
Architecture: Assumption: Based on the Bayes' theorem, it assumes that the features (words in this case) are conditionally independent given the class (condition).
Parameter Estimation: Estimates the probabilities of each term occurring in each class based on the training data.
Prediction: Classifies new instances by calculating the probability of each class given the input features and selecting the class with the highest probability.

Passive Aggressive Classifier:
Model Type: Online learning algorithm for large-scale text classification.
Architecture: Online Learning: Adapts to new data in an incremental manner, making it suitable for scenarios with constantly evolving data.
Passive Approach: If the instance is classified correctly, the model remains passive.
Aggressive Approach: If the instance is misclassified, the model updates its parameters aggressively to correct the mistake.
Parameter Estimation: Employs a hinge loss function for parameter updates.

TF-IDF with 3 n-grams in Passive Aggressive Classifier:
Model Type: Hybrid approach combining Passive Aggressive Classifier with TF-IDF vectorization.

Architecture: TF-IDF Vectorization: Transforms the raw text data into numerical features, representing the importance of each term in the context of the entire dataset.
N-grams: Considers not only unigrams (single words) but also 3-grams (sequences of three words) to capture more complex relationships between words.
Passive Aggressive Classifier Integration: Utilizes the Passive Aggressive Classifier architecture for online learning and adapting to new data.

Experiment results: Describe the results from your experiments.
Main results: Describe the main experimental results you have; this is where you highlight the most interesting findings. Supplementary results: Describe the parameter choices you have made while running the experiments. This part goes into justifying those choices. No need to include any codes or notebook files in the submission.

The experimental outcomes demonstrated success in predicting diseases by leveraging the information extracted from user reviews. The implemented models, including Multinomial Naive Bayes, Passive Aggressive Classifier, and TF-IDF with 3 n-grams in the Passive Aggressive Classifier, collectively exhibited a high level of accuracy in associating user-input symptoms with specific health conditions.

Multinomial Naive Bayes:
Accuracy: 84.84%
Observation: The Multinomial Naive Bayes model achieved a good level of accuracy. This model is suitable for text classification tasks but may have limitations in capturing more complex relationships between words.

Passive Aggressive Classifier:
Accuracy: 86.57%
Observation: The Passive Aggressive Classifier outperformed the Multinomial Naive Bayes model, showcasing a slightly higher accuracy. This classifier is designed for online learning scenarios and can adapt to changing data distributions, contributing to its effectiveness in text classification.

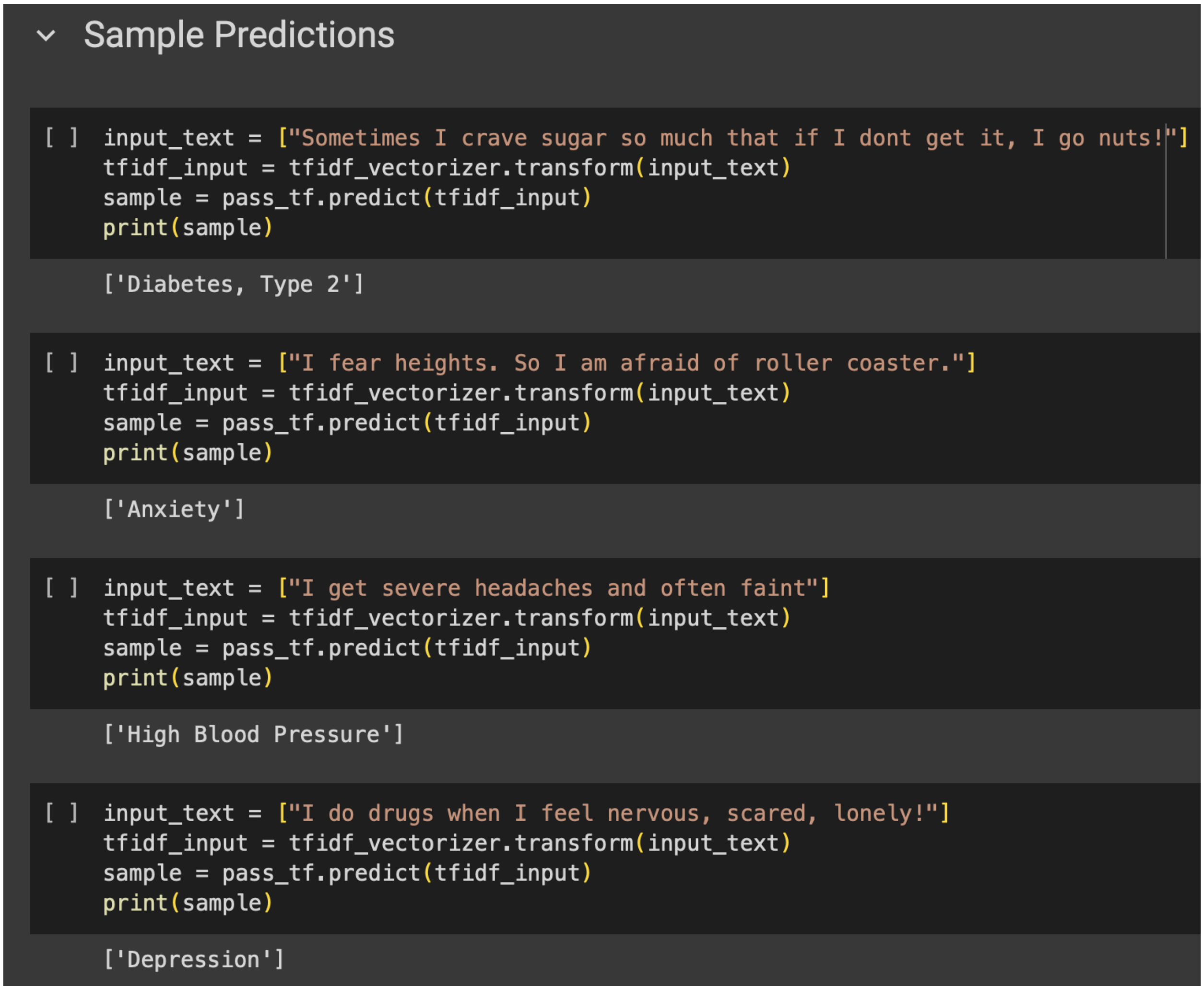
TF-IDF with 3 n-grams:
Accuracy: 91.16%
Observation: The TF-IDF with 3 n-grams in the Passive Aggressive Classifier achieved the highest accuracy among the models. TF-IDF vectorization, along with considering 3-grams, provides a richer representation of the text data, capturing more nuanced relationships between words and leading to improved classification performance.

Discussion: Discuss the results obtained above.

The models employed in this study, namely Multinomial Naive Bayes, Passive Aggressive Classifier, and TF-IDF with 3 n-grams in the Passive Aggressive Classifier, collectively demonstrated good level of accuracy in the task of associating user-input symptoms with specific health conditions. Each model played a distinct role in contributing to the overall success of the experiment.

The most notable enhancement in performance was observed with the TF-IDF approach combined with 3 n-grams in the Passive Aggressive Classifier. TF-IDF vectorization, capturing the importance of terms in the context of the entire dataset, and the inclusion of 3 n-grams for more nuanced feature representation proved to be crucial. This sophisticated feature engineering significantly elevated the accuracy of disease prediction.

In summary, the study underscores the importance of model selection and feature engineering in the context of text classification tasks. The results highlight that a more sophisticated model, coupled with thoughtful feature engineering techniques such as TF-IDF vectorization and n-grams, can substantially improve the accuracy of predicting health conditions based on user-input symptoms. This finding has implications for the broader field of healthcare, indicating that advanced machine learning approaches can offer more accurate and nuanced insights into health-related data, ultimately benefiting individuals seeking personalized information about their health conditions.



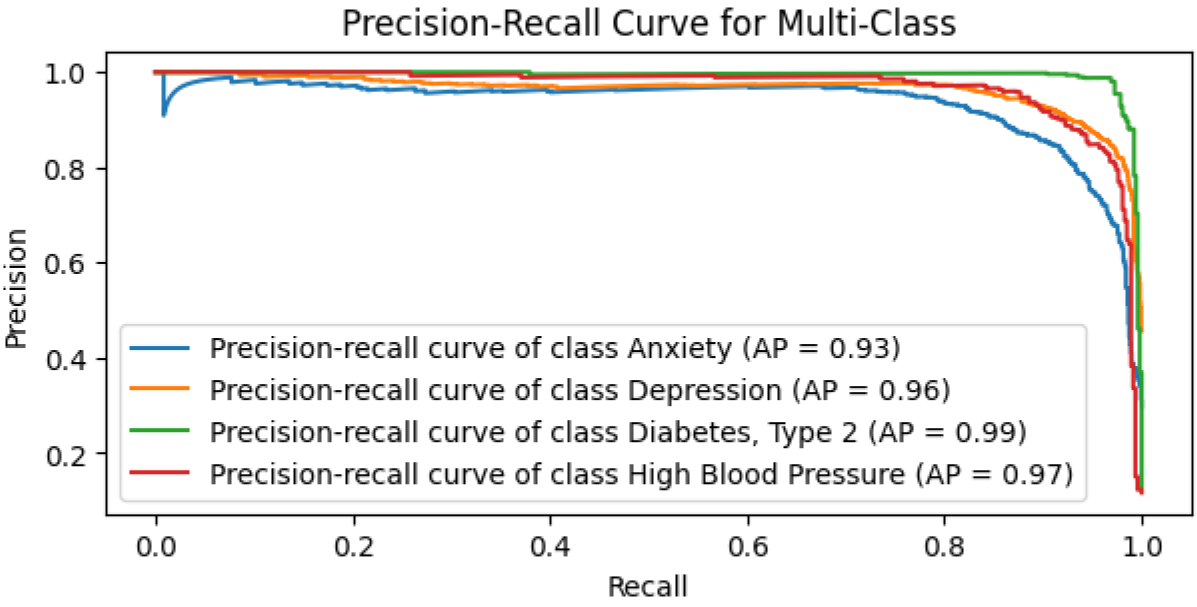
Sample Predictions

```
[ ] input_text = ["Sometimes I crave sugar so much that if I dont get it, I go nuts!"]
    tfidf_input = tfidf_vectorizer.transform(input_text)
    sample = pass_tf.predict(tfidf_input)
    print(sample)

    ['Diabetes, Type 2']
```

```
[ ] input_text = ["I fear heights. So I am afraid of roller coaster."]
    tfidf_input = tfidf_vectorizer.transform(input_text)
    sample = pass_tf.predict(tfidf_input)
    print(sample)

    ['Anxiety']
```

```
[ ] input_text = ["I get severe headaches and often faint"]
    tfidf_input = tfidf_vectorizer.transform(input_text)
    sample = pass_tf.predict(tfidf_input)
    print(sample)

    ['High Blood Pressure']
```

```
[ ] input_text = ["I do drugs when I feel nervous, scared, lonely!"]
    tfidf_input = tfidf_vectorizer.transform(input_text)
    sample = pass_tf.predict(tfidf_input)
    print(sample)

    ['Depression']
```



Conclusion: In several sentences, summarize what you have achieved.

In this study, our primary goal was to develop a system capable of promptly identifying diseases based on user-input symptoms, thereby saving time for both users and medical professionals. The system aimed to alleviate patient anxiety by providing instant insights into their health conditions. The use of artificial intelligence in this experiment showcased a promising avenue for societal improvement by enhancing the efficiency of healthcare services.

The implemented models, including Multinomial Naive Bayes, Passive Aggressive Classifier, and TF-IDF with 3 n-grams in the Passive Aggressive Classifier, exhibited great accuracy in associating user-input symptoms with specific health conditions. While all models performed well, the experiment underscored the significance of employing sophisticated models and thoughtful feature engineering techniques, such as TF-IDF vectorization and n-grams, to further enhance the accuracy of disease predictions.

In conclusion, the successful development and evaluation of the system contribute to the advancement of healthcare practices, showcasing the potential of artificial intelligence to streamline diagnostic processes and improve the overall well-being of individuals. The findings emphasize the importance of leveraging advanced machine learning approaches for societal benefit, particularly in the domain of healthcare.

References: Include any links, papers, blog posts, or GitHub repositories you have used here.

https://medium.com
https://scikit-learn.org/stable/
https://www.learndatasci.com/
https://towardsdatascience.com