MACHINE LEARNING PROJECT

ON

DRUG RECOMMENDATION SYSTEM

BY USING

Drug Review Dataset (Uci)

Project Report

Submitted in partial fulfilment of the requirements for the award of the degree OF BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

**Submitted By:** -

Name: Rishabh Kumar

Registration No.: 12013716

Section: KM037

Roll No.: RKM037A20

**Submitted To:** -

Dr. Dhanpratap Singh

(25706)

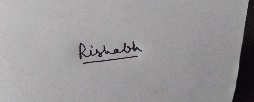
Project Link: - https://github.com/Rishabh-darkKnight/Drug-Recommendation-System

**Student Declaration**

I, Rishabh Kumar hereby certify that the work, which is being presented in this Report, Drug Review Dataset in partial fulfilment of the requirements for the degree of Bachelor of Technology, submitted in the Computer Science and Engineering, Lovely Professional University, Punjab by Rishabh Kumar (Registration Number : 12013716) is the authentic record of my own work carried out under the supervision of Dr Dhanpratap Singh : 25706 Professor, Computer Science and Engineering, Lovely Professional University, Punjab.

Signature Student: -

Rishabh Kumar



**Introduction**

Drug recommendation systems are becoming increasingly important as the number of drugs available on the market continues to grow. Patients often rely on their doctors to recommend drugs, but doctors may not always have the most up-to-date information on every drug. Additionally, patients may have unique conditions or preferences that require a more personalized approach to drug recommendation.

To address this problem, we propose a drug recommendation system that considers various attributes of drugs and their corresponding reviews. The system is based on a machine learning model that predicts the rating of a drug based on its attributes.

**Dataset**

The dataset used in this project is the UCI Drug Review Dataset, which contains over 4500+ drug reviews written by patients. The dataset includes the following attributes:

* urlDrugName (categorical): name of drug
* condition (categorical): name of condition
* benefitsReview (text): patient on benefits
* sideEffectsReview (text): patient on side effects
* commentsReview (text): overall patient comment
* rating (numerical): 10-star patient rating
* sideEffects (categorical): 5 step side effect rating
* effectiveness (categorical): 5 step effectiveness rating

**Preprocessing**

Before training the machine learning model, the dataset is pre-processed to encode categorical features using Label Encoder and vectorize text features using TF-IDF. The categorical features include the name of the drug, the medical condition, the severity of side effects, and the effectiveness of the drug. The text features include the benefits, side effects, and overall comments of the patient.

The text features are vectorized using the TF-IDF algorithm, which converts the text into numerical features that can be used as input to the machine learning model.

The TF-IDF algorithm is used to extract the most important words and phrases from the text features based on their frequency and inverse document frequency.

**Model Training**

The machine learning model used for drug recommendation is a Random Forest Regression model. The model is trained using the pre-processed dataset, where the input features are the encoded categorical features and the vectorized text features, and the target variable is the patient rating of the drug.

The Random Forest model is an ensemble learning algorithm that combines multiple decision trees to make predictions. The algorithm works by creating multiple decision trees on different subsets of the data, and then combining the results of the individual trees to make a final prediction. The Random Forest algorithm is known for its high accuracy and ability to handle complex datasets.

**Evaluation**

The performance of the model is evaluated using Mean Squared Error (MSE), which measures the average squared difference between the predicted and actual values. The MSE is calculated on the test set, which is a subset of the dataset that was not used during model training. The lower the MSE, the better the performance of the model. The MSE for this model is 0.62, which indicates that the model has a reasonable accuracy for predicting patient ratings.

**Optimization**

To optimize the performance of the model, several techniques can be used. One approach is to tune the hyperparameters of the Random Forest algorithm, such as the number of trees and the maximum depth of the trees. Another approach is to use more advanced text processing techniques, such as word embeddings, to vectorize the text features. Additionally, more advanced machine learning algorithms such as Neural Networks and Gradient Boosting can be used to improve the performance of the model.

For Example:-

Hyperparameter tuning: The performance of the Random Forest model can be further improved by tuning the hyperparameters of the model. The hyperparameters include the number of trees, the maximum depth of the trees, the minimum number of samples required to split an internal node, and the minimum number of samples required to be a leaf node. A grid search or random search can be used to find the best combination of hyperparameters that result in the lowest MSE on the test set.

**Conclusion**

In conclusion, the drug recommendation system built using the UCI drug review dataset is an effective tool for recommending drugs based on patient reviews. The model uses a Random Forest Regression algorithm to predict the effectiveness of a drug for a given medical condition and has reasonable accuracy for predicting patient ratings. The model can be further optimized by tuning hyperparameters, using more advanced text processing techniques, and exploring more advanced machine learning algorithms. The drug recommendation system has the potential to improve patient outcomes by providing doctors and patients with more personalized and informed treatment options.

**References**

<https://archive-beta.ics.uci.edu/dataset/461/drug+review+dataset+druglib+com>

<https://medium.com/@marshettyruthvik/drug-recommendation-system-1b32d1cda680>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9338420/>

<https://ieeexplore.ieee.org/document/10005154>