Healthcare Analytics

PROJECT REPORT

Group 10

Chinmay Chamoli

Jenson Puthenpeedikayin Jacob

Yashu Bhati

Nahush Krishna

Contents

[Abstract 3](#_Toc141391711)

[Related Work 3](#_Toc141391712)

[Dataset Selected 4](#_Toc141391713)

[Methods 4](#_Toc141391714)

[Results 4](#_Toc141391715)

[Discussion 7](#_Toc141391716)

[Conclusion 10](#_Toc141391717)

[Contributions 10](#_Toc141391718)

[References 11](#_Toc141391719)

[Appendices 11](#_Toc141391720)

## Abstract

This project deals with analyzing different medication brands and manufacturers in Bangladesh. Since we have multiple manufacturers with a slew of products for different health related issues, the plan to is understand how many different brands are available and the kind of dosage form that each manufacturing company prefers. We intend to show the top brands and medication types that are being used in the country. We will be using various visualization methods and an analysis model to try and find out what kind of dosage forms the company might create based on the medication strength, manufacturer size and brank being used.

The final aim of the project is to provide patients with a list of the best manufacturers who are selling their brand in the market, describe the reasons why they are on top, and help the patients find the best medication with the preferred dosage form.

Introduction

Any major Illness or health issue would need an individual to take some or the other form of medication. Throughout history medicine has been administered in many forms starting from home remedies, strange concoctions provided by shamans, communal healers, and discovery of medication over the past few centuries by scientific innovation. There were many different categories of medication like herbal, allopathy, naturopathy, ayurveda etc. which have been used to treat a range of aliments and diseases. Innovation in medicine is a continuous process, and by constantly improving the quality of the medication and finding the ways to remove side effects will ensure that the company maximizes its profits while providing safer medication to patients in the long term.

The team will try to analyze the most recommended medications, the companies selling such medications and analyze the reason and motivations for such a trend. The team will provide a list of the top medications and provide analysis on why they are on the top as compared to other companies and provide insights in which areas the companies lagging could concentrate on. The team will try to analyze the benefits and risks associated with using various medications and recommend the best dosage form based on company and brand.

## Related Work

Data science can play an important role in the analysis of medication usage by companies and there have been several research papers out which have highlighted the need for medicine analytics.

We got our inspiration from reports like ‘Big data analytics in medicine and healthcare’ [1] which talks about how medicine analysis is leveraged using the 6 Vs of big data analysis. It also talks about the various advantages of using big data analysis and how it could be leveraged in the future.

Another interesting read is the report about ‘Big Data analysis in preventative medicine’ [2], which talks about how we can use prescriptive analysis, descriptive analysis and forecasting to improve the quality of patient care.

Both reports mentioned above helped us understand the approach that we needed to take and in addition, we went through a few datasets to understand the data and analysis approach.

## Dataset Selected

We have selected Medicine analytics | EDA [3] from Kaggle. The link to the dataset is as follows:

<https://www.kaggle.com/code/muhammedtausif/medicine-analytics-eda>

We chose to analyze this dataset since it had multiple files which encompassed a large volume of data and the team felt that they could perform effective analysis since we had many features to choose from. We felt that the dataset would help us perform effective predictive analysis. We will try to use the term forecasting instead of predictive analysis since we, like our professor believe that we can never truly predict outcomes but can only try and forecast them.

## Methods

We used multiple models to perform our analysis. We tried to analyze the data using Logistic regression, XG Boost and RF classifier. We tried this to try and find which model would provide the highest prediction and the context in which the data could be analyzed.

## Results

Logistic regression:

We obtained the following values (fig 1) after performing the analysis. From the confusion matrix we can see that the model has considerably low accuracy using this method. The model accuracy is 76%, which can be considered a low accuracy score.

A reason for this accuracy is that the model is not sophisticated enough to analyze the dataset that we are using, and we will need ensemble models and boosting classifiers to analyze the data.

fig 1


Accuracy: 0.76102

Precision: 0.64826

Recall: 0.76102

F1 Score: 0.69717

Fig 1.

XG Boost:

We found that among the 3 analysis models we used, XG Boost gave us the highest accuracy of 97% which led us to select this model for testing our data. Fig 2 shows the accuracies and the output metrics obtained for the 5 groups after running the model.

fig 2


Accuracy: 0.9711948790896159

Classification Report:

precision recall f1-score support

0 0.95 0.86 0.90 398

1 0.99 1.00 0.99 164

2 0.98 1.00 0.99 202

3 1.00 1.00 1.00 201

4 0.97 0.99 0.98 1847

accuracy 0.97 2812

macro avg 0.98 0.97 0.97 2812

weighted avg 0.97 0.97 0.97 2812

fig 2.

RF classifier:

We wanted to be sure about the model we would want to use to measure accuracy and so we ran our dataset through the random forest classifier model as well. And just like we theorized the RF classifier model gave lower accuracy than XG Boost which is evident from fig 3. The overall accuracy was 95%, which was a high score. But since the accuracy of XG Boost was higher we opted for that model for analysis.

A diagram of a graph

Description automatically generated

Accuracy: 0.9598

Precision: 0.9709

Recall: 0.95297

F1 Score: 0.9602

## Discussion

After performing analysis, we tried to visualize the dataset after performing EDA and data cleaning we tried to analyse the data by visualizing it on Tableau.

Fig 4 shows a list of the top 5 manufacturers which the total number of brands they own. This gives us a fair idea of how big each manufacturing company is, and which brands each company owns.

A graph with colorful bars

Description automatically generated

Fig 4.

Fig 5 shows the dosage forms of each of the top 5 brands in the market. We could find that most of the brands had tablet as their dosage from. (Except Bexitrol F, since this was a medication for respiratory issues and offered inhalers)

A graph with multiple colored bars

Description automatically generated

Fig 5.

Fig 6 shows the list of medications the companies are offering based on the category. We observed that manufacturing companies had multiple brands of medication which improved overall health and immunity (Herbal and Nutraceuticals). This helped us realise that the companies believed that they could maximize profits since these were everyday over the counter meds that any individual could use irrespective of whether they were healthy or sick.

A group of colorful circles

Description automatically generated

Fig 6.

Fig 7 provides a list of top 10 manufacturers and the dosage forms that they offer. Through this we could analyse which dosage form the companies gave highest priority to and how many brands used the dosage forms per company.

A graph with different colored squares

Description automatically generated fig 7.

Fig 8 shows the count od the dosage form of all brands. From the image we observe that most of the brands use tables as the dosage form followed capsule and powder. Surprisingly we had assumed that syrup be higher than powder, but the analysis showed us otherwise.

A screenshot of a computer

Description automatically generated

Fig 8.

## Conclusion

We initially set out to find the top manufacturing companies and brands and were planning to help patients get the best options of medication for their ailments, but we felt that we could perform further analysis and predict which brand would likely use one of the different dosage forms. We then performed predictive analysis on our dataset using 3 different models. We compared the accuracies of the 3 models – Logistic regression, XG Boost and RF classifier. We found that analyzing the dataset using XG Boost model provided the highest accuracy and we decided to implement this model for our prediction. We were successfully able to predict the dosage form of the drug based on the manufacturing company and the brand.

We have only done a predictive analysis. In the future we believe that by performing medication analysis we can help patients gain access to medications that can effectively help them based on their body type, gender, and age. Also, we could help analyze the long term and short-term side effects of each medication and customize it for the patient to help them minimize the consequences of these side effects. We also have not been able to perform cost analysis of each brand since we could not find the price for each medication in any of the columns. Doing an analysis of price will also help the patients get affordable medication for their medical issues.

## Contributions

Chinmay Chamoli: worked on data preprocessing and basic EDA.

Jenson Puthenpeedikayin Jacob and Yashu Bhati: worked on data modeling.

Nahush Krishna: worked on documentation and visualization and overall management of team.

## References

Medicare Part D Spending by Drug

<https://catalog.data.gov/dataset/medicare-part-d-spending-by-drug-401d2>

*How drug marketing may influence prescriptions*. (2017, May 18). National Institutes of Health (NIH). <https://www.nih.gov/news-events/nih-research-matters/how-drug-marketing-may-influence-prescriptions>

*10 Essential Facts About Medicare and Prescription Drug Spending*. (2019, February). The Henry J. Kaiser Family Foundation. <https://www.kff.org/infographic/10-essential-facts-about-medicare-and-prescription-drug-spending/>

## Appendices

[1] Ristevski, B., & Chen, M. (2018). Big data analytics in medicine and healthcare. *Journal of Integrative Bioinformatics*, *15*(3). <https://doi.org/10.1515/jib-2017-0030>

[2] Razzak, M. I., Imran, M., & Xu, G. (2019). Big data analytics for preventive medicine. *Neural Computing and Applications*, *32*(9), 4417–4451. <https://doi.org/10.1007/s00521-019-04095-y>

[3] Muhammedtausif. (2023). Medicine Analytics | EDA. *Kaggle*. <https://www.kaggle.com/code/muhammedtausif/medicine-analytics-eda>