Analyzing and Predicting Hypertensions using National Healthcare Insurance Claims Dataset

Sukyung Seo[[1]](#footnote-1)\*

New York University, ss13969@nyu.edu

Identification of patients’ potential diseases is crucial in the healthcare industry and patients’ lives. Early prevention of diseases can lead to a patient's happiness and reduce healthcare costs. Since hypertension is one of the high occurrence diseases and can lead to life-threatening conditions, this study focused on prediction of hypertension using Korean National health claim dataset. First, I downloaded two kinds of data from the national healthcare database and pre-processed data such as merging based on the patient's ID and removing rows with missing values. Second, I analyzed the data and visualized the characteristics of data. Third, I applied eight machine learning techniques and obtained performance metrics. Naive k-Nearest Neighbor (kNN) model showed around 0.5 in accuracy. However, after adopting different algorithms and hyperparameter tunings, the results generally showed above 0.6 and achieved above 0.7. Finally, I reported the feature importance in the XGBoost model which showed the best performance scores.

**Additional Keywords and Phrases:** Machine Learning Algorithms, National Healthcare Insurance Claims Datasets, Performance Metrics

1. Introduction

According to Dieleman et al., health care expenses are increasing, and they account for almost 20 % of the US economy [1]. Since the onset of some diseases can be a financial burden to patients, predicting potential diseases can be helpful in reducing health care costs. Among them, hypertension is one of the most critical diseases in the world because it is often comorbid with other diseases such as diabetes and obesity and can further lead to life-threatening conditions including cardiovascular diseases [2]. 33% of US adults already have hypertension and pay nearly $2000 more per year for medical care than patients who do not have hypertension [3]. Thus, the identification of hypertension in its early stages is a key to preventing further, more deadly complications [4].

Hypertension prediction not only benefits patients, but also the government and insurance companies. Because identifying the risk of the disease helps to avoid or minimize its occurrence, it can affect the patient’s quality of life and potentially reduce future health care costs. In addition, thanks to the rapid advancement in the healthcare industry, researchers can utilize different kinds of data and tools for early diagnoses [5]. By using diverse data and recent tools proactively, doctors can contribute to the improvement of a patient's life and the health care system.

Many studies that predicted the development of diseases using health data and machine learning have already been conducted. Vaughn et al., predicted the risk of Inflammatory Bowel Disease (IBD) using insurance claims [6]. Since IBD is a serious disease that requires hospitalizations and expensive medication, this research helped to improve patient quality of life and reduce future expenses. Using 110 variables from insurance data, they predicted IBD-related hospitalization using topic modeling and showed that this prediction could prevent future hospitalization and insurance costs. Unlike the previous study, Ma et al., used Electronic Health Records (EHR) to predict the risk of potential diseases [7]. In this study, they analyzed prior medical knowledge including disease and risk factor relationships. Not only did they integrate the knowledge already available to them, but they also improved the performance of the model itself.

Recently, studies that have utilized both health insurance data and machine learning techniques are emerging in fields that have not yet been explored with these complex areas that have not been the target of data studies in the past. Krishnamurthy et al., predicted the occurrence of Chronic Kidney Disease (CKD) using Taiwan’s National Health Insurance Research Database [8]. They developed a machine learning model using medication and comorbidity information from the database to find correlations and aggregate health data. According to their analysis, convolutional neural networks showed the best prediction results and predictors such as diabetes mellitus, age, gout, and some specific medication played an important role in predicting the onset of CKD. These authors claim this model could be beneficial for public health policymaking in the future by identifying and forecasting disease trends.

In this study, I proposed a machine learning approach for the prediction and identification of hypertension by leveraging Korea’s national health insurance claim data. The hypothesis of this study holds that patients that have the same diseases share similar traits such as blood test results and vitals (e.g., weight, height). First, I downloaded two kinds of data from Korea's national health insurance homepage: basic examination data and disease information data. Second, I merged these two data sets of data using patients’ identifier codes and removed all missing values. Based on pre-processed data, I then analyzed the patient’s recorded information such as diseases distribution based on sex and age. Finally, I predicted the potential for the onset of hypertension based on the patients’ personal information using eight machine learning algorithms including Decision Trees, Random Forest, Naive Bayes, k-Nearest Neighbor (kNN), Stochastic Gradient Descent (SGD), Logistic Regression, Support Vector Machine (SVM), and eXtreme Gradient Boosting (XGBoost).

1. Methods
   1. Dataset Collection

I launched my data search by utilizing the Korean National Healthcare Insurance Claims Database in February 2022. On this site, researchers uploaded a variety of insurance data such as cost of insurance, basic examination results, eyesight results, and patients’ disease data. In this analysis, I used the patients’ basic examination data (e.g., aspartate aminotransferase (AST)/cholesterol blood tests, blood pressure, height, weight measures) and patients’ disease data (e.g., disease codes assigned based on what they are diagnosed with).

* 1. Dataset Preprocessing

The webpage provides the patients’ basic examination and disease data as two separate files. Since these two data’s column names were recorded in Korean, it caused an encoding error when I read the files in Jupyter Notebook. I changed the column names to English to avoid errors. Then, I merged the two data files based on the patient’s unique code. While organizing variables, I found that some fields of data contained a lot of missing values. In order to prevent any further errors and ensure the reliability of my analysis, I conducted the necessary statistical tests to accurately remove rows that had any missing values.

* 1. Dataset Visualization

In this process, I visualized diverse statistics such as disease types based on age, and the average number of prescription days by age. By illustrating disease types based on age, patients and researchers alike can notice which diseases are more threatening according to age demographics more easily. Similarly, by constructing a visualization that denotes the number of patients who have certain diseases, viewers can easily grasp just how prevalent some illnesses are. In the context of this study, I utilized these visualizations to narrow down the project’s scope. Since there were so many diseases present in this dataset, it was difficult to formulate a model and predict a variety of afflictions since there were several discrepancies and gaps between available patient disease data. As such, I decided to focus on predicting hypertension for the following reasons: it was in the top 4 diseases in both female and male samples; was highly impacted by blood test results; and presented many relationships with patients’ vitals and physical information such as weights and heights.

* 1. Adopting Machine Learning Techniques

After removing rows with missing values, the remaining data contained 1,735,182 rows. In this dataset, there were 48, 683 patients who experienced high blood pressure. The rest of the 1,686,499 rows indicated patients who experienced other diseases such as mental disease and urinary diseases. In order to yield the most accurate results, I focused only on patients experiencing hypertension, I excluded their other diseases and left the record with hypertension. I defined these as positive samples. In addition, if one patient had several records with different disease codes, I kept only one record. I defined these data as negative samples. Since the data were highly imbalanced, I randomly sampled the same number of samples from the negative dataset. Using 27 columns, I applied eight machine learning algorithms such as k-Nearest Neighbor (kNN), Decision Tree, Naive Bayes, Random Forest, Stochastic Gradient Descent (SGD) classifier, Logistic Regression, Support Vector Machine (SVM), and eXtreme Gradient Boosting (XGBoost) to this data [9-16]. Additionally, I also applied parameter tunings on various machine learning algorithms and got performance metrics such as F1, accuracy, precision, and recall.

1. Results

In this section, I reported the data visualization and performance metrics from different machine learning algorithms.

* 1. Data Visualization

The following figures display descriptive statistics regarding the national healthcare dataset. In Figure 1, the top diseases, as segregated by gender, are displayed.

Chart, bar chart

Description automatically generated

Figure 1: Top diseases by sex

Figure 2 displays the relative number of prescription days by age. Age brackets are in increments of 5 years, and the size of the circle shows the relative number of prescription days.

Chart, bubble chart

Description automatically generated

Figure 2: Average number of prescription days by age

Figure 3 and 4 are population histograms of weight and height, respectively.

Chart, bar chart

Description automatically generated

Figure 3: Number of people by weight

Chart, bar chart, histogram

Description automatically generated

Figure 4: Distribution of the number of people by height

* 1. Performance metrics

In this part, I reported the four performance metrics of machine learning: accuracy, precision, recall and F1. Performance metrics without adopting any hyperparameter tuning were recorded in Table 1. In Table 2, performance metrics with hyperparameter tuning were reported.

Table 1: Performance metrics without any hyperparameter tuning

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1** |
| --- | --- | --- | --- | --- |
| SGD | 0.5559 | 0.7911 | 0.1520 | 0.2550 |
| kNN | 0.6000 | 0.5931 | 0.6371 | 0.6143 |
| Decision Tree | 0.6652 | 0.6690 | 0.6537 | 0.6613 |
| Naive Bayes | 0.7515 | 0.7109 | 0.8475 | 0.7733 |
| Logistic Regression | 0.7518 | 0.7324 | 0.7936 | 0.7618 |
| SVM | 0.7547 | 0.7049 | 0.8762 | 0.7813 |
| Random Forest | 0.7565 | 0.7098 | 0.8678 | 0.7809 |
| XGBoost | 0.7621 | 0.7111 | 0.8828 | 0.7877 |

Table 2: Performance metrics after adopting hyperparameter tuning

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1** |
| --- | --- | --- | --- | --- |
| SGD | 0.7526 | 0.6863 | 0.9305 | 0.7899 |
| kNN | 0.6708 | 0.6465 | 0.7537 | 0.6960 |
| Decision Tree | 0.7619 | 0.7121 | 0.8791 | 0.7869 |
| Naive Bayes | 0.7542 | 0.7061 | 0.8708 | 0.7799 |
| SVM | 0.6316 | 0.6535 | 0.5603 | 0.6033 |
| Random Forest | 0.7590 | 0.7039 | 0.8941 | 0.7877 |
| XGBoost | 0.7620 | 0.7109 | 0.8830 | 0.7877 |

1. Discussion
   1. Analysis of the visualization results

As shown in Figure 1, the top diseases differed by sex. Females experienced dramatically more urinary problems than males. Although acute bronchitis was a common disease for both genders, I didn’t choose it for this analysis because this disease is generally caused by infectious agents in the environment and is difficult to predict using a patient’s medical history. Hypertension was the top disease for both genders and was chosen as the basis for this analysis.

In Figure 2, as the age increases, the size of the circle increases. It clearly shows that older people are susceptible to diseases, and they take a lot of medications. As indicated in Figures 3 and 4, Koreans are usually 60 kg and 160 cm, consistent with previous literature [17].

* 1. Analysis of the performance metrics

As shown in Table 1, kNN’s performance metrics in three categories were below 0.6. After adopting many different algorithms, they generally achieved almost above 0.6. This showed that kNN was not enough to capture the characteristics of this data, but other algorithms performed better. As depicted in Table 2, hyperparameter tuning in almost all cases improved performance. After hyperparameter tuning, most metrics achieved above 0.7, and some scores achieved over 0.8. Scores improved in all areas except for some models including XGBoost and SVM. This indicated that hyperparameter tuning was helpful to improve the model.

However, the best model was XGBoost without hyperparameter tuning. It achieved well-rounded results in accuracy, recall, and F1 score. It showed that the XGBoost model was good at correctly classifying the data and identifying true positives. Since it also achieved a good F1 score, performance of recall and precision in this model fitted the data well. Using this model, I reported feature importance in Figure 4.

Shape

Description automatically generated

Figure 4. Feature importance in XGBoost model

Based on Figure 4, the most influential feature was age, and it was twice as important as the next influential feature, sex. It has already been shown that age is an important risk factor of high blood pressure [18]. Gender is also shown as a key factor in Reckelhoff et al., [19].

**5. Conclusion**

In this study, I proposed a method to identify and predict hypertension using Korea’s national health claim data and machine learning algorithms. This study especially focused on the hypothesis that people who share key characteristics with hypertension patients might potentially experience hypertension in the future.

By leveraging a variety of information related to personal examination results, I achieved approximately 0.55 accuracy score in kNN model. After adopting various models and hyperparameter tuning, I obtained 0.76 accuracy, 0.88 recall, 0.71 precision, and 0.78 F1 in XGBoost model. Therefore, the scores improved 1.37x in accuracy.

This study proposes utilizing healthcare data as a way to predict future risk of a disease. Although I applied this method to hypertension, a similar approach could be taken to many different diseases. The limitation of this method is that if there is a disease with a small number of patients, it is hard to apply machine learning algorithms because it cannot fully exploit the characteristics of machine learning. However, since healthcare data is accumulating quickly, this limitation will likely be overcome in the near future.

REFERENCES

1. Dieleman, J. L., Cao, J., Chapin, A., Chen, C., Li, Z., Liu, A., ... & Murray, C. J. (2020). US health care spending by payer and health condition, 1996-2016. *Jama*, *323*(9), 863-884.
2. Mawaw, P. M., Yav, T., Mukuku, O., Lukanka, O., Kazadi, P. M., Tambwe, D., ... & Luboya, O. N. (2017). Prevalence of obesity, diabetes mellitus, hypertension and associated risk factors in a mining workforce, Democratic Republic of Congo. *Pan African Medical Journal*, *28*(1).
3. Kirkland, E. B., Heincelman, M., Bishu, K. G., Schumann, S. O., Schreiner, A., Axon, R. N., ... & Moran, W. P. (2018). Trends in healthcare expenditures among US adults with hypertension: national estimates, 2003–2014. *Journal of the American Heart Association*, *7*(11), e008731.
4. Lackland, D. T. (2017). Early-Life Detection of Hypertension Risks: Implications for Clinical Practice and Research. *Hypertension*, *70*(3), 486-487.
5. Hossain, M. E., Khan, A., Moni, M. A., & Uddin, S. (2019). Use of electronic health data for disease prediction: A comprehensive literature review. *IEEE/ACM transactions on computational biology and bioinformatics*, *18*(2), 745-758.
6. Vaughn, D. A., van Deen, W. K., Kerr, W. T., Meyer, T. R., Bertozzi, A. L., Hommes, D. W., & Cohen, M. S. (2018). Using insurance claims to predict and improve hospitalizations and biologics use in members with inflammatory bowel diseases. *Journal of biomedical informatics*, *81*, 93-101.
7. Ma, F., Gao, J., Suo, Q., You, Q., Zhou, J., & Zhang, A. (2018, July). Risk prediction on electronic health records with prior medical knowledge. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1910-1919).
8. Krishnamurthy, S., Ks, K., Dovgan, E., Luštrek, M., Gradišek Piletič, B., Srinivasan, K., ... & Syed-Abdul, S. (2021, May). Machine learning prediction models for chronic kidney disease using national health insurance claim data in Taiwan. In *Healthcare* (Vol. 9, No. 5, p. 546). Multidisciplinary Digital Publishing Institute.
9. Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. *Annals of translational medicine*, *4*(11).
10. Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*(1), 81-106.
11. Raschka, S. (2014). Naive bayes and text classification i-introduction and theory. *arXiv preprint arXiv:1410.5329*.
12. Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-14
13. Ketkar, N. (2017). Stochastic gradient descent. In *Deep learning with Python* (pp. 113-132). Apress, Berkeley, CA.
14. Wright, R. E. (1995). Logistic regression.
15. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.
16. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).
17. Jin, M. J., Yoon, C. H., Ko, H. J., Kim, H. M., Kim, A. S., Moon, H. N., & Jung, S. P. (2016). The relationship of caffeine intake with depression, anxiety, stress, and sleep in Korean adolescents. *Korean journal of family medicine*, *37*(2), 111.
18. Gurven, M., Blackwell, A. D., Rodríguez, D. E., Stieglitz, J., & Kaplan, H. (2012). Does blood pressure inevitably rise with age? Longitudinal evidence among forager-horticulturalists. *Hypertension*, *60*(1), 25-33.
19. Reckelhoff, J. F. (2001). Gender differences in the regulation of blood pressure. *Hypertension*, *37*(5), 1199-1208.

1. \* Place the footnote text for the author (if applicable) here. [↑](#footnote-ref-1)