Mercedes Joseph

D209

Professor Straw

Classification Analysis Task 1

July 20th, 2023

**Part 1: Research Question**

**A1.**

**What are the major predictor variables in determining or predicting the Readmission of patients?**

**A2.**

My goal in conducting data analysis for medical readmission is to identify the key predictors or risk factors associated with patient readmission. By analyzing the available data, I aim to determine which variables have a significant impact on the likelihood of a patient being readmitted. This analysis will provide me with valuable insights into the factors that healthcare providers and systems should focus on to reduce readmission rates and improve patient care. By understanding these important factors, I can develop strategies and interventions that specifically target them, ultimately working towards the goal of reducing readmissions and providing better healthcare outcomes for patients.

**Part II: Method Justification**

**B1.**

The classification method I chose for analyzing the selected data set is the K-nearest neighbors (**KNN)** algorithm. KNN is a supervised learning algorithm that can be used for classification tasks. In the case of medical readmission analysis, KNN would analyze the selected data set by using the available features or variables to classify patients as either likely to be readmitted or not. The algorithm works by calculating the distances between a new data point **(a patient’s readmission)** and the existing data points in the training set. The "K" in KNN refers to the number of nearest neighbors to consider when making predictions.

To analyze the data using KNN, the algorithm would first require a labeled training set, where each data point is associated with a known readmission outcome (**readmitted or not readmitted**). The algorithm would calculate the distances between the new patient and the existing patients in the training set based on the selected features. The KNN algorithm would then assign a label to the new patient based on the most frequent class among its K nearest neighbors.

The expected outcome of applying KNN to the data set would be a predictive model that classifies patients as likely to be readmitted or not, based on the identified key predictors or risk factors. The model would then learn from the patterns and relationships in the training data and apply that knowledge to make predictions on new, unseen patients. The accuracy of the model's predictions will be evaluated using performance metrics such as accuracy, precision, recall, or F1 score, which will help me further assess how well the model is able to correctly classify patients based on their readmission status.

**B2.**

One assumption of the K-nearest neighbors **(KNN)** classification method is that if two data points are close to each other in the feature space, they are likely to belong to the same class. So KNN basically assumes that similar things are grouped together. This assumption is the basis of how KNN works. It calculates the distance between a new data point and existing data points to determine its classification. The idea is that if two data points are close to each other, they probably have similar characteristics and should be classified in the same way. However, despite its effectiveness it's important to remember that KNN's accuracy depends on this assumption being true. If the assumption doesn't hold true in the given data set, the predictions made by the KNN algorithm may not be very accurate. So, it's always important to carefully examine the data and ensure that the assumption of similarity among neighboring data points is reasonable before using KNN for classification.

**B3.**

**In my analysis, I utilized various R packages to perform different tasks. Here's a breakdown of how I used each package:**

1. **tidyverse**: Tidyverse is a package helped provide tools for working with data, such as cleaning, transforming, and visualizing it. It makes the data analysis process more efficient and organized.
2. **stats:** Stats has more of a fundamental package in R that offers various statistical functions and algorithms. The package helps me perform calculations and tests to understand the data better and draw meaningful conclusions.
3. **ggplot2:** This package is specifically for creating visualizations. With ggplot2, I can make customize plots like bar charts, line graphs, and scatter plots to explore and present the data visually.
4. **vcd:** vcd’s package is useful when working with categorical data. It helps provide functions to create visualizations and analyze patterns and relationships within categorical variables.
5. **gmodels**: This package helps me generate descriptive statistics and contingency tables. Allowing me to summarize and explore categorical data, making it easier to understand the relationships between different variables.
6. **class:** This package played a vital role that includes the k-nearest neighbors KNN algorithm for classification tasks. It helps me build models that predict outcomes based on similarities with other data points.
7. **caret**: This package is a comprehensive toolkit for machine learning and predictive modeling. It offers functions for data preprocessing, feature selection, model training, and performance evaluation, making it easier to build accurate models.
8. **kknn:** Is another vital package that provides an implementation of the KNN algorithm with additional options for customization. It helps me classify data based on similarities with other data points.
9. **FNN:** FNN’s package offers implementation of the K-nearest neighbors KNN algorithm. It's designed to handle large datasets and provide faster computation times.
10. **pROC:** pROC’s package is used for evaluating the performance of classification models. It helps me calculate metrics like sensitivity, specificity, and area under the curve AUC to assess how well the model predicts outcomes.

By utilizing these packages, I had a comprehensive set of tools at my disposal to perform data cleaning, manipulation, visualization, and model building. This enabled me to conduct a thorough analysis and derive meaningful insights from the data.

**Part III: Data Preparation**

**C1.**

In preparing the data for KNN analysis, my preprocessing goal was to identify and handle outliers effectively. Outliers are data points that significantly deviate from the rest of the data and can have a substantial impact on the performance of the KNN model. By detecting and addressing outliers, I can ensure the accuracy and reliability of the model's predictions. Removing or adjusting these extreme values will help prevent them from influencing the distance calculations used in KNN, allowing the model to make more accurate classifications based on the nearest neighbors. This preprocessing step is crucial for creating a quality KNN model that can provide valuable insights and reliable predictions for patient readmission. By preprocessing the data to potentially handle outlier values appropriately, I increased the quality and reliability of the classification results.

**C2.**

**Continuous Variables:**

* **Income** - Represents the patient's income level.
* **TotalCharge** - Represents the total charges incurred by the patient for their medical treatment.
* **Additional\_charges** - Refers to any additional charges associated with the patient's medical treatment, beyond the total charges.

**Categorical Variables:**

* **HighBlood** - Indicates whether the patient has high blood pressure (1 = Yes, 0 = No).
* **Stroke** - Indicates whether the patient has a history of stroke (1 = Yes, 0 = No).
* **Overweight** - Indicates whether the patient is overweight (1 = Yes, 0 = No).
* **Arthritis** - Indicates whether the patient has arthritis (1 = Yes, 0 = No).
* **Diabetes** - Indicates whether the patient has diabetes (1 = Yes, 0 = No).
* **Hyperlipidemia** - Indicates whether the patient has high cholesterol (1 = Yes, 0 = No).
* **BackPain** - Indicates whether the patient experiences back pain (1 = Yes, 0 = No).
* **Anxiety** - Indicates whether the patient has anxiety (1 = Yes, 0 = No).
* **Allergic\_rhinitis** - Indicates whether the patient has allergic rhinitis (1 = Yes, 0 = No).
* **Reflux\_esophagitis** - Indicates whether the patient has reflux esophagitis (1 = Yes, 0 = No).
* **Asthma** - Indicates whether the patient has asthma (1 = Yes, 0 = No).
* **ReAdmis** (Outcome) - Indicates whether the patient was readmitted (1 = Yes, 0 = No).

**C3.**

I used the CSV file provided for this class and imported it into Rstudios. During the data cleaning process, I checked first for any duplicates and missing values in the dataset, but on evaluation and using code I found there were none. Next, I decided to improve the readability of the column names by renaming Item 1-8 to respective names. I will provide a snip of the code that was used to change the column names below so that the evaluator can see the item’s respective names.

A screenshot of a computer screen

Description automatically generated

I then wanted to identify any potential outliers in the data. I ran a code to detect outliers in the dataset, and it revealed that approximately **0.35%** of the data points were outliers, while **99.65%** of the data remained within acceptable ranges. These outliers could potentially impact the analysis and model performance, so I decided to investigate them further. Then to gain a better understanding of the outliers, I focused on some of the variables that had outliers, such as **"Lat”, “Lng", "Population", "Income", "TotalCharge","Additional\_charges", "Timely admission", "Timely treatment”, “Timely visits", "Reliability”, "Options", and "Hours of treatment"**. I created visualizations like boxplots to visualize the distribution of these variables and better grasp the presence of outliers.

By visualizing the data with graphs, I was able to gain insights into the spread of the data and the potential impact of outliers. This informed my decision-making process, and I could determine whether to treat or remove the outliers based on their impact on the analysis. Overall, the data cleaning and outlier detection processes were crucial in ensuring the quality and reliability of the data, providing a solid foundation for analysis and modeling tasks.

To see each segment of code I have provided Rscript that includes the clean data named **“D209 CD.R”.**

**C4.** I have attached the clean data set named **“md1\_data4.csv”.**

**Part IV: Analysis**

**D1.**

**Provided test and training csv file in the attachments:**

**train\_data.csv**

**test\_data.csv**

**D2.**

I conducted a classification analysis using the K-nearest neighbors (KNN) algorithm to address my research question: **"What are the major predictor variables in determining or predicting the readmission of patients?"**. Knn is a vital algorithm that makes data “easy to understand and performs at a relatively high performance” due to its sheer simplicity (Deng, Z., 2016). I started the analysis by splitting the dataset into two sets: a training set comprising 70% of the data and a test set comprising the remaining 30%. Both sets included the predictor variables which included medical conditions and other factors as well as the outcome variable “**ReAdmis**”. I created separate CSV files for the training and test sets, which I provided in D1.

To prepare the data for KNN, I first leveled the training and test sets, ensuring that the categorical variables were appropriately encoded for the analysis. This step made sure that the data was in a suitable format for performing the KNN analysis. After leveling, I applied the KNN algorithm to the training set and utilized the trained model to make predictions on the test set. By comparing the predicted values with the actual values in the test set, I created a confusion matrix to assess the performance of the KNN model. The confusion matrix provides insights into the number of correct and incorrect predictions for each class **(readmitted or not readmitted)**.

**The results from the confusion matrix and Stats:**

**A screenshot of a computer

Description automatically generated**

In my analysis, I examined the confusion matrix to assess the performance of the classification model in predicting patient readmission. It provided valuable insights into the accuracy of the model's predictions. Looking at the matrix, I observed that the model correctly predicted **1694** instances as not readmitted **(0)** and **1048** instances as readmitted **(1).** However, there were **66** cases where the model incorrectly predicted readmission when it was not readmitted **(false positives),** and **192** cases where it incorrectly predicted no readmission when it was readmitted **(false negatives).**

I decided that further evaluation may be needed for the model, I calculated several metrics. The accuracy of the model, which measures the overall correctness of the predictions, was found to be **0.914**. The sensitivity **(recall),** which indicates the model's ability to correctly identify readmitted instances, was **0.8982**. The specificity, which measures the model's ability to correctly identify not readmitted instances, was **0.9408**. The positive predictive value **(precision)** of the model was **0.9625**, indicating the proportion of predicted readmitted instances that were readmitted. The negative predictive value, which represents the proportion of predicted not readmitted instances that were not readmitted, was **0.8452**.

Overall, I am satisfied with the overall performance of my model in accurately predicting patient readmission. The high accuracy, sensitivity, specificity, and PPV values seem to demonstrate the effectiveness of the selected predictor variables and the K-nearest neighbors analysis process in capturing the underlying patterns in the data.

**F1 Score:**

****

In my KNN analysis of the classification model for predicting patient readmission, the F1 score plays a crucial role in evaluating its effectiveness. With an F1 score of 93%, the model demonstrates a good balance between precision and recall. Meaning that the model accurately identifies patients who are likely to be readmitted while minimizing both false positives and false negatives. The F1 score indicates that the model is reliable and capable of making accurate predictions based on the data patterns.

**D3.**

I have attached the rscript named **“D209 KNN Task 1. R”** to for your evaluation.

**Part V: Data Summary and Implications**

**E1.**

The accuracy of my K-nearest neighbors (KNN) classification model on the test set is 0.914. This means that the model correctly classified approximately 91.4% of the instances in the test set. A higher accuracy indicates that the model is performing well in predicting whether a patient will be readmitted or not. Also, the area under the curve (AUC) is 0.9195. The AUC is a measure of the model's ability to measure between the positive and negative classes. A higher AUC value suggests that the model has good idea on how to classify patients into the correct readmission category.

**Roc Curve**

A graph of a curve

Description automatically generated

I created a ROC curve graph to create a visual for the model. The model's accuracy is high, with a confidence interval of **0.898** to **0.941**, suggesting strong predictions. The model is based on “the notion of a separator" (or decision) variable” (Hajian-Tilaki, K., 2013). The AUC value of **0.92** means that the model is effective at classifying between readmitted and not readmitted patients. Overall, these results show that the model performs well in predicting patient readmission. Although, I can always perform more calculations to help improve the model “learning a predictive model is to optimize some performance metric” (Yang, T., & Ying, Y., 2022). I am overall satisfied with the results of my KNN analysis especially in its ability to perform well in predicting patient readmission.

**E2.**

The results of the classification analysis using K-nearest neighbors (KNN) provided valuable insights into predicting patient readmission. With an accuracy score of **91.4%** and an area under the curve (AUC) of **0.9195**, meaning the model demonstrated a strong ability to correctly classify patients as either likely to be readmitted or not. The findings have important implications for healthcare providers and systems. By identifying key predictors or risk factors associated with readmission, with the chosen predictor variables **("HighBlood", "Stroke", "Overweight", "Arthritis", "Diabetes", "Hyperlipidemia", "BackPain", "Anxiety", "Allergic\_rhinitis", "Reflux\_esophagitis", "Asthma", "Income", "TotalCharge", "Additional\_charges").** Having a high accuracy and AUC values that indicate the chosen predictor variables and the KNN algorithm effectively captures the pattern and relationship in the data. The results offer actionable information for healthcare providers to prioritize interventions and improve patient care, that ultimately leads to better outcomes and resource utilization.

**E3.**

If I must determine one limitation of my data analysis it would be that I lack specialized knowledge in the healthcare field to create a more effective model. While I have applied standard techniques and used the available data to the best of my ability, not having expertise in healthcare may impact the accuracy and reliability of the model.

To overcome this limitation, it would be beneficial to involve a healthcare expert in the analysis. Their deep understanding of the healthcare domain and specific factors influencing patient readmission, such as **“HighBlood”, “Stroke”, “Overweight”, “Arthritis”, “Diabetes”, “Hyperlipidemia”, “BackPain”, “Anxiety”, “Allergic\_rhinitis", “Reflux\_esophagitis”, “Asthma”, “Income”, “TotalCharge”, and “Additional\_charges”** can provide valuable insights. They can help identify relevant variables, refine the model's structure, and ensure that the analysis captures the complexities of patient readmission accurately.

By collaborating with a healthcare expert, we can develop a more accurate model that considers important nuances and factors specific to the healthcare context. Their expertise will contribute to a better understanding of the relationships between variables and improve the model's predictive capabilities. In summary, the involvement of a healthcare expert, who is knowledgeable about the specific variables and intricacies of patient readmission, would greatly enhance the quality and relevance of the data analysis. Their specialized knowledge will help address limitations and improve the overall effectiveness of the analysis. Collaborating with a healthcare expert is essential to develop a more accurate and relevant predictive model for patient readmission. Their expertise and understanding of specific variables will greatly enhance the analysis, leading to valuable insights and improved model performance. Their involvement is crucial in optimizing patient care and resource allocation, making the analysis more effective for real-world healthcare decisions.

**E4.**

To help solve the issue of patient readmission, I would need to focus on important factors that are the likely factors that lead to patient readmission. Research shows that “just about 25% of all readmissions are avoidable/preventable “meaning if the right precautions are taken readmission could be avoided” (Kansagara, D., 2011). By paying attention to these factors, we can identify patients who are more likely to be readmitted and provide them with extra support. We should also closely monitor high-risk patients by checking up on them regularly and creating personalized care plans. Where we would need the help of healthcare experts, so I can enhance the accuracy and reliability of the model for predicting patient readmission. The health expert’s in-depth knowledge and expertise in healthcare will provide valuable insights into the specific variables and nuances relevant to patient readmission. The collaboration will ensure that the analysis considers important factors specific to the healthcare context, leading to a more effective model. This approach helps our process in catching any issues early and take necessary steps to prevent readmission. Also, by developing specific care plans or interventions for patients with high blood pressure, stroke history, overweight, arthritis, diabetes, and high cholesterol. Continuously monitor and assess the outcomes of these interventions to evaluate their effectiveness and make necessary adjustments.

By doing so, we can optimize patient outcomes, allocate resources efficiently, and improve the overall quality of care. This data-driven approach enables us to make informed decisions and drive improvements in patient management and hospital operations. This data-driven approach allows us to optimize patient outcomes, allocate resources efficiently, and improve the overall quality of care. We may even need to investigate other models that will better complement and express the data properly.

**References**

Deng, Z., Zhu, X., Cheng, D., Zong, M., & Zhang, S. (2016). Efficient kNN classification algorithm for big data. *Neurocomputing*, *195*, 143-148.

Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., & Kripalani, S. (2011). Risk prediction models for hospital readmission: a systematic review. *Jama*, *306*(15), 1688-1698.

Yang, T., & Ying, Y. (2022). AUC maximization in the era of big data and AI: A survey. *ACM Computing Surveys*, *55*(8), 1-37.