Introduction - Describing the problem

In the ever-evolving landscape of healthcare, the ability to predict patient outcomes accurately holds immense value and utility for the improvement of patient care, health and hospital resource allocation. In this report, we will present our solution to an exceedingly critical problem: predicting hospital readmission for patients diagnosed with diabetes. For our solution, we leveraged the use of a comprehensive dataset that spans 10 years and incorporates data from 130 different hospitals. Our task was to develop a classification model capable of identifying individuals at risk of readmission. Additionally, we explore the application of K-means clustering to uncover a non-trivial set of patients that could offer valuable insights to the healthcare industry. This report aims to outline our approach, methodologies, findings, insights and recommendations aimed at enhancing patient care and operational efficiency within healthcare organisations.

What data will we use to solve this problem?

The dataset we will base our classification model on is available here.
There are 50 columns with 101766 rows of data (excluding the headers). Upon raw visual inspection, the data looks very dense. The column which appears to have the most missing data is “weight” (It is very prevalent for this to be missing in medical data sets).

The first thing that needed to be gained was an understanding of EVERY column. After some research, this is how we defined our columns.

- **encounter_id**: Just an identifier, no contribution to predictive power.
- **patient_nbr**: The same, but data leakage must be avoided to ensure the model doesn’t learn from multiple admissions of the same patient, or the same patient is present in train AND test data.
- **race, gender, age**: Important for predictive power.
- **weight:** Even if this data was present, this data doesn’t take into account height. BMI would be more appropriate. Most of this data is missing anyway.
- **admission_type_id, discharge_disposition_id, admission_source_id**: This can be informative, reflecting the severity of the admission, patient condition at discharge and the context of the hospital visit.
- **time_in_hospital**: Direct indicator of the severity of the last visit. Can be informative.
- **payer_code, medical_specialty**: Not directly related to patient health. But these may provide context about the quality of the care and the focus of the care. This may affect risk. However, lots of missing data.
- **num_lab_procedures, num_procedures, num_medications**: Indicators of the complexity and severity of the patient's condition(s).
- **number_outpatient, number_emergency, number_inpatient**: These counts from the year preceding the encounter provide a picture of the patient's recent healthcare utilisation, potentially indicating chronic condition instability or severity.
- **diag_1, diag_2, diag_3**: Diagnosis codes are crucial for understanding the patient's condition. The primary diagnosis is particularly important, but secondary diagnoses

provide depth to the patient's health status. These come as “ICD” codes. More specifically ICD9. We decided to categorise these, you’ll see how later on.

- **number_diagnoses**: Reflects the overall burden of illness, which could impact readmission risk.
- **max_glu_serum, A1Cresult**: Key laboratory test results that can indicate the control of diabetes, a central factor in readmission risk. Again lots of missing data. Most people do not do these tests.
- **change, diabetesMed**: Indicators of whether there was a change in diabetic medications and whether diabetes medications were prescribed can reflect the management and severity of diabetes.
- **Specific diabetes medications (e.g., insulin, metformin, etc.)**: While insulin might be the most directly relevant given its critical role in diabetes management, other medications can also provide insight into the patient's treatment regimen's complexity and type. The specific use patterns of these medications (increased, decreased, held, no usage) can reflect changes in clinical status or responses to treatment.
- **readmitted**: The target variable.

Data Cleansing & Visualisation

Data Cleansing

Data cleansing is essential for large datasets to ensure accuracy, reliability, and consistency in analysis by removing errors, inconsistencies, and duplicates, thus enhancing the quality and integrity of the data for meaningful insights and decision-making.

For the integrity of the analysis purpose, the data was cleaned, the cleansing included the following steps:

1. Replacement of the missing entries, which were denoted by “?” in the dataset.
2. Removal of the columns where more than 50% of data was missing and columns where 95% of the entries were the same.
3. Transformation of ‘age’ into a numerical representation by calculating the mean age of the range provided.
4. Replacing the missing values in the diagnosis data columns (diag_1, diag_2 and diag_3) with 0.
5. Removal of duplicates based on the ‘patient_nbr’ to avoid data leakage and ensure the uniqueness of the data.
6. Dropping of less informative columns like 'payer_code' and 'medical_specialty', which had a significant portion of missing data (to be precise: ‘payer_code’ having 39.56% missing data and ‘medical speciality’ with 49.08%). These two columns do not carry important data into our model, so removing them permanently does no damage.
7. Conversion of numerical columns into their appropriate numeric data types and exclusion of outliers by removing data points that lie beyond 3 standard deviations from the mean.
8. Transforming the ‘readmitted’ column into a binary variable for the ease of the analysis.

The processed dataset is then referred to as 'filtered_data'. This dataset was then used to generate visualisations to assess the impact of various factors on hospital readmission specifically for patients with diabetes-related issues. There were some missing values in the "race" column. To increase the amount of data at our disposal we imputed these cells with random values.

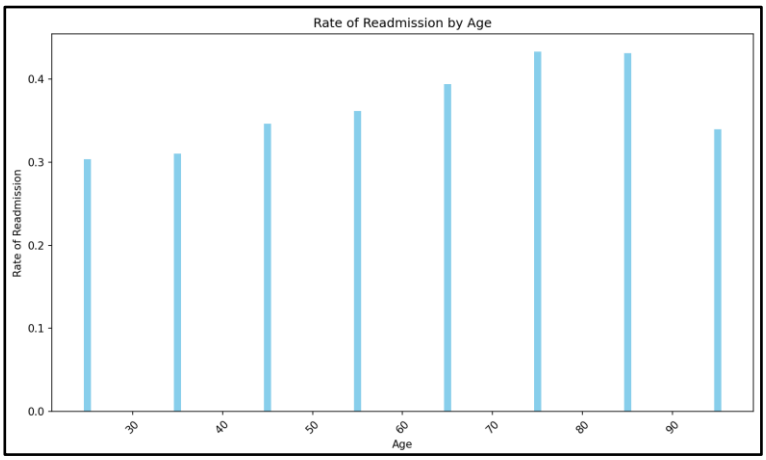
Data Visualisation

Before conducting the detailed data visualisation and analysis, several hypotheses were posed to guide the investigation. They are as follows:

- Age has a higher impact on readmission.
- African Americans are more likely to be readmitted than other ethnic groups.
- Women patients are more likely to be readmitted than men.
- Diagnosis types have a higher impact on readmission rates.

These hypotheses are the starting point for our analysis. We want to understand the factors that lead to patients being readmitted to the hospital.

Age vs Readmission

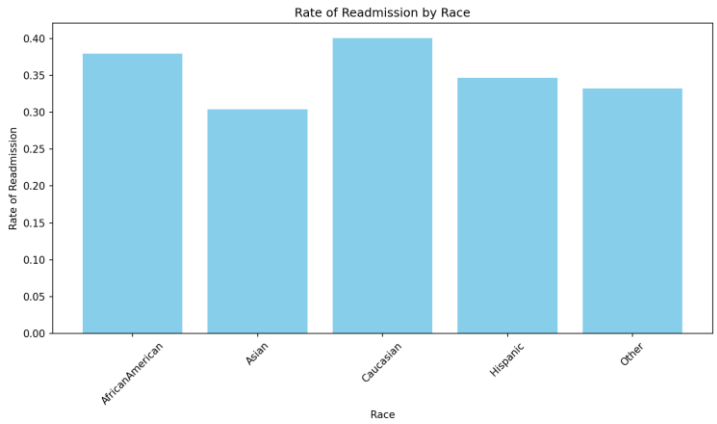


The bar chart depicts that the readmission rates are not uniformly distributed across the age groups, which indicates that age has a complex relationship with readmission likelihood. The graph shows a notable increase in readmission for individuals aged 60 and above. This trend suggests that people starting from 60 years old, are more likely to be readmitted to the hospital. This could be due to a variety of factors, such as the increased prevalence of chronic diseases, decreased resilience to health setbacks, or complications related to the ageing process.

Chi-square test for age: p=2.495066779305635e-82
Cramer's V for effect size between age: 0.07904422940779149

The chi2 test's result confirms a strongly statistically significant difference in readmission rates across different age groups, while the Cramer's V value, despite not being large, is significant enough to suggest that age has a noticeable association with readmission rates.

Race vs Readmission



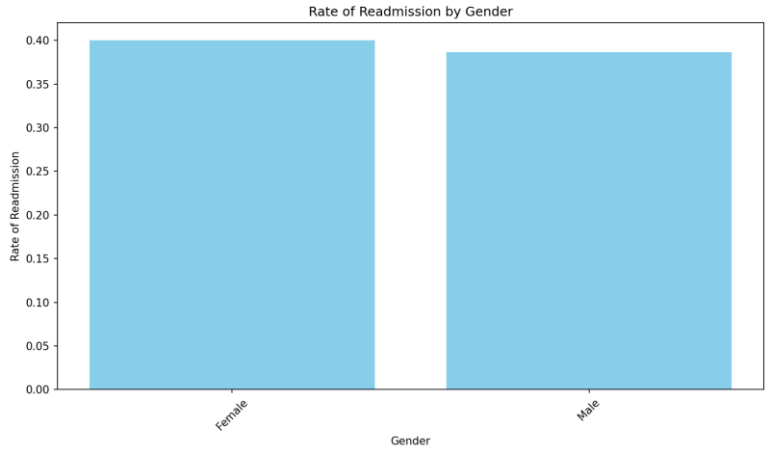
Upon examining the rate of readmission by race, the data indicates noticeable variability among different racial groups. The data underscores that Caucasian patients present the highest rate. Following them are African American patients who exhibit the second-highest rate of readmission, albeit marginally lower than that of Caucasians. However, it is crucial to note a distinct pattern wherein African Americans demonstrate a notable increase in readmission rates compared to other ethnic minorities. This observation suggests potential disparities or underlying factors influencing healthcare outcomes within the African American community.

This discrepancy, although small, suggests that there may be unique challenges and or systemic barriers specific to the African American community that contribute to their increased likelihood of readmission.

Chi-square test for race: p=8.838779083206117e-14
Cramer's V for effect size between races: 0.03240262967315212

The extremely low value in the chi2 test indicates a very significant association between race and rate of readmission. However, the Cramer's V value indicates a weak effect between race and readmission status. While statistically significant, the practical impact of this association is relatively small, suggesting that race alone explains only a small portion of the variability in readmission rates. Overall the statistical significance underscores race as a factor in readmission rates that should not be ignored. Its interaction with other variables needs to be studied further.

Gender vs Readmission



A comparison of readmission rates between the two genders shows no distinct difference, This observation disproves the hypothesis that women may be more likely to be readmitted than men.

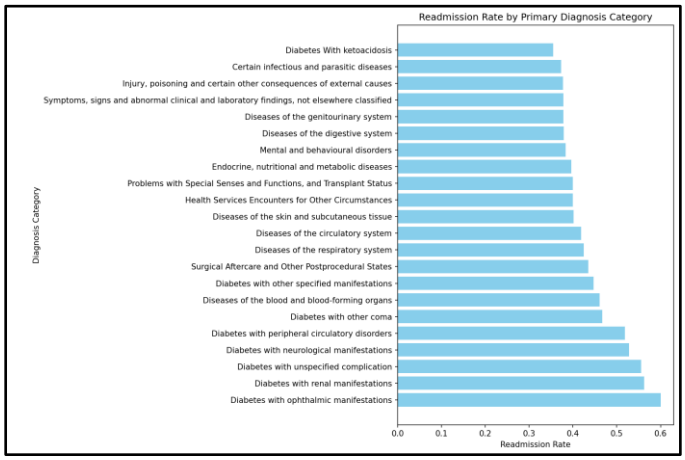
Chi-square test for gender: p=0.0010782787626037 66

Cramer's V for effect size between genders: 0.014611370541337489

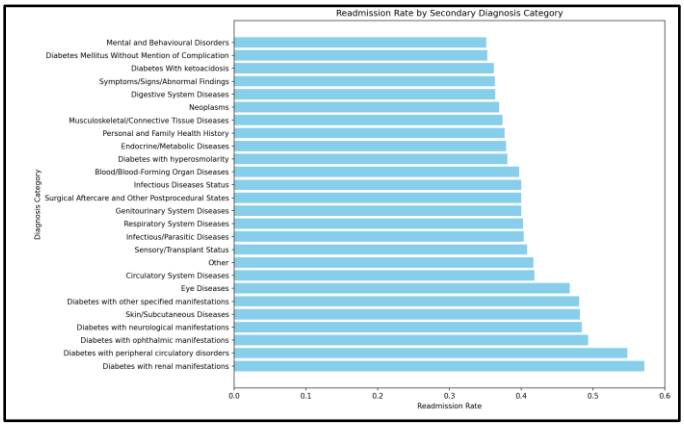
Our chi2 test shows that there is a statistically significant association between gender and readmission status. However, the effect size being close to 0 indicates that the practical significance of this association is minimal. This means that, despite the statistical significance, the difference in readmission rates between genders may not be large enough to warrant clinical or policy changes based on gender alone.

Diagnoses vs Readmission

In our dataset, we encountered an extensive range of medical conditions as ICD codes. To facilitate a more streamlined analysis, we consolidated these ICD codes into 33 broader disease categories. This categorisation was crucial for our visual and analytical examination of the data. However, upon initial inspection, it was evident that many of the categories presented a significantly low number of admissions. To ensure our analysis was focused on the most impactful data, we decided to refine our visual representation to include only the disease categories that accounted for more than 0.35% of admissions.

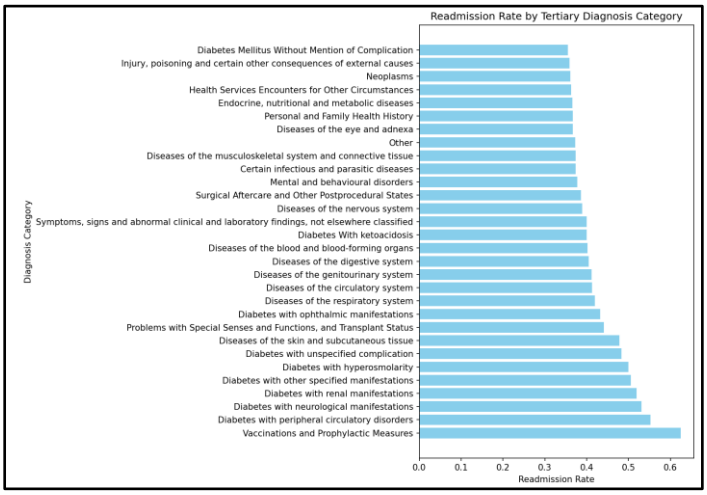


The primary diagnosis chart displays that “Diabetes with ophthalmic manifestations ” has the highest rate of readmissions among the diagnosis categories. This is followed by “Diabetes with renal manifestations”, “Diabetes with unspecified complications”,” Diabetes with neurological manifestations” and “Diabetes with peripheral circulatory disorders ”. These 5 diagnosis types are directly associated with various complications and manifestations of diabetes and emerge as significant contributors to readmission rates (meaning the patients with these diseases are more likely to be readmitted). This underscores the importance of focusing on specific diagnosis categories to effectively address and mitigate factors contributing to readmissions.



The Secondary Diagnoses chart underscores a clear pattern in which diabetes-related complications dominate the highest readmission rates. "Diabetes with renal manifestation" emerges as the primary concern with the highest readmission rate, followed closely by "Diabetes with peripheral circulatory disorders." Notably, these conditions exhibit considerably higher readmission rates compared to other diagnoses. Additionally, the cluster of diagnoses including "Diabetes with ophthalmic manifestations," "Diabetes with neurological manifestations," "Skin/Subcutaneous

Diseases," "Diabetes with other specified manifestations," and "Eye diseases" all demonstrate above-average readmission rates, further emphasising the impact of diabetes-related complications on readmission rates.



Similarly, the Tertiary Diagnosis chart highlights the prominence of diabetes-related issues in contributing to readmission rates. "Vaccinations and Prophylactic Measures" stand out as having the highest readmission rates by a significant margin, indicating potential issues with preventive care management. Following this, several categories related to diabetes complications, such as "Diabetes with peripheral circulatory disorders," "Diabetes with neurological manifestations," and "Diabetes with renal manifestations," demonstrate notably high readmission rates.

Commenting on these findings, it becomes evident that diagnosis types associated with diabetes have a substantial influence on readmission rates (considering we are using diabetic patients' data). Understanding the heightened risk associated with certain diagnosis types can inform targeted interventions and care plans aimed at reducing readmission rates and improving overall patient outcomes. Thus, healthcare providers should prioritise comprehensive management approaches tailored to address the specific needs and challenges posed by diabetes-related complications to mitigate the burden of readmissions on patients and healthcare systems.

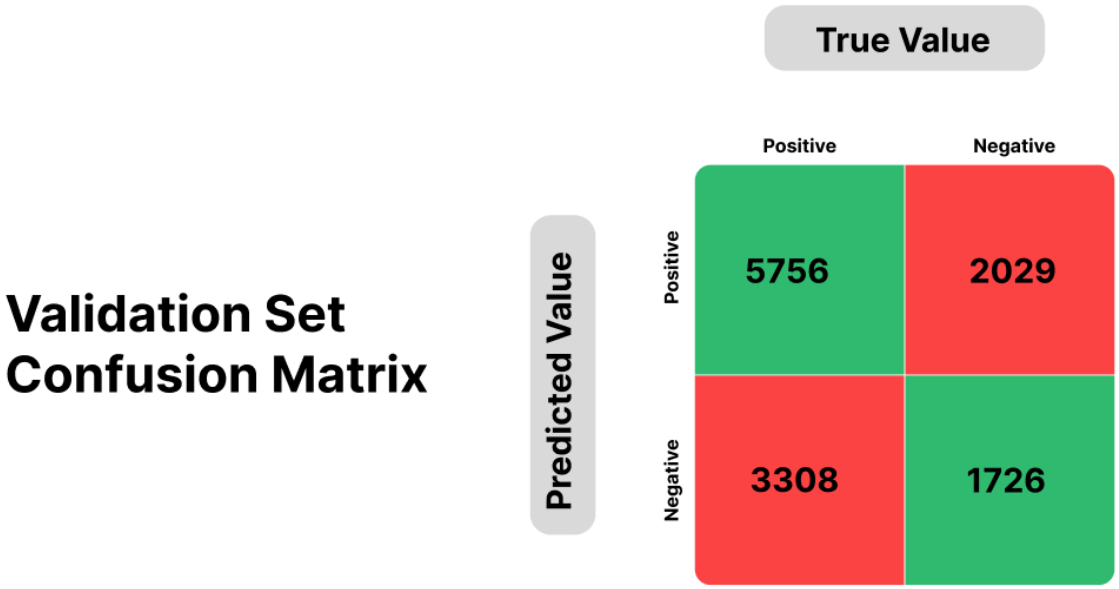
Basic Predictive Model

The classification model which was constructed to predict hospital readmission uses a subset of the following variables: *'num_medications', 'number_outpatient', 'number_emergency', 'time_in_hospital', 'number_inpatient', 'encounter_id', 'age', 'num_lab_procedures', 'number_diagnoses', 'num_procedures', 'readmitted'.* Before training the model, the previously mentioned data cleansing was done. Most importantly, the 'readmitted' variable was binary encoded, distinguishing between patients who were not readmitted (0) and those readmitted within or after 30 days (1).

We have chosen to build a **Random Forest** Classifier, which is known for its robustness and ability to handle non-linear relationships. The dataset was split into a training set and a test set, with 80% of the data used for training the model and the remaining 20% for testing its predictive capabilities. When evaluating the model, we found that it performs exceptionally well on the training data, however, it does not generalise effectively to the validation data. This is indicative of **overfitting**, where the model learns the training data and its noise too well but fails to predict accurately.

Here is a look at the scores we achieved between the training set and the validation set:

Scores	Training Set	Validation Set
Accuracy	98.9%	58.4%
Precision	99.0%	46%
Recall	98.3%	34.3%
F1 Score	98.6%	39.3%
ROC AUC Score	98.8%	54.1%



The **contrast** in these results is a clear sign that the model's ability to generalise is limited. The confusion matrix for the validation data also shows a relatively high number of both **false positives** and **false negatives**. This suggests that the model struggles to correctly classify readmission cases in unseen data. The cross-validation scores support these findings with:

- **Mean Accuracy**: 57.5%. This states that the model correctly predicts the outcome slightly more than half of the time. Suggesting that the model has learned only to some extent from the data.
- **Mean Precision**: 44.8%. This indicates that the model has a moderate tendency to correctly identify positive instances among all instances it labels as positive. However, the low score indicates a notable amount of false positives.
- **Mean Recall**: 32.8%. This implies that it misses a considerable number of actual positive instances, identifying less than one-third correctly.

- **Mean F1 Score**: 37.8%. This suggests that the model doesn't strike an optimal balance between precision and recall. The score clearly indicates that the model struggles both with identifying the relevant instances and capturing a large portion of the actual data.
- **Mean ROC AUC Score**: 55.2%. This shows that the model has a slightly better than random chance at distinguishing between positive and negative classes. It highlights that the model's ability to differentiate between classes isn't strong and could be improved

To enhance the model we can try to improve it by fixing the issue regarding overfitting. Hyperparameters can be used to optimise the model (such as 'n_estimators', 'max_depth', 'min_samples_split' and 'min_samples_leaf'), by adjusting the hyperparameters of the model we can get better control over the model's complexity and improve its ability to generalise to unseen data. We can also use the technique called Pruning, which simplifies the decision trees in our Random Forest by removing sections that provide little value. Moreover, enhancing our cross-validation approach, such as by increasing the number of folds in cross-validation can provide a more accurate estimate of the model's performance. As it will ensure a more thorough and balanced evaluation process.

Improved Model: What data is useful to us?

To enhance the accuracy and precision of the improved model, it's crucial to ensure that only relevant data is fed. Therefore, we must identify which dataset columns impact the likelihood of readmission. It's important to determine the factors contributing to a patient's medical volatility

Here's the final list of columns in our dataset that will be used:

race - After looking at our plots, we can see that this has an effect on the probability of readmission.
gender - Men have a much higher chance of having diabetes generally. However, our plots show that women seem to have a higher probability of readmission.
age - The older one gets, the more susceptible they are to disease.
admission_type_id - If one was last admitted in as an emergency for example, then this may affect their chance of readmission.
discharge_disposition_id: Knowing about the nature of one's discharge could be an important insight when it comes to the chance of readmission.
admission_source_id: More knowledge on the nature of one's condition.
time_in_hospital: If the patient spent a long time during their last visit it may indicate a degree of instability. Affecting the probability of readmission.
num_lab_procedures, num_procedures, num_medications: Can be an indicator of medical burden.
number_outpatient, number_inpatient, number_emergency: This can be an indicator of medical burden.
diag_1, diag_2, diag_3: These are important, as the more diagnoses, the more likely one is to be medically unstable. However, in its raw format, this data is NOT useful.
number_diagnoses: The above diagnoses are included in this number, this is a further indicator of medical volatility.

Information on the drugs that remain after data cleaning: Namely metformin, glimepiride, glipizide, glyburide, pioglitazone, rosiglitazone and insulin.
change, diabetesMed: Gives us direct insights on the patient's diabetes itself.

Our extensive dataset offers significant insights. To further improve our model's performance, we can engineer additional features. Here's are the ones we decided to create:

number_diagnoses_squared: The idea here is that there difference in medical instability goes up exponentially based on the amount of diagnoses one possesses. The difference between 1-6 diagnoses seems less drastic than say 6-12. We drop the original column to avoid multicollinearity as these values are correlated. We may also do this for **num_medications** or **age**. Not used for our cluster analysis (added after).

diabetes_medication_count: Counts the number of diabetes medications one is on. This is done BEFORE any data cleaning. So all 18 drugs are considered. Note: Combination drugs count as 1.

medications_increased, medications_decreased, total_medication_changes: These can be strong indicators of if one's diabetes is becoming more/less manageable.

medical_burden: This is essentially a "volatility" measure. The assumption is that the more one is using medical services the more likely their conditions are to be unstable.

There are many ways of implementing this, however. We could even add "time_in_hospital" as it appears relevant. Also, is simply adding these values together the best way? As this kind of medical data tends to be skewed we could take the LOGS instead. This is an attempt to make this variable linear.

We use np.log1p (log(1+x)) so that we don't perform a log(0) operation.

Diagnosis code categories: In assessing the utility of the raw data within these columns, it becomes evident that categorization is imperative for meaningful analysis. There are 3 types of ICD codes, the types of codes which may affect readmission probability are the plain codes and the V codes, which represent diseases and conditions/supplementary encounters respectively. E.g. V65.4 regards counselling on diet and exercise, which if done properly should reduce the severity of diabetes in a person.
A possible route to explore could be linking these categories to a comorbidity index. E.g Elixhauser index. Or for a more specialised and applicable index maybe the (Adapted) Diabetes Complications Severity Index. [1]

age_group: This aims to reduce the complexity of the model. By categorising rows by age group.

We find that this reduces precision but increases recall. This is omitted in our final model.

readmitted_binary: Our target variable is in binary form. This is the most important variable.

Finally, to address class imbalance we use SMOTE. This oversamples the minority class and has this effect:

```
Class weights before smote:  [0.82448796 1.27044462]
Class weights after smote:  [1. 1.]
```

This will make our measure of accuracy more useful. Since there is a 50/50 chance for **readmitted_binary** to be 1 or 0. This can introduce artificial noise into the data. So on the other hand we could undersample the majority instead. In a medical scenario maybe this is more appropriate.

So this leaves us with circa 64000 rows of data in total, ready to be analysed.

The improved model

Here we have a chance to be creative and innovative. Since the random tree/gradient boost models weren't so good at our scoring metrics we moved onto a more advanced type of model. A neural network. To do this we used tensorflow, which is great at identifying both linear and non-linear patterns. The data we used is from the "What data is useful to us?" section. Note: While testing our model, removing the **age** column improved our accuracy by a couple of percent. So our cluster analysis does not account for this variable.

We started with a simple model and kept increasing its complexity/tuning certain parameters until we saw diminishing returns. We ended up with this:

The model starts at 512 nodes, and funnels down (256, 128, 32, 16, 8, 4 to 1). This shape seems logical for a classifier model. Our optimiser is Adam. We tried to run the model with all of them once and Adam was the best. We used dropout layers, l2 regularisation and batch normalisation to reduce overfitting. **Maybe we used it too much?**

Here's a decent result derived from our stratified K-Fold testing of said model.

Firstly we see that the difference between the training and validation results is very small. This indicates a model that is not overfit and is generalising to the data well.

The accuracy tends to hit ~70%. This means the model is learning something useful. The question is, how much room for improvement is there? Could we possibly hit scores of over 80% with this data?

The precision here hits 80%. But in this context, is the cost of a false-positive high? In a medical scene maybe it would be better to increase the recall at the cost of this metric. This can be done by reducing the threshold for the model to determine what/who to flag or not. It would be better to avoid false negatives. In my opinion.

Our recall hits 60%. Going back to the previous point, a higher recall can be critical because missing out on a high-risk patient (a false negative) can have serious implications.

The ROC AUC score hits 80%. An AUC of 80% indicates the good discriminative ability of our model. It suggests that there's a high chance that a randomly chosen positive instance is ranked more likely to be positive than a randomly chosen negative instance.

Our F1 score hits 70%. F1 is simply the harmonic mean between precision and recall. Here we need to understand if we are striking an appropriate balance between these two metrics. This result shows us our model is fairly accurate and sensitive.

It's important to note that, while the **accuracy** and **ROC AUC** stay consistent. The **precision** and **recall** fluctuate. We have even hit a precision of 1.0 a few times and if you run our code you will almost definitely see it happen. When the **precision** is higher our **recall** tends to be lower and these differences are not proportional either. Our **F1 score** tends to be lower, the higher the **precision** is. The result above strikes the middle ground between these two values perfectly. This is why we created a callback to save the model with the best **F1 score**.

Interpretation

The precision and AUC are high, this shows that our model is good at distinguishing between the classes and ensuring that when it predicts high risk, it is correct a majority of the time. **The model has a high discriminative ability**.
The low recall may not be ideal in a medical context. It's important to assess the implications of a false negative in this context. But for this, **we need to consider the opinion of medical experts**.

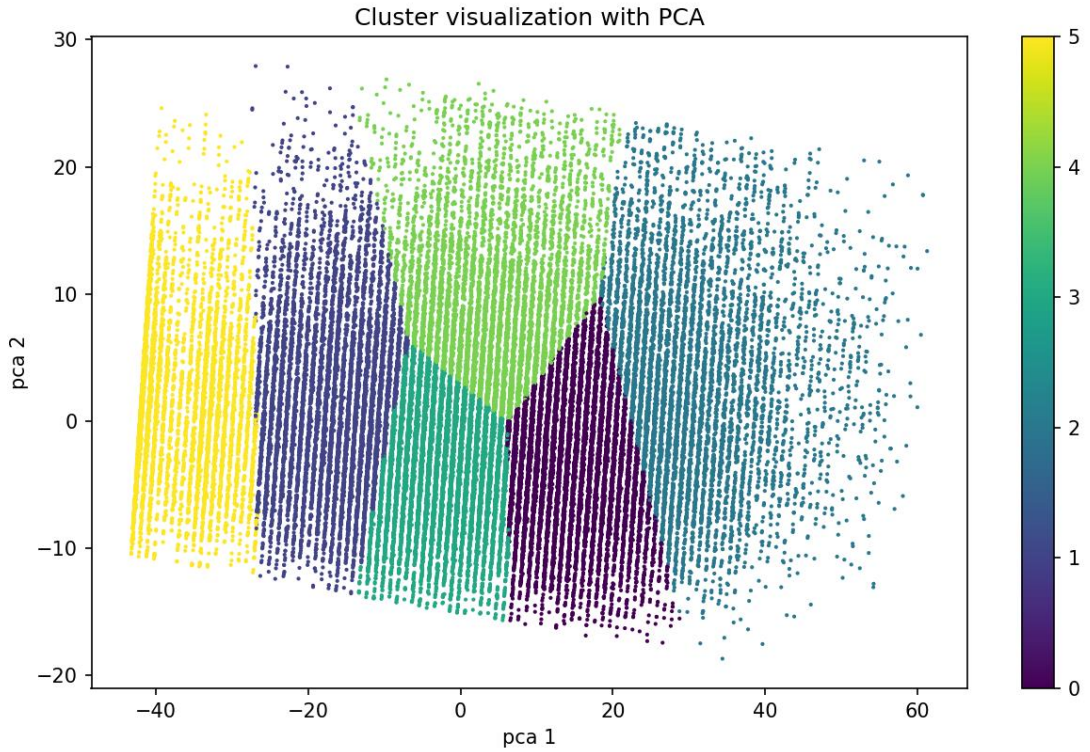
Moving forward

How could we improve this model even further? We already did some feature engineering, we already used some resampling techniques (SMOTE).

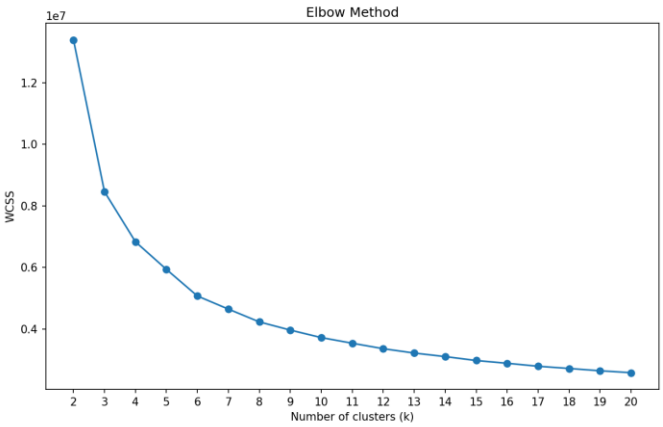
We tried some hyperparameter tuning: changing the amount of layers, epochs, the batch size, the learning rate, and even trying Bayesian optimization. Didn't really do much to our results but maybe we were doing it improperly. It appears the default values for the components we used were appropriate enough.

We saved one of our best models "70a70p70r80auc70f" if you would like to load the weights of it and test it.

Cluster and further analysis



Here are our clusters of the **X_train_smote** data. The number of clusters (6) was decided using the elbow method. The acceptable values appear to be 5,6 or 7. We used PCA for dimensionality reduction. This captures the essence of the data while putting it into a graphable form. Despite PCA being a linear technique, the clusters looked a lot better and clearer than when using (supposedly better-suited) alternatives like T-SNE.



After training our improved model we then make it predict our full SMOTEd dataset.

Then we group these predictions by cluster.

This generates the results in the image below:

The “mean” column represents the average risk score of an individual in the named cluster. Patients in clusters 1, 3 and 5 seem to be at a significantly elevated risk of readmission. With anyone in cluster 5 seemingly guaranteed to be readmitted. Cluster 2 has the lowest average risk score (mean = 0.366), indicating that, on average, patients in this cluster are much less

likely to be flagged as at risk of readmission. The “count” column just indicates the amount of patients in that cluster.

	Cluster	mean	count
0	0	0.552891	13389
1	1	0.937055	10497
2	2	0.366295	7814
3	3	0.859967	15232
4	4	0.551035	8157
5	5	0.993303	7015

Simply down to chance, our even-numbered clusters have a lower average risk and our odd-numbered clusters have a higher average risk. Keep that in mind.

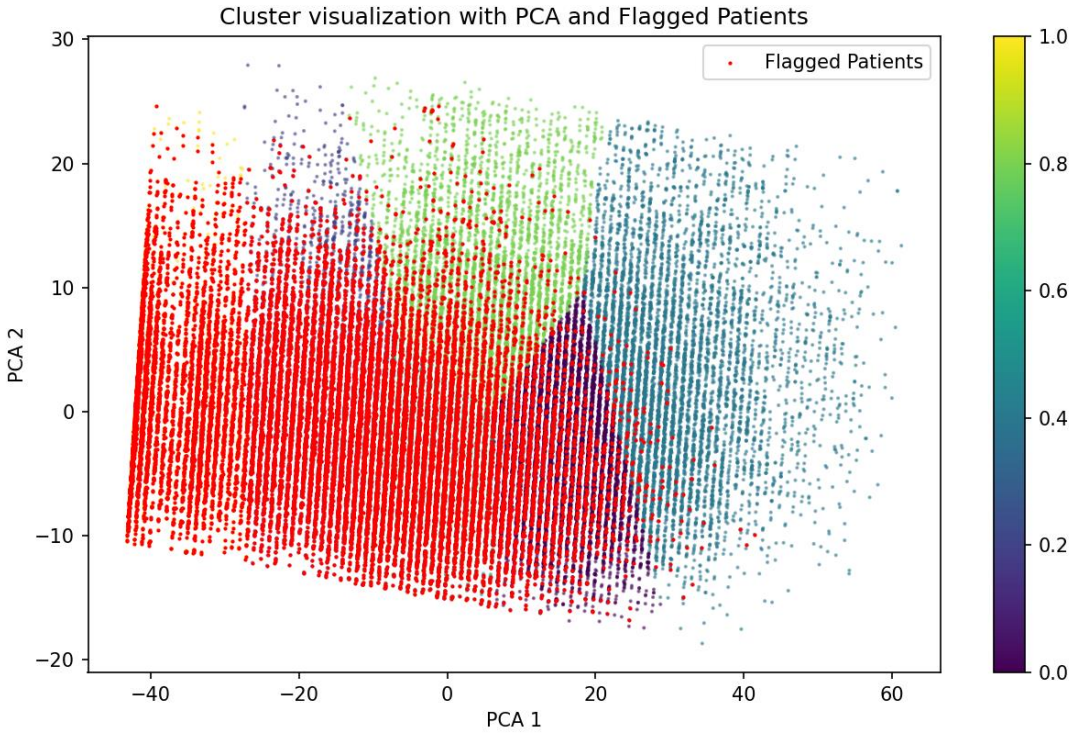
So what are the characteristics of these at-risk patients?

We can define high and low-risk patients like this:

```
high_risk_patients = data_with_clusters_and_predictions[data_with_clusters_and_predictions['AtRiskPrediction'] > 0.8]
low_risk_patients = data_with_clusters_and_predictions[data_with_clusters_and_predictions['AtRiskPrediction'] < 0.4]
```

Anything less than 0.4 for the low-risk patients does not give us enough data for us to be confident with our conclusions. We chose 0.8 to be an appropriate level to consider one at high risk.

Let's flag these high-risk patients on our cluster chart:



Then we can use some descriptive statistics to understand what's happening between the low-risk and high-risk patients, to try and see what are the main factors that set these groups apart.

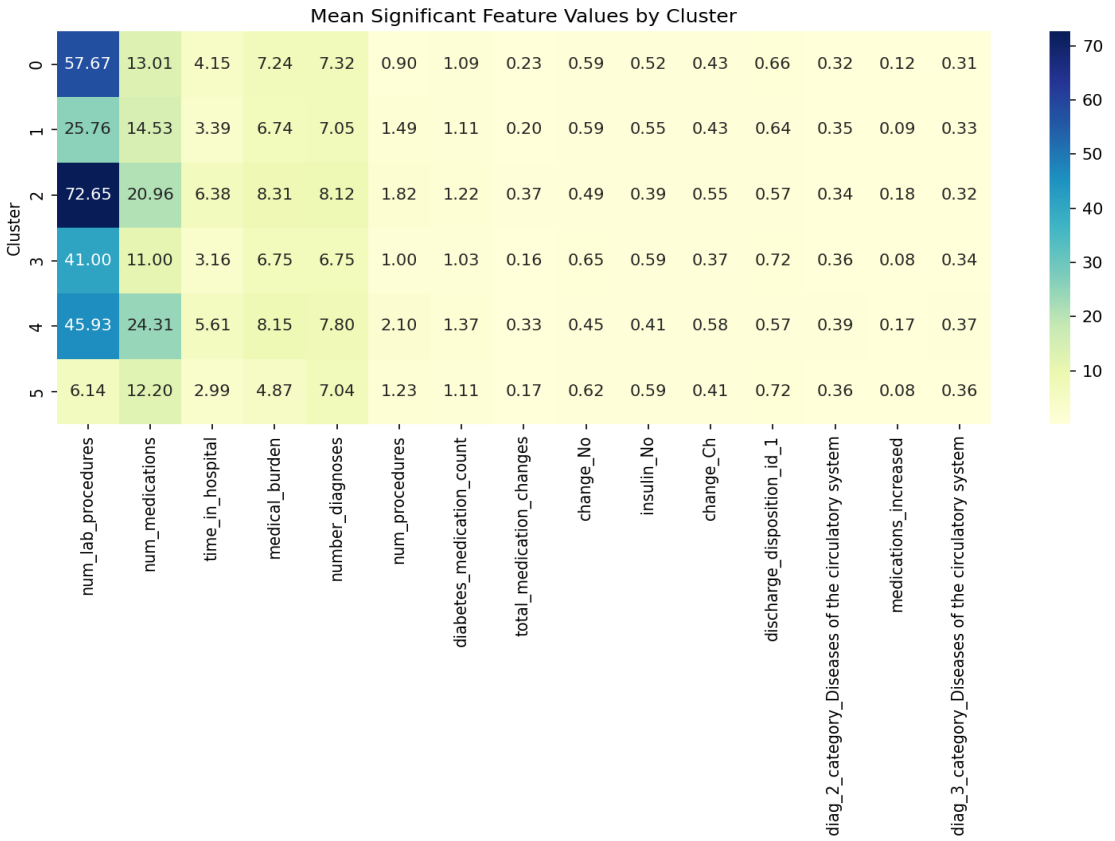
We take the absolute values of these means so that we can see differences going both ways. We also did this for the binary columns, for insights regarding the categorical variables.

```
Top 10 Significant Features by Mean Difference:
num_lab_procedures           35.518610
num_medications               8.993193
time_in_hospital              3.083076
medical_burden                1.734882
number_diagnoses              0.610642
num_procedures                0.498315
diabetes_medication_count     0.466087
total_medication_changes      0.297789
change_No                     0.278096
insulin_No                    0.270866
```

```
Most Significant Binary Features by Proportion Difference:
change_No                                          0.278096
insulin_No                                         0.270866
change_Ch                                          0.247091
discharge_disposition_id_1                         0.230636
diag_2_category_Diseases of the circulatory system 0.189896
diag_3_category_Diseases of the circulatory system 0.177797
diabetesMed_No                                     0.152047
admission_source_id_1                              0.148568
diag_1_category_Diseases of the circulatory system 0.147693
diabetesMed_Yes                                    0.124503
```

From this data, we can see that higher-risk patients tend to have a much higher amount of lab procedures and tend to be on a significantly higher amount of medications. Our **medical_burden** indicator also seems to be useful.

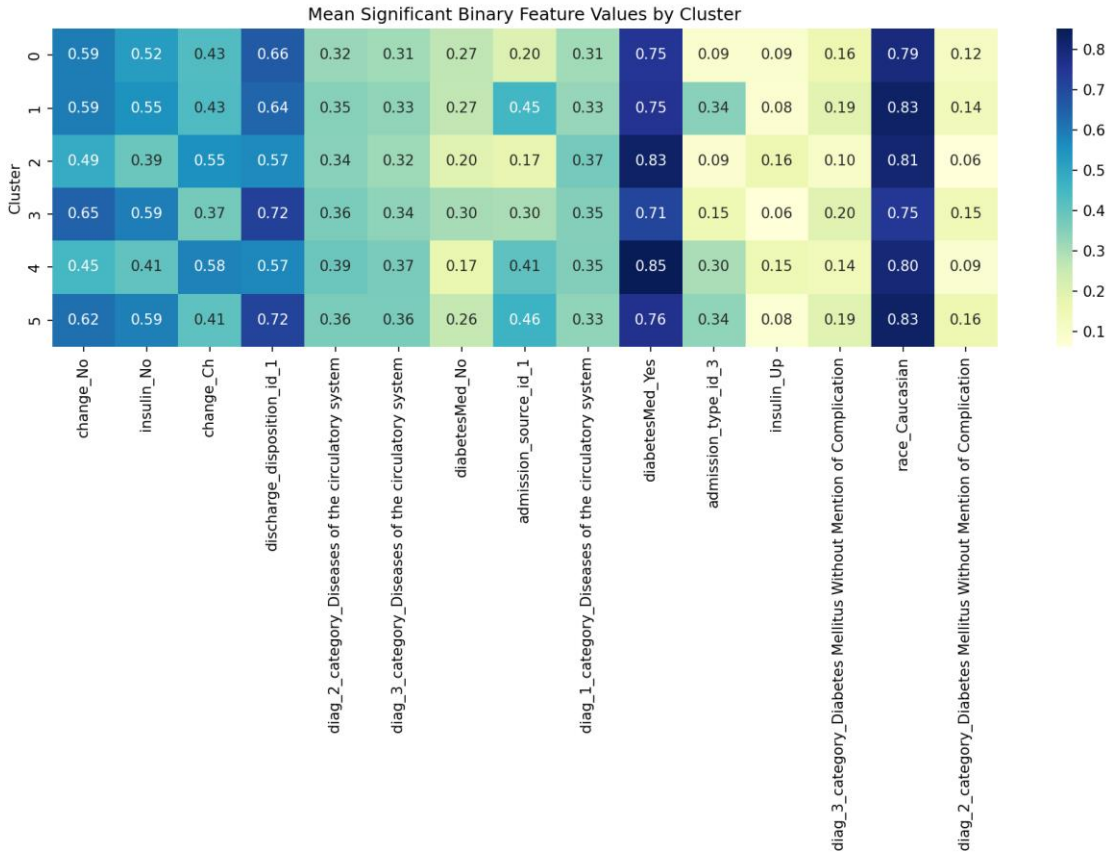
Now that we have the most significant features. Let's group the patients in each cluster and calculate the mean of each feature in said cluster.



Ignoring the categorical columns (the last 7) we now see in detail the characteristics of the patients in each cluster.
Cluster 1 and 5, with the highest average risk, also seem to have the significantly lowest mean **num_lab_procedures.** Indicating some potential mismanagement of patient conditions is causing increased readmission.

On the other hand, cluster 2, with a 0.366 average chance of readmission, has a higher amount of these two variables. Showing that these patients are very mindful/well disciplined when it comes to monitoring their health. Hence their risk is low.

We also do the same thing for the binary columns:



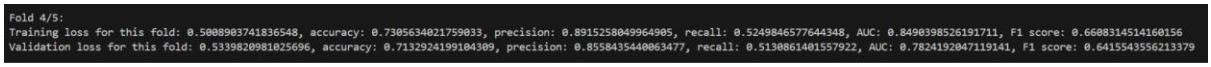
Just from the first column, we see that the high-risk clusters (odd-numbered, remember) have a higher chance of **no change in medication**. Again another sign of mismanagement. They also have a much **higher chance of NOT being on insulin**. Which is a key drug in diabetes treatment. On the contrary, we see that the even-numbered columns (specifically 2 and 4) have a **higher chance of being on insulin.** They also have a much higher chance of **change_Ch** again indicating strong and consistent management of their condition. Moreover, they tend to be on ANY diabetes medication more than the other clusters, indicated by the low values on **diabetesMed_No** and the higher values on **diabetesMed_Yes**. In cluster 1 and 5 we see that **admission_source_id_1** is also very prevalent. This ID corresponds to a referral by a physician, potentially showing that what was initially meant to be a routine check uncovered some negative truths, the patient has let something get out of hand.

So a conclusion we can draw out of these results is that **medical_burden** isn't the correct name for that variable as it isn't necessarily a negative thing. It should be called something more descriptive like **medical_service_utilisation_score** and with this general name we can incorporate more data: **time_in_hospital** for example. So far really we have only learnt that people who don't go to the doctor will probably get readmitted. What about those who do though? Well that seems to be cluster 3. Low and behold, their **medical_burden,** despite having lots of **num_lab_procedure** contributing to it, is **still** as low as the **medical_burden** of cluster 1! They're the most likely to not have any change to their medication, most likely to not be on insulin, joint-least likely to have any **medications_increased,** another feature we engineered. For these people, simply going to the doctor for a lab procedure often was NOT

enough. Further investigation is needed to see if that is cause of the patient or a weakness of the doctors.
It is also interesting looking at the differences in the **diag_3** and **diag_2_category_Diabetes Mellitus Without Mention of Complication**.

Across all high risk clusters there appears to be a significantly higher chance of not having any mention of complications. In the worst case scenario this suggests incomplete medical checks and incompetency causing the overlooking of key diagnosis signs. Further investigation is needed as this could be **a particularly urgent and expensive problem**. This is a symptom of an institutional problem that requires potential re-training and re-evaluation of various health services.

Could this be down to the quality of medical care?
We re-ran the model. This time including **payer_code** and found that yes, accuracy was improved slightly.



But as you can see the differences in test and validation scores indicate that overfitting has increased, likely due to the fact that to include **payer_code** in our data we must shave off 20,000 entries of data (33% of **filtered_data**). We would need access to **more data** to ensure our model is accurately predicting results. It seems like **payer_code** and **medical_specialty** may be key points of data.

Profiling

By analysing the clusters along with the model predictions, we can develop general profiles for patients at high and low risk of hospital readmission. Understanding the characteristics of these patient groups can provide valuable insights for healthcare providers.

So now let's build a profile for each:

High-Risk Patient Profile

The cluster analysis revealed certain clusters (1,3 and 5) with significantly higher average risk scores for hospital readmission. Patients in these high-risk clusters exhibit the following characteristics:

- A lower number of lab procedures were performed, indicating potential gaps in monitoring their condition.
- A lower number of medications were prescribed, suggesting simple to no treatment regimens at all.
- Lower medical burden scores reflect a lower overall utilisation of healthcare services.
- Lower likelihood of being prescribed insulin, a key medication for diabetes management.
- Higher likelihood of no changes in their diabetes medication regimen, which could signal a lack of proactive adjustments to their treatment plan.
- Higher likelihood of being diagnosed with diabetes without mention of complications.

The high-risk patient profile suggests the individual who may benefit from closer monitoring, medication reviews, and more frequent follow-ups to ensure that their diabetes is being managed effectively and to address any potential issues before they can escalate.

Low-Risk Patient Profile
On the other hand, cluster 2 exhibits the lowest average risk score for hospital readmission. Patients in this low-risk cluster generally displayed the following traits:

- A higher number of lab procedures were performed, indicating more consistent monitoring of their condition.
- Higher medical burden scores suggest a heavier overall health burden and more frequent use of healthcare services.
- Higher likelihood of being prescribed insulin and other diabetes medication, indicating proactive diabetes management.
- Higher likelihood of changes in their diabetes medication regimen, suggesting that their treatment is being actively adjusted based on their health status.
- More likely to have medication increases or decreases, suggesting their condition is being closely monitored and managed.

This low-risk patient profile represents individuals who are actively engaged in managing their diabetes, adhering to recommended monitoring and treatment plans, and making necessary adjustments to their care regimen as needed.

By understanding these distinct patient profiles, healthcare providers can tailor their approach and allocate resources more effectively. High-risk patients may require more intensive support and interventions, while low-risk patients may benefit from continued encouragement and reinforcement of their existing management strategies.

Conclusion

In this coursework, we have embarked on a comprehensive analysis of the provided dataset from 130 US hospitals focusing on diabetic patient readmission. The objective was to develop a predictive model to identify patients at risk of readmission and to propose meaningful decisions or actions that we gained through insights from our improved classification model as well as the obtained K-means clusters from the application.

The initial **Random Forest** Classifier model performed exceptionally well on the training data with high accuracy, precision, and recall scores. However, it suffered from **severe overfitting** issues, failing to generalise effectively to the validation data where the performance metrics dropped significantly. This was evident from the high number of false positives and false negatives in the confusion matrix. To improve the model's generalisability, strategies like hyperparameter tuning, pruning, enhancing cross-validation, and additional feature engineering can be explored to better capture the complex relationships in the data and reduce overfitting tendencies.

Our improved **neural network** model has demonstrated promising performance, with an accuracy average of 70%, a precision of 80%, and an AUC of 80%, indicating its **ability to effectively discriminate between high-risk and low-risk patients**. However, it is essential to strike a balance between precision and recall, ensuring that potential high-risk patients are not overlooked, as missing such cases could have severe consequences. However it seems like the model knows exactly what it wants when doing it's predictions.

The cluster analysis revealed distinct patient profiles, with clusters 1, 3, and 5 exhibiting significantly elevated risks of readmission. These clusters are characterised by **lower medical utilisation**, higher medication counts, and a higher likelihood of no medication changes, suggesting potential mismanagement or inconsistent monitoring of their conditions. Furthermore, we suggest implementing standardised protocols and decision support tools that incorporate our model's predictions and cluster insights. This could assist healthcare professionals in identifying high-risk patients early and initiating appropriate preventive measures to reduce the likelihood of readmission.

The final analysis suggests that the term "medical_burden" may not accurately reflect the variable's nature, proposing "medical_service_utilisation_score" as a more descriptive alternative. Furthermore, the findings reveal discrepancies in readmission risk among patients based on their utilisation of medical services, highlighting potential areas for further investigation regarding patient care and institutional practices.

While our analysis has yielded valuable insights, it is important to acknowledge the limitations of our study and the need for further research. Incorporating additional data sources, such as socioeconomic factors and patient-reported outcomes, could enhance the model's predictive power and provide a more comprehensive understanding of the underlying drivers of readmission. A crucial point of information would be the patient's BMI, there is a feeling that this data would drastically improve the performance of any model.

Ultimately, our findings underscore the importance of a data-driven approach to addressing the challenges of diabetes management and hospital readmissions. By leveraging the power of machine learning and data analysis, combined with the expertise of healthcare professionals, we can pave the way toward more efficient resource allocation, improved patient outcomes, and a more sustainable healthcare system.

References

[1] Validating the adapted Diabetes Complications Severity Index in claims data

https://pubmed.ncbi.nlm.nih.gov/23198714/