OPTIMIZING PATIENT CARE AND REDUCING READMISSIONS AT CAREPLUS HOSPITAL THROUGH PREDICTIVE ANALYTICS

AI & ML
ASSIGNMENT

Name: Khushi Malik
Roll no. 15A
Submitted to: Dr. Nitin Mehr

Problem Statement

HealthFirst Multispeciality Hospital is concerned about the increasing number of diabetic patients returning for readmission within a short time after discharge. The hospital aims to identify patterns and predictors of patient readmission to implement preventive measures, improve patient care, and optimize resource allocation.

Objective and Rationale

The objective of this analysis is to predict whether a diabetic patient is likely to be readmitted to the hospital after treatment. Early prediction will help the hospital management to take proactive actions such as follow-up consultations, personalized care, and patient education programs, thereby reducing operational costs and improving patient outcomes.

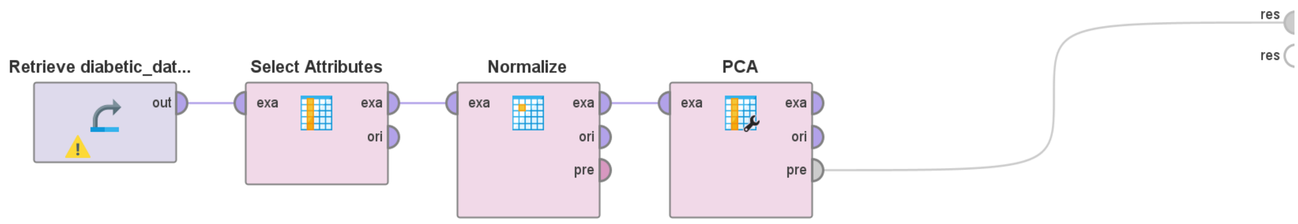
Methodology Adopted

1. **Data Collection:** A dataset containing information on diabetic patients' hospital visits, treatments, demographics, and outcomes was procured.
2. **Data Understanding and Cleaning:**
    - Cleaning and formatting were performed in Microsoft Excel.
    - The "Find and Replace" function was used to reformat categories.
    - Excel formulas were used to compute average values such as age ranges.
    - Rows with missing values were removed entirely.
3. **Data Models**
    - **Principal Component Analysis (PCA)** is used for dimensionality reduction, where the original dataset is transformed into a smaller set of uncorrelated variables, called principal components, which capture the most variance in the data. This helps in simplifying the model and reducing computational cost without losing critical information.
    - **Random Forest** is an ensemble learning technique that builds multiple decision trees and combines their predictions. It enhances accuracy by reducing overfitting and variance, especially in complex datasets, making it a robust choice for classification tasks.
    - **K-Means Clustering** is an unsupervised learning algorithm used to partition the dataset into K clusters based on feature similarities. It assigns each data point to the cluster with the nearest mean, helping in grouping similar data points and uncovering hidden patterns.
    - **Logistic Regression** is a statistical model used for binary classification, where the relationship between the independent variables and the dependent variable is modeled using a logistic function. It's effective for predicting the probability of class membership, especially in cases with linear relationships between predictors and the outcome.

Data Set and Source

- **Dataset Used:** Diabetic patient hospital visit records.
- **Source:** UC Irvine Machine Learning Repository (customized and downloaded independently for academic analysis).

Principle Component Analysis (PCA)



Component	Standard Deviation	Proportion of Variance	Cumulative Variance
PC 1	1.540	0.264	0.264
PC 2	1.235	0.169	0.433
PC 3	1.037	0.120	0.552
PC 4	0.986	0.108	0.660
PC 5	0.950	0.100	0.761
PC 6	0.824	0.075	0.836
PC 7	0.771	0.066	0.902
PC 8	0.711	0.056	0.958
PC 9	0.612	0.042	1.000

The first five components (PC1 to PC5) still explain 76.1% of the total variance, justifying their selection as the top five components.

Attribute	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9
Age median	0.272	-0.186	0.630	0.405	-0.006	-0.047	0.474	0.299	0.120
time_in_hospital(days)	0.468	0.008	-0.241	0.081	-0.043	0.599	0.323	-0.273	-0.420
num_lab_procedures	0.310	0.076	-0.574	0.432	-0.141	-0.400	-0.041	0.434	-0.111
num_procedures	0.368	-0.102	0.141	-0.682	0.171	-0.264	0.014	0.330	-0.401
num_medications	0.530	0.053	-0.156	-0.258	0.044	0.112	-0.027	-0.022	0.781
number_outpatient	-0.012	0.620	0.037	-0.162	-0.325	-0.387	0.489	-0.305	0.002
number_emergency	-0.034	0.398	-0.054	0.154	0.893	-0.027	0.123	0.004	-0.006
number_inpatient	0.022	0.628	0.246	0.005	-0.209	0.413	-0.337	0.462	-0.054
number_diagnoses	0.439	0.090	0.326	0.248	0.022	-0.275	-0.548	-0.482	-0.151

PC1: num_medications (26.4% Variance)

- **Loading**: 0.530 (positive, moderate).
- **Interpretation**: PC1 is driven by the number of medications, reflecting treatment complexity. Higher values indicate patients with multiple prescriptions, often linked to chronic conditions (e.g., diabetes, ICD-9: 250) or polypharmacy risks.

PC2: number_inpatient (16.9% Variance)

- **Loading**: 0.628 (positive, high).
- **Interpretation**: PC2 is dominated by the number of inpatient visits, indicating a history of severe or frequent hospitalizations (e.g., heart failure, ICD-9: 428).

PC3: age_median (12.0% Variance)

- **Loading**: 0.630 (positive, high).
- **Interpretation**: PC3 is strongly influenced by the median age, suggesting it represents the age-related risk profile of patients. Higher values indicate older patients, who are more vulnerable to complications.

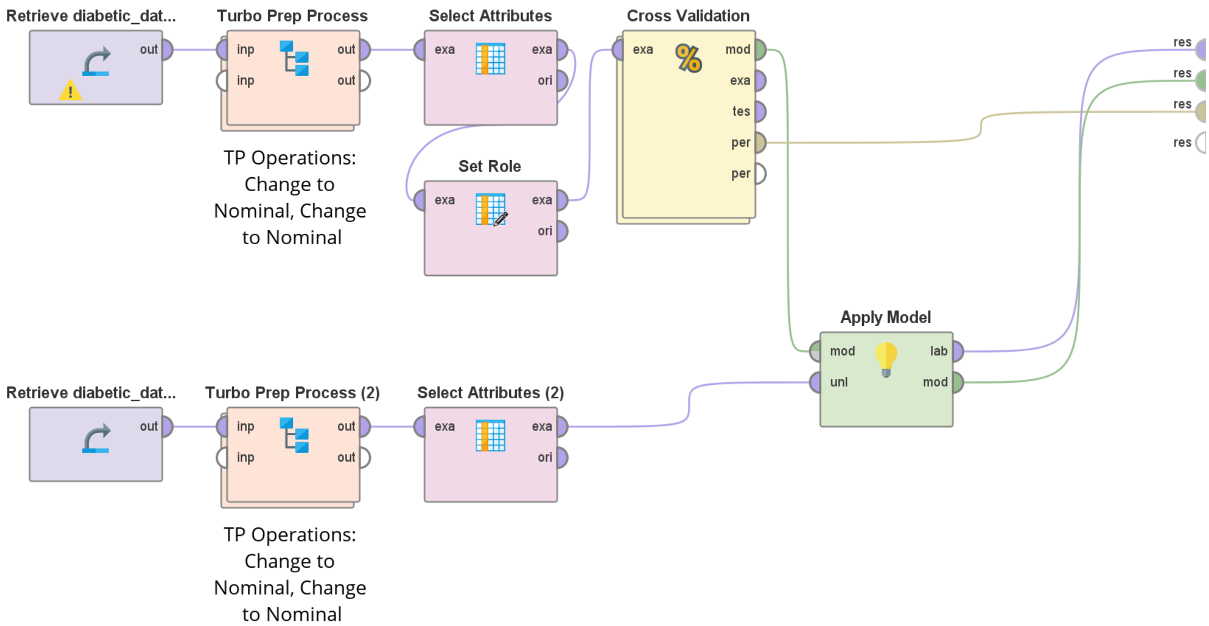
PC4: num_procedures (10.8% Variance)

- **Loading**: -0.682 (negative, high).
- **Interpretation**: PC4 is inversely related to the number of medical procedures, indicating that fewer procedures align with this component. This may reflect less invasive treatment or stable conditions.

PC5: number_emergency (10.0% Variance)

- **Loading**: 0.893 (positive, very high).
- **Interpretation**: PC5 is heavily weighted by the number of emergency visits, representing acute care utilization and unstable health states.

Random Forest



accuracy: 48.25% +/- 2.65% (micro average: 48.25%)

	true NO	true >30	true <30	class precision
pred. NO	417	359	72	49.17%
pred. >30	51	49	16	42.24%
pred. <30	3	1	2	33.33%
class recall	88.54%	11.98%	2.22%	

Accuracy: 48.25% ± 2.65%

This means the model predicts the correct readmission status in about **48 out of 100 cases**, which is just slightly better than random guessing for a 3-class problem.

Interpretation

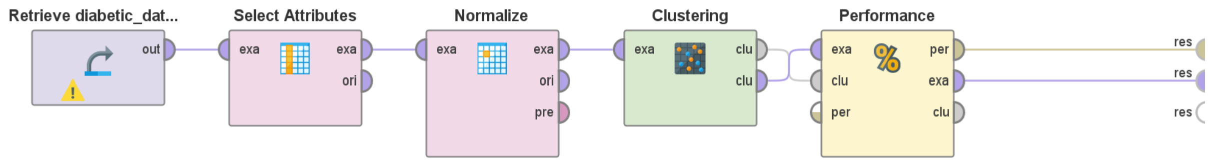
The Random Forest confusion matrix shows classification performance for predicting readmission timing: "NO" (not readmitted), ">30" (readmitted after 30 days), and "<30" (readmitted before 30 days). Accuracy is 48.25% (±2.65%), indicating moderate performance.

- **True NO (88.54% recall)**: 417 correctly predicted as not readmitted, but 359 misclassified as ">30" and 51 as ">30", showing over-prediction of readmission.
- **True >30 (11.98% recall)**: 49 correctly predicted, but 359 misclassified as "NO", reflecting poor sensitivity for late readmissions.
- **True <30 (2.22% recall)**: 16 correctly predicted, but 72 misclassified as "NO", indicating low detection of early readmissions.
- **Precision**: 49.17% ("NO"), 42.24% (">30"), 33.33% ("<30"), suggesting moderate reliability in predictions. The class imbalance (88.54% "NO" vs. 11.98% ">30" and 2.22% "<30") likely biases the model toward "NO".

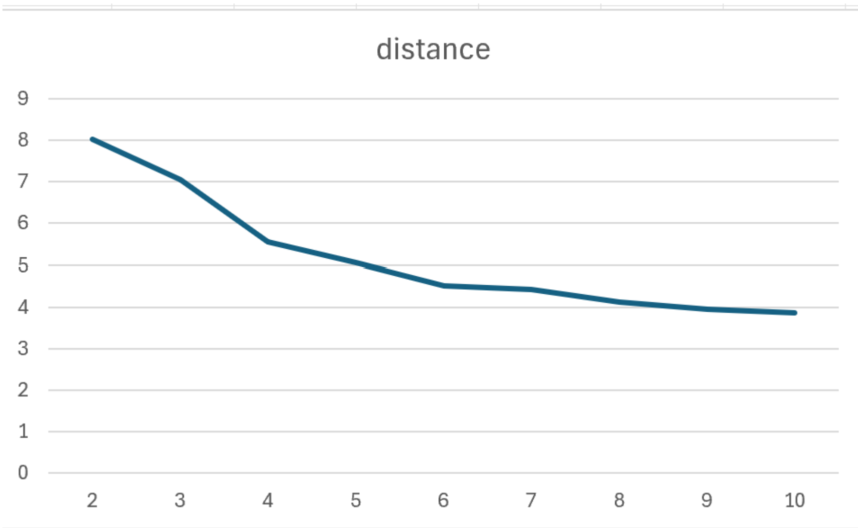
Inferences

- The model excels at identifying non-readmitted patients but struggles with readmission timing, especially "<30" days, due to low recall
- Model needs better sensitivity for minority classes: Readmitted patients (especially within 30 days) are under-represented in predictions, likely due to class imbalance in the dataset.
- Improvement needed before real-world deployment: Without better recall for readmitted cases, relying solely on this model could lead to missed interventions and poor patient outcomes.

K-Mean: Cluster Distance Performance



k	distance
2	8.023
3	7.042
4	5.546
5	5.062
6	4.519
7	4.431
8	4.113
9	3.948
10	3.855

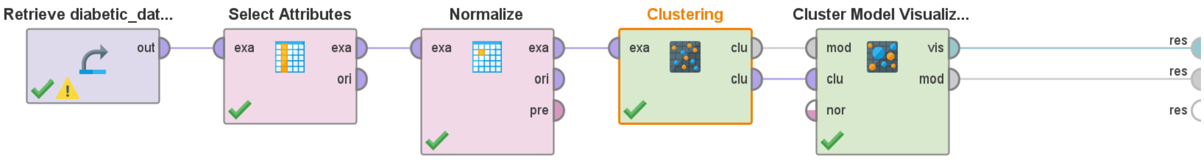


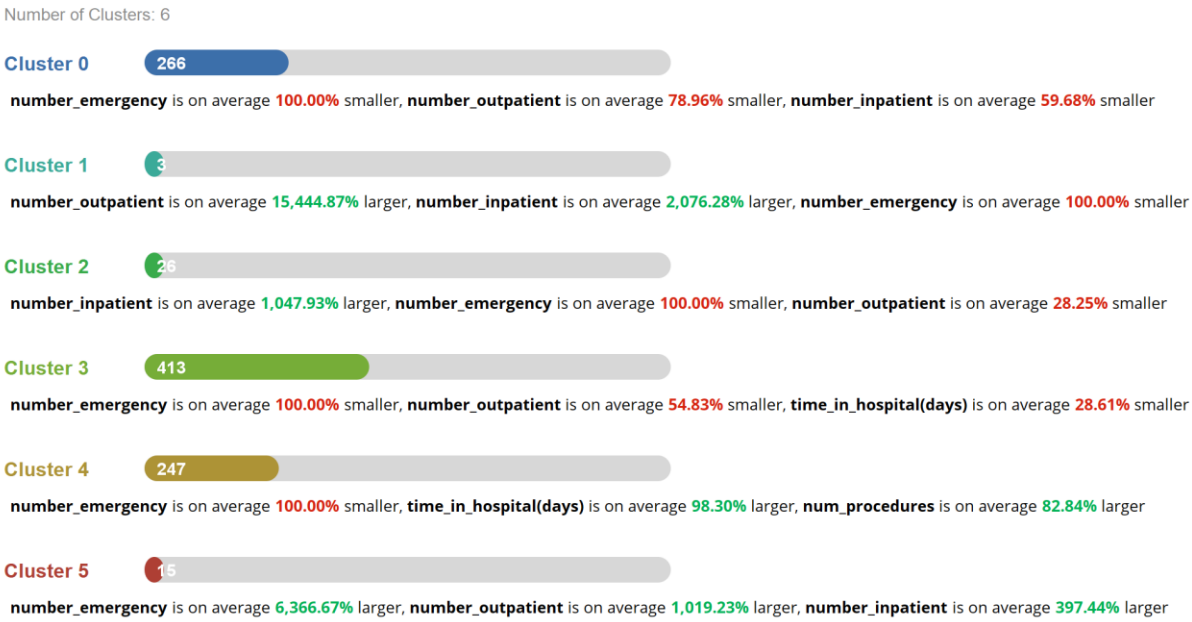
To identify meaningful groupings in the dataset, K-Means clustering was applied. The elbow method was used to determine the optimal number of clusters. The distance (inertia) values for different values of *k* were plotted, showing a steep decline until *k* = 6, after which the curve begins to flatten, indicating diminishing returns in variance reduction. This inflection point suggests that 6 is the optimal number of clusters.

Elbow Method Observations:

- The distortion score drops significantly from k = 2 to k = 6.
- Beyond k = 6, the decrease in inertia becomes minimal.
- Hence, k = 6 is chosen as the ideal number of clusters.

Cluster Interpretation





Cluster 0 (n = 266)

- **Emergency visits:** 100% lower (no emergency visits)
- **Outpatient visits:** 78.96% lower
- **Inpatient visits:** 59.68% lower

  Interpretation:
  These patients are relatively healthy, with significantly fewer hospital interactions across all types of visits. They might represent low-risk or well-managed chronic patients.

Cluster 1 (n = 3)

- **Outpatient visits:** 15,444.87% higher
- **Inpatient visits:** 2,076.28% higher
- **Emergency visits:** 100% lower (none)

  Interpretation:
  This small group represents **intensive-care or high-frequency patients**, visiting hospitals very frequently but avoiding emergency services. Likely under scheduled or long-term treatment.

Cluster 2 (n = 26)

- **Inpatient visits:** 1,047.93% higher
- **Emergency visits:** 100% lower
- **Outpatient visits:** 28.25% lower

  Interpretation:
  Patients in this cluster are primarily treated within hospitals (inpatients) but do not

frequently use emergency or outpatient services. Possibly long-term or post-operative inpatients.

Cluster 3 (n = 413)

- **Emergency visits:** 100% lower
- **Outpatient visits:** 54.83% lower
- **Time in hospital:** 28.61% shorter
  Interpretation:
  This largest cluster suggests **moderate-need patients**. They use fewer hospital resources, have shorter stays, and no emergency visits, indicating mild conditions or effective care.

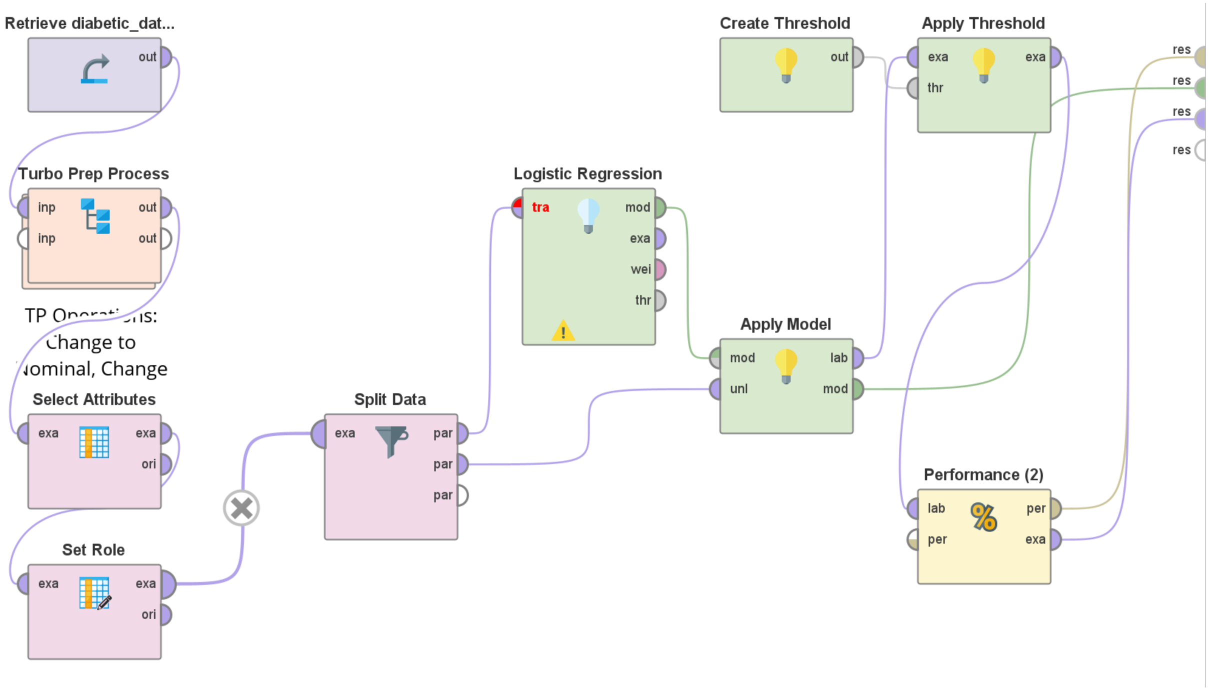
Cluster 4 (n = 247)

- **Emergency visits:** 100% lower
- **Time in hospital:** 98.30% higher
- **Procedures:** 82.84% higher
  Interpretation:
  Patients here undergo **long hospital stays and many procedures**, without emergency visits. Likely suffering from chronic or complex conditions requiring consistent monitoring and interventions.

Cluster 5 (n = 15)

- **Emergency visits:** 6,366.67% higher
- **Outpatient visits:** 1,019.23% higher
- **Inpatient visits:** 397.44% higher
  Interpretation:
  These are **critical patients** who require frequent attention across all types of hospital services, especially emergency. They likely have acute conditions or repeated complications.

Logistics Regression



accuracy: 0.00%

	true NO	true YES	true Low	true High	class precision
pred. NO	0	0	0	0	0.00%
pred. YES	0	0	0	0	0.00%
pred. Low	141	150	0	0	0.00%
pred. High	0	0	0	0	0.00%
class recall	0.00%	0.00%	0.00%	0.00%	

A Logistic Regression model was tested for CarePlus Hospital's readmission prediction, using features like age_median and number_emergency. The confusion matrix shows 0.00% accuracy, with all precision and recall at 0.00%, as 141 "NO" and 150 "YES" cases were misclassified as "Low," indicating complete failure to differentiate classes.

Interpretation

The poor performance likely stems from class imbalance, linear assumptions not fitting the data, or inadequate feature handling, contrasting with Random Forest's 48.25% accuracy.

Inferences

Logistic Regression is unsuitable due to its inability to capture non-linear patterns, suggesting a preference for ensemble methods like Random Forest.

Actionable Business Inferences for Strategic Decision Making.

Random Forest Insights

- The Random Forest model shows moderate accuracy, excelling at identifying non-readmitted patients but struggling with early and late readmission detection due to class imbalance.

Action: Enhance follow-up care for misclassified readmitted patients using telehealth, focusing on key risk factors, and apply balancing techniques to improve detection of high-risk cases.

PCA Insights

- The PCA analysis highlights treatment complexity, hospitalization history, age-related risk, treatment intensity, and acute care utilization as key drivers of readmission.

Actions:

- Optimize medication regimens for patients with complex treatments through pharmacy reviews.
- Provide coordinated care for patients with frequent hospitalizations.
- Offer tailored support for older patients to address age-related risks.
- Monitor patients with low procedure counts for adequate post-discharge care.
- Implement emergency triage for patients with high acute care use to prevent readmissions.

Integrated Strategic Insights

- Prioritize Random Forest and PCA over Logistic Regression for robust predictions, reallocating resources to refine these models.
- Focus interventions on high-risk patient groups to improve outcomes and reduce penalties.
- Use real-time risk scoring from PCA and Random Forest to continuously adjust strategies and sustain readmission reductions.

Suggestions to the Healthcare Industry (CarePlus Hospital)

1. **Enhance Proactive Patient Care Programs**: Invest in comprehensive post-discharge care initiatives, such as telehealth monitoring and home health services, to support patients with chronic conditions, reducing readmission rates and improving long-term health outcomes across the industry.
2. **Leverage Advanced Data Analytics**: Adopt cutting-edge data analytics and predictive tools to identify at-risk patients early, enabling personalized interventions and optimizing resource allocation, which can set a new standard for healthcare efficiency and patient satisfaction.
3. **Foster Collaborative Care Networks**: Build strong partnerships with community healthcare providers, social services, and insurers to create integrated care pathways, ensuring seamless transitions and addressing social determinants of health, thereby elevating overall patient care quality and reducing costs.

Excel Sheet

encounter_id	patient_nbr	race	gender	age	Age median	admission_type_id	discharge_disposition_id	time_in_hospital(days)	num_lab_procedures	num_procedures	num_medications	number_outpatient	number_emergency	number_inpatient	number_diagnoses	max_glu_serum	A1Cresult	metformin	repaglie
2278392	8222157	Caucasian	Female	[0-10)	5	6	25	1	41	0	1	0	0	0	1	None	None	No	No
149190	55629189	Caucasian	Female	[10-20)	15	1	1	3	59	0	18	0	0	0	9	None	None	No	No
64410	86047875	AfricanAmerican	Female	[20-30)	25	1	1	2	11	5	13	2	0	1	6	None	None	No	No
500364	82442376	Caucasian	Male	[30-40)	35	1	1	2	44	1	16	0	0	0	7	None	None	No	No
16680	42519267	Caucasian	Male	[40-50)	45	1	1	1	51	0	8	0	0	0	5	None	None	No	No
35754	82637451	Caucasian	Male	[50-60)	55	2	1	3	31	6	16	0	0	0	9	None	None	No	No
55842	84259809	Caucasian	Male	[60-70)	65	3	1	4	70	1	21	0	0	0	7	None	None	Steady	No
63768	114882984	Caucasian	Male	[70-80)	75	1	1	5	73	0	12	0	0	0	8	None	None	No	No
12522	48330783	Caucasian	Female	[80-90)	85	2	1	13	68	2	28	0	0	0	8	None	None	No	No
15738	63555939	Caucasian	Female	[90-100)	95	3	3	12	33	3	18	0	0	0	8	None	None	No	No
28236	89869032	AfricanAmerican	Female	[40-50)	45	1	1	9	47	2	17	0	0	0	9	None	None	No	No
36900	77391171	AfricanAmerican	Male	[60-70)	65	2	1	7	62	0	11	0	0	0	7	None	None	No	No
40926	85504905	Caucasian	Female	[40-50)	45	1	3	7	60	0	15	0	1	0	8	None	None	Steady	Up
42570	77586282	Caucasian	Male	[80-90)	85	1	6	10	55	1	31	0	0	0	8	None	None	No	No
62256	49726791	AfricanAmerican	Female	[60-70)	65	3	1	1	49	5	2	0	0	0	8	None	None	No	No
73578	86328819	AfricanAmerican	Male	[60-70)	65	1	3	12	75	5	13	0	0	0	9	None	None	No	No
77076	92519352	AfricanAmerican	Male	[50-60)	55	1	1	4	45	4	17	0	0	0	8	None	None	No	No
84222	108662661	Caucasian	Female	[50-60)	55	1	1	3	29	0	11	0	0	0	3	None	None	No	No
89682	107389323	AfricanAmerican	Male	[70-80)	75	1	1	5	35	5	23	0	0	0	9	None	None	No	No
182796	63000108	AfricanAmerican	Female	[70-80)	75	2	1	2	47	0	12	0	0	0	8	None	None	No	No
183930	107400762	Caucasian	Female	[80-90)	85	2	6	11	42	2	19	0	0	0	8	None	None	No	No
216156	62718876	AfricanAmerican	Female	[70-80)	75	3	1	3	19	4	18	0	0	0	6	None	None	No	No
221634	21861756	Other	Female	[50-60)	55	1	1	1	33	0	7	0	0	0	3	None	None	Steady	No
236316	40523301	Caucasian	Male	[80-90)	85	1	3	6	64	3	18	0	0	0	7	None	>7	Steady	No

chlorpropamide	glimepiride	acetohexamide	glipizide	glyburide	tolbutamide	pioglitazone	rosiglitazone	acarbose	miglitol	troglitazone	tolazamide	examide	citoglipton	insulin	glyburide-metformin	glipizide-metformin	glimepiride-pioglitazone	metformin-rosiglitazone	metformin-pioglitazone	change in medication	diabetesMed
No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
No	No	No	No	No	No	No	No	No	No	No	No	No	No	Up	No	No	No	No	No	Yes	Yes
No	No	No	Steady	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Yes
No	No	No	No	No	No	No	No	No	No	No	No	No	No	Up	No	No	No	No	No	Yes	Yes
No	No	No	Steady	No	No	No	No	No	No	No	No	No	No	Steady	No	No	No	No	No	Yes	Yes
No	No	No	No	No	No	No	No	No	No	No	No	No	No	Steady	No	No	No	No	No	No	Yes
No	Steady	No	No	No	No	No	No	No	No	No	No	No	No	Steady	No	No	No	No	No	Yes	Yes
No	No	No	No	Steady	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Yes
No	No	No	Steady	No	No	No	No	No	No	No	No	No	No	Steady	No	No	No	No	No	Yes	Yes
No	No	No	No	No	No	No	Steady	No	No	No	No	No	No	Steady	No	No	No	No	No	Yes	Yes
No	No	No	No	No	No	No	No	No	No	No	No	No	No	Steady	No	No	No	No	No	No	Yes
No	No	No	No	Up	No	No	No	No	No	No	No	No	No	Steady	No	No	No	No	No	Yes	Yes
No	No	No	No	No	No	No	No	No	No	No	No	No	No	Down	No	No	No	No	No	Yes	Yes
No	No	No	No	No	No	No	No	No	No	No	No	No	No	Steady	No	No	No	No	No	No	Yes
No	No	No	No	No	No	No	No	No	No	No	No	No	No	Steady	No	No	No	No	No	No	Yes
No	No	No	No	No	No	No	No	No	No	No	No	No	No	Up	No	No	No	No	No	Yes	Yes
No	No	No	Steady	No	No	No	No	No	No	No	No	No	No	Steady	No	No	No	No	No	Yes	Yes
No	No	No	No	Steady	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Yes
No	No	No	No	No	No	No	No	No	No	No	No	No	No	Steady	No	No	No	No	No	No	Yes
No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
No	No	No	Steady	No	No	No	No	No	No	No	No	No	No	Steady	No	No	No	No	No	Yes	Yes
No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Yes
No	No	No	No	Steady	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Yes	Yes
No	No	No	No	No	No	No	No	No	No	No	No	No	No	Steady	No	No	No	No	No	No	Yes