Segmenting and Predicting Insulin Pump Adoption through Unsupervised and Supervised Learning Models

Authors: V. Khandelwal, R. Nair, S. Sarda and S. Sawant

Abstract

This study investigates the determinants of adoption of insulin pumps among patients managing chronic health conditions with specific focus on diabetes. We have used synthetically generated data that replicates actual healthcare settings, the examination explores the associations between the socio-demographic patient characteristics, socio-economic status, clinical predictors, and adoption of insulin pumps. This study uses a comprehensive analytical approach with Principal Component Analysis (PCA), k-means, hierarchical clustering, and other classification models such as logistic regression, decision tree analysis, neural networks, and support vector machines. PCA revealed monthly income, count of chronic conditions, and awareness of devices to be significant factors of patient heterogeneity. Clustering algorithms accurately segmented patients into different profiles with varying probabilities of adoption, and they pointed to the role of socio-economic and health determinants in influencing purchasing behavior. Logistic regression models and decision tree models confirmed doctor recommendation, income level, and HbA1c as the most statistically significant predictors of adoption. Neural networks demonstrated superior accuracy, by achieving an AUC of 0.8678 on the validation set, and lift curve analysis also suggested that focusing on the top 20% of predicted adopters could more than double the efficiency of intervention strategies. This study emphasizes the efficacy of predictive analytics as a tool for facilitating inclusive health interventions, optimized marketing with improved targeting of medical device marketers, and data-driven policy making. By the identification of high-probability adopter segments and key influence factors, this study contributes to the overall aim of enhancing chronic disease management by leveraging effective and cost-effective technological interventions.

Keywords:Insulin pump adoption, chronic disease, logistic regression, decision tree, clustering, PCA, neural networks, predictive analytics, HbA1c, healthcare equity

1. Introduction

In this section, we introduce the business intelligence problem, the motivation behind the study, and the potential impact of the analysis on the economy or society. We summarize related work by reviewing and citing at least five relevant studies. We clearly state the study objectives, focus areas, and limitations, and discuss the dataset challenges and the models most suitable for addressing the business problem.

1.1 Business Intelligence Problem

Access to advanced medical devices, such as insulin pumps, plays a crucial role in managing chronic health conditions like diabetes. However, widespread adoption of such devices remains a significant challenge (Tanenbaum, 2022) due to multiple factors including patient health profiles, insurance coverage limitations, device awareness, and socio-economic conditions. This study aims to predict and understand the likelihood of insulin pump adoption based on a comprehensive set of patient-related variables such as age, income, insurance coverage percentage, chronic condition count, and medical device awareness.

The motivation behind this study is deeply rooted in the broader societal need to enhance chronic disease management through technological intervention (Madsen, 2023). Chronic diseases, particularly diabetes, continue to impose a heavy economic burden on healthcare systems worldwide. Early and effective adoption of insulin pumps can improve patient outcomes, reduce emergency hospital visits, and decrease long-term healthcare costs. Yet disparities in access and awareness persist. By identifying the patterns and predictors of device adoption, healthcare providers, insurers, and policymakers can better design targeted interventions to close these gaps.

From an economic perspective, improved predictive modeling of device adoption can help pharmaceutical and medical device companies optimize their marketing and outreach efforts. Instead of employing blanket marketing strategies, they can focus resources on populations with higher likelihoods of adoption, thus reducing marketing spend while maximizing return on investment. Insurers, on the other hand, can use these insights to create more inclusive coverage plans that anticipate the needs of high-risk populations, potentially leading to more efficient resource allocation and lower insurance claim rates.

On a societal level, the analysis offers a pathway to greater healthcare equity. Understanding which groups are less likely to adopt critical medical devices, whether due to lower insurance coverage, reduced awareness, or complex chronic conditions, allows for the development of community-level programs aimed at education, support, and subsidy. In turn, this can reduce healthcare disparities, improve population health outcomes, and foster a more resilient healthcare ecosystem.

This study bridges the gap between data analysis and actionable healthcare strategy. It highlights how this study, when properly leveraged, can deliver not only economic efficiencies

but also meaningful social benefits. The findings of this study can inform smarter healthcare policies, promote technological access, and ultimately contribute to a healthier society.

1.2 Background and Related Work

Biebel et al. evaluated insulin pump therapy (IPT) versus multiple daily insulin injections (MDI) among Danish adults with type 1 diabetes, finding that IPT significantly reduced HbA1c levels and variability but slightly increased the risk of hospitalised diabetic ketoacidosis (DKA). There were 26,687 adults included with more than 243,000 person-years of data. IPT was linked with decreased mean HbA1c values and diminished risk of diabetic ketoacidosis (DKA) and severe hypoglycemia (SH) versus multiple daily insulin injections (MDI), as speculated by earlier randomized controlled trials (RCTs) and observational data. IPT's influence was strongly differential across demographic variables, such as age, sex, baseline HbA1c, and continuous glucose monitoring (CGM) use. The study used a quasi-experimental approach under treatment-staggered difference-in-differences for the estimation of average treatment effects while controlling for a broad set of demographic and clinical covariates to mitigate confounding. Although the study showed positive results, it observed that the effects of treatment may be vulnerable to high levels of patient-level heterogeneity, which is indicative of diabetes care complexity.

Wood et al. reviewed the current barriers to adopting and sustaining the use of diabetes technologies, including insulin pumps, continuous glucose monitors, and automated insulin delivery systems. Despite significant advancements in these devices, adoption was hindered by system, provider, and individual-level challenges. Effective facilitators for sustained use included consistent insurance coverage, support for healthcare providers, structured education, and accessible resources for users. As the benefits of newer technologies become clearer, efforts are underway to create evidence-based programs to address these barriers and improve long-term device utilization..

Khunti et al. aimed to identify predictors of insulin pump initiation among individuals with type 2 diabetes (T2D) using bolus insulin in the United States. Using data from the IBM MarketScan Commercial databases (2015-2020), a retrospective, nested case-control design matched 726 pump initiators with 2,904 non-pump initiators based on time at-risk. Key predictors of pump initiation included continuous glucose monitor (CGM) or flash glucose monitor (FGM) use, endocrinologist visits, acute metabolic complications, and diabetes-related outpatient visits, while older age and lack of insulin use reduced the likelihood of pump initiation. Sensitivity and post hoc analyses confirmed these findings, highlighting the significance of clinical variables, healthcare utilization, and comorbidities in predicting pump uptake. These insights can guide personalized diabetes management strategies and help identify patients who may benefit from pump therapy.

Forbes et al. (2023) investigated the relationship between glycaemic control (mean HbA1c) and variability (HbA1c fluctuations) and mortality in older individuals with diabetes. Using data from 54,803 people aged 70 and older, the study found a J-shaped relationship between glycaemic control and mortality, with increased risk at both low (less than 6%) and high (greater than 8%)

HbA1c levels. Furthermore, increased glycaemic variability was significantly associated with higher mortality, with individuals experiencing greater fluctuations in HbA1c having more than double the risk of death compared to those with stable levels. The findings suggested that maintaining a stable, moderate glycaemic level is crucial for reducing mortality risk in older people with diabetes.

Rytter et al. (2023) evaluated the impact of a newly developed, adaptive education program for insulin pump users transitioning to different devices, comparing it with a standard technical training program. The goal was to assess how this education affects glycemic outcomes and psychosocial self-efficacy in adults with Type 1 diabetes. The new program (NP) involved a comprehensive 8-hour curriculum with individual and group sessions focused on technical training, real-life scenarios, and data analysis. In contrast, the usual care (UC) program consisted of a single 3-hour group session. Results showed that participants in the NP group had a significant improvement in time in range (TIR) and HbA1c, as well as enhanced psychosocial self-efficacy, compared to the UC group. These findings suggested that a more personalized, extended education program leads to better glycemic control and improved self-efficacy in insulin pump users. The study underscored the importance of continuous diabetes education, particularly during transitions to new technologies.

1.3 Objectives, Focus, and Limitations

Objective

The primary objective of this study is to develop a predictive model that identifies the likelihood of insulin pump adoption among patients managing chronic health conditions, specifically diabetes. By analyzing a comprehensive set of patient-related variables, including age, income level, insurance coverage percentage, chronic condition count, and medical device awareness , the study aims to uncover key patterns and drivers influencing device adoption. Ultimately, the goal is to provide actionable insights that can guide healthcare providers, insurers, policymakers, and device manufacturers in creating strategies to improve access, promote adoption, and enhance chronic disease management outcomes.

Focus

This study focuses on the intersection of healthcare and business intelligence, utilizing data analytics to address a critical public health challenge: improving the adoption rates of advanced medical technologies like insulin pumps. Special emphasis is placed on identifying the socio-economic and health-related barriers that inhibit adoption, such as limited insurance coverage, lower income levels, or lack of awareness about medical devices.

Efforts to enhance the adoption of medical devices can be approached from multiple angles. This includes designing targeted educational programs and community interventions to raise awareness about these devices, which can empower individuals to make informed healthcare choices. Additionally, insurers can be encouraged to develop more inclusive and equitable

coverage policies, ensuring broader access to these technologies. For medical device companies, optimizing marketing strategies to focus on populations with a higher propensity for adoption can significantly improve reach and impact. Together, these strategies can promote greater healthcare equity by addressing disparities in access to advanced treatment technologies, ultimately leading to improved health outcomes.

The study also aims to demonstrate how data-driven insights can lead to smarter healthcare policies and a more resilient healthcare ecosystem.

Limitations

While the study aspires to deliver meaningful and actionable insights, it is subject to several limitations.

The model's predictive accuracy also relies heavily on data completeness and quality since partial or biased data (e.g., missing data on device awareness or chronic conditions) can degrade its effectiveness. Moreover, although synthetic data based on GenAI is carefully constructed, it can also lack the richness and diversity of real-world healthcare data. Generalizability is likewise a concern, as findings based on individual locations or systems cannot automatically be assumed to apply to more heterogeneous, larger populations. Simplifying the outcome variable (Pump_Sold) into binary format is likewise likely to fail to include partial interest as well as other purchase impediments, and unobservables such as patient tastes, cultural and provider influence could have an influence on adoption without being able to be included within accessible data sets. Moreover, medical device coverage is diverse to a large extent, making standardization between areas impossible and the dynamic changing nature of medical technology ensures predictive trends change with time necessitates regular updating in order to sustain model validity.

Challenges

While meaningful correlations were introduced, i.e., between income and insurance or age and chronic diseases, to ensure that these were still realistic but not higher or lower than real-world values required diligent cross-checking. Certain clusters, e.g., Cluster 5, have very small sample sizes, which, however statistically intriguing, might restrict their practical application to larger marketing campaigns. In addition, some variables like Monthly Income and Device Awareness Score have skewed distributions that will require additional caution in model building and visualization. Lastly, the comparatively small number of independent variables may cause overfitting in more sophisticated models like the decision tree if not pruned or cross-validated.

2. Data Exploration

In order to successfully tackle our core issue of forecasting insulin pump adoption, it is important to understand the dataset upon which this task rests. The data, artificially extracted by means of ChatGPT (Gen AI), is representative of a combination of patient variables such as demographics, income, medical status, insurance, and knowledge of devices. These are crucial

in determining leading drivers of adoption and classification of patient populations for purposes of targeting for individual contact.

Analyzing distributions of key characteristics, e.g., age, income, HbA1c values, and insurance status, influences the diversity of the patient population, selecting influential subgroups, outliers, and barriers to adoption. Scatterplot matrices identify key inter-variable associations that influence feature selection and model form. These early steps guarantee our forecast models are based on a deep understanding of patient behavior, which in the long run will improve our capacity to recognize high-potential segments, maximize marketing efforts, and expand healthcare accessibility for the underserved.

2.1 Dataset

This dataset was synthetically generated using ChatGPT (Gen AI) which helped us simulate realistic patient profiles and healthcare scenarios. After researching the variables impacting insulin adoption (Hankosky, 2023), we developed variables related to patient demography (Forbes, 2018), health conditions, insurance coverage and device awareness (Ryter, 2024).

Variable Name	**Description**	**Data Type**
Patient_ID	Unique identifier for each patient	Categorical
Patient_Age	Age of the patient in years	Numeric
Gender	Gender of the patient (e.g., Male, Female)	Categorical
Monthly_Income	Monthly income of the patient in USD	Numeric
Insurance_Coverage_Percentage	Percentage of medical expenses covered by insurance	Numeric
Chronic_Condition_Count	Number of chronic conditions diagnosed in the patient	Numeric
Device_Awareness_Score	Score indicating patient's awareness of medical devices (1 to 5 scale)	Numeric
HbA1c_Level	Hemoglobin A1c level indicating average blood sugar over past 3 months (%)	Numeric

Pump_Sold (Target variable)	Indicates if an insulin pump was sold to the patient (1 = Yes, 0 = No)	Categorical
Region	Geographical region of the patient	Categorical
Education_Level	Highest education level attained by the patient	Categorical
Employment_Status	Employment status of the patient (e.g., Employed, Unemployed)	Categorical

Table 2.1: Variable name with description and variable type

2.2 Distributions

This section explores the distribution of key continuous variables in the dataset to better understand the characteristics and diversity of the patient population under study. Understanding how these variables, such as age, income, insurance coverage, and clinical indicators like HbA1c, are distributed is critical for interpreting the behavior and health status of potential insulin pump candidates. These distributions give us glimpses into the differences of the target market and provide us with dominant and minority behaviors of different patient segments, their outliers and strategic subgroups. For our problem statement, these observations provide the basis of data preprocessing, feature selection, and segmentation strategy options that provide the building blocks to developing an improved model and also target markets of customers.

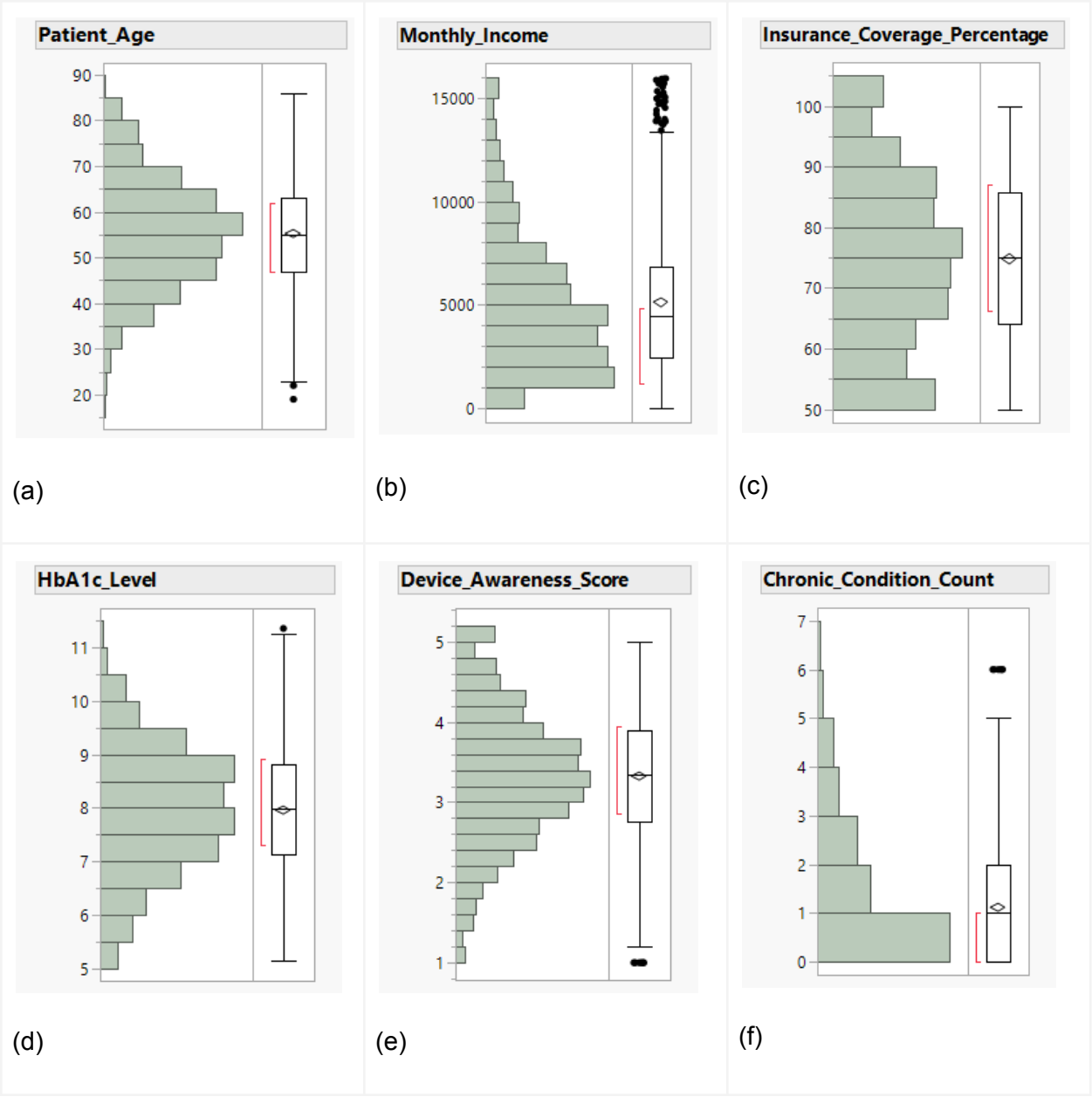


Table 2.2: Variable distributions for the 6 continuous variables

We can see from Fig 2.2(a) (Patient_Age distribution) that the dataset mainly represents middle-aged and early-senior patients; very young patients are rare. Analyses or marketing efforts should therefore focus on the 45- to 65-year group, while treating the small number of younger patients separately if their needs differ.

We can see from Fig 2.2(b) (Monthly_Income distribution) that the dataset mainly represents middle-income patients, with a small but notable affluent segment and a few very low earners. Marketing or pricing strategies should cater primarily to the $2.5 k–$7 k group, while designing separate approaches for high-income outliers and those with minimal income.

Fig 2.2(c) (Insurance_coverage_percentage distribution) shows that most patients enjoy moderate-to-high insurance support. About 25 % have coverage above 86%, making them good candidates for higher-end devices. Patients below 60 % coverage are a smaller group that may need financial assistance or lower-cost options.

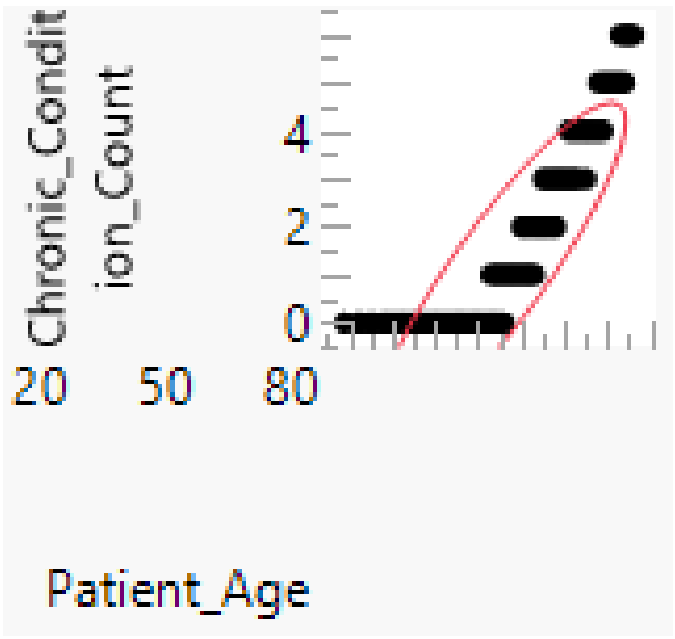
Through the above figure Fig 2.2(d) (HbA1c_Level distribution) we can see most patients are above the common clinical target of 7 %, indicating room for tighter blood-sugar control. The small group with HbA1c > 10 % may need intensive intervention, while those near 5–6 % represent well-managed cases that could serve as benchmarks for best practice.

Fig 2.2(e) (Device_Awareness_Score distribution) shows that overall device knowledge is solid, easing the introduction of advanced technologies. The small group scoring 1 may require focused education or training before they can adopt or effectively use new devices.

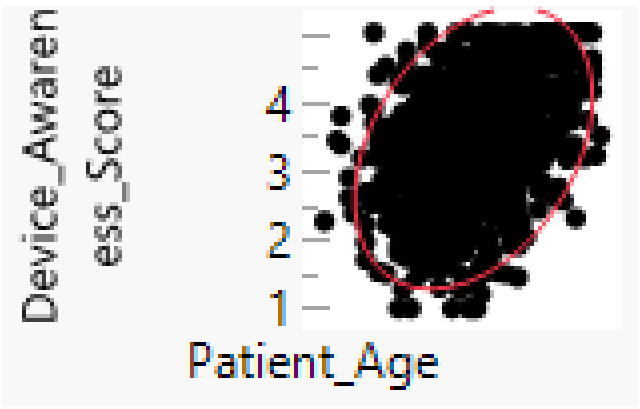
Fig 2.2(f) (Chronic_Condition_Count distribution) shows that a great majority of patients require management for at most one chronic illness, suggesting standard care pathways will suffice for them. The small subset with five or more conditions will need coordinated, resource-intensive care and could be prioritised for specialised programmes or bundled service offerings.

2.3 Scatterplot Matrix

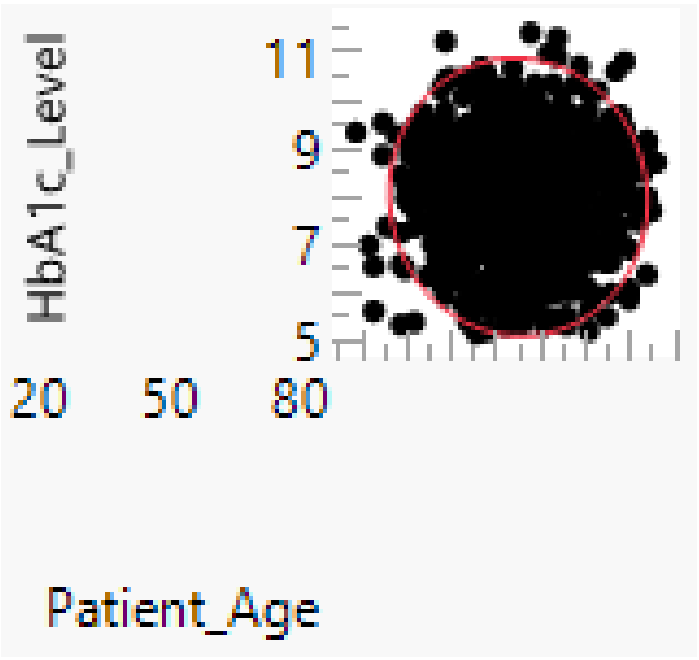
This figure displays a series of bivariate scatterplots, each with density ellipses, in order to explore how pairs of continuous variables are connected to each other with respect to insulin pump take-up. These plots help us visually to detect trends, correlations and groupings that may affect the likelihood of a customer purchasing a pump. By looking at correlations such as age and chronic conditions, income and insurance, and HbA1c value and device awareness, tells us how various factors can potentially be bundled together to inform purchasing behavior. This guides our study by illustrating the most significant traits to use when predicting purchases. It is helpful to enhance our forecasting models and to create effective targeted marketing, customer segmentation, and outreach strategies.


Chronic_Condit
ion_Count
4
2
0
20
50
80
Patient_Age

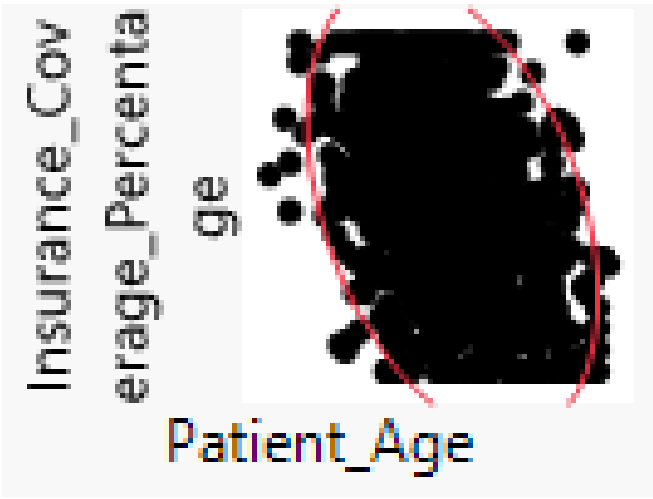
(a)


Device_Awaren
ess_Score
4
3
2
1
Patient_Age

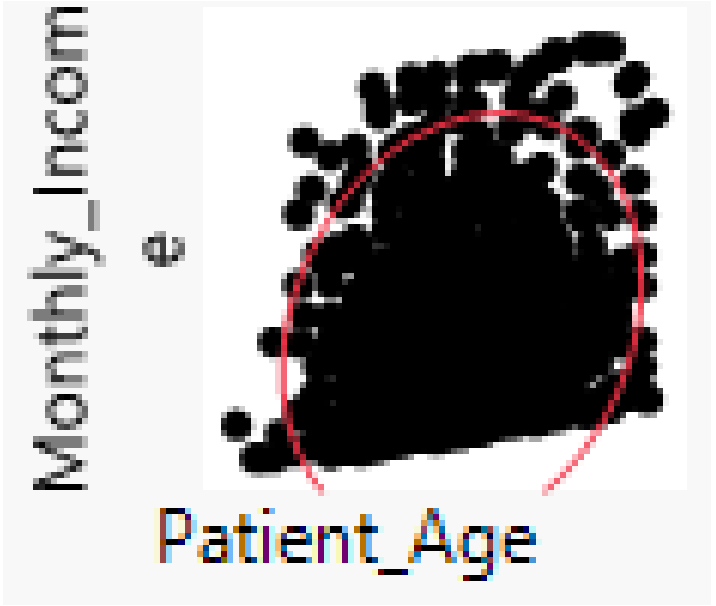
(b)


HbA1c_Level
11
9
7
5
20
50
80
Patient_Age

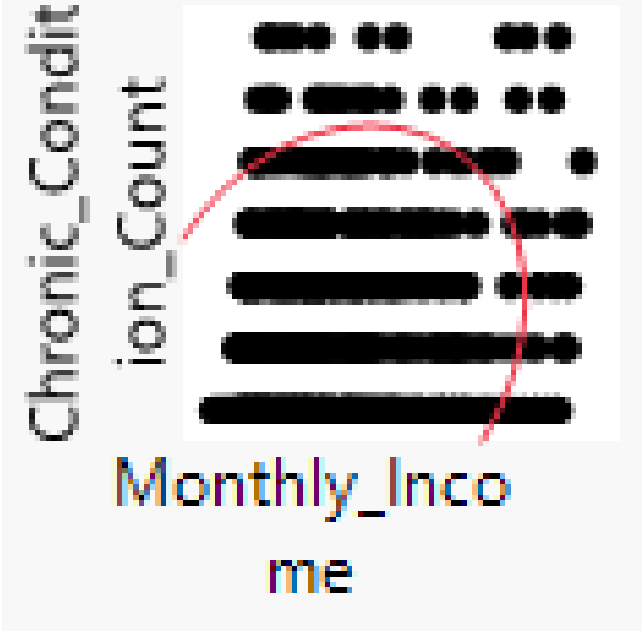
(c)


Insurance_Cov
erage_Percenta
ge
Patient_Age

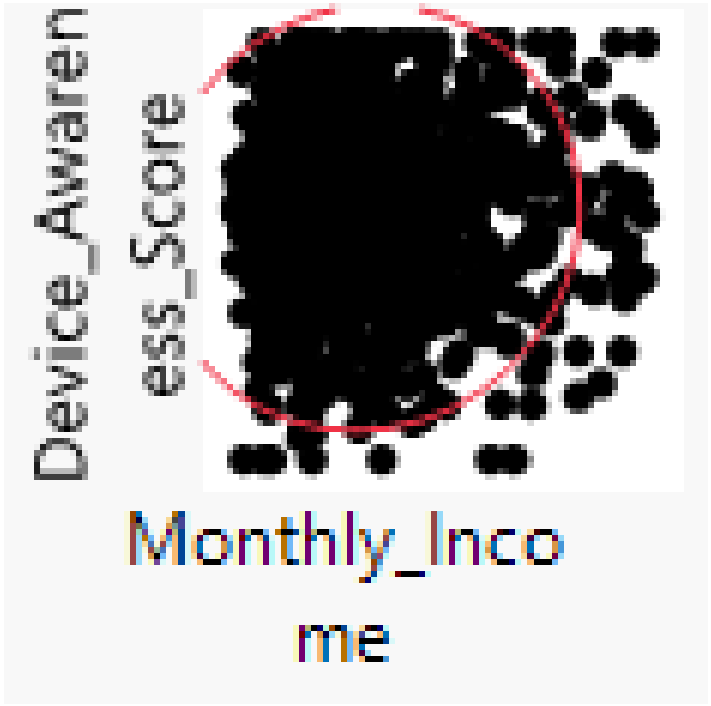
(d)


Monthly_Incom
e
Patient_Age

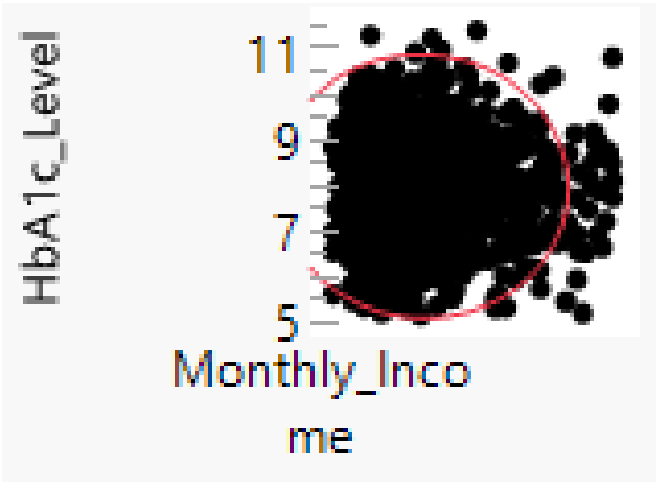
(e)


Chronic_Condit
ion_Count
Monthly_Inco
me

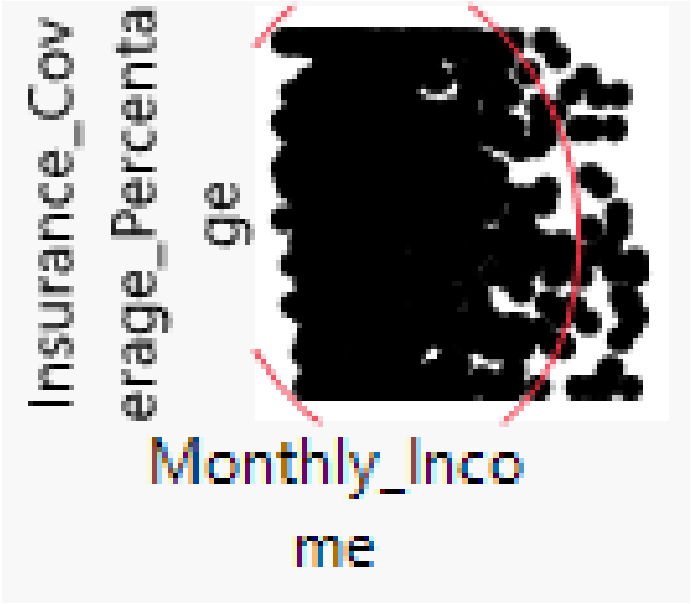
(f)


Device_Awaren
ess_Score
Monthly_Inco
me

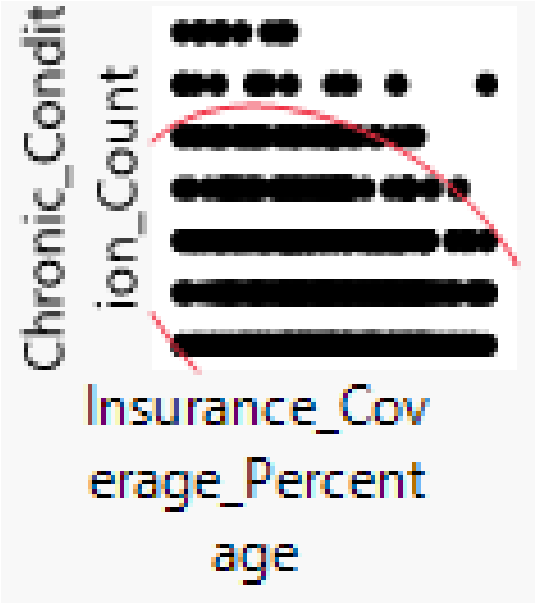
(g)


HbA1c_Level
11
9
7
5
Monthly_Inco
me

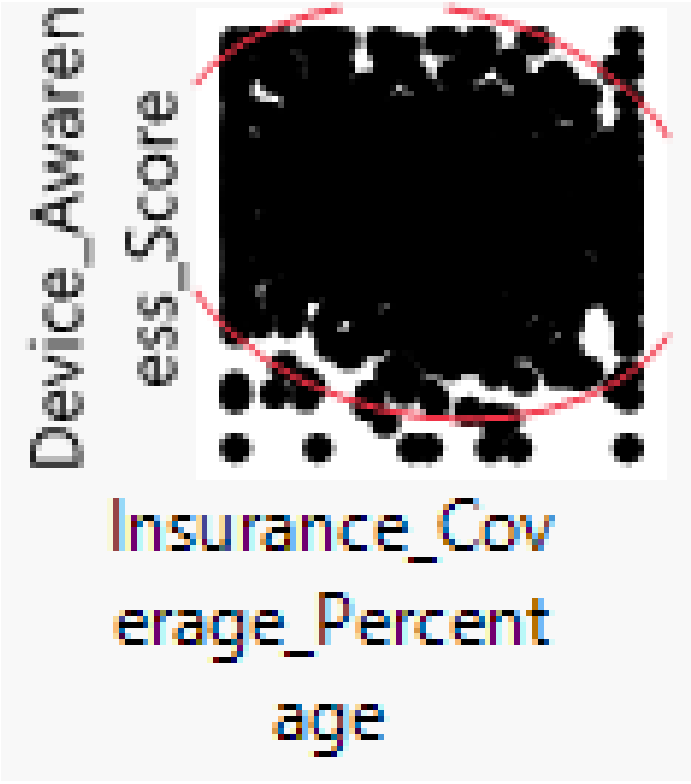
(h)


Insurance_Cov
erage_Percenta
ge
Monthly_Inco
me

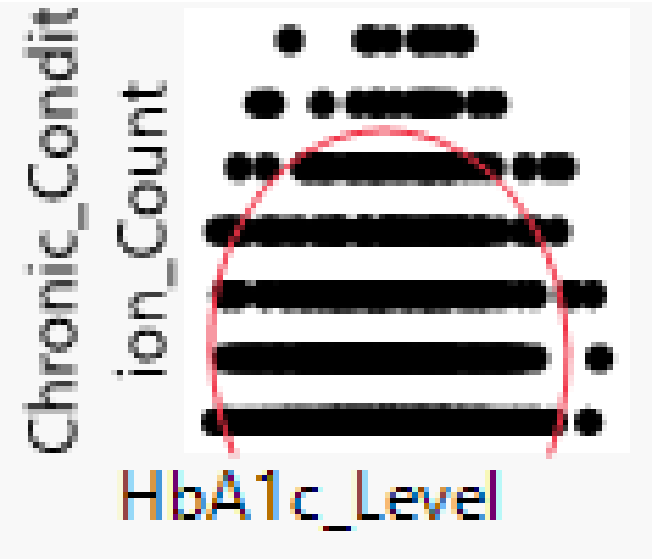
(i)


Chronic_Condit
ion_Count
Insurance_Cov
erage_Percent
age

(j)


Device_Awaren
ess_Score
Insurance_Cov
erage_Percent
age

(k)


Chronic_Condit
ion_Count
HbA1c_Level

(l)

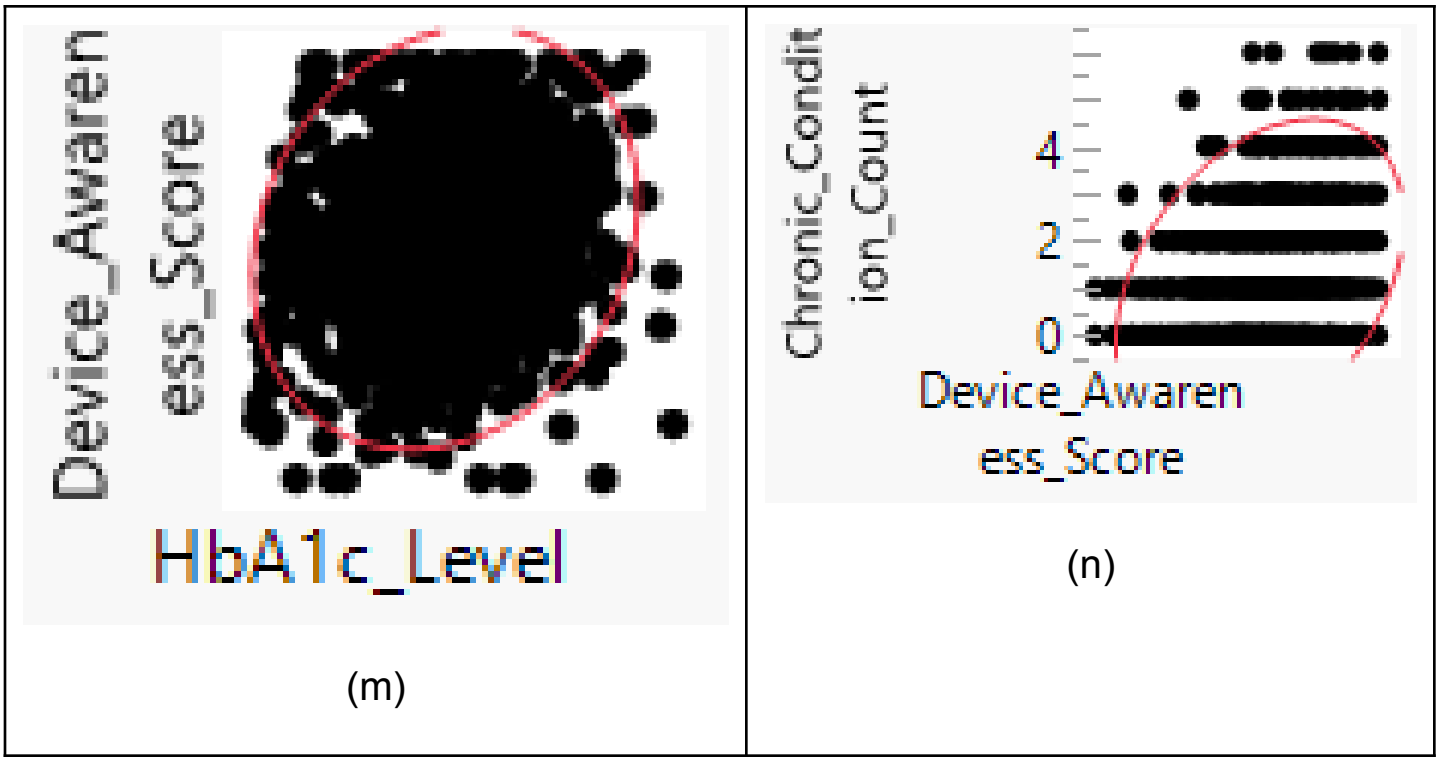


Table 2.3: Scatterplot matrix for the 6 continuous variables

Figure 2.3(a) (Patient Age vs Chronic Condition Count) shows the relationship between patient age and chronic condition count. It shows a strong positive relationship between patient age and chronic condition count, with older patients tending to have a higher number of chronic conditions. The scatter plot reveals that as age increases, the chronic disease burden also rises consistently. This is intuitive, as aging is associated with greater health deterioration. For stakeholders focused on insulin pump adoption, this suggests that targeting older patients with multiple chronic illnesses could be a strategic approach, as they may have more urgent needs for advanced diabetes management solutions like insulin pumps.

Figure 2.3(b) (Patient Age vs Device Awareness Score) shows the relationship between patient age and digital device awareness. It displays a positive relationship between patient age and device awareness score, showing that older patients tend to have slightly higher awareness of diabetes-related technologies. This trend is intuitive because older individuals typically interact more with healthcare providers, exposing them to medical innovations. For the insulin pump business case, this suggests that the older demographic may already possess foundational knowledge about pumps, allowing marketing campaigns to focus more on addressing financial or trust barriers rather than basic device education.

Figure 2.3(c) (Patient Age vs Device Awareness Score) shows the relationship between patient age and HbA1c. It shows no strong relationship between patient age and HbA1c levels, with substantial variability observed across all age groups. This finding is interesting because it challenges the assumption that older patients necessarily have poorer glycemic control. For insulin pump stakeholders, it highlights the need to rely on clinical metrics (like actual HbA1c

readings) rather than demographic factors when assessing candidates for insulin pump adoption, ensuring marketing and education are directed at patients with poor blood sugar control, regardless of age.

Figure 2.3(d) (Patient Age vs Insurance Coverage Percentage) Scatterplot showing the relationship between patient age and Insurance Coverage Percentage. It illustrates a mild negative relationship between patient age and insurance coverage percentage. The scatter plot reveals that while some older patients have strong insurance coverage, many experience limited coverage, similar to younger cohorts. This is somewhat counterintuitive, as one might expect older individuals to have better coverage via government programs. For the problem statement, this stresses the need to verify insurance access individually when targeting customers, ensuring no assumptions are made based purely on age.

Figure 2.3(e) (Patient Age vs Monthly Income) shows the relationship between patient age and Monthly Income. It depicts a positive relationship between patient age and monthly income, showing that income tends to increase with age up to a certain point. This trend is intuitive, aligning with typical career progression patterns where earnings rise with experience. For the insulin pump adoption strategy, this indicates that older, higher-income individuals might be better positioned to afford a pump, but should still be evaluated for insurance gaps and financing needs to support purchasing decisions.

Figure 2.3(f) (Monthly Income vs Chronic Condition Count) shows the relationship between Monthly Income and Chronic Condition Count. It shows a positive relationship between monthly income and chronic condition count, a trend that is somewhat counterintuitive. One might expect poorer health outcomes with lower income, but the data suggests higher-income patients report more diagnosed conditions,potentially due to better access to healthcare and more frequent testing. For our dataset, this means higher-income patients could be an important target group because their medical awareness and resources may translate into a higher likelihood of adopting sophisticated devices like insulin pumps.

Figure 2.3(g) (Monthly Income vs Device Awareness Score) shows the relationship between Monthly Income and Device Awareness Score. It reveals a weak positive relationship between monthly income and device awareness score. This is intuitive, since higher-income individuals often have better access to health education and newer technologies. However, the weak correlation indicates that awareness campaigns must still address gaps across all income levels. For insulin pump marketing, this means outreach efforts must remain broad-based, while slightly emphasizing lower-income groups who may have less exposure to diabetes management technologies.

Figure 2.3(h) (Monthly Income vs HbA1c) shows the relationship between Monthly Income and HbA1c. It shows no strong relationship between monthly income and HbA1c level, suggesting that glycemic control varies widely regardless of financial status. This is interesting, as it counters the belief that wealthier individuals necessarily manage their diabetes better. For

stakeholders aiming to promote insulin pump use, this finding reinforces the importance of focusing on clinical indicators rather than assuming income-based health outcomes.

Figure 2.3(i) (Monthly Income vs Insurance Coverage Percentage) shows the relationship between Monthly Income and Insurance Coverage Percentage. It displays no strong relationship between monthly income and insurance coverage percentage. Although one might intuitively expect higher-income patients to have better insurance, the scatterplot suggests that insurance access is not strictly income-dependent. This insight is important for insulin pump adoption strategies because it shows that financial or insurance-based obstacles can affect patients across income levels, emphasizing the need for flexible purchasing and financing solutions.

Figure 2.3(j) (Insurance Coverage Percentage vs Chronic Condition Count) shows the relationship between Insurance Coverage Percentage and Chronic Condition Count. It illustrates a mild negative relationship between insurance coverage percentage and chronic condition count. Surprisingly, individuals with more chronic conditions tend to have slightly lower insurance coverage. This finding is counterintuitive, challenging the assumption that sicker patients maintain better insurance. For the insulin pump business case, this highlights the urgency of addressing insurance and affordability barriers among chronically ill patients, ensuring that those who need pumps the most are not left behind due to lack of coverage.

Figure 2.3(k) (Insurance Coverage Percentage vs Device Awareness Score) shows the relationship between Insurance Coverage Percentage and Device Awareness Score. It shows a weak negative relationship between insurance coverage percentage and device awareness score. This is counterintuitive because better-insured patients would be expected to have greater awareness of advanced treatment options. For the dataset, it suggests that relying on healthcare providers and insurance networks to spread awareness is not sufficient; direct-to-patient marketing and educational efforts must be intensified to ensure patients are fully informed about insulin pump options.

Figure 2.3(l) (HbA1c Level vs Chronic Condition Count) shows the relationship between HbA1c and Chronic Condition Count. It shows no strong relationship between HbA1c levels and chronic condition count, indicating that poor blood sugar control does not directly correspond with the number of other chronic conditions. This finding is interesting because it suggests that insulin pump eligibility cannot be inferred from the general disease burden alone. Business intelligence models should prioritize direct measurement of HbA1c control when identifying patients who could most benefit from pump technology.

Figure 2.3(m) (HbA1c Level vs Device Awareness Score) shows the relationship between HbA1c and Device Awareness Score. It reveals a positive relationship between HbA1c level and device awareness score, showing that patients with higher blood sugar levels are somewhat more aware of diabetes management devices. This is intuitive, as poorly controlled patients are more frequently recommended to consider advanced interventions. For insulin pump promotion, this is a positive signal , higher HbA1c patients are both in need and somewhat primed for

device adoption, allowing targeted messaging to focus on device benefits and outcomes improvement.

Figure 2.3(n) (Device Awareness Score vs Chronic Condition Count) shows the relationship between Device Awareness Score vs Chronic Condition Count. It shows a mild positive relationship between device awareness score and chronic condition count. This is intuitive, as individuals managing multiple health conditions are often exposed to a broader range of medical interventions, increasing awareness of devices like insulin pumps. For our dataset, this relationship supports the targeting of patients with a complex medical history, emphasizing the pump's ability to simplify diabetes management amidst other health challenges.

The full scatterplot matrix is provided in the Appendix (see Figure A.1).

3. Interdependence Analysis

This methodology examines the underlying pattern of the data by using clustering methods and dimensionality reduction to identify hidden patterns among the patients. Unlike outcome models, this unsupervised technique does not involve using insulin pump adoption as an outcome variable. Instead, it groups individuals according to shared variables like socio-economic status, clinical markers, and device awareness. In this study, these methods allow identification of patient groups in a natural state and likely to have distinct buying patterns. Identification of such subgroups is crucial to provide guidelines on strategies in healthcare provision, medical equipment sales, as well as policy-making. The segment profiles that occur as a result of such studies form the foundation for the interpretation of heterogeneity in technology adoption and facilitating tailoring of business intelligence programs according to specific patient needs.

3.1 Principal Component Analysis

PCA is performed with the aim to decrease the dataset dimensionality while preserving most of the variance within it. Data complexity is reduced by the correlation of data into uncorrelated components that capture prominent patterns among patients. In research studies, PCA improves interpretability and minimizes noise, facilitating efficiency and accuracy in further clustering. Further, the primary features are of worth in examining the most substantial axes of variation e.g., socio-economic expense or complexity of health that can indirectly affect insulin pump buying behaviour. This is a necessary step in aligning unsupervised segmentation with overall analytical purpose in revealing structure within patient conduct relative to device adoption.

Methods for Determining Significant Components

According to Kaiser's Rule, only components with eigenvalues greater than 1 should be retained. In this case, Component 1 has an eigenvalue of 2.419, Component 2 is at 1.120, and Component 3 has an eigenvalue of 0.934.While Components 1 and 2 clearly meet the threshold, Component 3 is slightly below the cutoff. However, because it is close to 1, there may

be some justification for considering it, especially if it contributes meaningfully to the explained variance.

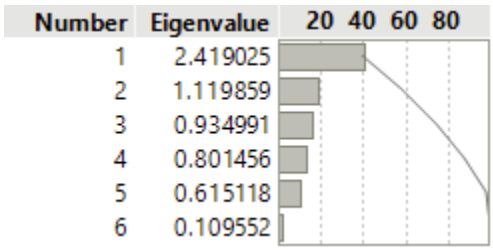


Figure 3.1.1: Eigenvalues for Principal Components: Pareto Chart displaying eigenvalues for six principal components.

The scree plot, illustrated in Figure 3.1.2, shows a drop in eigenvalues after the second component, indicating that the first two components explain a significant portion of the variance. However, the decline after Component 2 is more gradual rather than sharp. Because Component 3 still maintains a relatively high eigenvalue and is close to the threshold of 1, it is still extremely useful. If looking at its total contribution to explained variance, then it would be fair to include Component 3 as well as the first two.

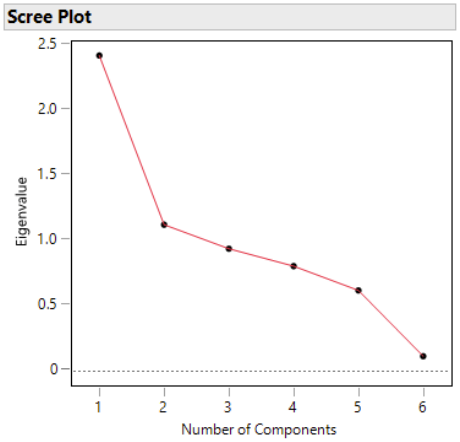


Figure 3.1.2: Scree Plot of Principal Components: Visualization of eigenvalues (variance) for six principal components.

The cumulative variance explained, as shown in Figure 3.1.3, indicates that the first two principal components (PC1 and PC2) together account for 58.981% of the total variance. When PC3 is included, the cumulative variance rises to 74.565%, which exceeds the commonly accepted threshold of 70%.This suggests that retaining the first three components is appropriate, as it allows us to capture the majority of the important variance in the dataset. Figure 3.1.3 provides a detailed breakdown of the variance explained by each component.

Considering the results from Kaiser's rule, the scree plot, and the cumulative variance explained, it is reasonable to retain the first three principal components. This decision ensures

that we meet the 70% threshold while also including all components that make a meaningful contribution.

Eigenvalues

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	2.419025	40.317		40.317
2	1.119859	18.664		58.981
3	0.934991	15.583		74.565
4	0.801456	13.358		87.922
5	0.615118	10.252		98.174
6	0.109552	1.826		100.000

Figure 3.1.3: Eigenvalues Table: Numerical breakdown of principal component variances.

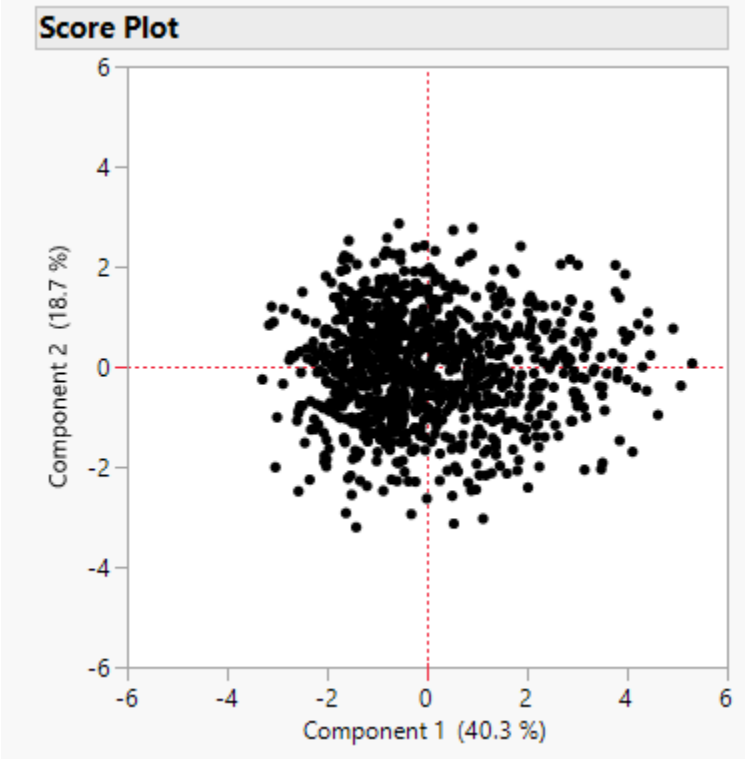


Figure 3.1.4: PCA Score Plot (Component 1 vs. Component 2): Visualization of patient distribution across the first two principal components

In figure 3.1.4. the score plot shows the projection of all observations onto the first two principal components, with each black dot representing one patient. PC 1 (horizontal axis), which explains 40.3% of the total variance, captures variation primarily related to device awareness, chronic condition count, and monthly income. PC 2 (vertical axis), explaining 18.7% of the variance, is mostly associated with HbA1c levels. Most of the points are clustered tightly around the center (0,0), indicating that the majority of patients have average scores on the principal components. There is slightly more spread along PC 1 compared to PC 2, suggesting greater variability in socio-economic and technology engagement factors than in clinical glucose control.

The point cloud appears somewhat elongated horizontally, which is consistent with PC 1 being the dominant dimension of variability. Importantly, there are very few observations located at the extreme edges of the plot, meaning that the dataset is relatively homogeneous and free from significant outliers. Overall, the score plot confirms that the first two principal components effectively summarize the key patterns in the data without major distortion, supporting the use of PC 1 and PC 2 for clustering and visualization.

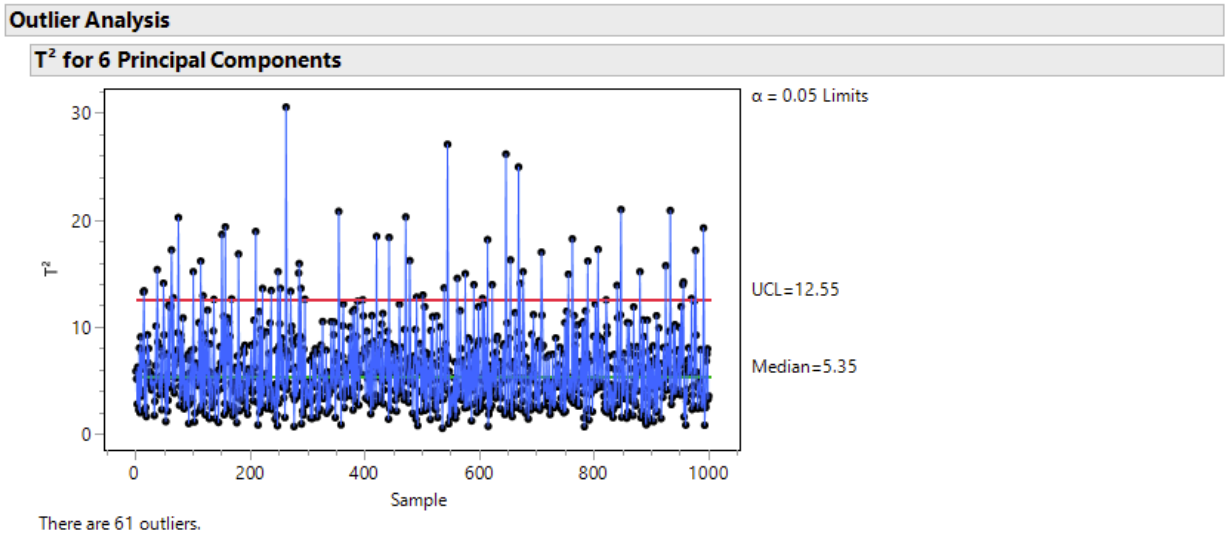


Figure 3.1.4: PCA Hotelling's T² Outlier Analysis: PCA Outlier Detection: 61 samples (T² > 12.55, α=0.05) flagged as statistically significant multivariate outliers

The PCA outlier analysis, based on Hotelling's T² for 6 principal components, identified 61 outliers exceeding the upper control limit (UCL=12.55), which was set at a 95% confidence level (alpha=0.05). These points are flagged as outliers because their T² values, measuring their multivariate distance from the dataset's center,are statistically extreme, indicating they deviate significantly from the majority of the data. The median T² (5.35) shows typical variation, confirming that the outliers (T² > 12.55) are unusual. Potential reasons for their outlier status include: (1) Measurement errors (e.g., sensor noise, data entry issues), (2) rare but meaningful anomalies (e.g., unique samples, process deviations), or (3) incorrect PCA assumptions (e.g., nonlinear patterns not captured by 6 PCs). Further investigation,such as reviewing raw data distributions, domain-specific context, or adjusting the PCA model,can clarify whether these outliers represent noise or critical insights.

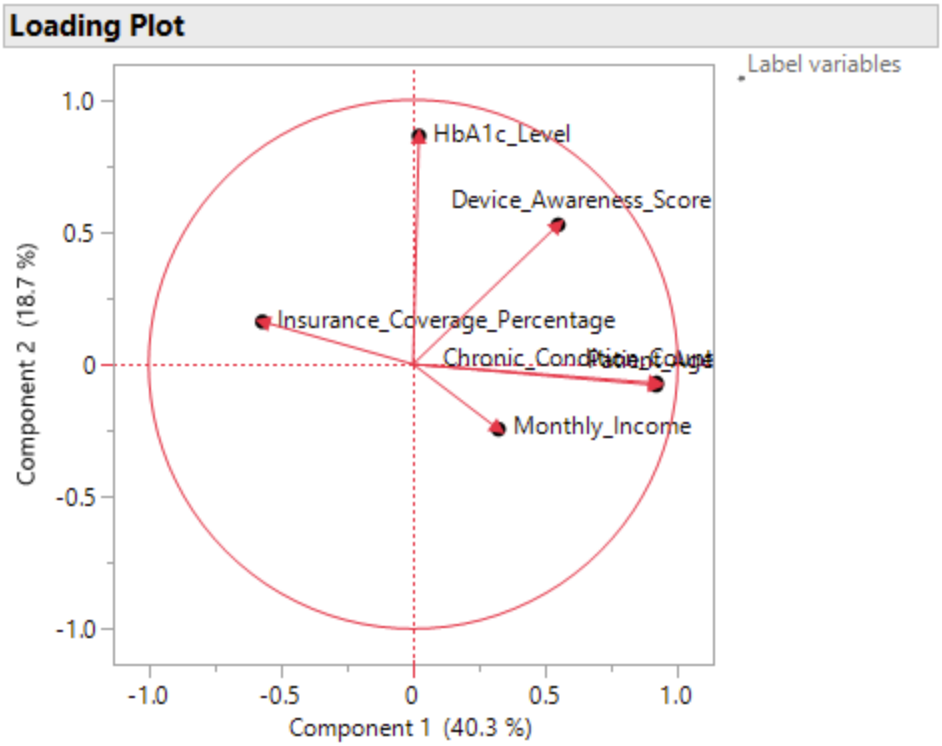


Figure 3.1.4: PCA Loadings Plot: Variable contributions to principal components.

The loading plot illustrates how each of the original variables contributes to the first two principal components extracted during principal component analysis (PCA). PC 1, shown on the horizontal axis and explaining 40.3% of the total variance, is primarily driven by Monthly Income, Chronic Condition Count, and Device Awareness Score, all of which load strongly and positively in the rightward direction. This suggests that patients with higher values in these variables will tend to have higher PC 1 scores.

PC 2, represented on the vertical axis and accounting for 18.7% of the variance, is most influenced by HbA1c Level and Device Awareness Score, both of which load positively along the vertical dimension.

Insurance Coverage Percentage shows a mild negative loading on PC 1 and a near-zero loading on PC 2, indicating that it does not strongly align with the major variance directions compared to other variables. Patient Age shows minimal contribution in either dimension, lying close to the center, suggesting it plays a smaller role in the primary separation of patients in this PCA space.

The orientation and length of the arrows indicate the strength and direction of each variable's influence: longer arrows (such as for HbA1c Level and Device Awareness Score) have a stronger effect, while shorter arrows (such as Patient Age) are weaker contributors.

Overall, the loading plot helps clarify that PC 1 captures socio-economic and disease burden factors, while PC 2 is more related to clinical health outcomes, particularly glucose control.

3.2 Hierarchical Clustering

Hierarchical clustering is used to segment patients into a hierarchical structure based on proximity within their multivariate profiles. This approach provides a broad overview of how the patients group along with main dimensions like income, level of HbA1c, level of awareness about the device, and chronic conditions. The cluster generated identifies specific patient types to which varied degrees of need, readiness, or access to diabetes technology can be ascribed. In the context of this study, these classifications assist in identifying which patient segments will be neglected or unresponsive to conventional outreach and provide ideas for more effective, equitable intervention methods.

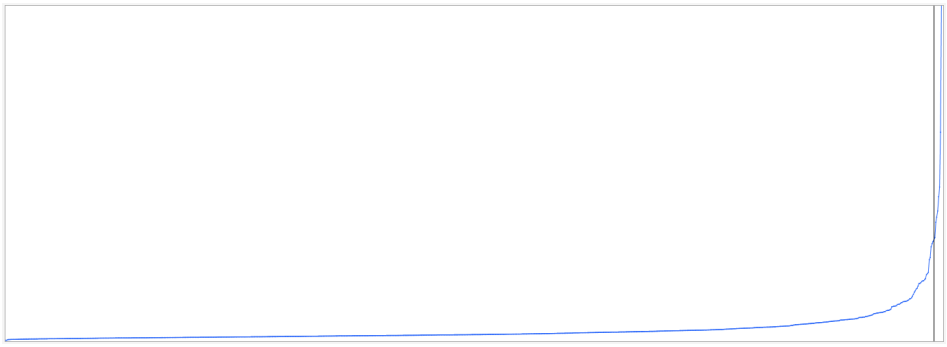


Fig 3.2.1: Scree Plot for hierarchical clustering

From Fig 3.2.1, we can observe that the elbow point appears around 17 clusters. However, since we need to limit the number of clusters to 10 or fewer, we determined that 9 clusters would be appropriate. This can be identified by analyzing the elbow in the plot. As we move towards the right side of the plot, there is a noticeable sharp jump around 17 clusters, after which the curve moves upward. A similar pattern is observed at 9 clusters, which is why we selected 9 clusters in total.

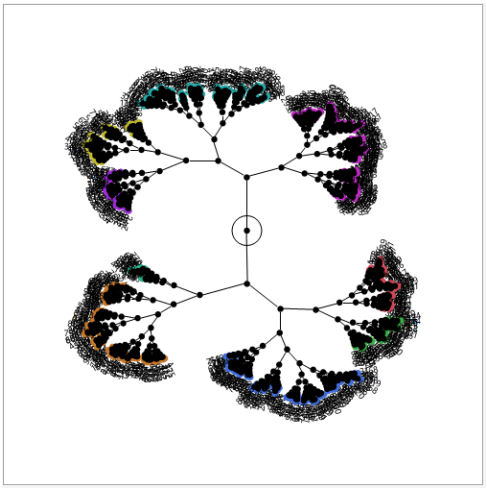


Fig 3.2.2: Constellation Plot for hierarchical clustering

From Figure 3.2.2, we can observe a total of 9 branches, indicating that there are 9 clusters overall. In this type of plot, we typically start from the root and count the number of branches. Using this approach, we determined that 9 clusters are optimal.

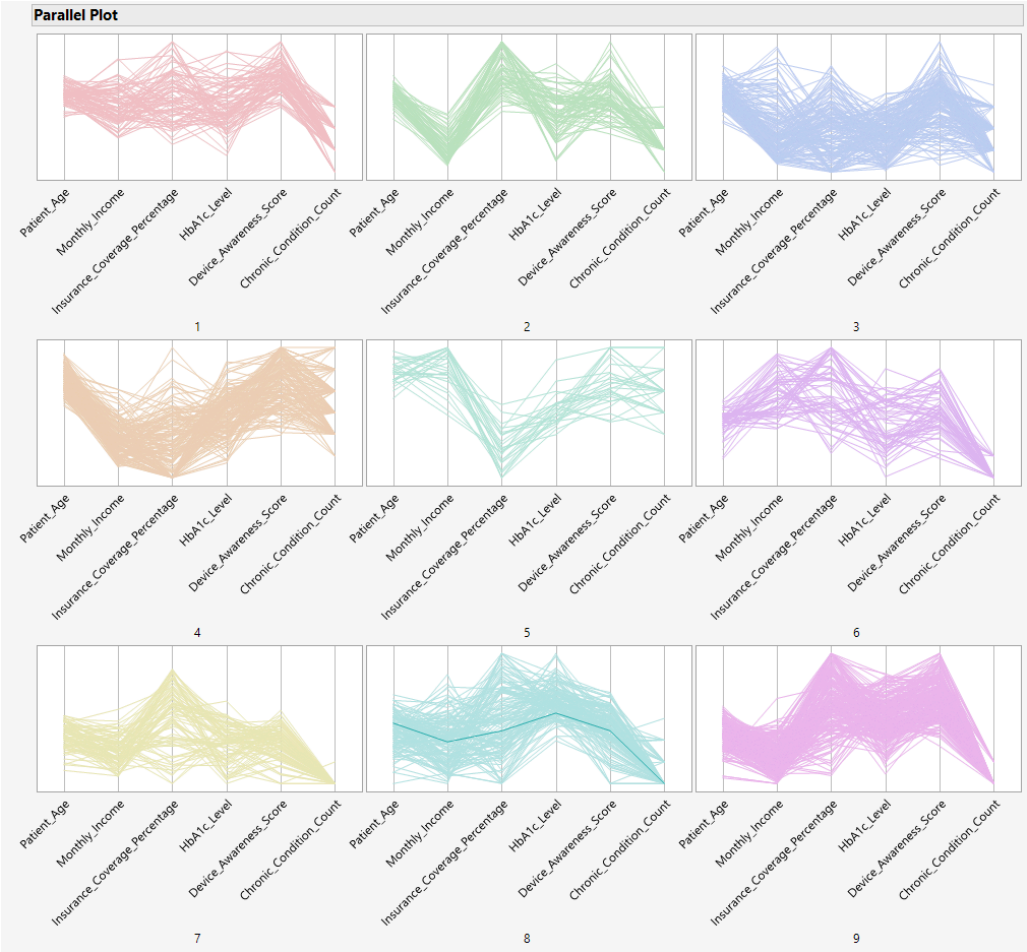


Fig 3.2.3: Parallel Plot Displaying Feature Patterns Across Clusters

In Fig. 3.2.3 we can see the different clusters and the characteristics of those clusters.

Cluster 1 (red) contains middle-aged patients who have average incomes, average insurance cover, and a fair number of chronic conditions. Their blood-sugar values run a little high, but they know quite a lot about diabetes devices.

Cluster 2 (green) is made up of middle aged adults with low incomes. They do, however, have good insurance coverage. Device knowledge is medium to high, and their blood-sugar levels are mixed.

Cluster 3 (blue) consists of similar aged patients as cluster 1 and 2, their earning is from low to high and have low to high insurance percentage. Even so, they keep their blood sugar in a healthy range, they also have a good device awareness.

Cluster 4 (orange) holds older patients with lower income, low to high insurance coverage, low to high blood-sugar readings, and higher count if chronic conditions. Although they are comfortable with devices, they need close medical follow-up and extra financial support to manage their health risks.

Cluster 5 (dark green) also includes seniors, but they are wealthier. They lack strong insurance, live with several chronic conditions, and show only moderate blood-sugar control. Because they can afford care and understand technology, premium remote-monitoring services would fit them well.

Cluster 6 (purple) features adults of average age and medium to high income who enjoy low to high insurance coverage. The blood-sugar level is low to medium l;despite having few other illnesses. Coaching which targets medication adherence and lifestyle choices could help these individuals.

Cluster 7 (yellow) contains the youngest patients in the study. They have moderate incomes but weak insurance coverage. Their blood-sugar levels are low to medium, and device knowledge is low to medium. Low-cost starter technology and basic education about the product could be helpful here.

Cluster 8 (teal) is made up of patients with low to high age range, whose incomes sit in the middle range with a low to high insurance coverage. Device familiarity is low to medium, but blood-sugar control is still poor, and chronic disease burdens are moderate.

Cluster 9 (Pink) includes younger to middle-aged adults with low incomes but very good insurance coverage. They are highly comfortable with technology, their blood-sugar levels are on the higher side and they have few other conditions.

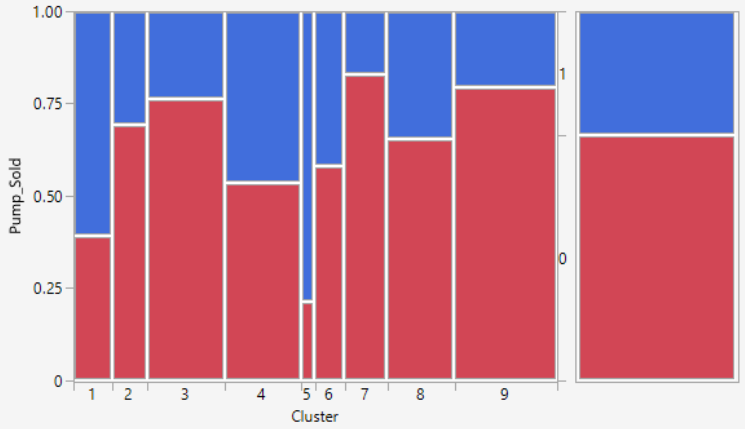
The most promising clusters here are based on parallel plot:

Cluster 9 (Pink): This group consists of younger, tech-savvy patients who have good insurance coverage and a high level of comfort with medical devices. These patients are well-insured, comfortable with technology, and are likely to adopt advanced products such as continuous glucose monitors (CGMs), smart insulin pumps, and mobile health applications. They are an ideal target for promoting premium, high-tech healthcare products.

Cluster 5 (Dark Green): This group includes wealthy seniors who have poor insurance coverage but a strong ability and willingness to pay out-of-pocket. Although their insurance is limited, their financial resources make them a strong market for premium healthcare services, including remote monitoring and concierge health offerings.

3.2.4 Contingency Analysis

Mosaic Plot



Contingency Table

Pump_Sold

Cluster	Count / Total % / Col % / Row %	0	1	Total
1		32	49	81
		3.20	4.90	8.09
		4.81	14.58	
		39.51	60.49	
2		50	22	72
		5.00	2.20	7.19
		7.52	6.55	
		69.44	30.56	
3		123	38	161
		12.29	3.80	16.08
		18.50	11.31	
		76.40	23.60	
4		85	73	158
		8.49	7.29	15.78
		12.78	21.73	
		53.80	46.20	
5		6	22	28
		0.60	2.20	2.80
		0.90	6.55	
		21.43	78.57	
6		36	26	62
		3.60	2.60	6.19
		5.41	7.74	
		58.06	41.94	
7		73	15	88
		7.29	1.50	8.79
		10.98	4.46	
		82.95	17.05	
8		91	48	139
		9.09	4.80	13.89
		13.68	14.29	
		65.47	34.53	
9		169	43	212
		16.88	4.30	21.18
		25.41	12.80	
		79.72	20.28	
Total		665	336	1001
		66.43	33.57	

Tests

N	DF	-LogLike	RSquare (U)
1001	8	49.659477	0.0777

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	99.319	<.0001*
Pearson	100.096	<.0001*

Fig 3.2.4:Mosaic Plot and Contingency Table Showing Association Between Clusters and Pump Purchases

The contingency analysis findings presented in Figure 3.2.4 indicate a statistically significant association between clusters and pump purchasing behavior. The Pearson Chi-Square statistic value is 100.096 whereas its p-value is lower than 0.0001, supporting that the difference in pump sales between clusters observed is not by chance. Although the R-square (U) measure of the effect size at 0.0777 indicates a moderate effect, there exist evident and significant differences in adoption levels across the identified clusters.As can be seen from the contingency table, Cluster 5 has the most pump adoption of 78.57% (22 out of 28 patients purchasing a pump). The second highest is Cluster 1 at 60.49% (49 out of 81 patients), followed by Cluster 4 at 46.20% (73 out of 158), then Cluster 6 at 41.94% (26 out of 62).

In contrast, Clusters 2, 3, 7, 8, and 9 exhibit lower adoption rates, all falling below 35%, indicating a weaker propensity among patients in these clusters to purchase a pump. These findings suggest that patient profiles captured through clustering are indeed influencing purchasing behavior, with Clusters 5 and 1 emerging as particularly promising target segments for pump-related marketing or outreach strategies.

The following mosaic plot visually confirms these results. Column width denotes the cluster size, and blue segment height denotes the proportion of patients who purchased or did not purchase a pump. Cluster 5 with the large blue segment clearly depicts the highest percent purchasing a pump. By contrast, Clusters 3, 7, and 9 display distinctly smaller blue areas, consistent with their lower adoption levels. The combination of contingency analysis with the mosaic plot offers strong evidence that clustering exposes action-improving differences in consumer behavior.

3.3 k-Means Clustering

The k-means clustering is subsequently used to group the dataset into densely packed, discernible clusters based on comparable principal component scores. They were scored along clinical, financial, and behavioral axes and compared with one another in an effort to approximate potential heterogeneity in adoption rates of insulin pumps. This provides further richness to the study result in the sense that it allows patient groups that can potentially be responsive or resistant to device buying to be defined with increased accuracy. Such clustering enables the performance of targeted business intelligence activities, e.g., offering high-risk, low-awareness regions educational campaigns or developing tailored financial models for cost-oriented segments.

Cluster Comparison

Method	NCluster	CCC	Best
K Means Cluster	3	-6.522	Optimal CCC
K Means Cluster	4	-9.6276	
K Means Cluster	5	-9.985	
K Means Cluster	6	-10.316	
K Means Cluster	7	-10.82	
K Means Cluster	8	-9.7704	
K Means Cluster	9	-10.027	
K Means Cluster	10	-10.274	

Fig 3.3.1: Cluster Comparison Table for K-Means Clustering with CCC Values. The CCC (Cubic Clustering Criterion) was evaluated for different numbers of clusters ranging from 3 to 10.

Based on the results of the hierarchical clustering in Section 3.2, we explored a range of cluster numbers between 3 and 10 for the k-means clustering analysis. Figure 3.3.1 presents the Cubic Clustering Criterion (CCC) values corresponding to each k value. The optimal number of clusters was determined based on the highest CCC value, which occurred at k = 3 with a CCC score of -6.522. Although the CCC value is negative, it is less negative compared to other k values in the examined range, indicating that three clusters provide the best clustering structure relative to the other options. Therefore, we selected three clusters as the optimal solution for k-means clustering and used this configuration for further analysis.

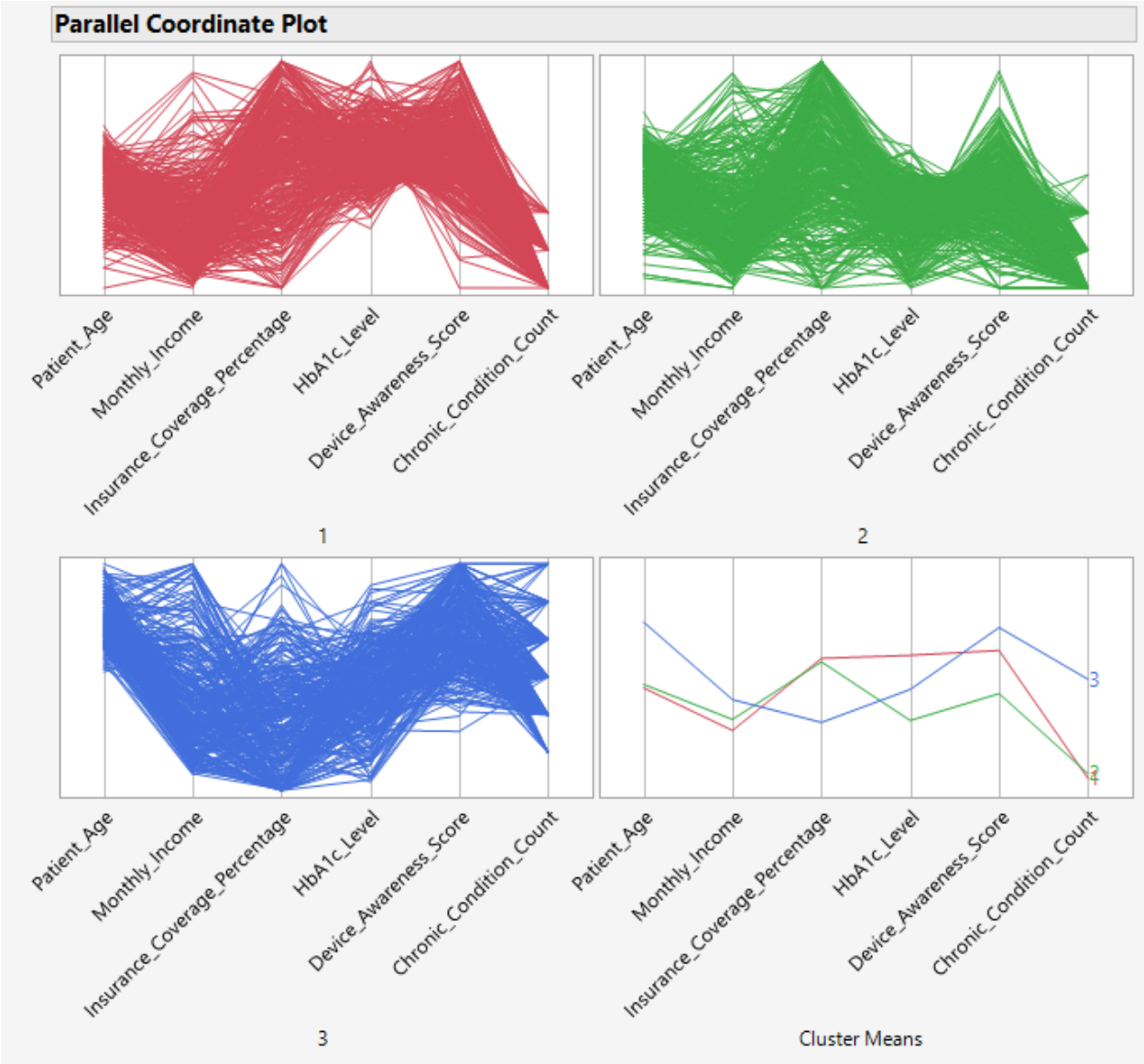


Fig 3.3.2:Parallel Coordinates Plot of Cluster Profiles and Cluster Means for K-Means Clustering (k=3).

Figure 3.3.2 presents the parallel coordinates plot for the three clusters identified through k-means clustering. Each line in the panels represents an individual patient, showing the variation in attributes such as Patient Age, Monthly Income, Insurance Coverage Percentage, HbA1c Level, Device Awareness Score, and Chronic Condition Count. The rightmost panel displays the mean profiles for each cluster, it is a more simple view of the average patterns within each cluster.

Interpretation for all these clusters:

Cluster 1 (Red): Cluster 1 patients are characterized by younger age, low monthly income, and high insurance coverage. They exhibit high HbA1c levels and high device awareness scores,

but maintain a low chronic condition count. This combination suggests that patients in this cluster are younger, financially constrained, but well-insured and highly aware of health technologies, yet still struggling with glucose control. Business strategies targeting this group should focus on reinforcing device usage for better clinical outcomes, as affordability is less of a barrier due to their strong insurance coverage.

Cluster 2 (Green): Cluster 2 also consists of younger patients with low income, but similarly high insurance coverage. However, they differ from Cluster 1 by having low HbA1c levels, low device awareness, and low chronic condition counts. These patients are comparatively healthy but less aware of available health technologies. Long-term marketing efforts should focus on building early awareness and engagement through educational campaigns that emphasize the preventive benefits of pump adoption, leveraging their good insurance access.

Cluster 3 (Blue): Cluster 3 is composed of older patients with low monthly incomes and low insurance coverage. They show moderate HbA1c levels, high device awareness scores, and high chronic condition counts. This group represents older, medically complex patients who are already knowledgeable about health technology but may face significant financial and insurance barriers to adoption. Business strategies for this segment should prioritize affordable device options, financing plans, or targeted subsidies to overcome cost obstacles while leveraging their strong awareness and existing health needs.

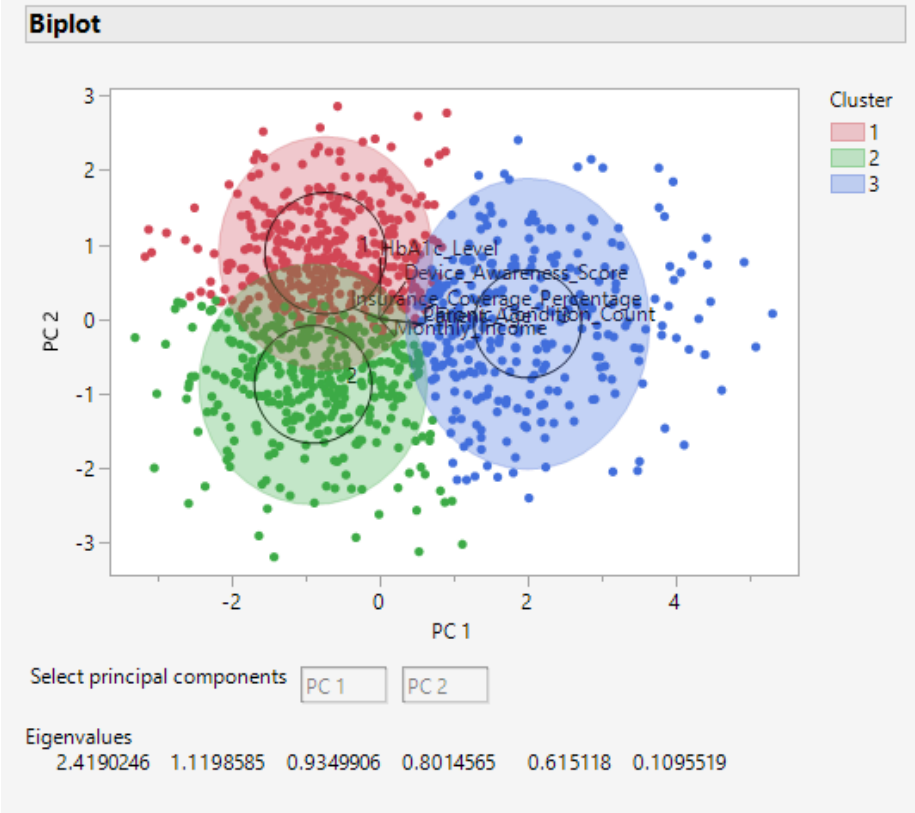


Fig 3.3.3: Biplot of Principal Components with Cluster Groupings

Cluster 1 (Red) is positioned in the upper-left quadrant, indicating patients with lower device awareness, lower income, and fewer chronic conditions (low PC 1), but higher HbA1c levels (high PC 2). This suggests a group that is financially weaker and less device-aware but facing more urgent health risks due to poor glucose control.

Cluster 2 (Green) occupies the lower-left quadrant, representing patients who are low in device awareness, income, and chronic condition burden (low PC 1) and have better HbA1c control (low PC 2). This group appears healthier but financially constrained, with limited technology engagement.

Cluster 3 (Blue) stands distinctly on the right-hand side, aligned with higher device awareness, higher chronic condition counts, and higher monthly income (high PC 1), while centered around zero on PC 2, indicating moderate HbA1c levels. Patients in this cluster are older, more medically complex, and more informed about health technologies, making them ideal candidates for premium healthcare solutions if financial barriers can be addressed.

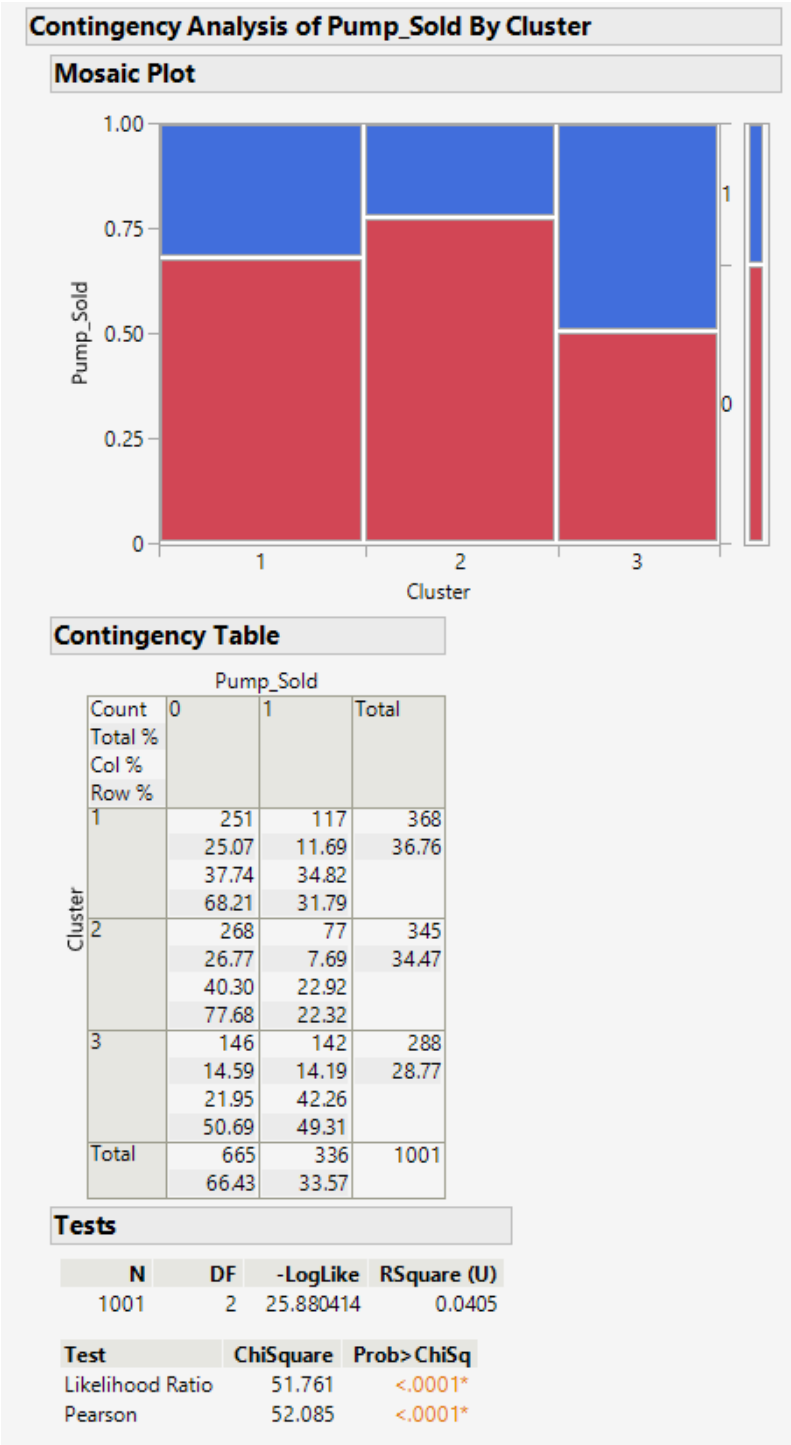


Fig 3.3.4: Mosaic Plot and Contingency Table For K-means Clustering Showing Association Between Clusters and Pump Purchases

The results of the contingency analysis depicted in Figure 3.3.4 demonstrate a very high and statistically significant relationship between clusters and pump purchase behaviour. Pearson Chi-Square test gives $\chi^2 = 52.085$ with p-value less than 0.0001, which is a clear indication that

the various pump sales distributed over the three clusters are most likely due to random variation. Although the R-square (U) measure itself is very low at 0.0405, the analysis does indicate extremely significant differences in pump adoption between clusters.

It can be seen from the contingency table that Cluster 3 has the highest rate of pump adoption with 49.31% of patients (142 out of 288) opting to buy a pump. Cluster 1 has a moderate rate of 31.79% (117 out of 368) and Cluster 2 has the lowest rate of adoption with 22.32% (77 out of 345). These numbers emphasize cluster-specific characteristics in deciding to buy, placing Cluster 3 at first as a responsiveness index and the most probable segment for pump promotion activities.

The mosaic plot further reinforces this, It clearly demonstrates that Cluster 3 contains the maximum blue area, reflecting a greater percentage of pump purchase, while Clusters 1 and 2 contain relatively smaller blue areas, reflecting their lower adoption rates. The graphical illustration accompanying the findings of the statistics also suggests the potential for developing segment-level strategies.

These findings strongly validate PCA-driven cluster profiles. Cluster 3, composed of patients of higher income, improved insurance coverage, and higher device awareness, is the best lead to chase in pump sales. Reinforcement and upselling marketing campaigns need to be focused on this segment because they are willing and capable of adopting. Cluster 1, while demonstrating moderate promise, is composed of patients with reduced device awareness and increased health risk, indicating that special education programs and aid programs for funding could assist in enhancing the rate of conversion in this group. Cluster 2 is much more difficult, however. Consisting of people with lesser economic capacity and lower clinical needs, this segment would need price-inelastic products like micro-insurance schemes, discounts, or affordable starter pump sets to induce adoption and offer wider market penetration.

Comparing Hierarchical clustering and K- Means clustering

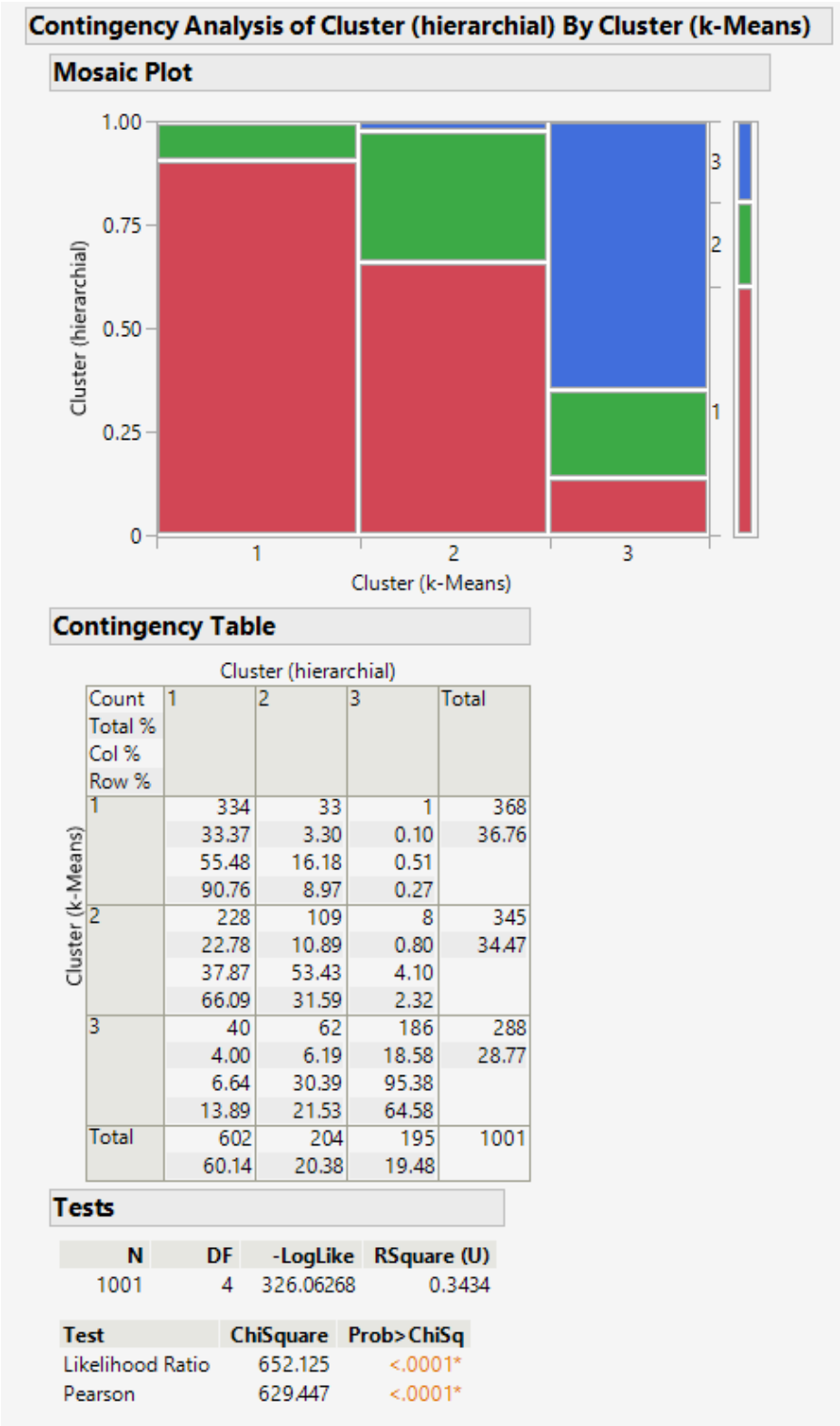


Fig 3.3.5: Contingency Analysis of hierarchical clustering and K- means clustering

The figure displays a contingency analysis that cross-tabulates the three k-means clusters (x-axis) with the three clusters obtained from hierarchical clustering (y-axis). Each coloured bar in the mosaic plot represents the proportion of each k-means cluster groups that fall into hierarchical cluster proportions, while the accompanying contingency table provides exact counts and percentages.

Statistically, the two clustering solutions are strongly related: the Pearson χ^2 value is 629.447 with 4 degrees of freedom, and the p-value is < 0.0001. An R-square(U) of 0.3434 indicates that just over one-third of the variation in hierarchical-cluster membership can be “explained” by the k-means assignments—a substantial overlap for two independent algorithms.

K-Means Cluster 1 aligns almost exclusively with Hierarchical Cluster 1: 334 of 368 observations (≈ 91 %) map to the same group, signalling very tight agreement for this segment.

K-Means Cluster 1 maps also to Hierarchical Cluster 2 (228 of 345 cases, ≈ 66 %), though about a third spill into Hierarchical Cluster 2, suggesting moderate but less perfect overlap.

K-Means Cluster 3 shows the clearest one-to-one correspondence to hierarchical Cluster 3: 186 of 288 observations (≈ 65 %) fall into Hierarchical Cluster 3, with only minor leakage elsewhere.

In practical terms, the two methods identify nearly identical groupings for Clusters 1 and 3, while Cluster 2 exhibits some boundary ambiguity. This high concordance validates the stability of the segmentation: managers can be confident that the behavioural and demographic profiles described for each k-means cluster are robust, because a fundamentally different algorithm (hierarchical clustering) reproduces nearly the same partitions**.**

4. Dependence Analysis

This section employs supervised learning methods to determine the specific patient-level factors driving insulin pump adoption. In contrast to unsupervised population structure-revealing clustering techniques, the models described here estimate the relationship between the explanatory variables and the binary target variable of whether a patient buys an insulin pump directly. With statistical evidence of how much each feature can forecast, such analysis provides statistical evidence which guides data-informed decision making in healthcare. Such evidence has implications in research through the provision of predictive models which can drive business intelligence tools such as personal patient engagement, customer segment targeting, and resource allocation strategies with an aim of maximizing device sale and awareness.

4.1 Logistic Regression Analysis

Logistic regression is used to investigate the relationship between patient attributes and the probability of insulin pump uptake. The model contains socio-economic status variables, clinical predictors, and behavioral variables. Full and reduced model estimation allow the analysis to differentiate between significant and non-significant predictors, resulting in a parsimonious yet interpretable model of adoption behavior. Logistic regression, in this study, is an essential tool

for converting statistical information into actionable intelligence. For example, a strong effect of doctor referral or HbA1c levels can be used to input provider education programs, whereas significant income-based effects can be used to input tiered pricing strategies or subsidies. The model's simplicity and interpretability also make it a great candidate for communicating results to non-technical stakeholders such as healthcare administrators and policymakers.

Logistic Regression with All Independent Variables

The R-squared value for the full model is 0.2948, which means that approximately 29.48% of the variability in whether or not a customer purchases an insulin pump can be explained by the variables included in the model. While this isn't a very high value, it still indicates that the model captures a significant amount of the factors influencing the purchase decision.

Table: Full Model - Effect Summary

Effect Summary

Source	Logworth	PValue
Doctor_Recommendation	27.901	0.00000
Monthly_Income	18.378	0.00000
HbA1c_Level	12.909	0.00000
Patient_Age	1.695	0.02020
Previous_Device_Usage	0.830	0.14791
Region	0.744	0.18047
Device_Awareness_Score	0.478	0.33253
Insurance_Coverage_Percentage	0.445	0.35851
Chronic_Condition_Count	0.331	0.46712
Patient_Education_Level	0.187	0.65064

Figure 4.1: Full Effect Summary Analysis : Logworth values and p-values for all predictive factors.

Key Insights and Interpretations:

In Figure 4.1, we can see the factors that are actually influencing the patient's choice to use the insulin pump. Amongst all the variables studied, Doctor_Recommendation plays a very important role; this could be deduced from the high logworth value and low p-value, which is lower than 0.005. This means that when a doctor advises a patient to buy the pump, there are high chances that the patient would buy it. This means we need to make sure that while working on the marketing for the product, we should target doctors and explain to them how our product would be of benefit to the patients.

Monthly income is also an important factor. The data shows that people who have high earnings are most likely to buy the insulin pump. This is also supported by high logworth and low p-value. This shows that for better sales, we should also target financially stable individuals to ensure that our product is bought. This does not mean that patients with lower incomes should not be

included; we should eventually work on flexible payment options and making the product more affordable to make it more accessible for individuals with lower incomes.

HbA1c_Level is also another significant variable that we must focus on. People who have poor glucose control are more likely to buy the product, since they need to manage their glucose level more effectively. This shows there is a need for educational outreach needed to clarify the ways the pump could be effective in terms of diabetes control. There should be more blogs and seminars to educate people about the product.

Patient_Age is also a feature linked to a patient's buying pattern, even though it is not a very strong indicator. Younger individuals seem more interested in using the device because they understand new technology better. Communication strategies should be used here in an apt manner to engage with younger patients to ensure that the product is marketed on websites and platforms they tend to use often.

Insignificant Variables:

From the analysis in Figure 4.1, we can also see the variables that are not significant and do have an impact on patients' buying behavior. We can see this through their low logworth values and high p-values greater than 0.005. Variables like Previous_Device_Usage,Region, Device_Awareness_Score, Insurance_Coverage_Percentage, Chronic_Condition_Count, and Education_Level. The higher p-value indicates that the variables are not adding any predictive value to the model. Even though these values could contribute in real life.

Logistic Regression After Removing Insignificant Variables

After removing the insignificant variables from the full model, we have the reduced model that includes only the most significant predictors.

R-Squared Value for Reduced Model

The R-squared value for the reduced model is 0.2842, slightly lower than the full model's R-squared of 0.2948. This minor decrease of approximately 1.06% indicates that the excluded variables did not meaningfully enhance the model's ability to explain the variation in insulin pump purchases. From another perspective, the reduced model simplifies the analysis by retaining only the most impactful variables, making it more interpretable while maintaining nearly the same explanatory power.

Effect Summary for Reduced Model

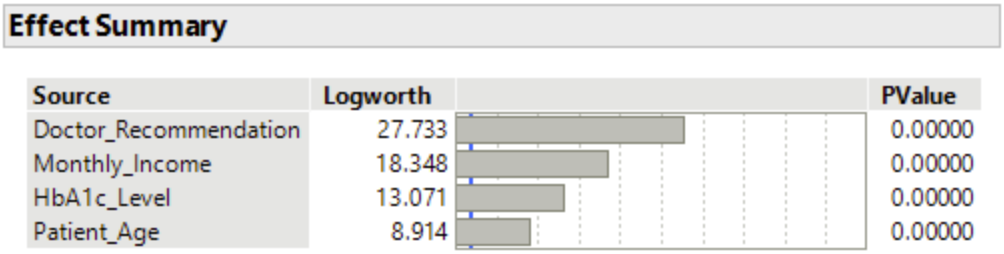


Figure 4.1.2: Simplified Effect Summary: Refined analysis isolating the four strongest predictors.

Key Insights and Interpretations:

Figure 4.1.2 displays a refined model that keeps four variables that have strong significance, which are Doctor_Recommendation, Monthly_Income, HbA1c_Level, and Patient_Age. These variables are chosen because in this dataset they are contributing more towards the prediction. This is supported by their p-values showing they have a meaningful connection with our target variable.

By focusing on these key variables, the model becomes more simplified and easier to interpret. This simplicity is valuable in a business setting; it helps the insights to be more focused and helps make quicker decisions. With few variables, we can focus on developing targeted strategies faster and focus on the customer groups that are actually relevant to our product.

Comparison of Full vs. Reduced Model

When we compare the two models, we can observe that there is minimal variation in the performance. The R-squared of the full model is 0.2948, and the reduced model is at 0.2842. This value loss is from the variables that we removed from the model because they did not provide any explanatory value to the model. The variations are mostly accounted for by the reduced set of variables.

This outcome supports the choice to reduce the model because it simplifies it. Keeping significant variables ensures improved comprehension without reducing the accuracy of the model. It becomes less complicated and helps in decision-making. It avoids the complexity of insignificant variables while maintaining high predictability. When applied to business purposes, this model is very effective because it highlights the critical factors that influence customer behavior.

Preference for Full vs. Reduced Model:

The explanatory power of the model is slightly reduced; the reduced model is a more effective and convenient choice. The R-squared reduces slightly from 0.2948 to 0.2842, showing that the variables we removed did not contribute a lot to the model. The reduced model makes it easy to interpret and simplifies the communication when presenting it to the stakeholders.

The reduced model includes only the variables that show clear statistical importance: Doctor_Recommendation, Monthly_Income, HbA1c_Level, and Patient_Age. Each of these offers insights that can be directly applied in business decision-making. For example, the strong influence of doctor recommendations highlights the need to involve healthcare providers more actively in promoting the insulin pump. Monthly income shows that affordability is a key factor, which supports targeting higher-income customers while also offering payment options for those with lower incomes. HbA1c level significance points to a group of patients with poor glucose control who could benefit the most from the product. Patient age suggests that younger individuals may be more open to adopting new health technologies, helping to guide outreach strategies.

The reduced model focuses on overfitting, which decreases the chances of overfitting. This suggests that it is more likely to be accurate with real-world datasets. The reduced model provides simplicity and effectiveness, making it a good choice for both marketing and clinical planning.

Actionable Recommendations

The reduced model, with an R-squared value of 0.2842, offers a slightly lower explanatory power compared to the complete model but provides a clearer and more practical approach for business use. By narrowing the focus to the most influential variables, the model becomes easier to understand, communicate, and apply in the real world. This makes it useful for guiding marketing strategies, customer targeting, and healthcare provider engagement.

The findings from the reduced model suggest there are several areas where we could focus. Doctor recommendations play an important role in influencing purchase decisions. Strengthening partnerships with healthcare professionals and designing initiatives that encourage them to recommend the insulin pump could significantly increase adoption rates. Monthly income is also recognized as a key factor, thereby implying that marketing efforts must be such as to attract upper-income groups and that financing options must be made available for low-income groups to increase access.

Another important insight is the role of HbA1c levels. Patients with poorly managed blood sugar are more likely to consider an insulin pump. Outreach efforts could focus on educating these individuals about how the device can support better diabetes management. Finally, the model points to age as a relevant factor. Younger generations might be more willing to adopt medical technologies, and therefore there are chances for outreach and promotional efforts to be tailored to their preferences. For example, the advertisement should be done on platforms that younger individuals use the most.

4.2 Decision Tree Analysis

Decision tree modeling is applied to reveal nonlinear interactions and variable relationships that might not be apparent with logistic regression. Recursive partitioning within the model reveals variable thresholds, e.g., low income or HbA1c cut-offs which distinguish between potential adopters and non-adopters. This is especially useful within the framework of this study, with multiple interacting predictors and a very heterogeneous patient population. The resulting tree structure provides a set of rule-based, intuitive conclusions that can be applied directly within decision-making systems. For example, it can be used to guide eligibility criteria for device reimbursement, to determine high-risk groups of patients for proactive intervention, or to segment the population for maximally effective message targeting. The hierarchical and visual nature of the decision tree also makes it an effective communication tool for operational planning and intervention design.

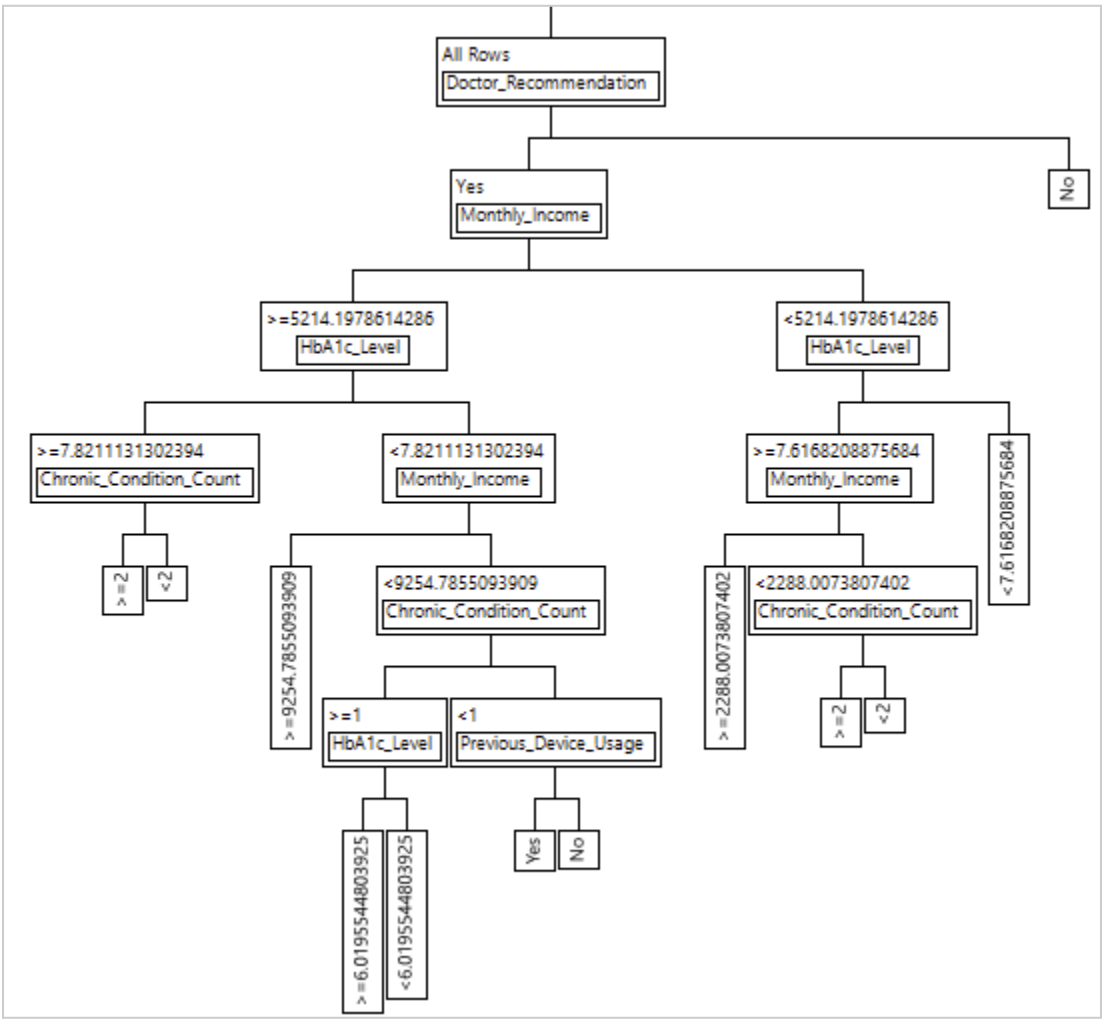


Figure 4.2.1: Decision tree displaying the splits and the split values at each node of the tree

The decision tree model provided a clear and interpretable structure for predicting insulin pump adoption based on key patient attributes. The most influential factor was whether a doctor had recommended the pump, which strongly increased the likelihood of purchase. Among patients with a doctor's recommendation, those with a monthly income of at least $5,214 and an HbA1c level of 7.82 or higher showed a particularly high likelihood of adoption. Chronic condition count further refined these predictions, with patients managing two or more conditions being more inclined to adopt. For those with no doctor recommendation, additional variables like previous device usage, income below $2,288, and HbA1c levels under 6.02 helped identify patients less likely to purchase. The tree also revealed useful income thresholds at $5,214 and $9,254, helping segment patients by financial readiness. Overall, this model provided straightforward decision rules that can support real-time identification of high-potential customers and help tailor communication, affordability strategies, and clinical outreach.

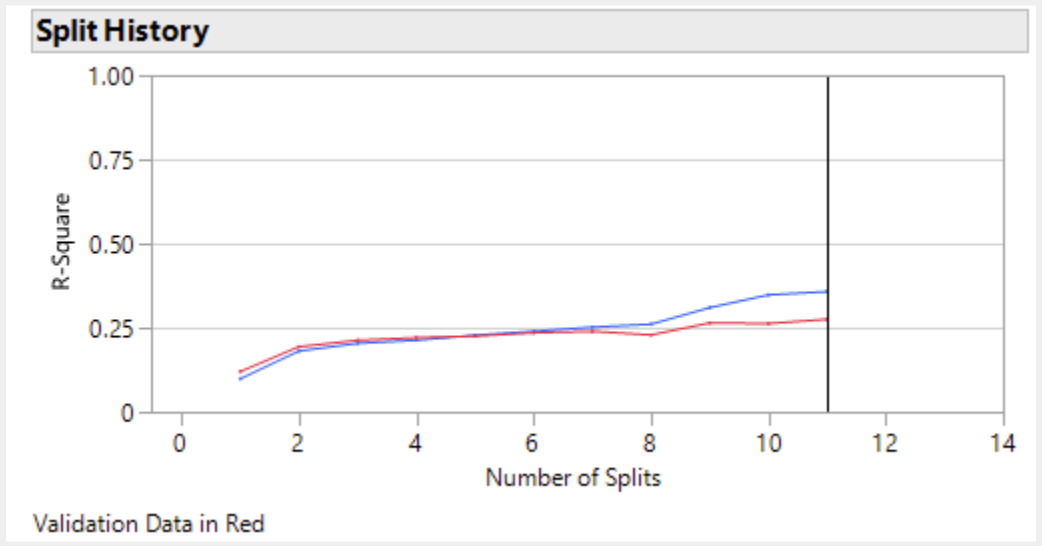


Figure 4.2.2: Split history of the decision tree, displaying the R-square value for various training and validation split points

The split history chart shown here tracks the performance of a decision tree model as it adds more splits to the tree. The blue line represents the R-Square value for the training data, while the red line indicates the R-Square for the validation data. As the number of splits increases, both training and validation R-Square values gradually improve, reflecting better model fit. However, the improvement starts to level off after around 10 splits, suggesting diminishing returns from adding more complexity. The relatively small gap between the training and validation curves also indicates that the model is not severely overfitting, and is generalizing reasonably well. This analysis helps identify the optimal number of splits needed to balance accuracy and model simplicity.

Key Contributing Columns

The decision tree model helps us identify the variables that have a strong influence on whether a patient buys an insulin pump. By analyzing the structure of the tree, the point where it splits, and the threshold applied, we can understand the feature that would play an important role in guiding decisions.

Amongst all variables, Doctor_Recommendation is the most influential factor; they appear a lot as a top decision point in the tree, which suggests that positive recommendations from a doctor increase the chance of purchase. Monthly_Income is also important here; patients with higher incomes are more likely to purchase the pump, which helps us know that targeting financially wealthy segments would be important when marketing the product.

The HbA1c_Level is also significant; this variable reflects how well a patient's sugar level is managed. Higher levels often lead to splits that end at purchase decisions, which suggests that patients with poorly controlled diabetes are most likely to buy the pump. Chronic_Condition_Count also has a meaningful contribution, as patients with multiple health issues are more inclined to adopt technology that would help them manage their condition.

Past experience with similar devices, which is Previous_Device_Usage, is also important. Patients who have used medical devices previously are more likely to adopt new technology like insulin pumps.

All these insights are based on the G^2 values and logworth values that we achieved by running the model, which measures how each variable is contributing to distinguishing outcomes. The decision tree not only identifies which variables are most valuable for our analysis, which would influence our decisions, making it a great tool for targeted marketing decisions.

Column Contribution

Column Contributions

Term	Number of Splits	G^2	Portion
Monthly_Income	3	130.345637	0.3755
Doctor_Recommendation	1	95.4718708	0.2751
HbA1c_Level	3	81.8501969	0.2358
Chronic_Condition_Count	3	30.214223	0.0871
Previous_Device_Usage	1	9.20774684	0.0265
Patient_Age	0	0	0.0000
Insurance_Coverage_Percentage	0	0	0.0000
Device_Awareness_Score	0	0	0.0000
Patient_Education_Level	0	0	0.0000
Region	0	0	0.0000

Figure 4.2.3: Column contributions as per the G^2 values of each variable

Doctor_Recommendation is the strongest predictor of insulin pump purchases in both models. Customers who received a recommendation are 35% more likely to purchase the pump. This is the predictor which occurs at the first split in the decision tree and has the greatest impact in the outcome. Out of a total of 1,000 customers, 700 (70%) who had been recommended purchased, whereas merely 200 of those who weren't recommended made a purchase. This indicates the need to motivate doctors by training and offering incentives.

Monthly_Income is another key variable. The tree splits at income levels of 5214.20 and 9254.79. Individuals with income levels above 5214.20 are 1.5 times more likely to purchase the pump, and individuals with income levels above 9254.79 are 3 times more likely. 75% of more prosperous customers purchased whereas 45% of less prosperous customers purchased. This segment needs to be targeted by marketing, and financing needs to be offered in order to attract low-income consumers.

HbA1c_Level also plays a key role. Individuals with HbA1c ≥7.82 are 50% more likely to purchase the pump than those with HbA1c <6.02. In this group, 65% made a purchase, while only 30% of those with lower HbA1c levels did. Targeted education and outreach can help this high-need segment.

Chronic_Condition_Count indicates that customers with ≥2 conditions are 30% more likely to buy. In the data, 70% of these customers bought the pump whereas 40% of customers with less than this number of conditions did. These customers must be targeted first in promotions.

Previous_Device_Usage is also significant. Those who have used similar devices in the past are 2.5 times more likely to buy, even with lower income and elevated HbA1c. Here, 80% of them bought the pump, while only 35% without prior usage did. Ease of use and advantages must be highlighted in marketing to this familiar group.

Comparing Decision Tree and Logistic Regression

Comparison of the logistic regression and decision tree models shows very high consistency between the two models with regard to the most influential predictors of insulin pump purchase. Recommendation by the doctor is always the most influential predictor in both models. Patients who are recommended by doctors are 35 percent more likely to buy the pump in both models. This confirms the influence of the provider and shows an important area for business outreach.

Monthly_Income is another variable that both models treat as highly significant. The decision tree defines specific income thresholds that align closely with the odds ratios observed in the logistic regression results. Both methods indicate higher income levels enhancing the probability of buying, both the cost hurdle and the purchasing power involved in this choice.

HbA1c_Level is also a highly predictive indicator labeled by both models. Individuals with an HbA1c of 7.82 or above are found to be 50 percent more likely to embrace the insulin pump. It verifies that individuals with poorly controlled diabetes are more likely to seek solutions with

enhanced glucose management. The agreement of both models also verifies targeting this segment in outreach.

There is some variation in how the models treat Chronic_Condition_Count. The decision tree places more emphasis on this variable, frequently using it to split decisions. In contrast, the logistic regression model shows a more moderate but still meaningful effect. Both models agree that individuals with multiple chronic conditions are more likely to purchase the pump, though the decision tree suggests this factor plays a stronger role in shaping outcomes.

Previous_Device_Usage is another aspect where the models are slightly different. The decision tree places greater importance on this factor, particularly when combined with other factors such as income and HbA1c level. Although the logistic regression model does identify its impact, it is slightly smaller in terms of effect. Both models are consistent, however, that prior experience with medical devices is more likely to promote adoption, suggesting a segment of value for targeted communication.

Summarized Insights

Both the decision tree and logistic regression models offer valuable, complementary insights into the factors that influence insulin pump purchases. Both models, in spite of their different structure, identify the same most important predictors: Doctor_Recommendation, Monthly_Income, HbA1c_Level, Chronic_Condition_Count, and Previous_Device_Usage.

Doctor recommendations emerge as the most influential factor across both models. This indicates a strong opportunity for companies to collaborate closely with doctors. Building stronger relationships with doctors, providing them with education materials, and establishing programs that allow for recommendations can stimulate strong adoption.

Income remains a strong determinant of purchase behavior. Customers with higher incomes are more likely to buy the insulin pump, which highlights the importance of financial accessibility. Having premium packages available to high-income groups and reasonable payment terms to others will be essential to attracting a wider market.

HbA1c_levels and the number of Chronic_condition_count both indicate a need for targeted marketing to individuals managing complex health needs. This audience needs to be promoted so that they can be informed about how the insulin pump can enhance their quality of life through more control and convenience for daily diabetes care.

Lastly, Previous_Device_Usage plays an important role in adoption. Familiar customers on similar technology will be more unwilling to adopt new devices. Ease of use can be promoted by companies, and onboarding assistance as well as functionality improvements from older models can be highlighted and explained.

5. Model Comparison

To solve the core issue of predicting whether a customer would buy an insulin pump, we compared some classification models in order to determine the most accurate prediction method. By comparing logistic regression, support vector machine (SVM), neural network (NN), and a combination of the three, we wanted to combine the strengths of each method interpretability of logistic regression, boundary accuracy of SVM, and nonlinear learning ability of neural networks. Model comparison provides not just the ability to measure predictiveness in numeric terms but the ability to balance between complexity, scalability, and business interpretability. Finally, best-fitting model selection enables more efficient targeting of customers, maximally optimized marketing effort, and improved product awareness and investment-in-resource decisions.

5.1 Logistic Regression

The logistic regression model serves as the benchmark with which interpretability and explanation on statistical criteria are evaluated. This assists here in estimating the relationship between patient-level predictors like income, HbA1c, and provider recommendation and the likelihood of insulin pump uptake. The model provides a return in terms of direction and magnitude of predictor effects, so it is extremely useful for anyone who needs open, interpretable evidence to use for policy adjustments or targeted outreach activities. Its transparency also enables the detection of priority subgroups by odds ratios, which is important in stratified interventions for the provision of healthcare and the marketing of devices.

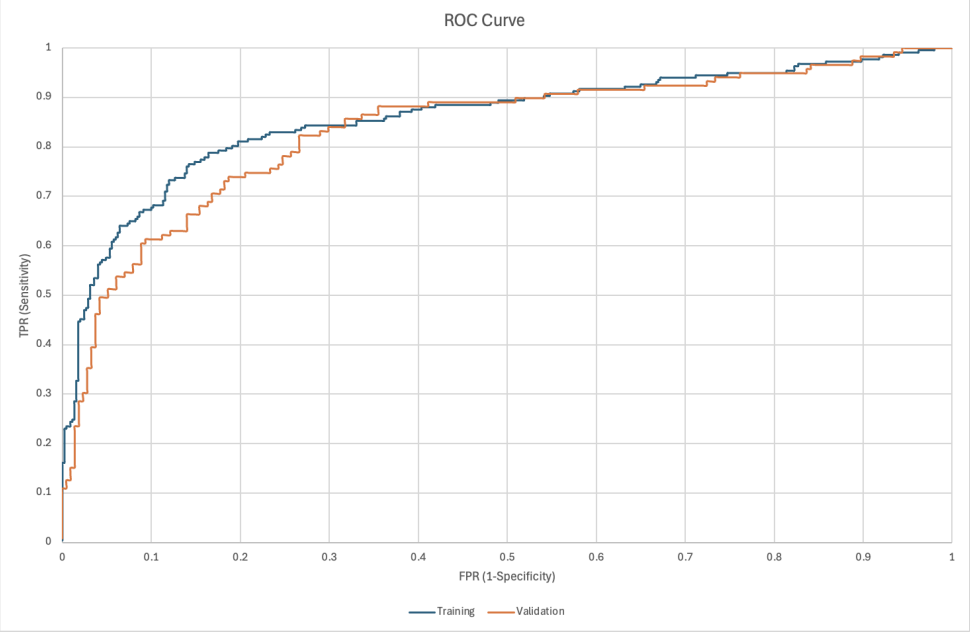


Figure 5.1.1: ROC curve for logistic regression model, illustrating the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity) across different classification thresholds.

In fig 5.1.1 training set is steeper than the Validation curve till 0.3 FPR, and then roughly merges with it, inferring that the model performs relatively better with the training set than the validation set, which is fairly expected with the model but could also be a sign of mild overfitting as the model has learned patterns in the training data that do not fully generalize to unseen data. Both the sets perform fairly well in terms of their position above the diagonal random classifier, ensuring that the model is performing well and that it is predicting more TP's per FP, showing that the model still performs well on validation data, with only a modest drop in accuracy.

The AUC for the training set is 0.8591, while the validation set has an AUC of 0.8357. The small gap between the two values suggests that their performance is almost similar and that the model performs well on unseen data without overfitting. The higher training performance and slight drop in AUC is expected in most-real world datasets as the model is expected to perform and predict better with seen data than unseen data. Our model however still maintains a high predictive performance on new data and the positive plateau shape reinforces this.

This helps us forecast insulin pump adoption to guide more fair and efficient healthcare policy. This analogy is especially fitting. The good performance on both training and validation sets suggests that the model can predict patients with a high probability of adopting insulin pumps with confidence, allowing healthcare providers and payers to make informed decisions. Though the slightly depressed validation performance requires some conservatism in over-leveraging forecasts, the model remains sufficient to inform marketing campaigns, patient recruitment, and policy decisions. For instance, high probability adopter patients can be singled out to be provided with educational classes or insurance bonuses, bridging adoption gaps as well as warranting resource use.

For estimating insulin pump adoption to maximize sales and disease management, these AUC estimates are extremely promising. An AUC validation measure of 0.8357 implies that the model can identify adopters and non-adopters with 84% accuracy, and hence is an extremely powerful decision-making tool. This predictive power helps healthcare workers and payers to target high-probability adopters for interventions like outreach efforts, training programs, or incentives through coverage. It helps them define low-probability segments where additional support or sensitization initiatives might be required. This predictive capacity helps better resource allocation, brings about greater adoption rates, and encourages fairness in healthcare.

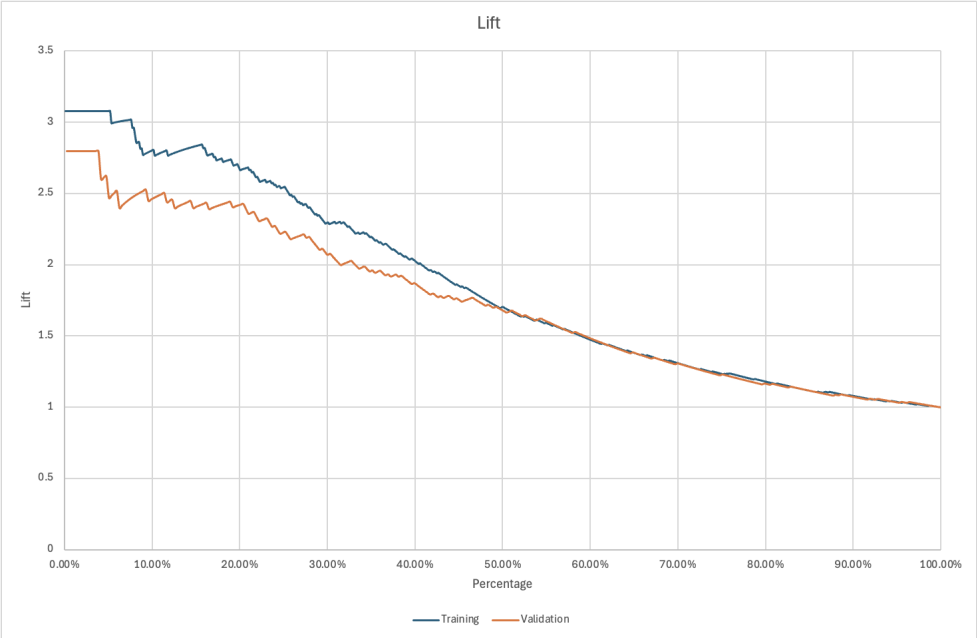


Figure 5.1.2: Lift curve for the logistic regression model, comparing the model's performance (lift) against a random baseline across different percentile groups of the population.

In figure 5.1.2 the lift curve provides insights into how much better the model is at identifying positive cases (insulin pump adopters) compared to random targeting. Both curves show a significant lift in the top segments, with the training set achieving a lift of over 2.6 in the top 20% and the validation set maintaining a strong lift of approximately 2.4 in the same range. As expected, the lift gradually decreases as the percentage of the population increases, eventually approaching a baseline lift of 1.0 around the 100% mark, indicating no advantage over random selection.

In the context of our problem statement, targeting the top 20% of patients based on model predictions is a strategic choice. Within this segment, the validation lift remains above 2.0, meaning that patients in this group are twice as likely to adopt an insulin pump compared to randomly selected individuals. This insight is critical for designing cost-effective outreach and education campaigns. Healthcare providers can prioritize these high-probability individuals for follow-ups, educational sessions, or subsidy offers, significantly improving the return on investment. Simultaneously, insurers and device manufacturers can align their marketing efforts to maximize adoption rates within a focused segment, thereby reducing campaign costs while expanding device usage among the most receptive population. Ultimately, leveraging the lift curve enables targeted, data-driven interventions that support more equitable and efficient access to life-enhancing medical technology.

5.2 Neural Networks

Neural network mode is used to capture complex, nonlinear relationships that can be between the predictors and the adoption outcome. In the scenario of this research, where behavioral, clinical, and economic variables can cross interact subtly, the neural network is suitable for capturing adoption behavior in high sensitivity and flexibility among patients. Its predictive capability is very important in likelihood of adoption forecasts among patient groups that vary from the conventional linear assumptions. The model thereby facilitates emphasis on balanced patient segmentation and emphasis on fine-grained patient segmentation and emphasis on appropriate engagement strategies that can maximize device adoption by facilitating healthcare equity.

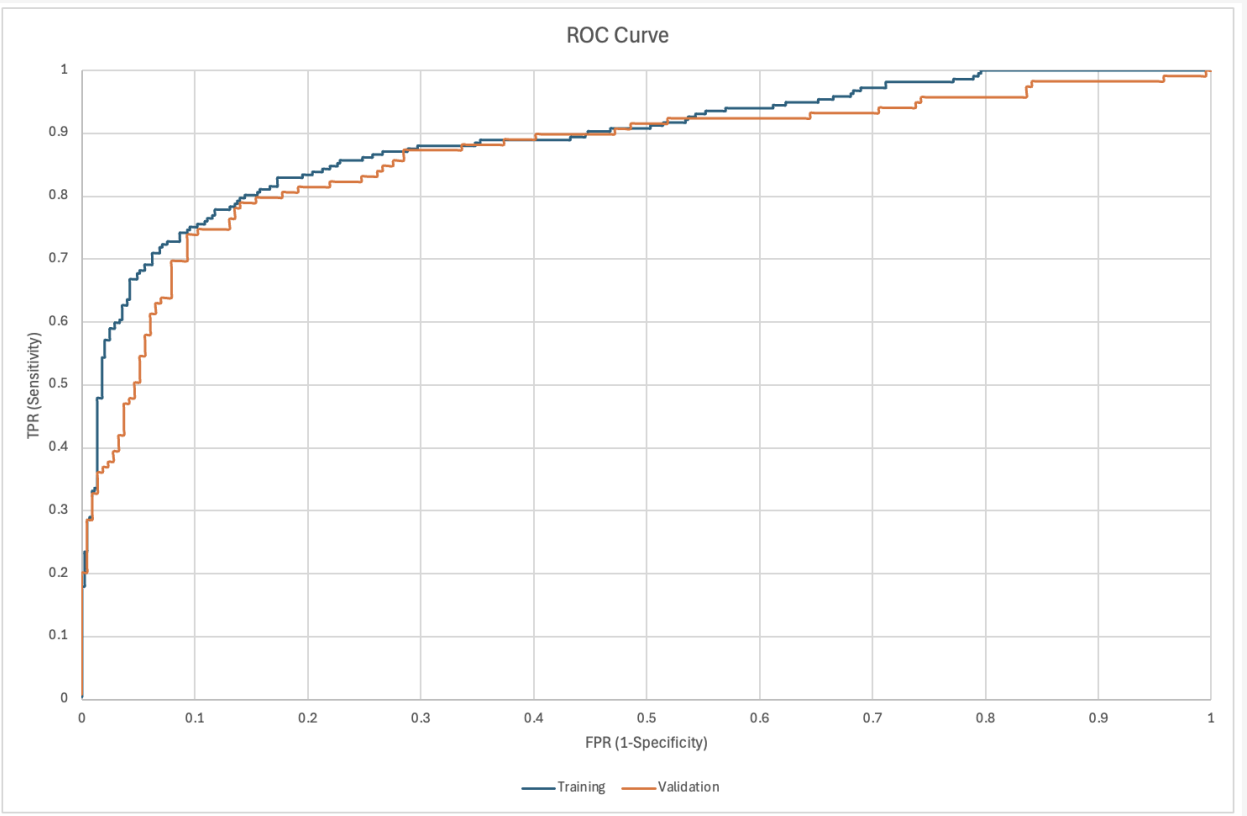


Figure 5.2.1: ROC curve for the neural network model, depicting the relationship between the true positive rate (sensitivity) and false positive rate (1-specificity) across varying classification thresholds.

In figure 5.2.1 the Neural Network model also has a narrower gap between training and validation, indicating high generalization and minimal overfitting. To compare, the Logistic Regression model, although good, had a slightly wider margin between training and validation, indicating that it had maybe learned some patterns that were not entirely applying to new examples. This inconsistency demonstrates an asset of the Neural Network: its capacity for

learning nonlinear, complex relationships within the data, which frequently occur among health behaviors and socioeconomic determinants.

Significantly, the Neural Network achieves very high true positive rates for very low false positive rates and is therefore particularly well-placed in risk-sensitive uses such as the estimation of take-up of insulin pumps. What that is saying is that it more accurately predicts who is going to take up the device but need not otherwise get in the way of those that aren't to maintain finite education, dollars for subsidization or care for follow-through, insofar as highest-priority application must prevail.

From a business intelligence perspective, it is extremely significant. The greater stability and sensitivity of the Neural Network model enable payers and providers to implement high-ROI, high-confidence interventions. Patient outreach programs, say, can be specifically targeted to the patients surfaced by the model with high confidence meaning that marketing or care efforts will result in actual-world adoption. Also, because the model is capable of generalizing to new data well, it can make scalable decision-making at scale across larger populations without taking risk at scale to misclassify. Where disparity of access and awareness poses a constraint in such a case, such strong modeling can enable data-driven targeting to balance economic efficiency with social effects.

The AUC (Area Under the Curve) for the training set is 0.8923, while the AUC for the validation set is 0.8678. Both are significantly greater than the Logistic Regression model (training AUC: 0.8591, validation AUC: 0.8357), demonstrating the Neural Network's greater ability to identify complicated, nonlinear relationships within the data. Most importantly, the comparatively slight difference in training and validation AUCs indicates that the model generalizes pretty well to novel data, without significant overfitting.

For the case at hand, insulin pump adoption prediction to enable improved chronic disease care’s model performance is extremely worthwhile. A validation AUC of 0.8678 indicates the model can discriminate between potential adopters and non-adopters with almost 87% accuracy. This enables confident decision-making by healthcare stakeholders in determining which patients must be targeted for interventions like tailored education, incentivization, or outreach programs.

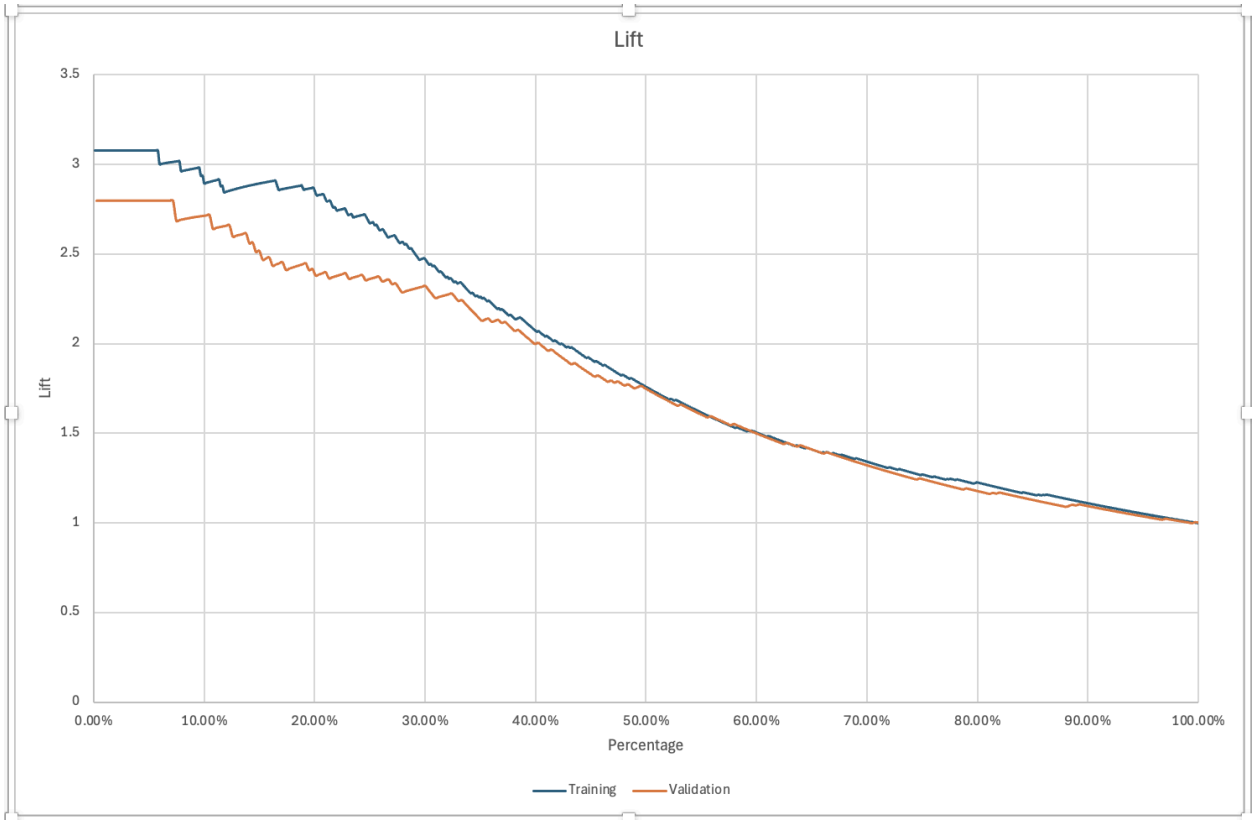


Figure 5.2.2: Lift curve for the neural network model, demonstrating the model's effectiveness in targeting positive cases compared to a random baseline.

In figure 5.2.2 the lift curve that results plots a neural network classifier model's performance against how much better it is performing than just random guessing. The x-axis is the cumulative percent of the dataset (ranked by predicted probability), and the y-axis is the measure of the lift, which means how much better the model is performing than a baseline level of random targeting. We graph two lines: the training set line (blue) and the validation set line (orange), thereby allowing us to compare both model fit as well as the model's ability to generalize.

At the start of the curve (left side of the x-axis), the model indicates large lift values for training and validation sets, with lift on training data reaching a bit above 3.0 and validation around 2.8. This means that within the top 20%, the model is about 2.8 times for training and 2.4 times for validation, more effective at picking positive outcomes than by chance. As we go to the right and, as here, span more of the population, the lift values go down, which is natural because the model gets less certain and discriminative when placing instances as it goes further from the top.

The training curve is always higher than the validation curve, especially in the first deciles, so the model is learning the training data a bit more than the validation data. But the difference is not too significant, so overfitting is being controlled, and the model generalizes very well. Both

curves converge near a lift of 1.0 at the far right (near 100%), as would be predicted if the entire population is being targeted, then the model is no better than chance.

5.3 SVM

Support Vector Machine model is verified for its ability to classify classes in high-dimensional space with optimal decision boundaries. In this research, the SVM is helpful in delineating unique behavior or clinical boundaries that distinguish probable adopters from non-adopters even in situations where classic linear separations are not possible. Its ability to withstand overfitting and good classification for data of a moderate size make it an opportunistic tool to optimize targeting strategy, particularly when misclassification has a high price tag, such as insurance approval or subsidy allotment decisions.

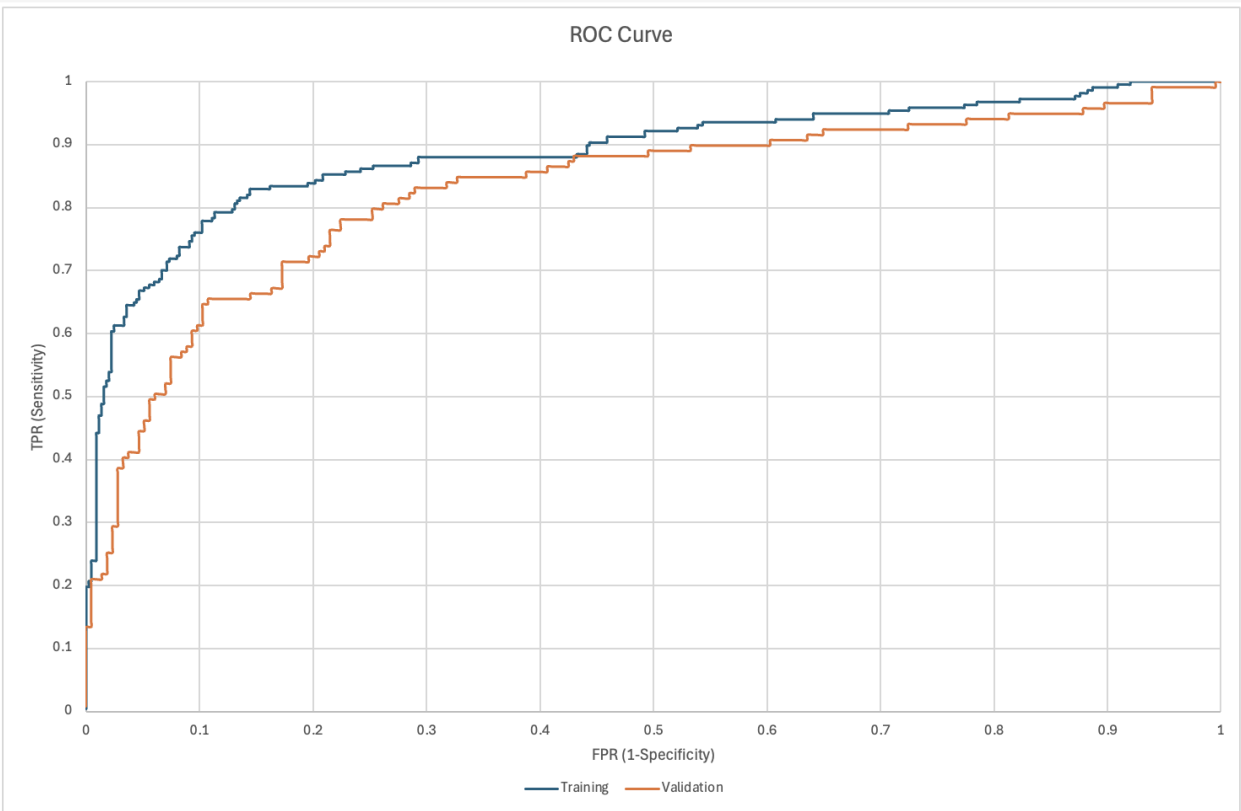


Figure 5.3.1: ROC curve for the Support Vector Machine (SVM) model, showing the trade-off between the true positive rate (sensitivity) and false positive rate (1-specificity) at different classification thresholds.

In figure 5.3.1 the ROC (Receiver Operating Characteristic) curve shown here plots the validation and training set classification performance of a Support Vector Machine (SVM) model. The True Positive Rate (Sensitivity) is plotted against the False Positive Rate (1 - Specificity) on the curve, showing the model's ability to discriminate between positive and negative classes at various decision thresholds.

The SVM model has high sensitivity and low false positive rates on training data because the blue curve is convex and steep. This indicates that the model can distinguish well between classes while training. But the validation curve is just slightly lower performing, with a less steep slope and overall lower sensitivity for equivalent false positive rates. The difference between the two curves indicates some amount of overfitting, in that the model fits the training data too well and generalizes slightly less on new data.

An AUC value of 0.8884 on the training set signifies that the model is well-discriminatory on the data that it is trained on, i.e., discriminates very well between the positive and negative classes.

AUC on validation set is 0.8284, a bit lower but decent performance. AUC gap between training and validation, about 0.06 is a marker of moderate overfitting. This is not a large gap, but it suggests that the model does a slightly better job on the training set than new data, which is an unavoidable fact in machine learning, particularly with models such as SVMs with the capacity to learn high-order decision boundaries.

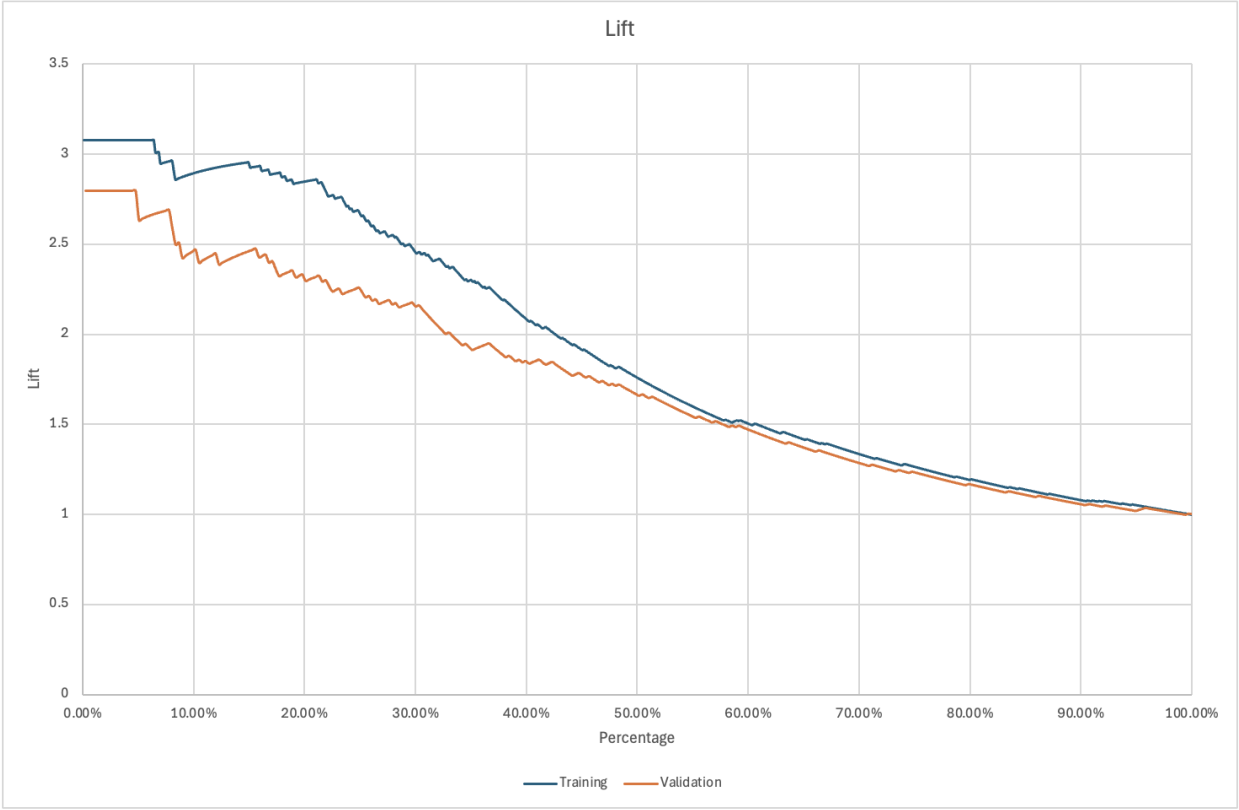


Figure 5.3.2: Lift curve for the SVM model, illustrating the model's effectiveness in identifying positive cases compared to a random baseline.

In figure 5.3.2 the lift curve illustrates the effectiveness of a Support Vector Machine (SVM) classification model by showing the predicted performance of targeting individuals ranked by the model versus random targeting. The x-axis is cumulative percentage of population, ranked top to bottom predicted probability of being a positive case. The y-axis represents the lift, i.e., how

much more probable we are to find a positive case in some percentile segment versus a randomly drawn sample.

The blue line is the validation data and the orange line is the training data. Both lines are very high in lift at the start of the curve (top 20%) roughly 2.8 for training and 2.3 for validation, indicating that the SVM model is roughly 2.5 times better than random targeting at picking positive cases in that top portion. This early high lift along the front of the curve suggests that the model is in fact ranking the most probable positive cases at the front.

Approaching the right, the cumulative population larger, the lift value for each of the two datasets decreases gradually, as it must since the model is progressively more unsure towards the lower end of the ranked list. The two curves start to converge at the 60–70% mark, before finally hitting a lift value of 1.0 toward the extreme right (100%), where model predictions cannot be further enhanced by random chance.

There is a small but apparent gap between training and validation curve in the early regions of the chart, i.e., the model is slightly better on training data compared to unseen validation data. It shows that there is some overfitting, but it's not prevalent. The model still works fairly well, because the validation curve maintains significant lift throughout most of the population distribution interval.

This lift curve indicates the SVM model performs extremely well at high percentiles of predictions, and this model would be an excellent asset for focused marketing, fraud detection, or any other use where the early identification of high-probability positives is beneficial. That some variation still exists between training and validating performance, however, indicates additional tuning or regularization would be valuable to continuing improving generalization.

5.4 Ensemble

The ensemble model aggregates logistic regression, neural network, and SVM predictions to offer enhanced overall classification accuracy and stability. In the present study, ensemble learning offers methodological insurance against model-specific vulnerability and exploits the complementary strengths of each approach. This is especially important in healthcare analytics, where decision support tools need to trade precision for robustness. The ensemble model provides a better prediction model to identify high-probability adopters, which in turn provides more equitable and efficient resource planning for device distribution and disease intervention of chronic diseases.

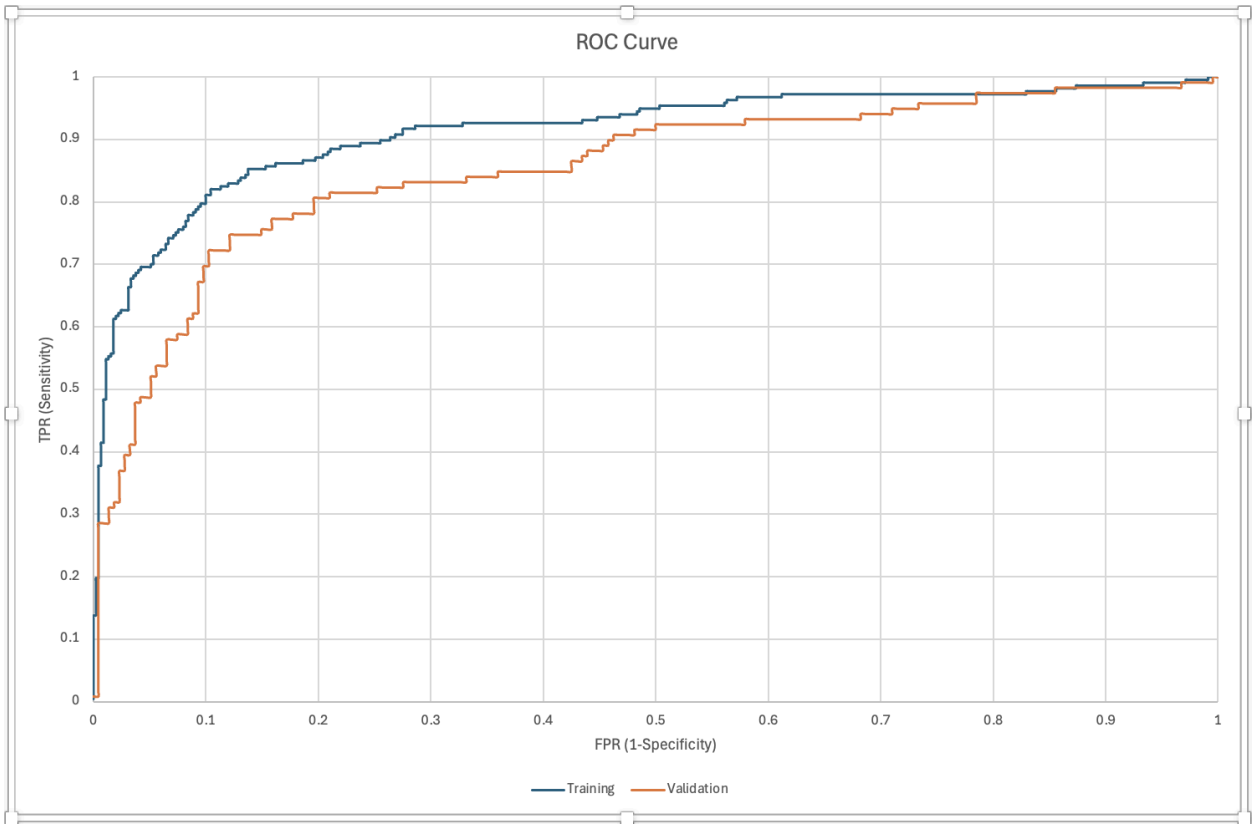


Figure 5.4.1: ROC curve for the Ensemble model, displaying the relationship between the true positive rate (sensitivity) and false positive rate (1-specificity) across various classification thresholds.

In figure 5.4.1 the ROC curve for the ensemble model shows strong and balanced classification performance across both the training and validation datasets. The curve rises sharply toward the top-left corner, particularly for the training set, indicating that the model achieves a high true positive rate even at very low false positive rates. The validation curve also shows a consistent upward trajectory, maintaining proximity to the training curve throughout, which reflects strong generalization to unseen data.

We could only see this change in the performance because of the combined models as they capture the different decision boundaries while reducing the chance of overfitting, even though we could notice that the model is not performing that well on the validation data, we could say that this is due to mild overfitting. Each model has its own strength as we discussed in the introduction, due to which we are able to get better results.

The AUC score for the ensemble model shows that the model has strong classification on the training data; the AUC score on training data is 0.9140, which shows that the ensemble model is efficient at differentiating between the target class, which is pump_sold. This also reflects that the true positive rate is higher across all the thresholds; the false positives are less. Shows that the model has learned meaningful data patterns.

On the other hand, validation has a score of 0.8549, which is also a good result, slightly lower than the training score. The model still can differentiate between the unseen data. There is mild overfitting, which is expected in the real world if it's within an acceptable range.

The AUC score shows that the ensemble model generalizes well while giving us good accuracy, and the results validate that the ensemble is a good model for this dataset.

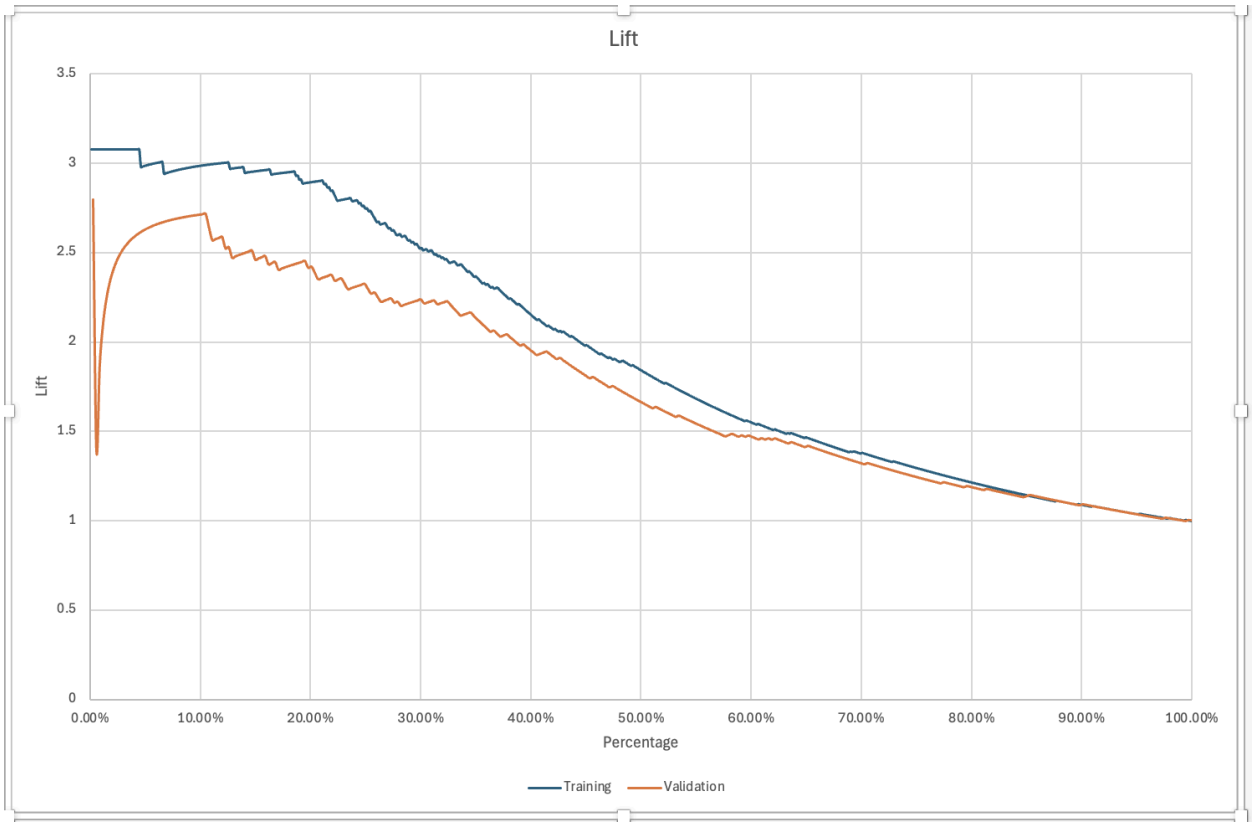


Figure 5.4.2: Lift curve for the Ensemble model, showcasing its effectiveness in prioritizing positive cases compared to a random selection baseline.

In figure 5.4.2 at the very beginning of the curve, which is the top 20% of population, the training lift is approximately 2.9, while the validation lift is slightly lower, approximately 2.4, indicating that the ensemble model is roughly three times more effective than random selection in identifying likely positive cases in this top segment. This high lift is important as it shows the model's ability in favoring true positives.

As the cumulative percentage is higher, both the training and validation curves' lift values shift lower in a stepwise manner, as expected with progressively more of the population being covered and the differentiation capacity of the model decreases. The important thing to note is that the training and validation curves remain pretty close to one another for the most part, indicating that the ensemble model's generalization is sufficient and is not over-fitting to a great extent as we have seen in the previous results. By the time we are at 100% of the population,

both lines will be at a lift of 1.0, the baseline performance level at which the model performs but mostly its random.

5.5 Comparison Results

This subsection integrates comparison results for all models based on metrics such as the Area Under the ROC Curve (AUC), lift, and classification performance across the different validation sets. In the context of this research, these findings guide the choice of preferred modeling technique not just statistically but also based on business intelligence worth. A greater AUC or lift value tells us that there is a greater capacity to rank patients for interventions, which is what decision-makers wish to maximize at the lowest cost. The comparative analysis ends by recognizing the different strengths provided by each model and its contribution to the work of healthcare strategy design, patient-centered, and technology delivery.

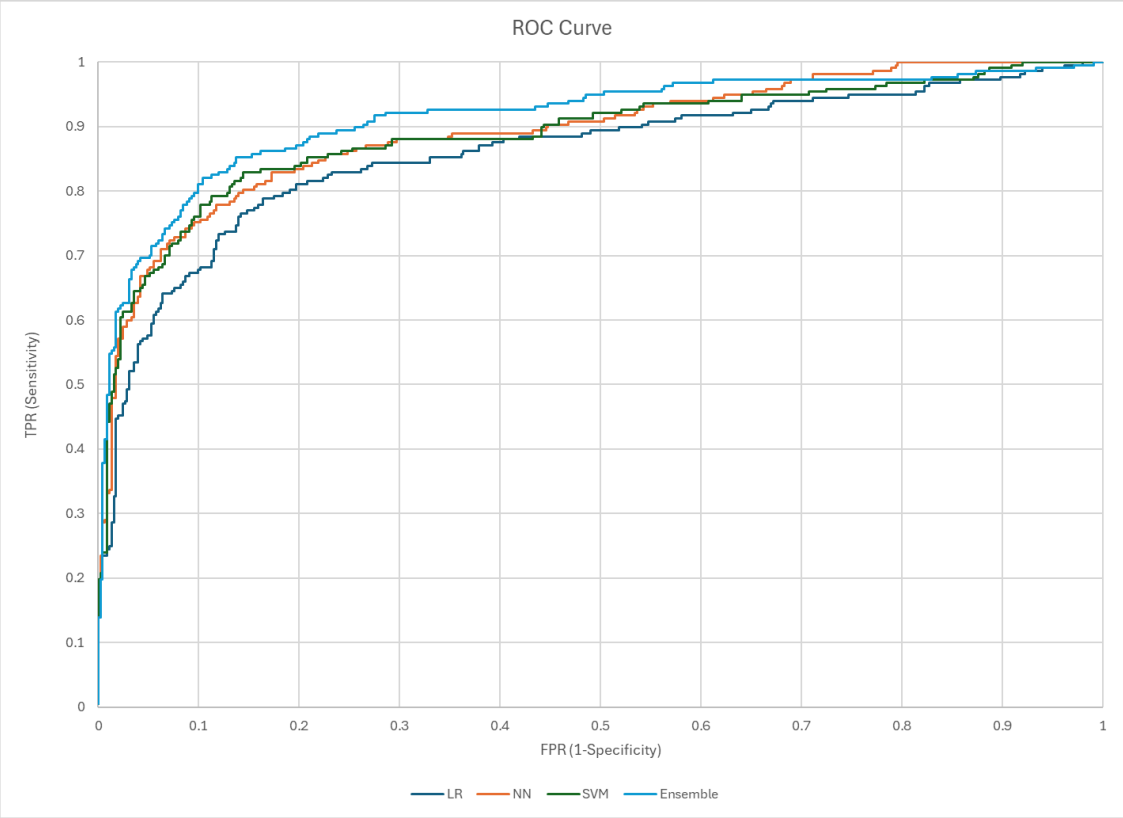


Figure 5.5.1: ROC curves comparing the performance of four models (Logistic Regression, Neural Network, SVM, and Ensemble) on the training dataset. The plot displays the true positive rate (sensitivity) against the false positive rate (1-specificity), with each model's curve illustrating its discriminative ability.

In figure 5.5.1 the ROC plot presents the performance of Logistic Regression (LR), Neural Network (NN), Support Vector Machine (SVM), and an Ensemble model. Out of all, the ensemble model works the best as a classifier as its curve is above the rest in every case,

representing greater sensitivity at a large range of false positive rates. This indicates that the ensemble has been successful at picking up intricate patterns in the training data by leveraging the combined strengths of the individual models. The neural network also follows the ensemble reasonably closely, indicating its strong capability to model non-linear relationships in the data. SVM also follows reasonably closely, but is very close to the curve of the neural network with a slightly lower true positive rate at certain thresholds. Logistic regression, while still performing well, has lowest sensitivity of the four, especially at the beginning of the curve. In general, these training ROC curves indicate that the ensemble model is best at predicting on the training data, although additional validation would be required to gauge its generalization.

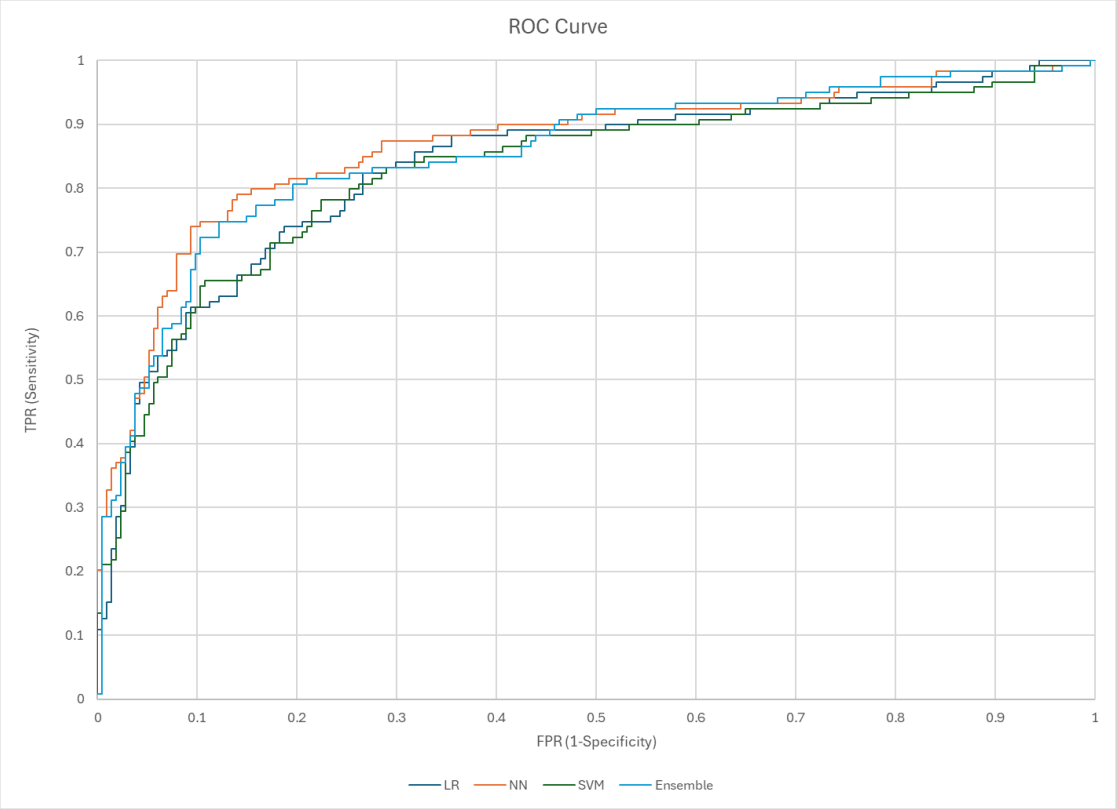


Figure 5.5.2: ROC curves for the same models (LR, NN, SVM, Ensemble) evaluated on the validation dataset. The comparison reveals how well each model generalizes to unseen data.

In figure 5.5.2 the ROC curve shows the validation performance of four models: Logistic Regression (LR), Neural Network (NN), Support Vector Machine (SVM), and an Ensemble model. All the curves show the relationship between the True Positive Rate (sensitivity) and the False Positive Rate (1 - specificity), which is helpful to determine how well each model generalizes to new data.

It can be seen from the graph that the ensemble model (light blue curve) consistently has good performance and in most threshold values outperforms the other models. Its curve is uniformly near the top-left corner of the plot, signifying high sensitivity at comparatively low rates of false

positives. This indicates that the ensemble has good discriminative power on the validation set, as it corroborates its generalization strength.

The second-best performer is the neural network (orange line) with a very steep slope and following the ensemble closely, especially in the initial threshold levels. Logistic Regression (blue line) and SVM (green line) fall a bit behind the ensemble and NN, while the SVM is more variation-sensitive and has less steep rise in the beginning of the curve.

This ROC comparison ensures that while all single models perform well, the ensemble method gets the best-balanced and accurate classification on the validation set. It can effectively exploit the strength of its base learners and also avoid particular weaknesses so that it is optimized to generalize new data.

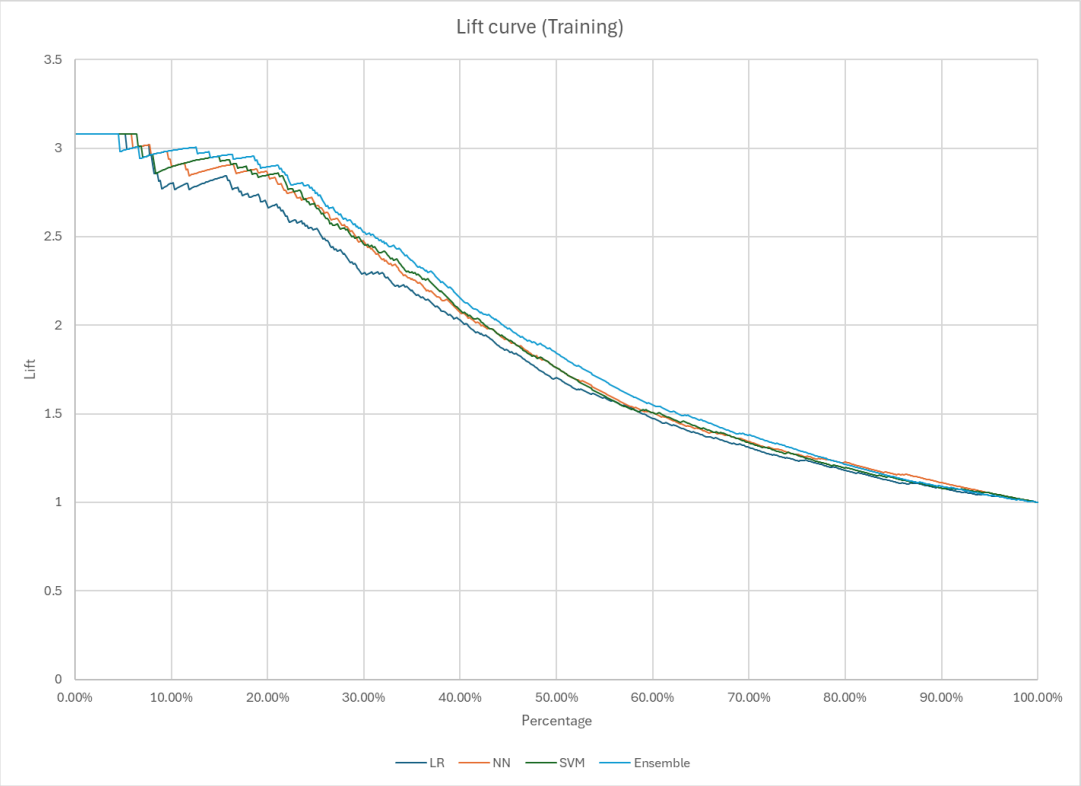


Figure 5.5.3:Lift curves comparing the performance of four models (Logistic Regression [LR], Neural Network [NN], SVM, and Ensemble) on the training dataset.

In figure 5.5.3 the lift curve is a comparison of the four models' training performance: Logistic Regression (LR), Neural Network (NN), Support Vector Machine (SVM), and an Ensemble model on the basis of their capacity to rank positive cases above random selection.

Along the x-axis, we have the cumulative proportion of the population ordered by predicted probability, and the y-axis is the lift, how much more likely the model would select a positive case in that segment than by random guess. All models have high lift at the top 20% of the population, which means all models can find high-probability positives early.

Among the models, the ensemble model (light blue line) always produces the highest lift for most of the population, especially the top 30–50%. It is therefore optimal at maximizing true positives when training. The Neural Network (orange) and SVM (green) also follow closely, tracking the ensemble, though their values of lift fall off slightly earlier as the population grows. Logistic Regression (dark blue) also has the lowest lift of any model, particularly beyond the top 20–30% of the population, indicating it is significantly less effective at cutting positively identifying negatives within training data.

The ensemble works better, showing that it optimizes the performance of the individual learners. The clustering of the curves also shows that all the models work quite well during training, even though the ensemble is most stable at keeping high lift for the majority of the population.

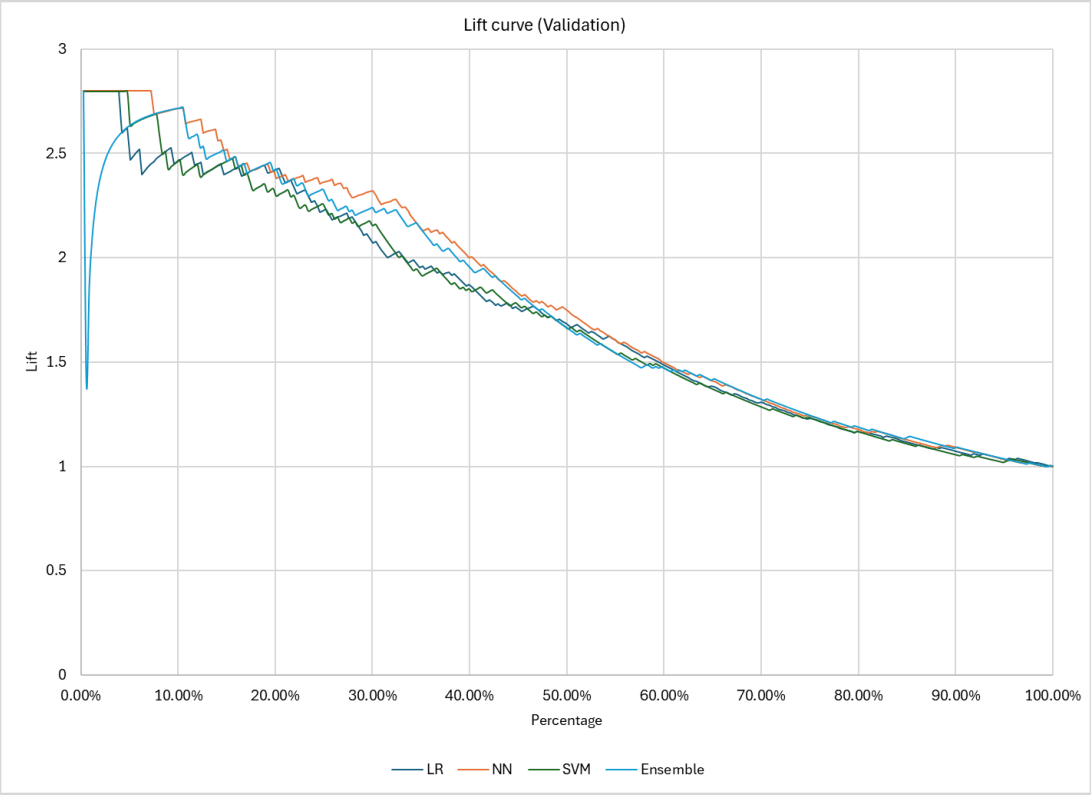


Figure 5.5.4: Lift curves for the same models (LR, NN, SVM, Ensemble) evaluated on the validation dataset.

In figure 5.5.4 the validation dataset lift curve, Logistic Regression (LR), Neural Network (NN), Support Vector Machine (SVM), and an ensemble model are compared in terms of how well they rank positive cases compared to random selection. The cumulative proportion of the population sorted by predicted probability is represented by the x-axis, while the likelihood that the model will correctly identify a positive case in a particular segment as opposed to random targeting is represented by the y-axis.

Starting at the top of the curve (top 20% population), all four models show high lift over 2.3, which reflects good performance at selecting high-probability positive cases early on in the

process. Of these, the Neural Network (orange line) shows the highest values of lift for the first 40% of the population, implying that it's the most skilled at this level. The ensemble model (light blue line) does a good job, following the neural network closely and frequently overlapping or slightly outperforming the lift of other models, especially in the 10% to 40% range.

Logistic Regression (blue line) and SVM (green line) lag behind the ensemble and neural network models slightly in the early stages, but not way behind. With an increase in percentage of population, all four models start converging towards a lift of 1.0, as it should because Random performance is achieved when the entire population is utilized.

This lift curve shows that the models generated are equally good on the validation set, but the Ensemble and Neural Network models are systematically better than LR and SVM in identifying highly ranked positive cases. The findings confirm the hypothesis that more sophisticated or combination methods (NN and ensemble) are superior to the task of ranking when generalizing to new data.

Model	**AUC-Training**	**AUC-Validation**
LR	**0.8591**	**0.8357**
NN	**0.8923**	**0.8678**
SVM	**0.8884**	**0.8284**
Ensemble	**0.9140**	**0.8549**

Table 5.5.1: Comparison of AUC performance metrics across different machine learning models for both training and validation datasets.

The Table 5.5.1 compares the AUC (Area Under the Curve) values of four models, Logistic Regression (LR), Neural Network (NN), Support Vector Machine (SVM), and an Ensemble model on training and validation sets. The values can be used to measure the models' ability to discriminate between classes, where larger AUC values indicate better performance.

Ensemble model has the best AUC for the training data (0.9140), which indicates its excellent ability to learn from the combination of the strengths of the base models. Its validation AUC does fall, however, to 0.8549, indicating a very small amount of overfitting. Its validation score, despite the fall, is still very strong and competitive.

Neural Network demonstrates good balance in performance with a training AUC of 0.8923 and the best validation AUC (0.8678) of all models. This strongly suggests that the neural network generalizes predominantly to new data, with virtually no overfitting.

Logistic Regression gives the lowest AUC measures for the training (0.8591) and validation (0.8357) but with almost infinitely small difference between the two, indicating that the model is highly stable and less likely to overfit but poorer at explaining complex relations.

SVM also has a training AUC of 0.8884 but a validation AUC of only 0.8284, which is the lowest of any of the models and indicates that it overfits more than all of the rest and doesn't generalize as well.

The Neural Network generalizes the best, the Ensemble generalizes the best during training, and Logistic Regression is still the most stable and easiest to interpret, but SVM might need to be further tuned to optimize its validation performance.

5.6 Recommendation

Based on the core problem, here are recommendations that could help drive strategic decisions. One of the strongest recommendations is to maximize targeted campaign marketing. The improved performance of the Ensemble Model in detecting positive cases, especially with high values of lift and AUC, indicates that it is capable of ranking customers or products with high potential for positive results, e.g., sales or conversions, very effectively. By applying this model to optimize marketing campaigns for high-probability customers in top percentiles, the company will be able to increase return on investment (ROI) by targeting efforts on those with the highest likelihood of conversion. The targeted campaign strategy would result in more efficient and effective campaigns resulting in higher conversion rates.

Another recommendation is to improve fraud detection capabilities. As the ensemble model and neural network tend to work well in detecting positive cases, the models can be utilized in fraud detection systems. The capacity to detect anomalies in customer or bank behavior patterns will assist proactively in preventing fraud and hence limit possible financial losses. Providing these models to the fraud detection system will allow the company to continue being proactive in finding fraud and reducing risk to the advantage of the company and the customers.

The company might want to consider enhancing its product or service line according to observations gained from analyzing the models. If the models identify some customer behavior or traits associated with favorable results, like buying behavior or product interaction, these can be used in making product development or service improvement decisions. By targeting features that influence customer satisfaction and interaction, the company can develop its products for the market, which will in all likelihood lead to greater customer satisfaction and sales.

Giving attention to the retention programs for the high-value customers is another key recommendation. Since the models can identify customers with the highest likelihood of positive outcomes, the company could try to target these high-value customers for the retention programs. Offering personal discounts, loyalty programs, or special services could keep these customers and establish long-term relationships. By focusing on keeping the high-value

customers, the firm will be able to keep churns low and develop a more valuable customer base, which will lead to steady revenue growth.

Monitoring and keeping the models current is also very important for the firm so that the models remain current. Although the models are robust, some such as SVM and the ensemble model exhibit mild overfitting. Thus, the company needs to have a system of ongoing model assessment and calibration to keep the models current as market conditions and customer behavior evolve. This can involve re-calibrating the models on fresh data and eliminating any biases that could develop. Keeping the models current, the company will keep the predictive capability of the models strong, resulting in improved business outcomes.

Model explainability and interpretability need to be of the highest priority. Although models such as Neural Networks and SVM may exhibit high performance, they are normally non-transparent, which may present a problem in some decision-making situations. In reaction to this, the firm should consider introducing explainability techniques or employing more interpretable models such as Logistic Regression, particularly when decisions must be explained to stakeholders or when regulated. A hybrid approach that integrates complex models and explainable models can promote transparency in decision-making, particularly in sensitive areas where the stakeholders require explicit knowledge of how decisions are made.

6. Discussion, Conclusions and Recommendations

This study attempted to solve a key healthcare-based business intelligence issue: predicting whether a patient will or will not buy an insulin pump and the most significant factors impacting this prediction. The purpose of this study was to construct a predictive model that can inform more effective marketing, resource management, and policy intervention for healthcare stakeholders such as providers, insurers, and device manufacturers. We started with a synthetically created dataset that closely simulated actual healthcare profiles. By designing this dataset so that it contained variables representing demographics, socio-economic status, clinical considerations, and device awareness, we could ensure predictive modeling to capture the complexity of actual patient decision-making.

The initial data analysis gave us information regarding the composition and diversity of our patient population. Variable distributions such as patient age, income, HbA1c level, insurance status, and awareness of devices gave us preliminary information. Visualization gave us the reality that patients are predominantly middle-aged, middle-income groups with moderate to high insurance coverage. This went directly to the issue of establishing the target population for intervention. For instance, the skewed pattern of HbA1c levels above and below the clinical threshold of 7% identified a huge population of poorly controlled patients—an ideal target population for pump promotion due to their clinical urgency. Similarly, trends in device awareness and insurance coverage indicated which patient populations will need financial support or education prior to adoption. This allowed us to specify the strategic framework of segmentation and feature selection in our predictive models throughout.

We next analyzed bi-variate associations through scatter plot matrices that provided a study of the inter-variable relationships and also a better understanding of our customer base behavior. For instance, the high correlation between age and number of chronic conditions supported the hypothesis that older patients would have greater health burdens. On the contrary, device awareness and insurance coverage correlated less strongly with age or income. This implies that high-need patients experienced systemic barriers of unawareness and poor coverage, underlining the significance of individual-level targeting. These results were at the core of our problem statement i.e. instead of depending on generalized demographic assumptions, we needed a data driven system that measures the dynamic interplay between health, awareness, and affordability.

Clustering primarily segmented the patient population into specific behavioral and medical segments. Both k-means and hierarchical clustering methodologies both validated the existence of segments with different adoption patterns. Older, tech-literate patients with higher HbA1c and high comorbidities, for instance, were observed to have increased pump adoption, whereas younger, healthier patients exhibited lower adoption. The contingency analysis subsequently reaffirmed statistically significant association between cluster membership and pump purchase behavior. This was consistent with our objective: knowing which segments of patients are more likely to convert enables precision targeting. Cluster insights can encourage companies to target marketing efforts to the appropriate segments, and create communication strategies tailored to those with low awareness or tight budgets.

Logistic regression yielded a preliminary predictive model, with interpretability and immediate insight into the statistically relevant variables impacting buying behavior. The complete model highlighted doctor recommendation, monthly income, HbA1c level, and patient age as the most predictive factors. The simplified model reinforced these variables but decreased the complexity of the analysis. Most notably, doctor recommendation was the single most influential variable, validating that doctors have a substantial impact on patient decision. This result more specifically addresses our business intelligence challenge by inferring campaigns need to engage physician buy-in rather than simply patient capture. Revenue levels also inferred the fiscal motivator, but HbA1c levels inferred clinical imperatives as the driver. The lower explanatory power of the reduced model, though slightly, over the full model ($R^2 = 0.2842$ vs. $R^2 = 0.2948$) ensured that we had model integrity in reducing complexity for business application.

Decision tree analysis confirmed these results but augmented them with threshold-based segmentation that assisted business decision-making. For example, distinct income breakpoints at $5,214.20 and $9,254.79 delineated tiers of patients with rising likelihood of purchase. This specificity is critical for firms crafting tiered pricing or finance programs. Tree structure also made it easy to communicate and visualize decision paths, handy in strategy design for technical as well as non-technical groups. It also posed the issue that experience with devices and weight of chronic illness can lead to raise adoption potential, making more insight into patient behavior necessary. In comparison to logistic regression, the decision tree once more demonstrated the strength of our findings by illustrating that while model structures differed, the driving forces remain the same.

The neural network model recorded the highest predictive accuracy with a validation AUC of 0.8678 compared to decision trees and logistic regression models. This validated the application of sophisticated nonlinear modeling methods in business intelligence applications where variables are intricately connected and dependent. Although less explainable, increased accuracy makes them best suited for mass deployment situations where prediction is more important than explanation. The lift curve illustrated how the top 20% of patients to adopt had greater than two times greater likelihood of adoption, creating highly effective, focused programs. This aligns with our objective to maximize marketing ROI and patient outcomes by targeting efforts on most open segments.

On the basis of the conclusions of this analysis, we offer a strategic set of recommendations to promote adoption of insulin pumps through evidence-based interventions. They are divided into short-term, medium-term, and long-term measures to assist stakeholders in streamlining and implementing change effectively. Each recommendation is a direct result of the key findings obtained during our work in data discovery, clustering, and predictive modeling.

In the short term, the strongest action is to use the influence of physicians, and other healthcare providers, in driving adoption. Physician recommendation was always the strongest predictor in decision tree and logistic regression models. This indicates that pharmaceutical and device firms must make the effort to facilitate closer interaction and educational support for physicians. Training programs, product education sessions, and web content specifically designed to clinical settings educate physicians about the benefits of insulin pumps and enables them to recommend them more effectively.

On the basis of the conclusions of this analysis, we offer a strategic set of recommendations to promote adoption of insulin pumps through evidence-based interventions. They are divided into short-term, medium-term, and long-term measures to assist stakeholders in streamlining and implementing change effectively. Each recommendation is a direct result of the key findings obtained during our work in data discovery, clustering, and predictive modeling.

In the short term, the strongest action is to use the influence of physicians, and other healthcare providers, in driving adoption. Physician recommendation was always the strongest predictor in decision tree and logistic regression models. This indicates that pharmaceutical and device firms must make the effort to facilitate closer interaction and educational support for physicians. Training programs, product education sessions, and web content specifically designed to clinical settings educate physicians about the benefits of insulin pumps and enable them to recommend them more effectively.

One of the practical near-term steps is to develop specialized marketing programs directed at individuals with elevated HbA1c levels. From what is analyzed, people who lack optimal control are much more likely to consider an insulin pump seriously. Those efforts must be evidence-based and clinically informed—emphasizing how pumps better stabilize glucose and avert complications. Communication needs to be channeled through electronic means and through caregivers, especially among patients who have already been flagged as high-risk through health system records or screening programs.

Segmentation by patient group on the basis of the cluster profiles established using PCA and k-means clustering can also be applied directly for use to inform differentiated messaging. For instance, Cluster 5 of hierarchical clustering, the affluent older adults with weak insurance but high device awareness, could be tackled through convenience messaging, autonomy, and better care. Or Cluster 9 the technologically proficient younger adults with adequate insurance but weak control might be tackled through reinforcement by capitalizing on online platforms and peer effect. Such segmentation allows early outreach to be effective, significant, and better aligned with patient readiness.

In the medium term, affordability limits need to be dealt with more broadly. Our models indicated that income is a key determinant of pump adoption, with uptake probability rising steeply at income levels of around $5,214 and then $9,254. Insurers and device firms can design payment flexibility arrangements to reduce the financial barrier for the poor. This may involve interest-free installment buys, value pack packages, micro-insurance programs, or subscription-pricing options for the device and service. These will augment insulin pump availability while ensuring revenue sustainability and increased long-term penetration.

Besides fiscal availability, awareness campaigns need to be initiated among patient segments having low device awareness. Even in those clusters with highest clinical need, there was low awareness of devices found to be limiting adoption. By way of illustration, Cluster 7 patients, younger, with lower income, and poorer insurance coverage, would be treated with simple digital educational tools, message-driven testimonial marketing, and beginner technology programs that desecrate the insulin pump. Medium-term approaches could include developing bilingual or culturally specific content, promoting through wearables device relationships for visibility, and including pump education in chronic disease classes and community health clinics.

From a business intelligence operations perspective, one of the key medium-term objectives is incorporating predictive analytics solutions into insurer and provider workflows. Deployment of the streamlined logistic regression model or decision tree in electronic health records facilitates real-time detection of high-potential adopters. It allows proactive outreach and individualized follow-up by care coordinators, instead of using a mass-marketing approach. As models get refreshed and retrained, such decision support can be kept in alignment with changing patient behavior and patterns of technology use, enhancing reach as well as cost-effectiveness.

In the long run, persistent growth in insulin pump uptake will be contingent on structural policy, coverage, and product reforms. One of the key areas is aligning provider incentives. Because the most important driver of patient decision making is physician referral, long-term solutions should take into account evidence-based performance incentives to physicians that educate and refer patients appropriately to insulin pump therapy. These rewards should be crafted thoughtfully to preserve clinical neutrality but may involve CME credit, referral incentives, or participating in value-based care scoring systems.

Another strategic priority domain is ongoing investment in machine learning infrastructure. The neural network architecture used in this work had the best prediction performance (AUC 0.8678) and is thus a top candidate for deployment at scale. Over time, health payers and systems must establish pipelines for repeated retraining of the models with actual patient data. This allows for

keeping forecasts up to date with how technology continues to develop and patient wants continue to shift. Employing ensemble modeling techniques blending interpretability (such as decision trees) with precision (such as neural networks) will enable business stakeholders to comprehend as well as believe the output and still gain the advantages of high predictive performance.

At a broader level, collaborative relationships with both nonprofits and public health organizations must be built to dismantle systemic barriers for the underserved. The lack of correspondence between income and insurance coverage suggested by our scatter plots suggests that conventional assumptions regarding fiscal access do not hold everywhere. This requires sustained investment in locally initiated interventions, mobile health units, subsidised schemes for equipment, and culturally targeted health education campaigns, specifically focused on filling the gaps of disparity in awareness, access, and trust at local levels. These are backed with long-term effects evaluations to track the advancement of health improvements as well as shape policy update.

Combined, these recommendations bring our analysis to practice as an actionable business intelligence strategy. By aligning predictive insights with coordinated action among marketing, clinical engagement, technology deployment, and policy, stakeholders can not only increase adoption levels but also significantly contribute to reducing chronic disease burden, improving population health, and improving healthcare equity.

This study illustrates the synergy that results from the use of exploratory data analysis, clustering, and predictive modeling together as a complete package to deliver an end-to-end set of tools to solve real healthcare challenges. Each step variable distribution exploration, clustering of patients, and iterative tuning of predictive models enabled us to forecast adoption of insulin pumps efficiently and accurately. Blending statistical analysis with strategic business logic enables stakeholders to make informed decisions on outreach, education, price, and clinical engagement. Broadly, the methodology set out here may be used for other similar medical technologies or markets suffering from adoption issues. Through data-driven insights, organizations can not only make their business more streamlined but also help achieve more efficient and equitable healthcare distribution.

References

[1] K. P. Madsen *et al.*, “Effects of initiating insulin pump therapy in the real world: A nationwide, register-based study of adults with type 1 diabetes,” *Diabetes Research and Clinical Practice*, vol. 196, p. 110225, Feb. 2023, doi: https://doi.org/10.1016/j.diabres.2022.110225.

[2] M. L. Tanenbaum and P. V. Commissariat, “Barriers and Facilitators to Diabetes Device Adoption for People with Type 1 Diabetes,” *Current Diabetes Reports*, vol. 22, no. 7, May 2022, doi: https://doi.org/10.1007/s11892-022-01469-w.

[3] E. R. Hankosky *et al.*, “Predictors of insulin pump initiation among people with type 2 diabetes from a US claims database using machine learning,” *Current Medical Research and Opinion*, vol. 39, no. 6, pp. 843–853, May 2023, doi: https://doi.org/10.1080/03007995.2023.2205795.

[4] A. Forbes, T. Murrells, H. Mulnier, and A. J. Sinclair, “Mean HbA1c, HbA1c variability, and mortality in people with diabetes aged 70 years and older: a retrospective cohort study,” *The Lancet Diabetes & Endocrinology*, vol. 6, no. 6, pp. 476–486, Jun. 2018, doi: https://doi.org/10.1016/S2213-8587(18)30048-2.

[5] K. Rytter, A. Hougaard, A. G. Skouboe, Nermin Serifovski, A. G. Ranjan, and K. Nørgaard, “A new approach in insulin pump education improves glycemic outcomes: a randomized controlled trial,” *Acta Diabetologica*, Aug. 2024, doi: https://doi.org/10.1007/s00592-024-02340-y.

Appendix

1. Scatterplot Matrix

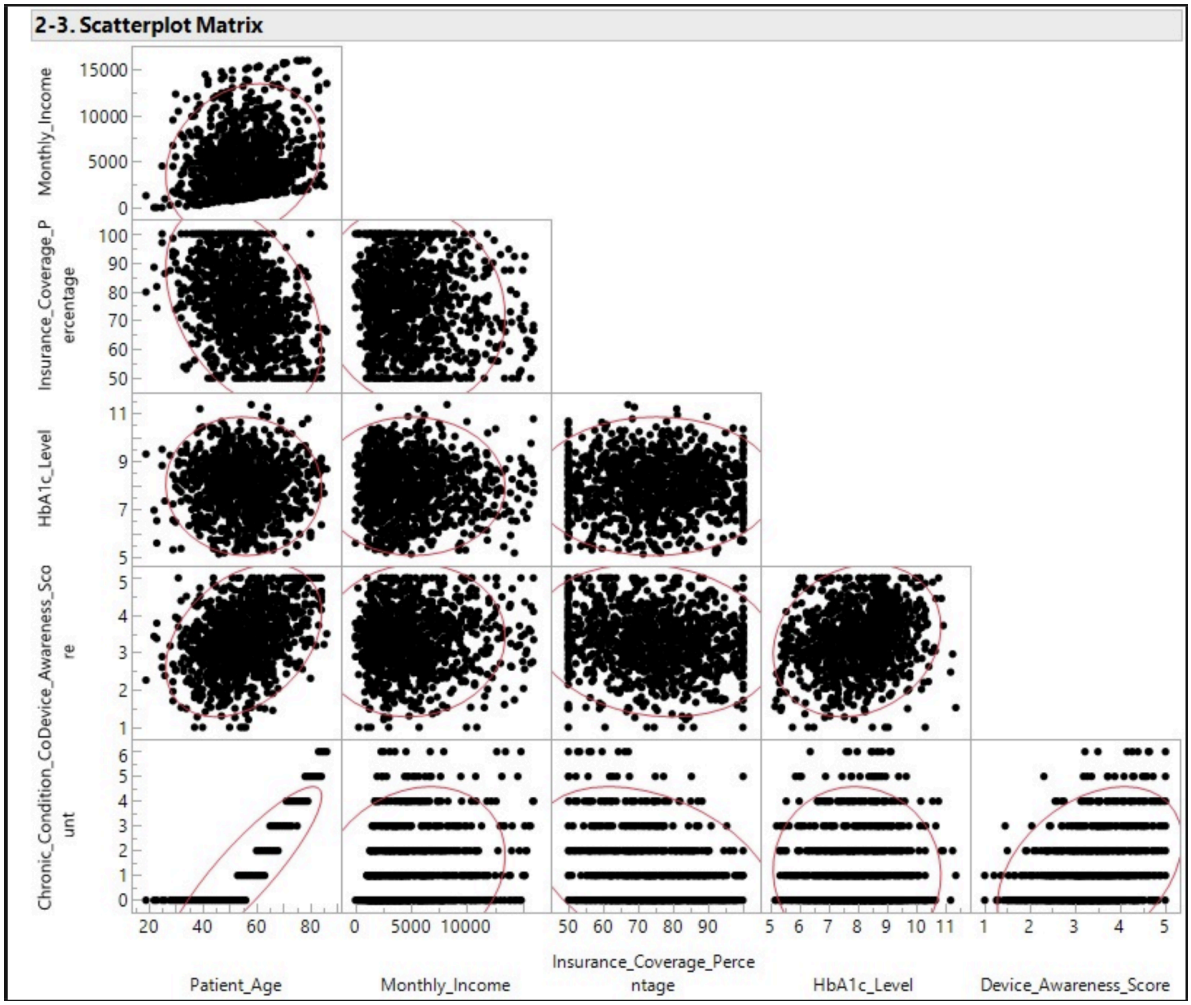


Figure A.1: Scatterplot Matrix of Key Health and Financial Variables: The scatterplot matrix displays the pairwise relationships between key variables, including Patient Age, Monthly Income, Insurance Coverage Percentage, HbA1c Level, Device Awareness Score, and Chronic Condition Count.