IMPLEMENTACIÓN DE MODELOS DE CLASIFICACIÓN BASADOS EN INTELIGENCIA ARTIFICIAL PARA APOYAR EN LA VENTA DE SEGUROS A LA COMPAÑÍA PRUDENTIAL

Integrantes

Mauricio Aguiar Gil

Carlos Zapata Arango

Presentado en la materia Introducción a la Inteligencia Artificial para Ciencias e Ingeniería al profesor Raúl Ramos Pollán

Universidad de Antioquia 2023-1

INTRODUCCION

Los modelos de aprendizaje de máquina basados en inteligencia artificial son modelos basados en aplicaciones de la teoría estadística en los que hay reglas de detección de patrones que se aplican a nivel computacional para terminar almacenando dicha detección en memoria y con esto permitirse el uso de ella para nuevas detecciones de patrones en conjuntos de datos que sean nuevos y congruentes. Es precisamente la mejora en capacidades de procesamiento y de almacenamiento volátil en los computadores modernos el que abre la puerta al uso cotidiano de este tipo de modelos y en el trabajo que se presenta a continuación se expone una implementación de modelos basados en inteligencia artificial para el reconocimiento de patrones en un conjunto de datos.

El conjunto de datos usado en este ejercicio fue creado por la compañía de seguros Prudential, y está alojado en la base de datos de concursos del portal Kaggle. Cuenta con 59.831 registros de aspirantes a comprar seguros a lo largo de 128 variables, entre categóricas y numéricas, que pretenden caracterizar a cada uno de estos aspirantes en un espectro de 8 etiquetas (los números del 1 al 8) que clasificaría a cada aspirante a comprar un seguro. Se busca crear un modelo que le pueda permitir a la compañía de seguros Prudential clasificar con la mayor precisión posible y con el menor sesgo posible a cualquier potencial cliente en esta escala de ocho niveles.

ORIGEN Y OBTENCIÓN DEL CONJUNTO DE DATOS

La compañía de seguros Prudential desea contar con un modelo automático de clasificación que le permita determinar el nivel de aseguramiento requerido para un aspirante a tomar un seguro según una serie de datos que este debe aportar. Estos datos incluyen valores relacionados con la edad y el estado físico del aspirante, su historia laboral, familiar y médica, así como el tipo de seguro que el aspirante desea adquirir.

El conjunto de datos fue tomado del portal de Kaggle y en los notebooks usados para la exploración y el entrenamiento de datos se implementaron bloques de código que permiten la descarga de los datos a un almacenamiento en la nube de Google Drive previo registro en Kaggle y la extracción de los archivos necesarios en dicho almacenamiento.

EXPLORACIÓN DESCRIPTIVA DEL CONJUNTO DE DATOS

De acuerdo con la información aportada por la compañía de seguros Prudential, el conjunto de datos cuenta con:

- Una columna de identificadores por aspirante. El equipo de trabajo consideró que esta columna debe ser removida del conjunto de datos pues el número de identificación de un aspirante no es predictivo de su nivel de riesgo para asegurarlo.
- 60 variables categóricas. Ellas relacionadas con:
    - Información del producto de aseguramiento requerido por el aspirante
    - Información del empleo del aspirante
    - Información particular tomada del aspirante para su aseguramiento
    - Información particular de la historia de aseguramiento del aspirante
    - Historia médica del aspirante
- 13 variables continuas. Ellas relacionadas con:
    - Información relacionada con la historia laboral del aspirante
    - Información relacionada con la historia familiar del aspirante
- 5 variables discretas. Ellas relacionadas con la historia médica del aspirante
- Una columna de clasificaciones realizadas con anterioridad a los aspirantes. Esta columna se consideró la variable respuesta para los procesos de entrenamiento supervisado que se llevaron a cabo.

Y un conjunto de variables dummy que reflejan la presencia o ausencia de ciertas palabras clave en el formulario de solicitud del aspirante.

Yendo más a fondo con la exploración de datos, se encuentran que el conjunto de datos de entrenamiento contiene 59.381 registros correspondientes a la misma cantidad de aspirantes a ser asegurados por la compañía Prudential.

ONE-HOT-ENCODING DE LOS DATOS

Dada la presencia de variables categóricas, se procede a codificarlas para que puedan ser manejadas como variables numéricas. Con este proceso se obtiene un nuevo conjunto de datos con 144 variables, entre los cuales hay variables discretas y continuas.

DISTRIBUCIÓN DE LA VARIABLE RESPUESTA EN EL CONJUNTO DE DATOS PROPORCIONADO

La columna de clasificaciones previamente realizada a los aspirantes representa la variable que los modelos deben predecir. Esta columna aloja 8 categorías en las que cae cada aspirante de manera individual según su perfil de aseguramiento y estas categorías están denotadas con los números del 1 al 8. Se realizó otro análisis exploratorio sobre esta variable para indagar cómo están distribuidas las categorías y el resultado se muestra en la Figura 1. Se observa que la compañía Prudential categorizó previamente a la mayoría de los aspirantes en el nivel 8 (32,8%), seguido del nivel 6 (18,9%) y del nivel 7 (13,5%) cayendo en estas tres categorías casi dos tercios de los aspirantes. Habida cuenta de esto, cabe pensar que hay un riesgo de que los modelos queden sobreajustados dado este sesgo en la distribución de los datos.

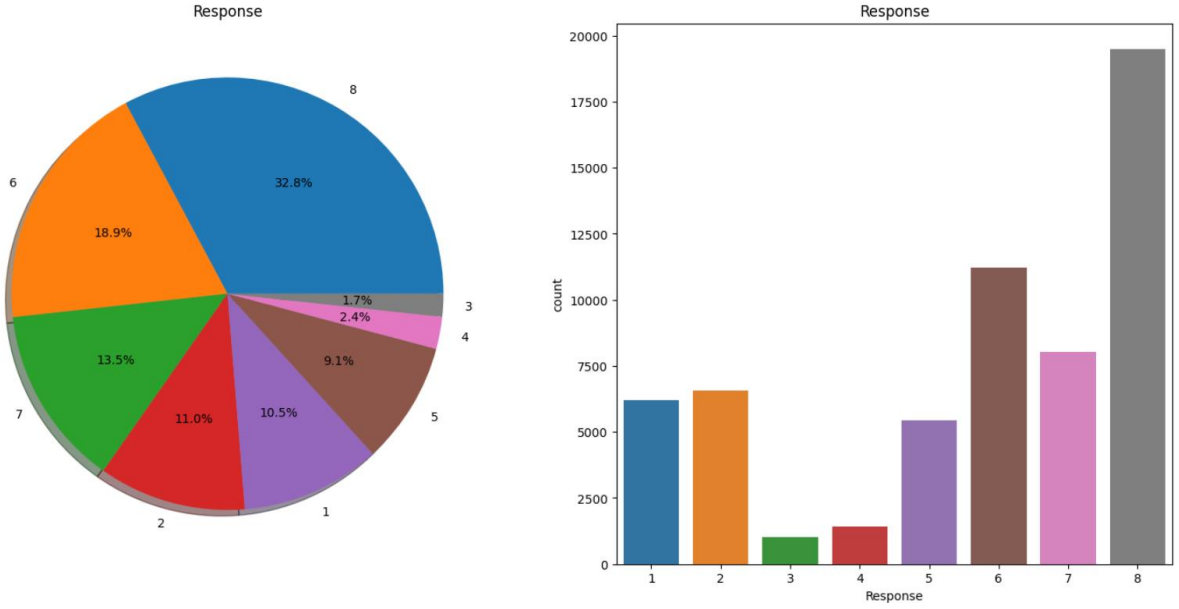

Figura 1. Distribución de las categorizaciones asignadas por Prudential a los 59.381 aspirantes que conforman el conjunto de datos

Se observa también en la Figura 1 que la compañía categorizó a pocos aspirantes en los niveles 3 y 4 (un poco más del 4% del total de aspirantes) este hecho también puede calificarse como sesgo y puede influenciar el desempeño de los modelos de clasificación que se van a crear.

AUSENCIA DE DATOS

En la exploración del conjunto de datos se indagó por la ausencia de datos en una o varias de las variables que lo conforman. Con ayuda de la gráfica mostrada en la Figura 2, la cual es un fragmento de la gráfica original que se encuentra en el notebook correspondiente a la exploración de datos, se logró determinar que el conjunto de datos carecía de valores en varias de sus columnas. Esta gráfica presenta en una escala de colores la presencia o ausencia de datos, siendo el color violeta presencia de datos y el color verde-amarillo la ausencia de datos. De esta indagación surgen las siguientes conclusiones:

- No es necesario hacer eliminación artificial de datos como se instruyó para este trabajo pues el conjunto de datos carece de ellos en varias columnas, las suficientes para cumplir con los requisitos establecidos para el trabajo
- Los datos faltantes deben llenarse con algún método. Se preferirá el reemplazo de datos usando la moda para las variables categóricas y el reemplazo de datos usando la mediana para las variables continuas.

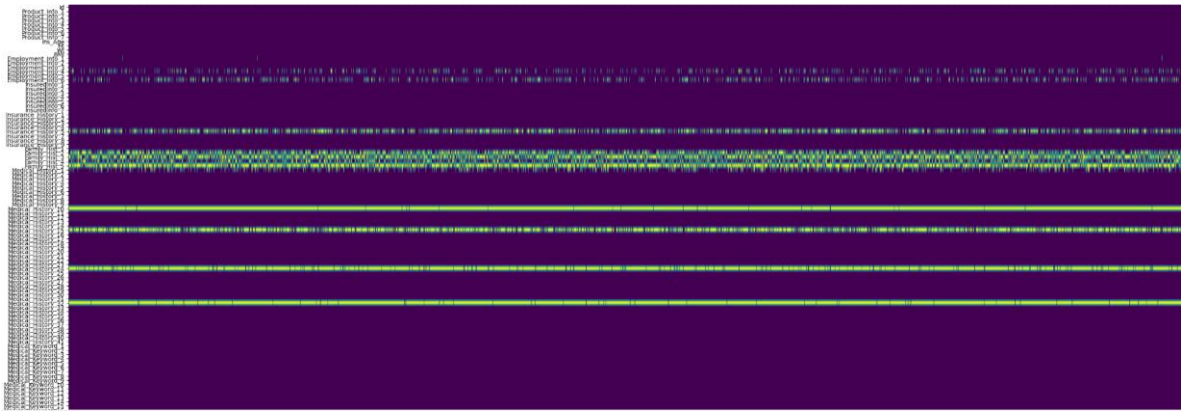

Figura 2 Visualización gráfica de la presencia y ausencia de valores en el conjunto de datos. La presencia de valores se representa con el color violeta y la ausencia de datos se representa con el color verde-amarillo. Cada línea de esta visualización corresponde a una columna (variable) del conjunto de datos

PRESENCIA DE CORRELACIONES

Se construye una matriz gráfica de correlaciones entre las variables del conjunto de datos, que se muestra en la Figura 3. Los colores más oscuros, hacia el violeta y el rojo, indican la presencia de altas correlaciones mientras que los colores más pálidos indican ausencia de correlación entre pares de variables. Se concluye de la figura que las variables no están fuertemente correlacionadas entre sí y que no hay lugar a un posible descarte de variables por encontrarse fuertemente correlacionadas con alguna otra. Se decide trabajar con todas las columnas del conjunto de datos en la tarea de creación de modelos de clasificación.

CONTEO DE DATOS AUSENTES

Se decide realizar un conteo de datos ausentes y se confirma que el conjunto de datos carece de valores en varias casillas. Como se mencionó anteriormente, este conjunto de datos cuenta con 59.381 datos y con aproximadamente 80 variables, lo que conformaría un total de 4’750.480 valores, de los cuales el proceso de conteo de valores ausentes arroja un resultado de 393.103 valores faltantes.

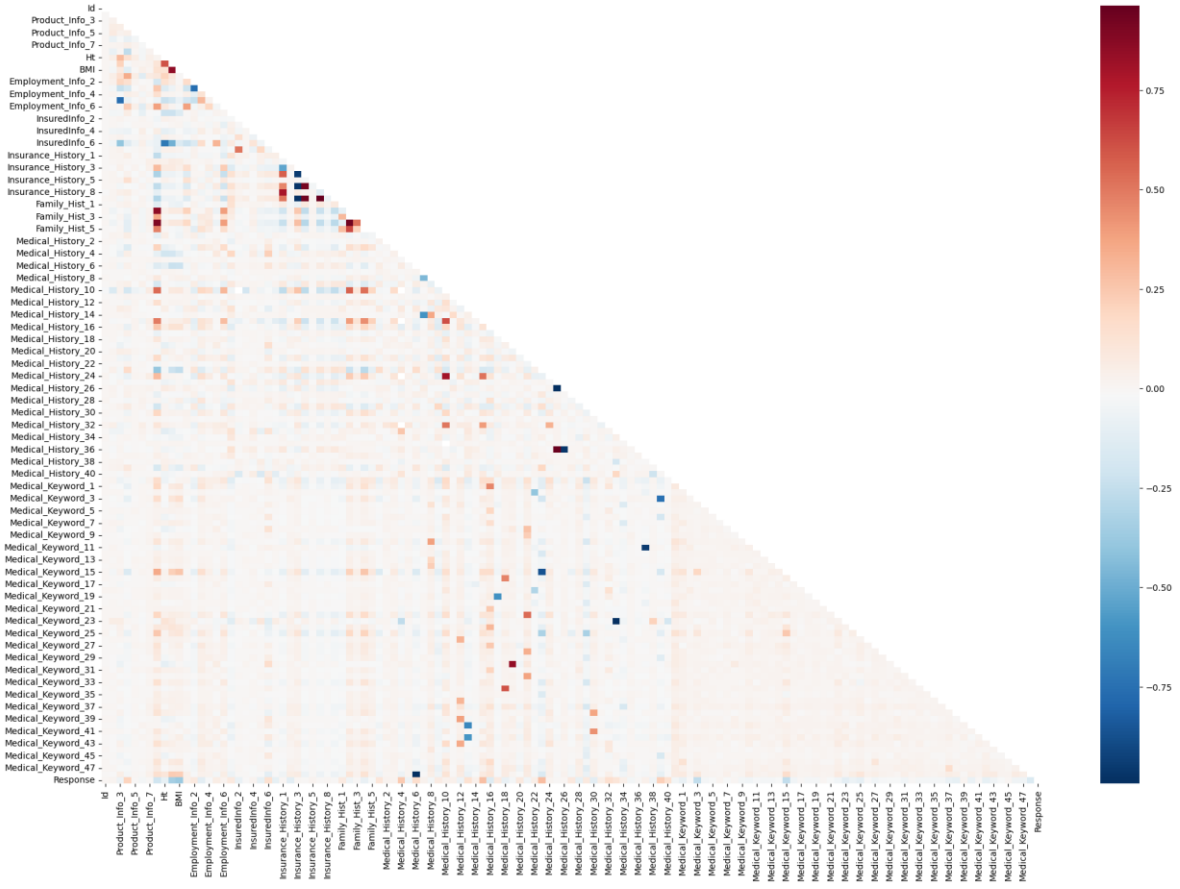

Figura 3 Matriz de correlación (positiva o negativa) para las variables del conjunto de datos. Se observa que en general la mayoría de las variables no están correlacionadas entre sí. Los colores oscuros (rojizo o azul) representan respectivamente correlaciones positivas o negativas.

PREPARACIÓN DE LOS DATOS

Antes de proceder a entrenar y validar modelos de clasificación (supervisados y no supervisados) se procedió al siguiente proceso de preparación de los datos:

- Llenado de valores faltantes
  En cada columna donde hubiese variables categóricas con valores faltantes, se llenaron estas casillas con la moda de los valores presentes en la columna. En cada columna donde hubiese variables discretas o continuas con valores faltantes, se hizo el llenado de casillas vacías con la mediana de los valores presentes en la columna.

- Codificación de las variables categóricas
  Se codificaron las variables categóricas usando un algoritmo de one-hot-encoding como el que ofrece la librería scikit-learn. Con esto, el número de columnas a procesar subió a 146, en las que las nuevas columnas contienen valores de cero o uno según la presencia o ausencia de estas variables categóricas en el caso específico de cada registro (fila) del conjunto de datos

- Separación de columnas del conjunto de datos
  Se desprendieron dos columnas del conjunto de datos: La primera columna, que contiene la identificación de los aspirantes a ser asegurados; como no se considera que la identificación de una persona sea predictiva de su calidad de aseguramiento, esta columna se desprende del conjunto de datos para no ser usada. Se desprendió también la última columna, que contiene la clasificación realizada a cada uno de los aspirantes a ser asegurados porque esta columna es la columna correspondiente a la variable de respuesta que se usa para entrenar modelos supervisados.

ENTRENAMIENTO Y VALIDACIÓN DE MODELOS SUPERVISADOS

De manera general, los modelos escogidos se entrenaron y validaron usando el método de partición y validación cruzada (k-fold cross validation) repetido 5 veces para todos los modelos. Excepto para el caso del modelo basado en redes neuronales, se calcularon métricas de desempeño en la clasificación para todos los modelos (contraste entre modelos entrenados y la partición de los datos de respuesta usados como test) en cada una de las cinco iteraciones de la validación cruzada Para este problema de clasificación se escogieron los siguientes modelos:

- Redes neuronales
- Máquinas de vectores de soporte para clasificación (SVC)
- Random forest
- Regresión logística

Un recuento de los pormenores básicos de las implementaciones para cada uno de estos cuatro clasificadores:

- Redes neuronales
  Se implementó una red neuronal de cuatro capas con la función tangente hiperbólica (tanh) como función de activación en las dos capas ocultas y la función sigmoide en la capa de salida. La cantidad de unidades para cada una de las dos capas ocultas se escogió de manera aleatoria como un número entero mayor o igual a 5. La función de costo calculada para esta implementación fue la función sparse categorical cross-entropy. El ajuste se realizó con un número de 3 épocas en cada iteración de la validación cruzada.

- Máquinas de vectores de soporte para clasificación (SVC)
  Consultando la documentación de la librería scikit-learn se encontró que para los problemas de clasificación multiclase es preferible el uso de la funcionalidad LinearSVC que viene configurada por defecto para enfrentar el problema de clasificación binaria (para el que está

diseñado el modelo de máquinas de vectores de soporte) a través del método “Uno contra el Resto” (OvR) en el que se estudia la separación de cada etiqueta respecto a las demás tantas veces como etiquetas haya. Dado que en ensayos previos la interfaz de Google Colab informaba que el ajuste del modelo no estaba convergiendo, se tomó la recomendación de aumentar la cantidad máxima de iteraciones desde su valor por defecto (1000) a 2000.

- Random Forest
  Se usó un modelo basado en 250 estimadores. No hubo ningún criterio particular que fundamentó la toma de esta decisión.

- Regresión logística
  Debido a que el modelo de regresión logística es un modelo de clasificación binario como el modelo basado en máquinas de vectores de soporte, también se debió recurrir a una estrategia de clasificación de “uno contra el resto” (OvR) A modo de hiperparámetros, se utilizó el optimizador lbfgs.

RESULTADOS OBTENIDOS CON LOS MODELOS DE CLASIFICACIÓN SUPERVISADOS

En las Tablas 1, 2 y en la Figura 4 se resumen los resultados obtenidos. De todo esto se concluye que el desempeño menos malo para este problema de clasificación se logró con los modelos basados en random forest y regresión logística, no obstante, el primer modelo cuenta con un accuracy promedio de apenas el 58% mientras que el modelo de regresión logística cuenta con un accuracy promedio de 52%

Acto seguido, se procedió a agregar un procesamiento adicional a los datos aplicándoles reducción de dimensionalidad con PCA a las siguientes dimensiones: 16, 48, 80, 112 y 144 (el conjunto de datos original tiene 145 dimensiones después de aplicar one-hot-encoding a todas las variables categóricas) y en cada una de estas iteraciones de PCA se entrenaron nuevamente los cuatro modelos ya mencionados, calculando solamente el accuracy promedio para cada uno de ellos. En la Tabla 2 se muestra el resultado en accuracies para cada uno de estos modelos y en cada una de las reducciones de dimensionalidad a 16, 48, 80, 112, 144 y ninguna reducción respectivamente. Se observa de este nuevo proceso de entrenamiento con datos de dimensionalidad reducida que los desempeños menos malos para este problema de clasificación se logran con los modelos basados en regresión logística y random forest, tal como se observó en el proceso anterior sin reducción de dimensionalidad con PCA; además se observa que las reducciones en dimensionalidad no se traducen en mejores accuracies para ninguno de los modelos usados.

Tabla 1 Métricas de desempeño de los cuatro modelos supervisados usados para clasificar aspirantes a aseguramiento por la compañía Prudential.

Modelo	Accuracy Promedio	Pérdida Promedio	Recall Promedio	F1-Score Promedio
Redes neuronales	0,26	2,41	N.D.	N.D.
Máquinas de vectores de soporte	0,51	N.D.	0,38	0,38
Random forest	0,57	N.D.	0,49	0,50
Regresión logística	0,52	N.D.	0,43	0,43

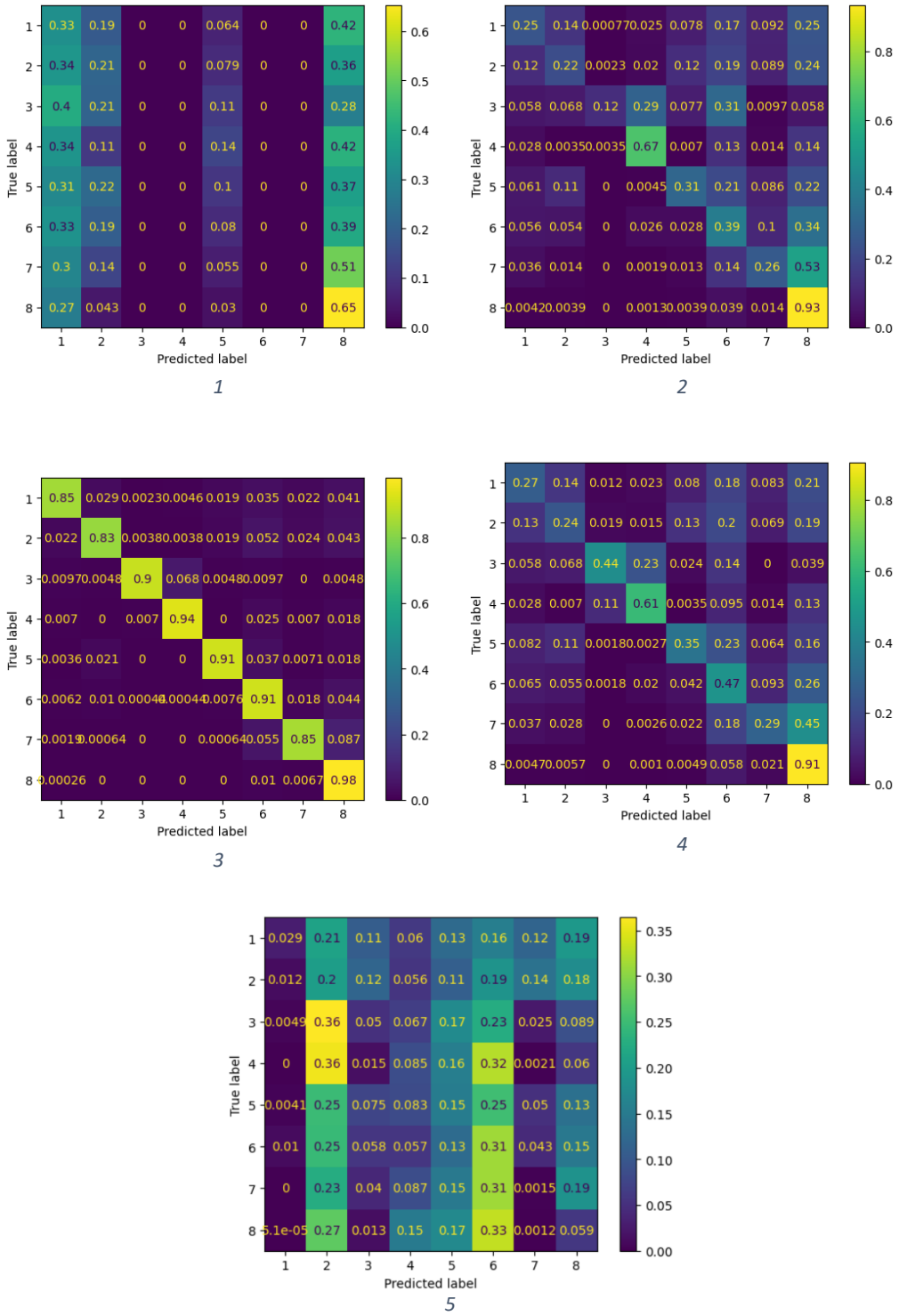

Figura 4. Matrices de confusión para los cinco modelos implementados para este trabajo, cuatro supervisados y uno no supervisado. De la izquierda y arriba hacia la derecha y abajo las matrices de confusión corresponden respectivamente a redes neuronales (1), máquinas de vectores de soporte (2), random forest (3), regresión logística (4) y K-means (5). Obsérvese que por lo general todos estos modelos tienden a predecir en exceso varias etiquetas y que los de mejor predicción son random forest y regresión logística.

Tabla 2 Accuracies para los cuatro modelos supervisados entrenados con nuevas variables determinadas con reducción de características PCA a 16, 48, 80, 112 y 144 componentes. Se compara con el espacio de características original.

Modelo	16 componentes	48 componentes	80 componentes	112 componentes	144 componentes	Sin reducción (145 componentes)
Redes neuronales	0,256	0,251	0,248	0,250	0,261	0,255
Máquinas de vectores de soporte	0,417	0,448	0,467	0,499	0,504	0,505
Random forest	0,449	0,464	0,465	0,479	0,479	0,572
Regresión logística	0,429	0,462	0,487	0,520	0,523	0,523

ENTRENAMIENTO Y VALIDACIÓN DE MODELOS NO SUPERVISADOS

Se escogió un modelo de tipo K-means para ser entrenado y validado con las entradas del conjunto de datos. Se escogió aplicar validación cruzada con k = 5 en el entrenamiento y el test para ganar certeza de la buena distribución de los datos y de comportamiento de los modelos entrenados independiente de la partición de los datos en conjuntos de entrenamiento y de test, también se generó un reporte de métricas (accuracy, recall, F1-score) en cada iteración de la validación cruzada para llevar dicho control del comportamiento del modelo frente a la diversidad de los datos y se encontró que este modelo de tipo K-means no tiene en principio buenas métricas de desempeño: Los valores de accuracy, recall y F1-score son muy bajos comparados con los obtenidos con los modelos supervisados, incluso menor que el modelo basado en redes neuronales cuyo desempeño fue el más bajo. Estos resultados se presentan en la Tabla 2.

Tabla 3 Tabla 4 Métricas de desempeño del modelo no supervisado basado en K - means para clasificar aspirantes a aseguramiento por la compañía Prudential.

Modelo	Accuracy Promedio	Pérdida Promedio	Recall Promedio	F1-Score Promedio
K - Means	0,12	N.D.	0,11	0,09

Esto puede deberse a que el modelo de tipo K-means etiqueta los aglomeramientos que se le pide identificar (en el caso de este problema, 8 aglomeramientos correspondientes a 8 etiquetas del 1 al 8) y los nombra de manera arbitraria, es decir, por ejemplo, el aglomeramiento que es etiquetado como “5” por el algoritmo de K-means puede tratarse en realidad de alguna de las etiquetas del problema (ser la etiqueta “2” en lugar de la “5” o ser la etiqueta “6” en lugar de la etiqueta “5”, …)

Para intentar dilucidar esta situación, se construyó una matriz de confusión para este modelo predictivo y no supervisado de clasificación. Dicha matriz es mostrada en la Figura 4. Los resultados de graficar dicha matriz muestran que el modelo de K-means no logra realizar una discriminación clara entre las etiquetas y por ende no goza de buena capacidad predictiva.

En la Figura 4 se muestran las matrices de confusión para los cinco modelos aplicados:

- Modelos supervisados sin previa reducción de dimensionalidad con PCA (Redes neuronales, máquinas de vectores de soporte, random forest y regresión logística) totalizando cuatro.
- Modelos no supervisados (K-means con K = 8 pues se trata de 8 categorías que se quieren caracterizar) totalizando uno

Se observa de todas las matrices de confusión (exceptuando la primera de la segunda fila, correspondiente a random forest) que los modelos tienden a excederse en la predicción de algunas etiquetas, particularmente las etiquetas 6 y 8 (aquellas que en la Figura 1 mostraban alta frecuencia de aparición entre los datos de respuesta para el entrenamiento)

CONCLUSIONES

- Se observa que en general los modelos implementados, sin las transformaciones con PCA y con distintas transformaciones usando PCA, no logran alcanzar un alto nivel de precisión, considerando “alto nivel de precisión” a un nivel superior al 70% o por lo menos al 60%
- De los modelos implementados, los que alcanzaron mejores métricas de clasificación (accuracy, recall, F1-score) fueron el modelo basado en random forest y el modelo basado en regresión logística
- El procesamiento previo de los datos con PCA de los modelos supervisados no redundó en mejoras en el accuracy de ninguno de ellos por lo que considera mejor abordar este conjunto de datos sin reducir sus dimensiones.
- Se observa en cuatro de los cinco modelos implementados que tienden a predecir en exceso dos o tres de las etiquetas buscadas (particularmente las etiquetas ‘2’, ‘5’, ‘6’ y ‘8’) que justamente están entre las etiquetas más prevalentes en el conjunto inicial de datos de respuesta (y) lo cual puede indicar o que los modelos han sido sobreentrenados o que el conjunto de datos adolece de un fuerte sesgo.
- El hecho anterior resalta un poco más si se ve desde la luz de la implementación basada en k-means, donde se busca encontrar aglomeraciones distinguibles en el conjunto de datos y definir esas aglomeraciones como las clases o etiquetas, sin embargo la matriz de confusión obtenida (la número 5) insinúa que no hay distinguibilidad entre aglomeraciones, apuntando a que posiblemente las categorías son difícilmente distinguibles dados los datos.
- Debido a que hubo tantos datos faltantes de varias columnas, cabe pensar que el proceso de imputación realizado a ellas basado en utilizar la moda en datos categóricos y la mediana en datos no categóricos pudo haber influido en el sesgado del modelo. Se considera importante considerar otras maneras de llenar estos valores vacíos tales como el ‘ffill’ del método de pandas ‘filllna()’
- A modo de reflexión, se plantea la cuestión de la calidad de los datos como elemento crucial para la obtención de modelos de calidad. El nivel de sesgo, de validez, de veracidad, de ruido o de incompletitud de los datos influyen en la satisfacción con los modelos implementados.