Predicción de gastos de la atención médica a partir de algoritmos de Machine Learining para la toma de decisiones basadas en datos en empresas de seguros de salud.

(Prediction of health care expenses from machine learning algorithms for data-driven decision making in health insurance companies)

Enrique Hernández-Laredo
Instituto Tecnológico del Petróleo y Energía / Facultad de Medicina Universidad Autónoma del Estado de México
Toluca de Lerdo, México
ehernandezl190@alumno.uaemex.mx

I. ENTENDIMIENTO DEL NEGOCIO:

A. Contexto del problema:

A partir de una base de datos publica (1) que contiene el registro de datos personales (edad, sexo, zona residencial, número de hijos), y datos relativos a la salud (índice de masa corporal, consumo de tabaco) se postula calcular los gastos médicos individuales facturados por un seguro de salud.

B. Objetivos de negocio:

El aumento en los costos de la atención médica en años recientes genera la necesidad justificada de tener modelos predictivos precisos, ya que desde el punto de vista económico las correctas predicciones en los gastos de servicios de atención medica podrían ayudar con la planificación general de los negocios (2,3), por ejemplo, a una aseguradora le permitiría estimar el valor de las primas de seguro a cobrar. Sin embargo, no solo las empresas aseguradoras o prestadoras de servicios de la salud podrían beneficiarse de los modelos predictivos de gastos médicos, puesto que los pacientes podrían realizar una compra informada del servicio de salud a partir del conocimiento de deducibles y primas apropiadas (2).

C. Objetivos de ciencia de datos:

Se plantea realizar un análisis exploratorio de los datos para entender que variables influyen en los gastos médicos generados por los clientes asegurados.
Además, se propone generar un modelo de regresión basado en algoritmos de *machine learning* (ML) para predecir los gastos de servicios de salud de los clientes a partir de sus datos personales y datos relativos a la salud.

D. Casos de Éxito.

Una revisión sistemática relacionada con modelos predictivos para estimar los costos de atención medica mostraron que los algoritmos de aprendizaje supervisado tienen mejores rendimientos en comparación con modelos basados en reglas fijas o en estadísticos (2).

Por otra parte, Mohamed H. realizó una comparación de diferentes modelos de ML y *deep learning* para predecir los gastos de atención medica en las aseguradoras, encontrando que el algoritmo Stochastic Gradient Boosting tuvo el mejor rendimiento para esta tarea (4). Contrario a lo reportado por Belisario P (5), quien sugirió que el mejor predictor se basó en un modelo de redes neuronales (6) generado a partir de los datos históricos del Hospital Tsuyama Chuo.
Desde otra perspectiva, Nataliya S. utilizó un ensable learning construido por múltiples modelos regresores (*K nearest neighbors, support vector machine, regression tree, linear regression, stochastic, gradient boosting, random forest*) probando mejorar la precisión y generalización con respecto a los modelos individuales(3).

E. Criterio de éxito:

Para este proyecto se plantea como un caso de éxito un valor de R-Squared (R^2) mayor a 0.7 (7). Sin embargo, se espera superar un R^2 de 0.755813 reportado en el estado de arte al ajustar un modelo de regresión lineal multivariable(4).

II. ENTENDIMIENTO DE LOS DATOS.

A. Recolección de datos:

Se utilizó la base de datos publica "*Medical Cost Personal Datasets - Insurance Forecast by using Linear Regression*" contenida en el repositorio Kaggle (1).

B. Descripción de datos:

El conjunto de datos utilizado para este proyecto se compone de 6 caracteristicas predictoras de entrada y una etiqueta de salida como se muestra en la tabla I.

Tabla 1 Descripción de los datos

Nombre	Tipo de variable	Descripción
Age	Variable de entrada de tipo continuo.	Valor en año de la edad de los asegurados.
Sex	Variable de entrada de tipo categórica dicotómica.	Valores relacionada con el sexo de los asegurados (male=hombre, male=mujer).
BMI	Variable de entrada de tipo continuo.	Valor del índice de masa corporal de los asegurados.
Children	Variable de entrada de tipo discreta.	Número de hijos de los asegurados.
Smoker	Variable de entrada de tipo categórica dicotómica.	Indicador del consumo de tabaco (no=no fumador y yes= fumador).
Region	Variable de entrada de tipo categórica.	Zona de residencia (southeast, southwast, northwest y northeast).
Charges	Variable de salida de tipo continuo.	Costos médicos individuales facturados por el seguro médico en dólares.

C. Exploración de datos

Para el análisis exploratorio de datos (DEA), primero se calcularon las medidas de tendencia central y de dispersión de las variables continuas (tabla 2).

Posteriormente mediante histogramas se visualizó la distribución de todas las variables, observando un sesgo a la derecha en los costos médicos individuales facturados por el seguro médico (Fig. 7).

Tabla 2 Estadística descriptiva de las variables continuas

	N	media	SD	min	Q1	Q2	Q3	max
Age	1338	39.20	14.04	18.00	27.00	39.00	51.00	64.00
BMI	1338	30.66	6.09	15.96	26.29	30.40	34.69	53.13
Charge	1338	13270.42	13270.42	1121.87	4740.28	4740.28	16639.91	63770.42

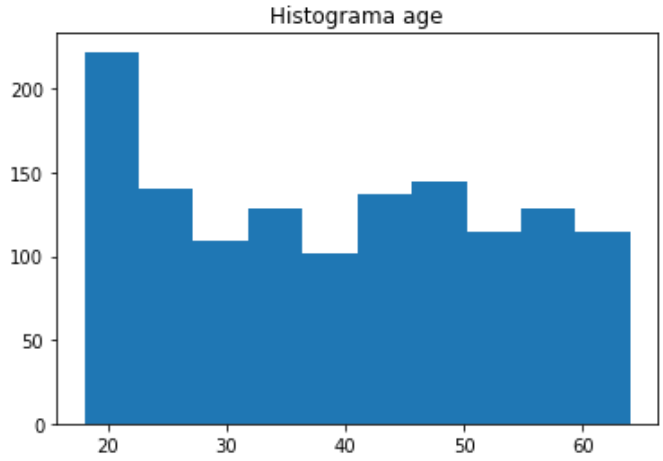

Fig. 1 Histograma que muestra la distribución de la variable age

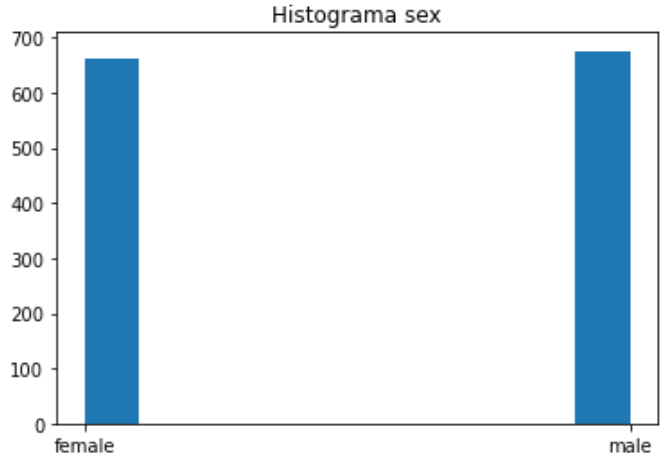

Fig. 2 Histograma que muestra la distribución de la variable sex

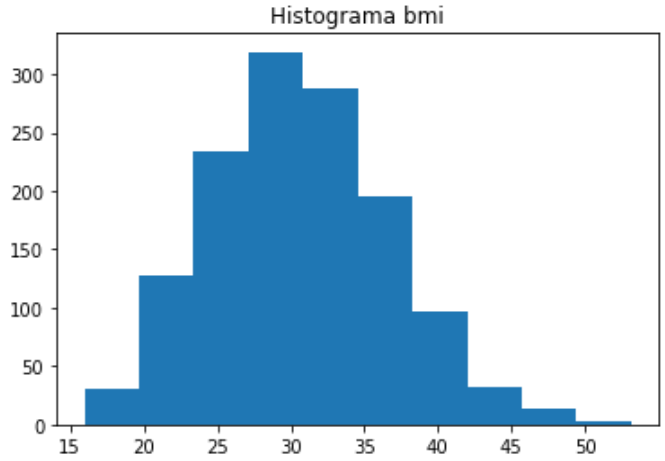

Fig. 3 Histograma que muestra la distribución de la variable

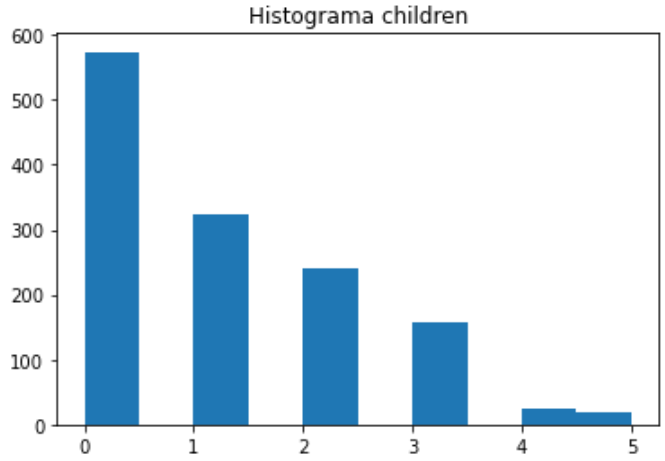

Fig. 4 Histograma que muestra la distribución de la variable children

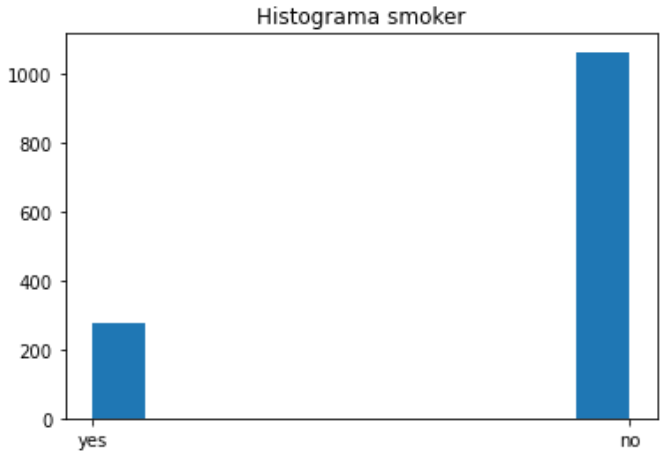

Fig. 5 Histograma que muestra la distribución de la variable smoker

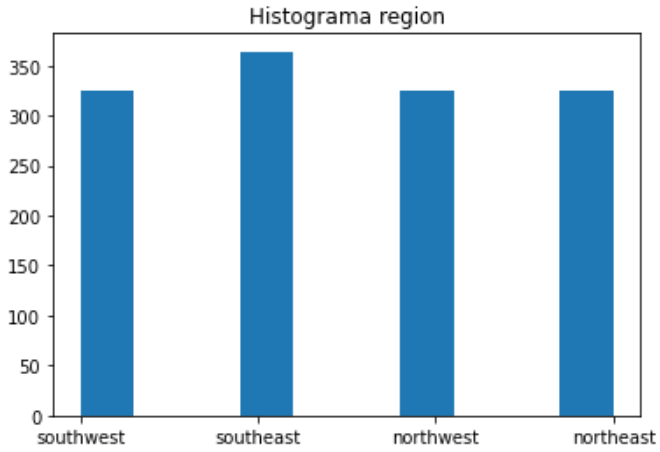

Fig. 6 Histograma que muestra la distribución de la variable region

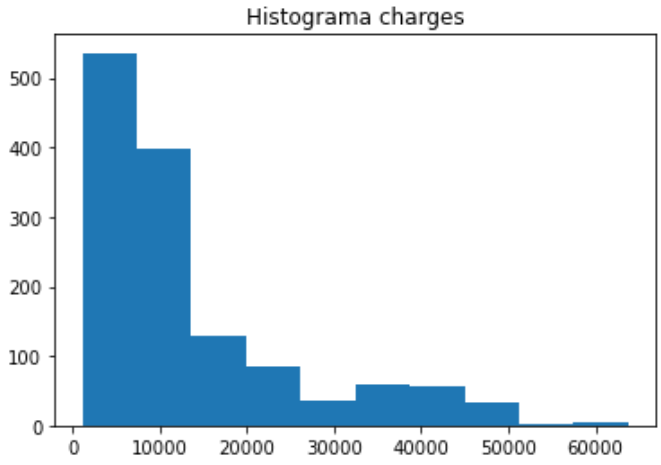

Fig. 7 Histograma que muestra la distribución de la variable charges

Como parte del DAE se calculó una matriz de correlación entre las variables, dando a conocer que existía una mayor correlación entre la variable age y charges (como se muestra en la Fig. 8)

Además, se observó que, segmentando nuestro conjunto de datos general, en dos subconjuntos más, agrupando a los fumadores y no fumadores se podía obtener una matriz de correlación que representara de mejor manera la relación lineal entre las variables como se muestra en la Fig. 9 y Fig. 10

	age	bmi	charges
age	1.000000	0.109272	0.299008
bmi	0.109272	1.000000	0.198341
charges	0.299008	0.198341	1.000000

Fig. 8 Matriz de correlación de todo el conjunto de datos

	age	bmi	charges
age	1.000000	0.122638	0.627947
bmi	0.122638	1.000000	0.084037
charges	0.627947	0.084037	1.000000

Fig. 9 Matriz de correlación para el grupo de no fumadores

	age	bmi	charges
age	1.000000	0.059674	0.368224
bmi	0.059674	1.000000	0.806481
charges	0.368224	0.806481	1.000000

Fig. 10 Matriz de correlación para el grupo de fumadores

D. Calidad de los datos.

La calidad en la base de datos *Medical Cost Personal Datasets - Insurance Forecast by using Linear Regression* se consideró como optima debido a que no contaba con valores faltantes o/y anómalos

III. PREPARACIÓN DE DATOS

A.Limpieza:

A.1) Valores *faltantes*: Se realizó un análisis de la calidad de los datos mediante la función isnull() de la librería pandas.

A.*2)* Valores *anómalos:* Se encontraron múltiples valores outliers en la variable charges, pero a pesar de que eliminarlos podría suponer mejorar el rendimiento de los regresores, se optó por no eliminarlos debido a que podría afectar de manera negativa al realizar predicciones "pobres" en un modelo de producción debido a falta en la robustez en los datos.

B. Generación de Atributos:

Se generó una nueva variable categórica a partir de los datos la variable BMI, esta nueva variable indica el estatus del peso de modo que se tienen cuatro nuevas categorías (Underweight, Healthy Weight, Overweight y Obesity) de acuerdo por lo reportado por el Center for Disease Control and Prevention (8). No obstante, agregar más características de entrada empeoró el rendimiento del modelo final, por lo cual se decidió no incluir la generación de nuevos atributos.

C Transformación:

Para los modelos de regresión lineal se normalizaron los datos mediante una transformación box-cox. Sin embargo, al ejecutar el test de normalidad (scipy.stats.normaltest) solo la variable BMI mostró tener una distribución normal. Por otra parte, para los regresores de tipo *Gradient Boosting* la normalización o estandarización no impactan en el rendimiento del regresor (9), por lo cual se decidió ocupar los datos sin aplicar transformaciones para los modelos *Gradient Boosting*.

D. Segmentación de los datos:

Después del análisis de las matrices de correlación se decidió separar el conjunto de datos en tres:

1. Grupo general: Conjunto con todos los datos.
2. Grupo fumadores: Subconjunto de datos que cumpliesen la condición ser fumadores.
3. Grupo no fumadores: Subconjunto de datos que cumpliesen la condición ser no fumadores.

IV. MODELADO

A. Seleccionar técnica de modelado:

Debido a que el tipo de tarea de ciencia de datos es predecir un valor continuo, se utilizaron modelos de ML capaces de realizar regresiones. Se optó por una regresión lineal con el motivo de obtener mayor entendimiento de los datos, para ello se ajustó un modelo multivariable para el conjunto de datos general; se ajustó otro modelo de regresión lineal para el grupo fumadores únicamente basado en el BMI, y por último se ajustó otro modelo de regresión lineal para el grupo de no fumadores basado en la edad.

Posteriormente se ajustó un modelo más robusto como lo es *Gradient Boosting* para mejorar el rendimiento del modelo final. Para ello se ajustó un modelo de *Gradient Boosting* para el grupo general y para el grupo de fumadores respectivamente, y se realizó un "*ensable* condicional" que siguiera las reglas de la Fig. 11.

Al final se comparó el rendimiento del modelo de regresión lineal para el grupo general, el modelo *Gradient Boosting* para el grupo general y el modelo de "*ensable* condicional"

B. Diseño de evaluación

Para la evaluación de los modelos posterior a realizar un train *test Split 80-20* con un *random_state=2,* se utilizarán las métricas R^2, *mean absolute error* (MAE), *root mean squared error* (RMSE), mean absolute percentage error (MAPE) y *mean percentage error* (MPE).

C. Entrenamiento:

Lo modelos de regresión lineal se ajustaron para cada uno de los subconjuntos de datos como se muestra en la Fig.12, Fig.13

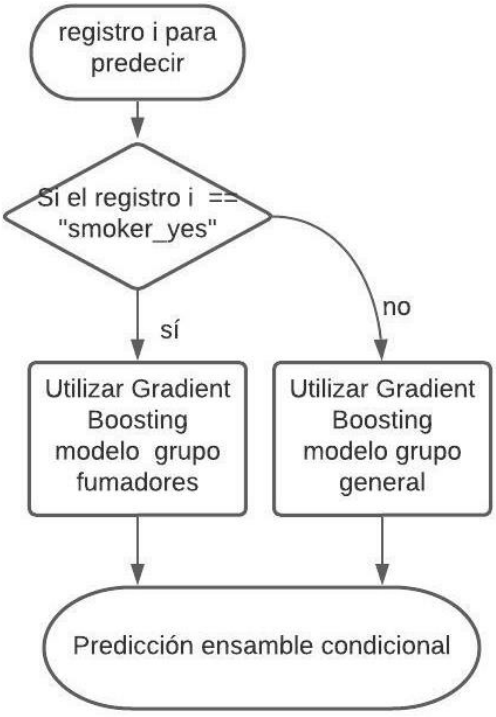

Fig. 11 Funcionamiento modelo ensamble condicional

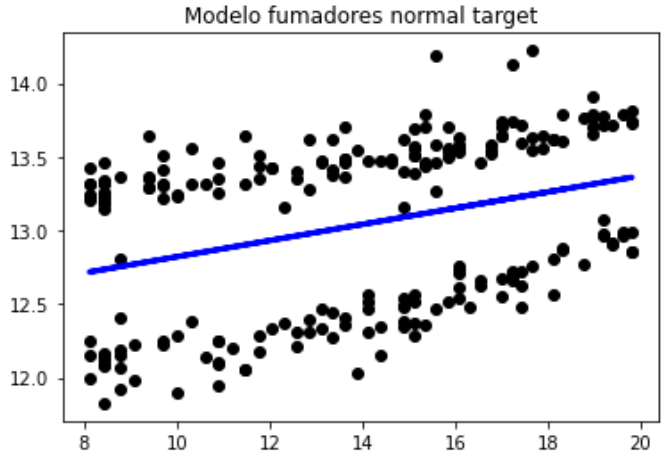

Fig. 12 Modelo de regresión lineal para el grupo de fumadores con datos normalizados.

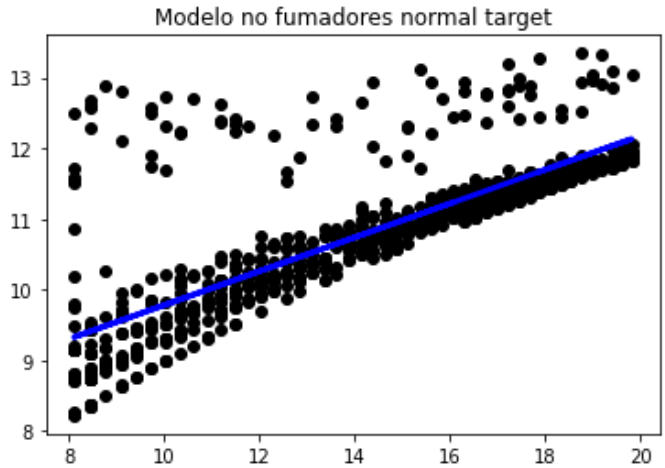

Fig. 13 Modelo de regresión lineal para el grupo de no fumadores con datos normalizados

El modelo de regresión lineal multivariable para el grupo general quedo parametrizado por:

$$\hat{y} = 0.195b1 + 0.125b2 + 0.142b3 + 0.053b4 - 0.053b5 - 1.160b6 + 1.160b7 + 0.132b8 + 0.016b9 - 0.053b10 - 0.056b11 + 7.99$$

Por otra parte, los modelos *Gradient Boosting* no presentan una función matemática que parametriza al modelo, a cambio genera un árbol de decisión, desafortunadamente el modelo *Gradient Boosting* no contiene una librería capaz de graficar su árbol de decisión. Por lo cual se plantea presentar los hiperparámetros utilizados en el modelo general y fumadores en la tabla 3, con el fin de que los resultados sean reproducibles.

Tabla 3 Hiperparametros modelos Gradient *Boosting*

Hiperparámetros	*Gradient Boosting* general	*Gradient Boosting* fumadores
n_estimators	100	162
max_depth	2	2
max_features	Auto	sqrt
min_samples_leaf	1	1
min_samples_split	2	5

Mientras que el modelo *ensable* condicional se puede representar de la siguiente manera:

$$\hat{y} = \begin{cases} if\ fumador = \hat{y}(Gradient\ Boosting\ \text{fumadores}) \\ else\ fumador = \hat{y}(Gradient\ Boosting\ \text{general}) \end{cases}$$

D. Evaluación de modelos.

Los resultados del rendimiento de los modelos finales se presentan en la tabla 4.

V. EVALUACIÓN

A. Evaluación de resultados:

El modelo que describió de mejor manera los datos fue el basado en *ensable* condicional, como se puede observar en la tabla 4, el modelo *ensable* condicional obtuvo una mejor métrica R^2 respecto a los demás modelos, lo que indica de manera general un mejor ajuste de los datos, del mismo modo, el modelo *ensable* condicional obtuvo el mejor rendimiento en las demás métricas con excepción al MAPE, donde el modelo basado en regresión lineal obtuvo mejor rendimiento, este pudo ser causado debido al tratamiento de transformación box-cox que puede mejorar la precisión en la predicción de los regresores (4,9) respecto a los modelo *ensable* condicional y Gradient Boosting donde no se realizó la transformación box-cox. Finalmente, el signo negativo en la métrica MPE indicó que existe una tendencia de los modelos a subestimar el valor a predecir.

Es importante mencionar que los resultados obtenidos en este trabajo superaron a lo reportado por M. Hanafy(4), quien utilizó la misma base de datos (1) y el algoritmo Stochastic Gradient Boosting, obteniendo un R^2=0.8582. Se sugiere que la utilización de técnicas de *ensable* (3) y optimización de parámetros permitieron obtener un mejor rendimiento respecto a lo reportado en la literatura.

B.Reflexión.

Los conocimientos adquiridos durante el "Diplomado de Ciencia de datos, ITPE" pueden aplicarse para resolver múltiples problemas a nivel descriptivo, diagnostico, predictivo e incluso prescriptivo en diferentes áreas de conocimiento.

Además, aprender a utilizar la metodología cross industry standard process for data mining permite crear una "guía" para resolver cada una de las tareas de ciencia de datos de manera ordenada.

Tabla 4 evaluación de modelos de regresión

	R^2		MAE		RMSE		MAPE		Signo MPE	
Modelo	train	test	train	test	train	test	train	test	train	test
Regresión lineal	0.6022	0.6203	4033.9212	3857.7767	7609.9139	7551.255	0.2655	0.2857	-	-
Gradient Boosting general	0.879	0.8702	2313.8296	2413.9343	4197.5102	4414.3822	0.2695	0.3161	-	-
Ensable condicional	0.8817	0.8739	2249.847	2314.8349	4149.7295	4352.2536	0.267	0.3109	-	-

Por último, hay que mencionar que durante el diplomado se desarrolla un pensamiento crítico que permite comprender la importancia de las decisiones basadas en datos, asimismo lo primordial de la relación entre el entendimiento del negocio y las tareas de ciencias de datos para lograr proyectos exitosos.

VI. ANEXO

Para mayores detalles sobre el código de programación y el análisis exploratorio de datos, puede consultar el siguiente enlace: https://github.com/enriquehdez98/Prediction_health_care_expenses_with_ML

VII. REFERENCIAS

1. Choi M. Medical Cost Personal Datasets | Kaggle [Internet]. 2017 [cited 2021 Dec 4]. Available from: https://www.kaggle.com/mirichoi0218/insurance

2. Morid MA, Kawamoto K, Ault T, Dorius J, Abdelrahman S. Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation. AMIA Annual Symposium Proceedings. 2017;2017:1312.

3. Shakhovska N, Melnykova N, Chopiyak V, Gregus Ml M. An ensemble methods for medical insurance costs prediction task. Computers, Materials and Continua. 2022;70(2):3969–84.

4. Mohamed H. Predict Health Insurance Cost by using Machine Learning and DNN Regression Models. IJITEE. 2021;10(3):137–42.

5. Panay B, Baloian N, Pino JA, Peñafiel S, Sanson H, Bersano N. Predicting Health Care Costs Using Evidence Regression. Proceedings 2019, Vol 31, Page 74. 2019 Nov 21;31(1):74.

6. Duncan I, Loginov M, Ludkovski M. Testing Alternative Regression Frameworks for Predictive Modeling of Health Care Costs. http://dx.doi.org/101080/1092027720151110491. 2016 Jan 2;20(1):65–87.

7. Jason F, Smith A, Perez Y. R-Squared [Internet]. 2021 [cited 2021 Dec 6]. Available from: https://www.investopedia.com/terms/r/r-squared.asp

8. Bhaskaran K, Douglas I, Forbes H, dos-Santos-Silva I, Leon DA, Smeeth L. Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5·24 million UK adults. The Lancet [Internet]. 2021 Aug 30 [cited 2021 Dec 16];384(9945):755–65. Available from: https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html

9. Javier F. How data normalization affects your Random Forest algorithm | by Javier Fernandez | Towards Data Science [Internet]. 2021 [cited 2021 Dec 17]. Available from: https://towardsdatascience.com/how-data-normalization-affects-your-random-forest-algorithm-fbc6753b4ddf