Llegar a la vejez

Análisis de métricas de la salud y edad

Proyecto de Ciencia de Datos
por Juan Pablo Quevedo

Acerca del proyecto

El avance tecnológico ha transformado prácticamente todos los aspectos de la medicina, desde el diagnóstico hasta la investigación, pasando por el tratamiento y la atención al paciente.

Este proyecto aborda otro uso de la tecnología, y tiene como objetivo buscar y analizar posibles relaciones entre distintas métricas de la salud, para así poder proyectar a que edad podría llegar una persona según su situación.

Los resultados a obtener pueden ser de utilidad para estudios de prevención de enfermedades, medicamentos, tratamientos médicos, entre otros. El público objetivo es toda persona que desarrolle una actividad de este ámbito.

- ¿Qué impacto genera una dieta especial en la longevidad y enfermedades de una persona?
- ¿Puede un mal hábito volver más propensa una enfermedad, a pesar de tener un buen historial familiar? ¿Cuándo?
- ¿Qué hábitos o enfermedades son más influyentes en la esperanza de vida?

Análisis exploratorio

El dataset está compuesto de 3000 filas y 26 columnas (13 numéricas y 13 categóricas).

Las variables abordan:

- Estadísticas físicas de la persona
- Hábitos de alimentación
- Estilo de vida
- Enfermedades
- Entorno e historial familiar

Target: ‘Age (years)’

Análisis exploratorio

Se observa que las variables más asociadas a la tercera edad no son producto de enfermedades, solo hablan del deterioro de las capacidades físicas de la persona. Próximamente ahondaré en las variables de hábitos y enfermedades que no se mencionan aquí para determinar si hay relaciones escondidas, y así descubrir si hay hábitos o enfermedades que empeoran la esperanza de vida y qué factores influyen en ellos.

Principales correlaciones con la variable **target** y sus respectivas influencias:

Blood Pressure (s/d): **0.75**

- Hearing Ability (dB): 0.53
- Vision Sharpness: -0.67
- Bone Density (g/cm²): -0.70

Hearing Ability (dB): **0.71**

- Blood Pressure (s/d): 0.53
- Vision Sharpness: -0.64
- Bone Density (g/cm²): -0.67

Vision Sharpness: **-0.90**

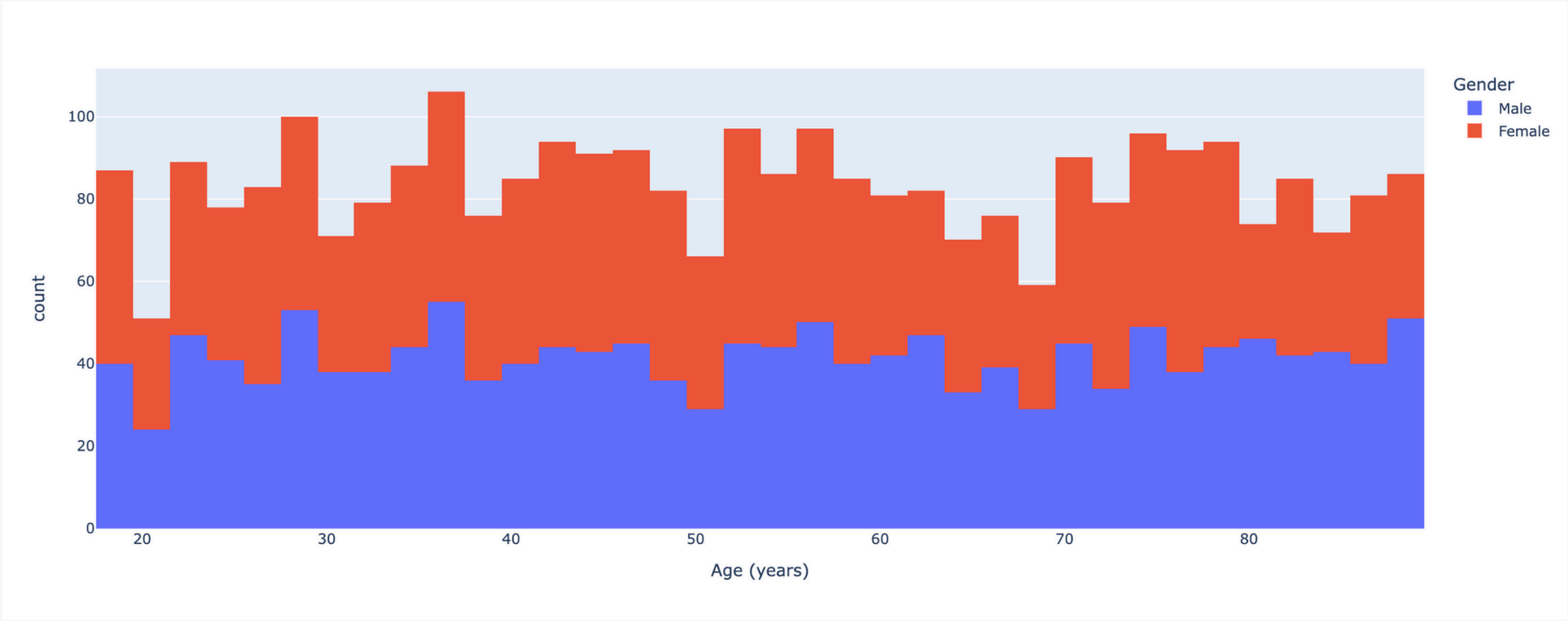
- Bone Density (g/cm²): 0.85
- Hearing Ability (dB): -0.64
- Blood Pressure (s/d): -0.67

Bone Density (g/cm²): **-0.94**

- Vision Sharpness: 0.85
- Hearing Ability (dB): -0.67
- Blood Pressure (s/d): -0.70

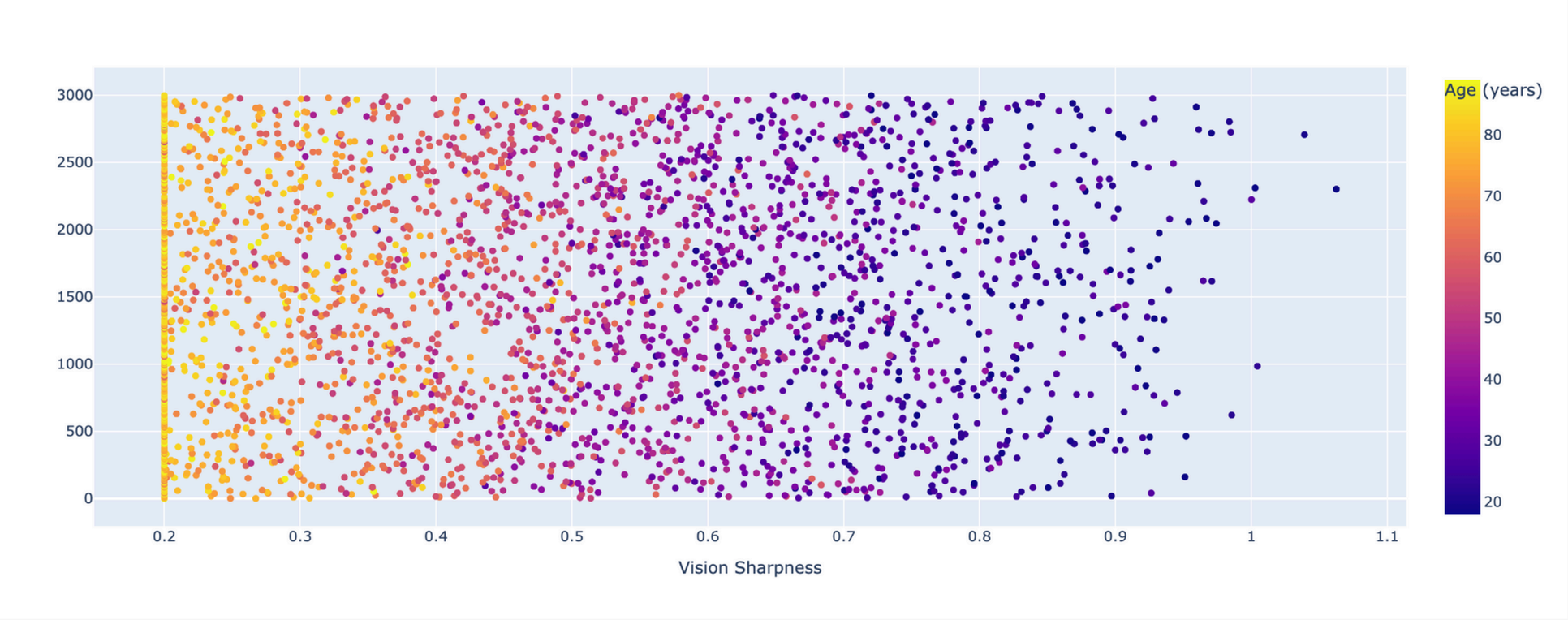
Análisis exploratorio

- Espectro de edad: 18 - 89, promedio 53 años.
- Distribución normal de género.



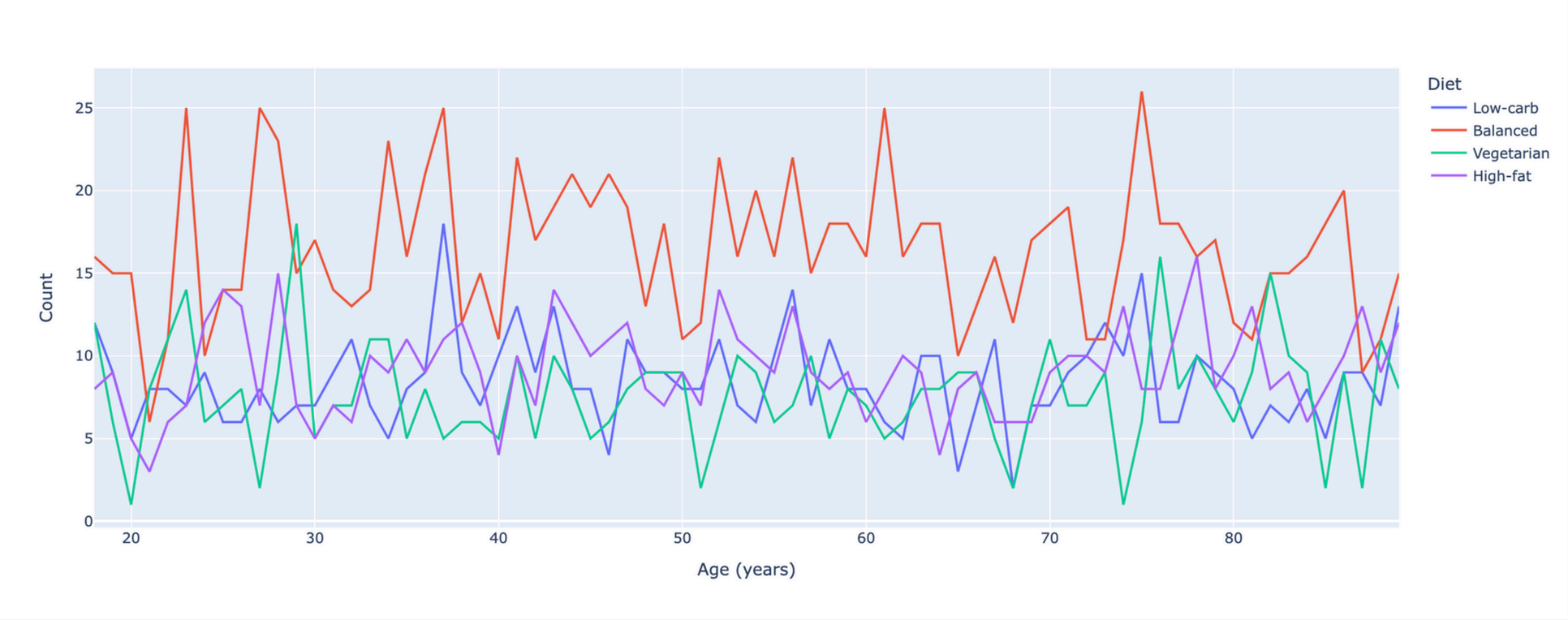
Análisis exploratorio

- La menor nitidez de visión es una condición común en la población de la tercera edad, y por lo general empieza a deteriorarse en edades tempranas, no hay hábitos o enfermedades en el dataset que demuestren relación directa con el empeoramiento de esta variable.



Análisis exploratorio

- La dieta preferida por la población es la balanceada, el resto de opciones manejan cuotas similares entre sí. Después de los 80 años, el promedio de la población vegetariana disminuye ligeramente más respecto las otras dietas.

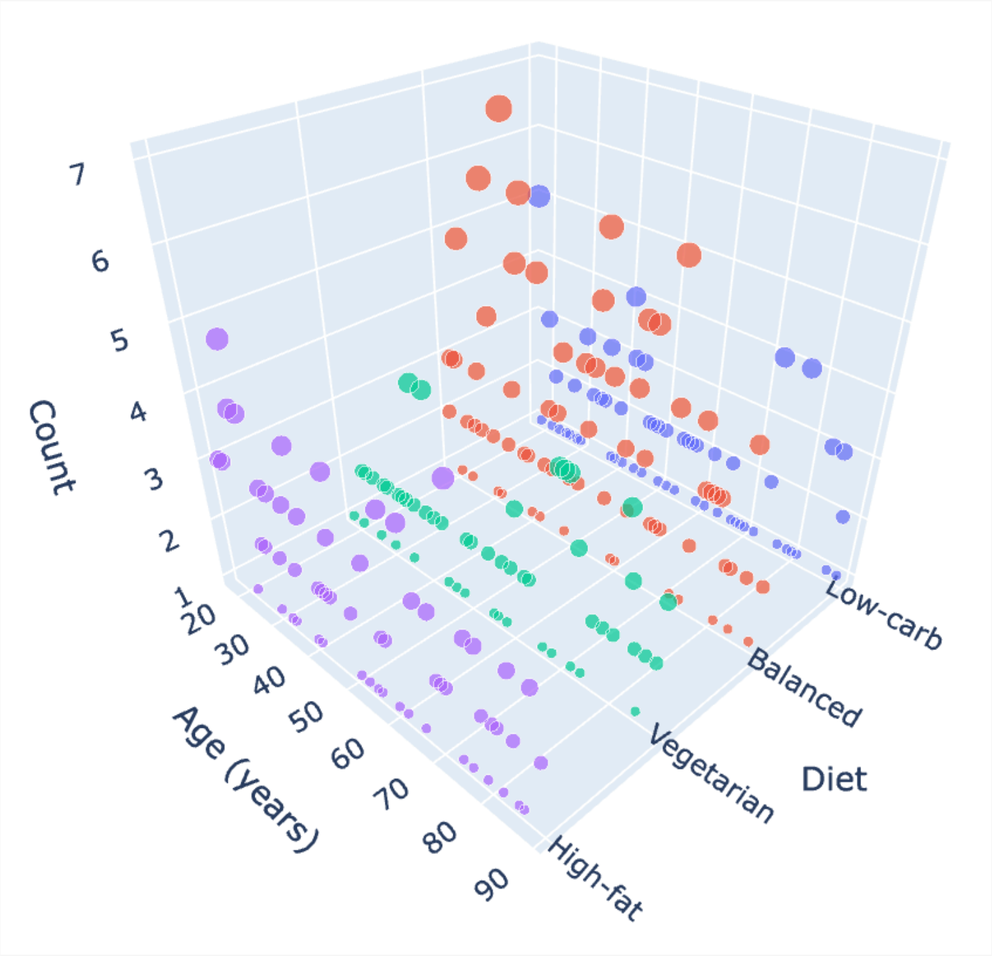


Análisis exploratorio

Porcentaje de diabéticos según dieta:

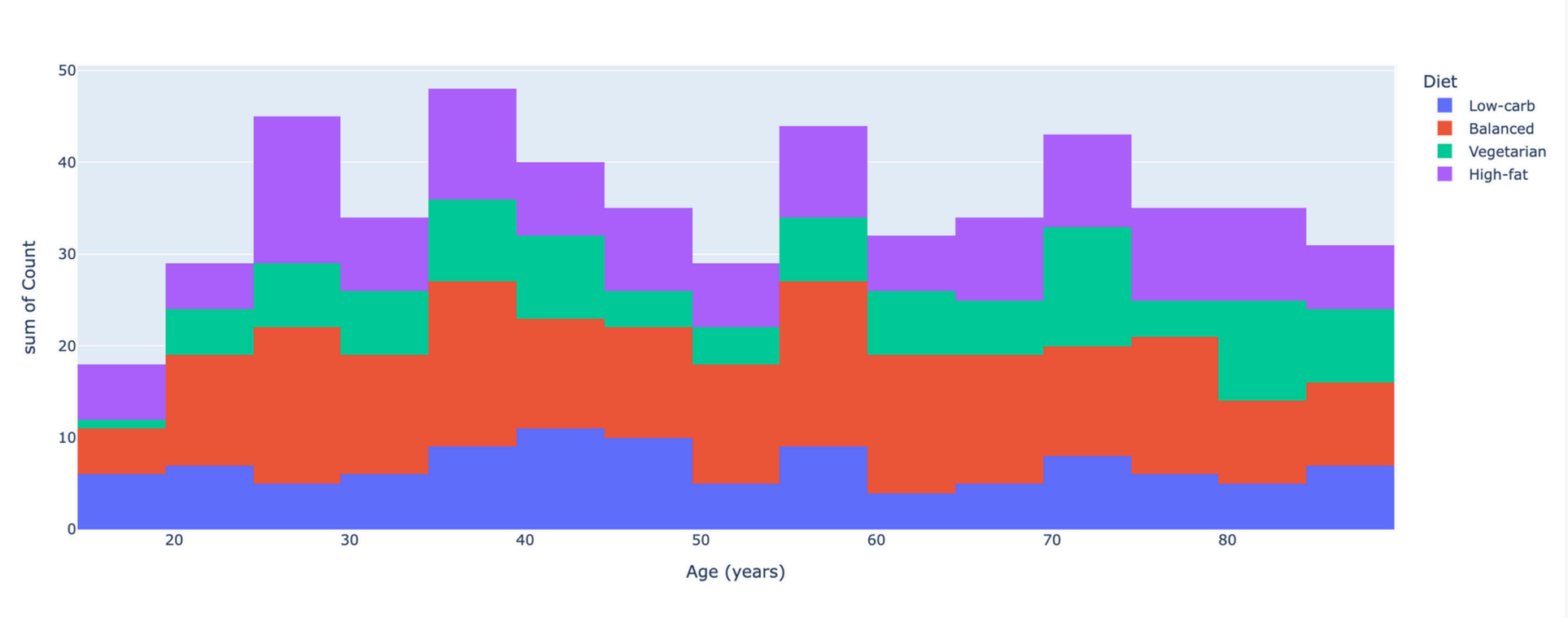
- Low-carb: 17%
- Balanced: 16%
- Vegetarian: 18%
- High-fat: 20%

Podemos asumir que no hay una dieta que disminuya considerablemente el riesgo de padecer diabetes.



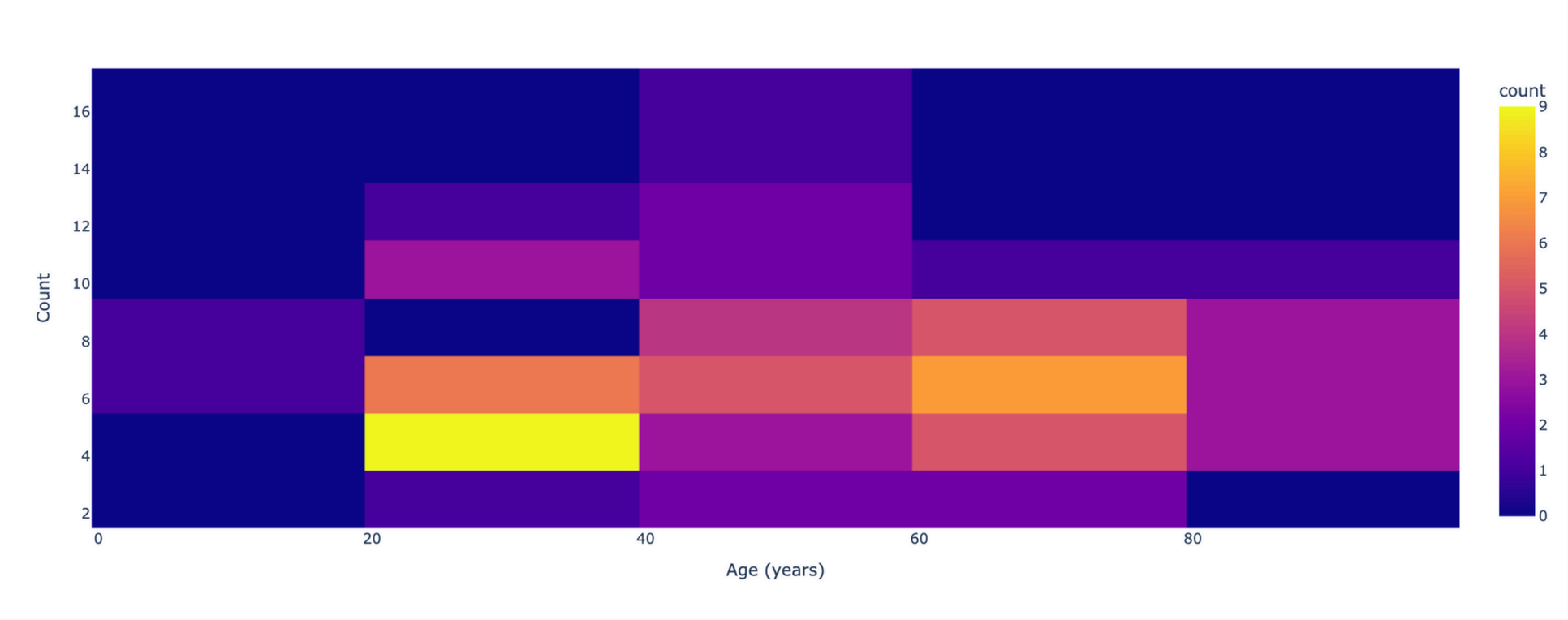
Análisis exploratorio

- Dentro del universo de las personas diagnosticadas con diabetes, la mayoría de personas llevan una dieta balanceada. El resto de dietas tienen una distribución pareja, al igual que en el total de la muestra.



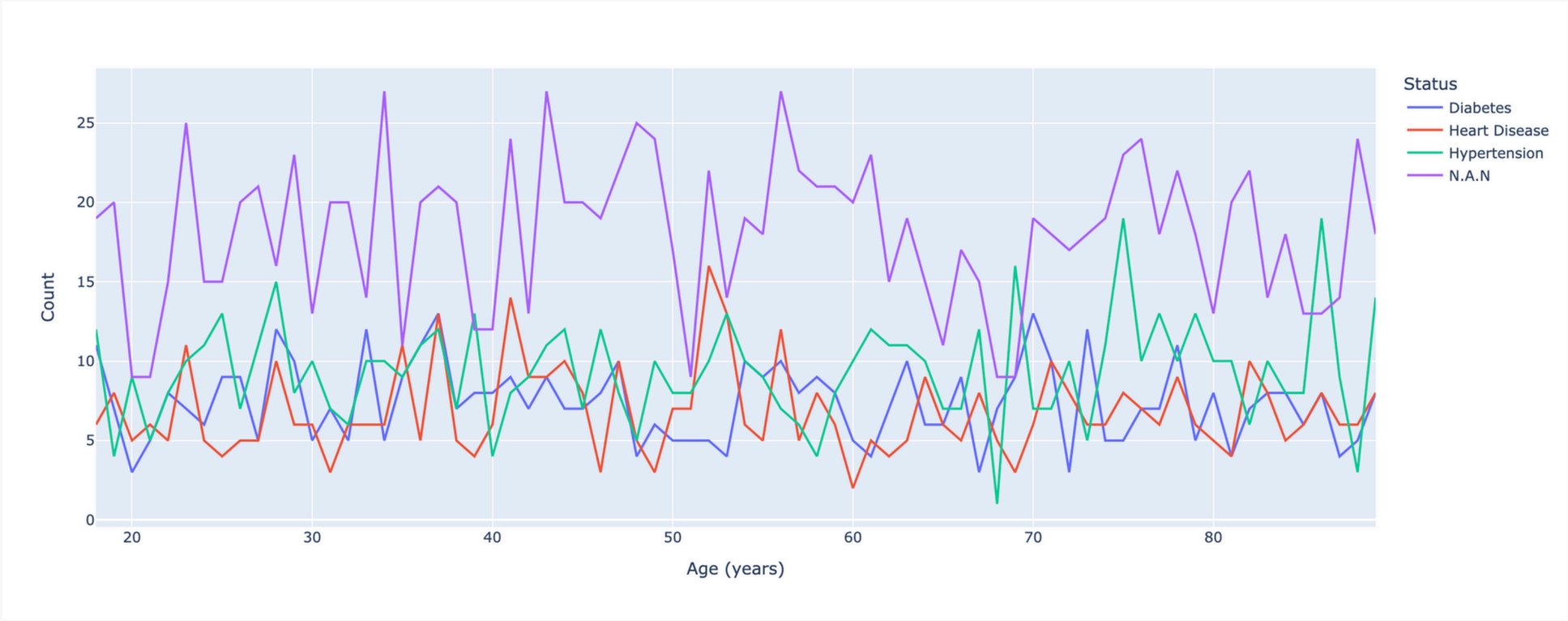
Análisis exploratorio

- El grueso de la población que padece diabetes no pertenece al grupo de la tercera edad. Podríamos hipotizar que a pesar de ser una enfermedad tratable, contribuye a la disminución de esperanza de vida.



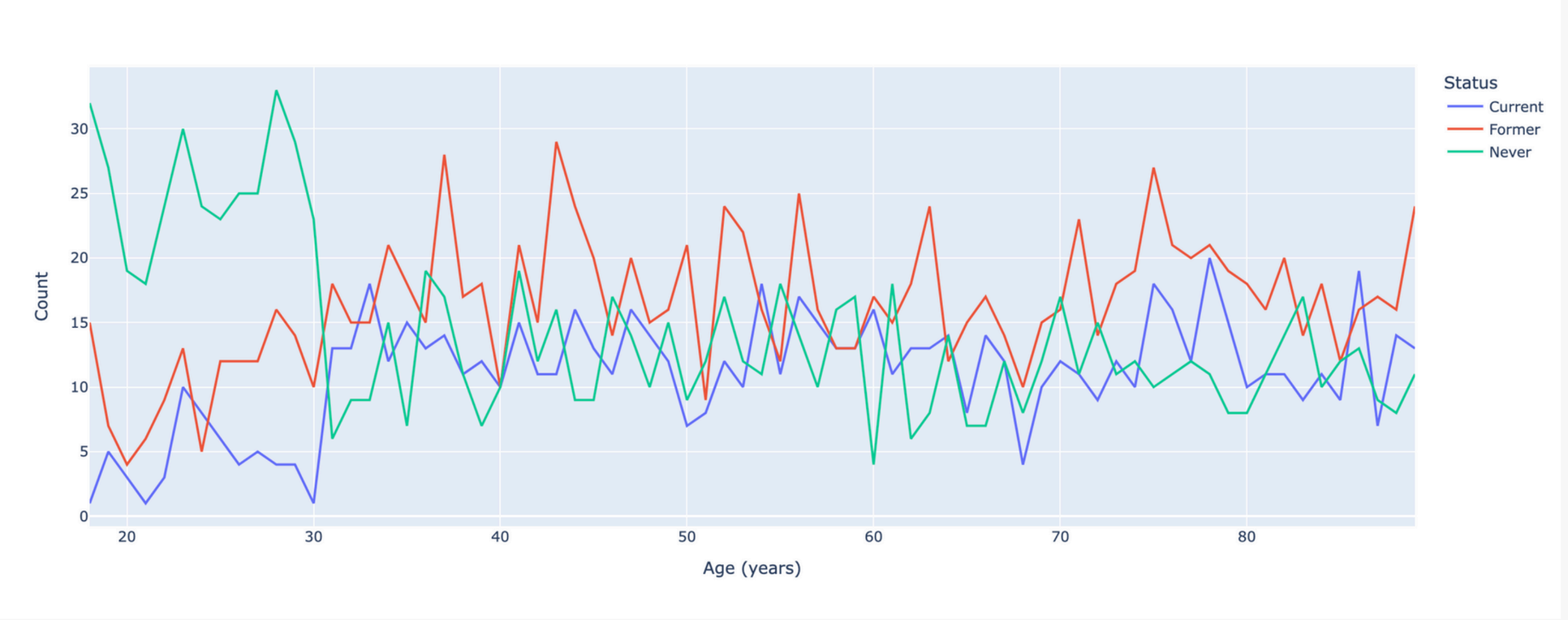
Análisis exploratorio

- Dentro de las enfermedades crónicas, la hipertensión suele ser la menos mortal.
- El historial familiar influye tanto en el padecimiento de enfermedades como el uso de medicamentos.



Análisis exploratorio

- La población se distribuye de manera bastante equitativa en el hábito de fumar, a diferencia de lo que se esperaría no se evidencia que el no fumar mejore la esperanza de vida.



Análisis exploratorio

Conclusiones del análisis

- Las variables con correlación más fuerte con la edad hacen referencia al desgaste de capacidades físicas, las cuales pueden explicarse con el envejecimiento y no muestran una dependencia clara a algún hábito o enfermedad crónica.

- La dieta balanceada predomina en el total de la muestra, mientras que el resto se distribuyen de manera equilibrada. Las personas con enfermedades crónicas tienen preferencias similares. No hay una dieta que disminuya considerablemente la posibilidad de contraer enfermedades ni de empeorar la esperanza de vida.

- Las enfermedades cardiovasculares y la diabetes son enfermedades menos vistas en la tercera edad. La hipertensión es una enfermedad que termina afectando una mayor cantidad de la población de adultos mayores. El historial familiar influye tanto en el padecimiento de enfermedades como el uso regular de medicamentos.

Variables menos influyentes en la edad: género, nivel de educación, situación socioeconómica, salud mental, patrones del sueño, consumo de alcohol y/o tabaco, entorno.

Modelos Analíticos

Elegí estos 3 modelos para el proyecto por los siguientes motivos:

- **Regresión Lineal**: Modelo simple y práctico.
- **Ridge**: Controla overfitting.
- **Random Forest**: Ofrece flexibilidad de parámetros (iniciales: n_estimators: 200, max_depth: 7).

Para cada modelo apliqué una validación cruzada inicial para comparar su rendimiento:

- **Regresión Lineal**: array([0.93551592, 0.94147853, 0.93475039, 0.93009375, 0.93076084])
- **Ridge**: array([0.9354937, 0.94135316, 0.93497244, 0.929842, 0.93083257])
- **Random Forest**: array([0.91209825, 0.91052202, 0.91507075, 0.89917219, 0.90431106])

En primera instancia, la **Regresión Lineal** tiene un mejor rendimiento, pero al tener resultados tan similares con Ridge y al haber tantas posibilidades de parámetros en el modelo de Random Forest, decidí someterlos a un Hyperparameter Tuning con GridSearchCV.

Modelos Analíticos

Sometí los modelos al Hyperparameter Tuning con la siguiente configuración:

- **Regresión Lineal**: sin cambios.
- **Ridge**: alpha: [0.1, 1.0, 10.0].
- **Random Forest**: n_estimators: [300, 400], max_depth: [5, 7], min_samples_split: [5].

En Ridge, probé distintos valores para alpha con el fin de controlar el overfitting. Respecto Random Forest, le di la posibilidad de ajustar la cantidad y tamaño para árboles más grandes, buscando un ajuste más complejo sin caer en overfitting. La Grilla trabajaría con 5 validaciones cruzadas, y usaría una porción reducida del set (1000 datos de 3000) para mejorar su rendimiento.

El resultado fue el siguiente:

- Pipeline(steps=[('estimator', **LinearRegression()**)])

Conclusión

Dado el resultado de GridSearchCV, realicé una predicción de Regresión Lineal. Posteriormente el coeficiente r2 midió la precisión del modelo, arrojando un resultado de **0.9353**, dejando en evidencia que el modelo explica muy bien la variable dependiente.

Teniendo el respaldo de la validación cruzada inicial, la evaluación de GridSearchCV y el coeficiente r2, puedo concluir que el modelo más adecuado para este trabajo es la **Regresión Lineal**.

Si el modelo es entrenado nuevamente en el futuro, se podrían probar alternativas de imputación a través de Pipeline, con el fin de monitorear, ajustar y/o validar las técnicas.