Data Summary

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
df = pd.read_csv('/work/diabetes_prediction_dataset.csv')
```

Reporte de Datos

El conjunto de datos esta compuesto por las siguientes columnas

['gender', 'age', 'hypertension', 'heart_disease', 'smoking_history', 'bmi', 'HbA1c_level', 'blood_glucose_level', 'diabetes'] Donde diabetes es la variable objetivo y el resto son las caracteristicas

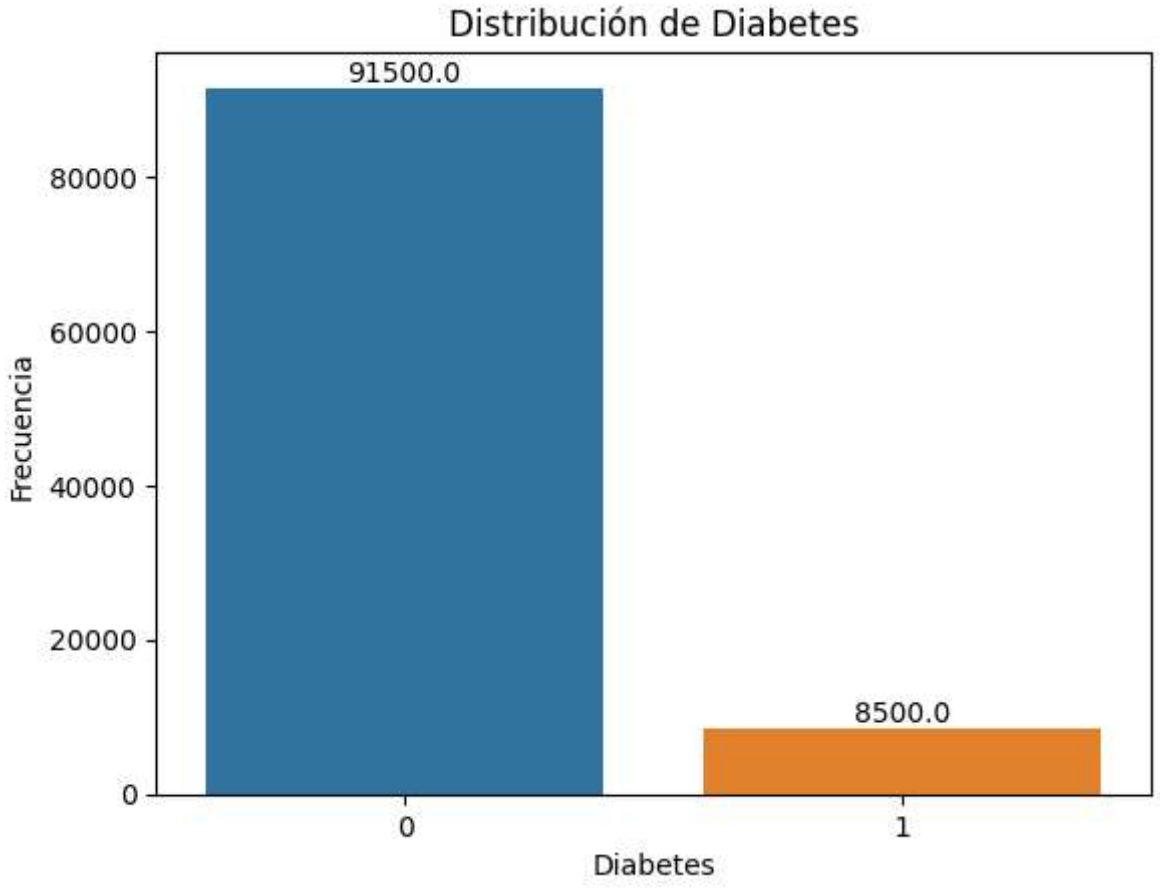
Por otra parte,se observa que es un df desbalanceado y no hay existencia de datos nulos

```
sns.countplot(x='diabetes', data=df, order=df['diabetes'].value_counts().index)

for p in plt.gca().patches:
    plt.gca().annotate(f'{p.get_height()}', (p.get_x() + p.get_width() / 2, p.get_height()
                       ha='center', va='bottom')

plt.xlabel('Diabetes')
plt.ylabel('Frecuencia')
plt.title('Distribución de Diabetes')

plt.show()
```

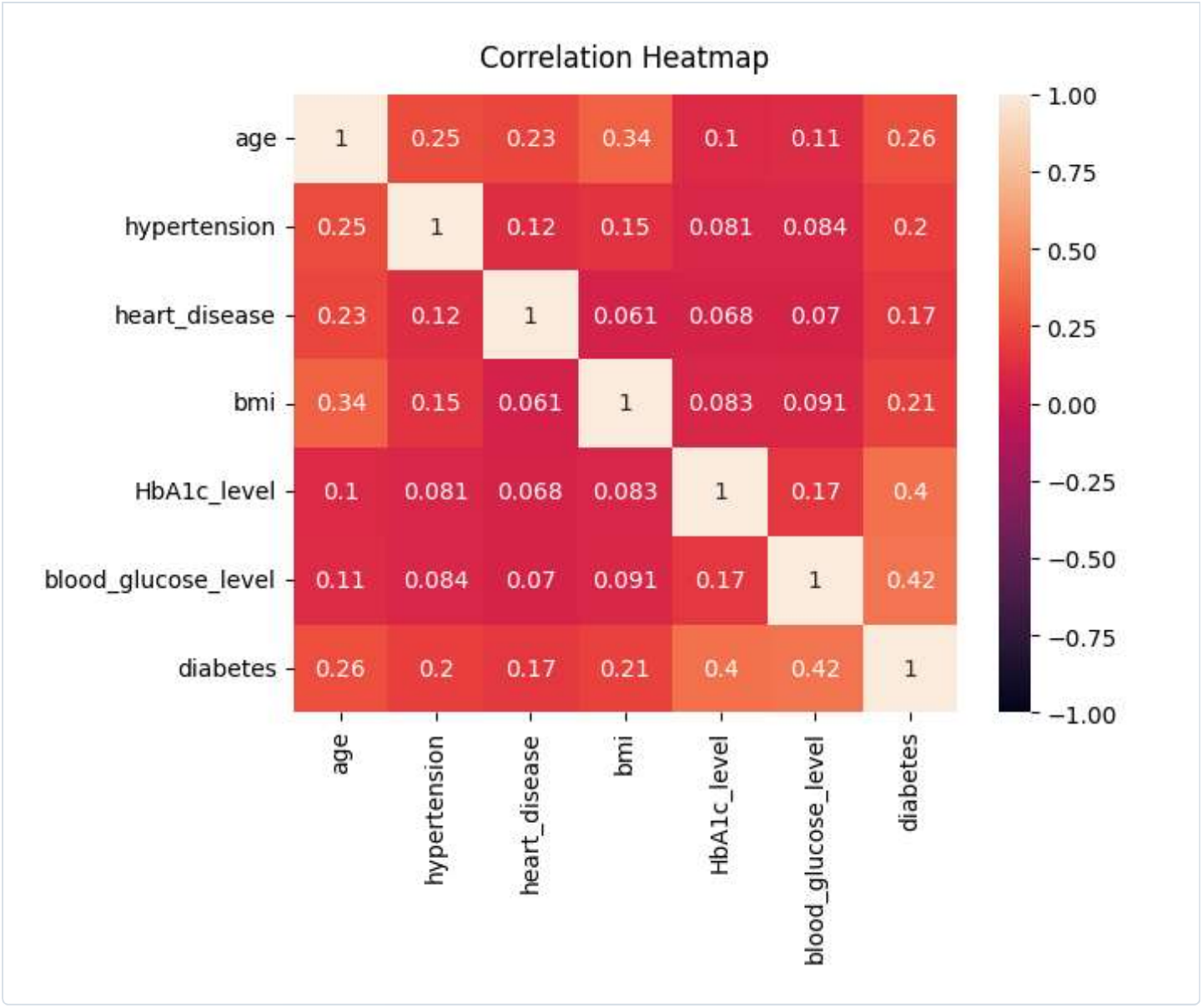


Correlacion e Importancia de Variables

Se observa que las variables que tienen una correlacion con respecto a diabetes son:

1. blood_glucose_level
2. HbA1c_level
3. age
4. bmi
5. hypertension
6. heart_disease

```
heatmap = sns.heatmap(df.corr(), vmin=-1, vmax=1, annot=True)
heatmap.set_title('Correlation Heatmap', fontdict={'fontsize':12}, pad=12);
```


Correlation Heatmap
age
hypertension
heart_disease
bmi
HbA1c_level
blood_glucose_level
diabetes
1 0.25 0.23 0.34 0.1 0.11 0.26
0.25 1 0.12 0.15 0.081 0.084 0.2
0.23 0.12 1 0.061 0.068 0.07 0.17
0.34 0.15 0.061 1 0.083 0.091 0.21
0.1 0.081 0.068 0.083 1 0.17 0.4
0.11 0.084 0.07 0.091 0.17 1 0.42
0.26 0.2 0.17 0.21 0.4 0.42 1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00