# Missing-entries-dataset-and-files-hospital

February 11, 2026

[1]:
```
# Since the previous interpreter state might have cleared the variables,
# I will quickly regenerate the files to ensure they are available for download.

import numpy as np
import pandas as pd
from sklearn.linear_model import LogisticRegression
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import OneHotEncoder
import joblib

# 1. Regenerate Data
n_samples = 100
data = {
    'Blood_Pressure': np.random.randint(110, 180, n_samples).astype(float),
    'BMI': np.random.uniform(18, 40, n_samples),
    'Smoking_Status': np.random.choice(['Non-smoker', 'Occasional', 'Regular'],␣
 ↪n_samples),
    'Family_History': np.random.choice([0, 1], n_samples)
}
df = pd.DataFrame(data)

# Logit logic
logit = (0.05 * (df['Blood_Pressure'] - 120) + 0.1 * (df['BMI'] - 25) + 0.8 *␣
 ↪df['Family_History'] - 2)
prob = 1 / (1 + np.exp(-logit))
df['Has_Disease'] = np.random.binomial(1, prob)

# Create raw with missing values
df_raw = df.copy()
df_raw.loc[df_raw.sample(frac=0.1).index, 'Blood_Pressure'] = np.nan
df_raw.loc[df_raw.sample(frac=0.1).index, 'BMI'] = np.nan
df_raw.to_csv('medical_data_raw.csv', index=False)

# 2. Preprocess
num_imputer = SimpleImputer(strategy='mean')
df[['Blood_Pressure', 'BMI']] = num_imputer.fit_transform(df[['Blood_Pressure',␣
 ↪'BMI']])
```

```
encoder = OneHotEncoder(drop='first', sparse_output=False)
smoking_encoded = encoder.fit_transform(df[['Smoking_Status']])
smoking_cols = encoder.get_feature_names_out(['Smoking_Status'])
smoking_df = pd.DataFrame(smoking_encoded, columns=smoking_cols)
df_processed = pd.concat([df.drop('Smoking_Status', axis=1), smoking_df],␣
 ↪axis=1)
df_processed.to_csv('medical_data_processed.csv', index=False)

# 3. Save Model
X = df_processed.drop('Has_Disease', axis=1)
y = df_processed['Has_Disease']
model = LogisticRegression().fit(X, y)
joblib.dump(model, 'medical_model.pkl')

print("Files generated successfully.")
```

```
Files generated successfully.
```

[ ]: