Additional:

* Dataset description: Specify the total sample size, demographic composition, and exact class distribution (stroke vs. non-stroke) before and after SMOTE.

Dataset yang digunakan dalam penelitian ini memiliki total sampel awal sebanyak 5.110 data, yang terdiri atas 249 data dengan label stroke dan 4.861 data dengan label non-stroke. Setelah melalui tahap pra-pemrosesan, jumlah sampel berkurang menjadi 4.909 data, dengan distribusi kelas yaitu 209 data berlabel stroke dan 4.700 data berlabel non-stroke. Ketidakseimbangan jumlah sampel antar kelas ini menunjukkan adanya dominasi data non-stroke yang jauh lebih besar dibandingkan data stroke, sehingga berpotensi menimbulkan bias pada model pembelajaran mesin yang dibangun.

Untuk mengatasi permasalahan tersebut, digunakan metode Synthetic Minority Oversampling Technique (SMOTE). Melalui pendekatan ini, jumlah data pada kelas minoritas (stroke) diperbanyak hingga setara dengan kelas mayoritas. Hasil penerapan SMOTE menunjukkan bahwa distribusi kelas menjadi seimbang, yaitu 4.700 data stroke dan 4.700 data non-stroke. Dengan demikian, total sampel setelah dilakukan resampling adalah 9.400 data.

• Pre-processing transparency: Report the number of records removed as duplicates, how missing values were imputed (mean, median, or other), and which features were dropped after correlation analysis.

Pada tahap pra-pemrosesan data, dilakukan beberapa langkah penting untuk memastikan kualitas dataset sebelum digunakan dalam proses pemodelan. Pertama, data yang memiliki nilai kosong (*missing values*) dihapus secara keseluruhan. Keputusan ini didasarkan pada pertimbangan jumlah sampel yang cukup besar, yaitu sekitar 5.000 data. Dengan demikian, penghapusan kurang lebih 200 data yang mengandung nilai kosong tidak memberikan dampak signifikan terhadap representativitas dataset secara keseluruhan. Pendekatan ini dipilih dibandingkan metode imputasi seperti rata-rata (*mean*) atau median, karena dinilai lebih tepat dalam menjaga integritas data asli.

Selanjutnya, dilakukan analisis korelasi antar fitur untuk mengidentifikasi atribut yang tidak memiliki keterkaitan yang cukup relevan dengan target variabel. Berdasarkan hasil matriks korelasi, terdapat beberapa fitur dengan nilai korelasi negatif yang sangat rendah, di antaranya *gender\_Other* (-0,003010), *work\_type\_Never\_worked* (-0,014149), dan *work\_type\_children* (-0,080971). Fitur-fitur ini kemudian dihapus dari dataset karena kontribusinya yang minimal terhadap proses klasifikasi. Dengan menghilangkan fitur-fitur tersebut, diharapkan model dapat bekerja lebih optimal karena hanya memproses variabel-variabel yang memiliki relevansi kuat terhadap prediksi.

• SMOTE application: Quantify the number of synthetic minority samples generated and present the final balanced dataset size.

Penerapan Synthetic Minority Oversampling Technique (SMOTE) bertujuan untuk menyeimbangkan distribusi kelas yang sebelumnya tidak seimbang antara label *stroke* dan *non-stroke*. Sebelum dilakukan SMOTE, jumlah data berlabel *stroke* hanya 209 sampel, sedangkan data berlabel *non-stroke* mencapai 4.700 sampel. Ketidakseimbangan ini berpotensi mengurangi kemampuan model dalam mengenali pola pada kelas minoritas.

Melalui penerapan SMOTE, sejumlah sampel pada kelas *stroke* dihasilkan untuk menyeimbangkan jumlahnya dengan kelas mayoritas (non-stroke). Total sampel sintetik yang dihasilkan adalah 4.491 data baru, sehingga jumlah data *stroke* yang semula 209 meningkat menjadi 4.700. Dengan demikian, distribusi kelas akhir setelah SMOTE menjadi seimbang, yaitu 4.700 data *stroke* dan 4.700 data *non-stroke*. Secara keseluruhan, ukuran dataset meningkat menjadi 9.400 sampel.

• Experimental settings: Provide details on random seed, train–test split, and any cross-validation to allow replication.

Pada pengaturan eksperimen, beberapa langkah dilakukan agar proses penelitian dapat direplikasi dengan hasil yang konsisten. Pertama, proses oversampling dilakukan menggunakan **SMOTE** untuk menyeimbangkan distribusi kelas. Selanjutnya, dataset hasil resampling dibagi menjadi data latih (*training set*) dan data uji (*testing set*) menggunakan fungsi train\_test\_split dengan proporsi 96% data latih dan 4% data uji (test\_size=0.04). Pada tahap ini digunakan **random seed 12** (random\_state=12) untuk memastikan data konsisten setiap kali kode dijalankan.

Setelah pembagian data, dilakukan normalisasi menggunakan **StandardScaler** agar setiap fitur memiliki skala yang sama. Data latih kemudian dibagi oleh Keras menjadi data latih dan data validasi melalui parameter validation\_split=0.4. Artinya, 40% dari data latih digunakan sebagai **validation set**, sehingga model dapat dievaluasi selama proses pelatihan (*training*).

Dengan pengaturan ini, replikasi eksperimen dapat dilakukan dengan cara yang sama, karena seluruh parameter penting seperti random seed, proporsi train–test split, dan validasi telah dijelaskan secara transparan.

• Figure 1: Ensure the caption fully explains each pipeline stage so the figure can stand alone.