

Details of deleted attributes from the mortality dataset with HealthDataPrep

Activity	Details
DC1. Deduplication	Deleted attributes: <i>NATURAL</i> , <i>DTNASC</i> , <i>ESC</i> , <i>CAUSABAS_O</i> AND <i>ESCFALAGR1</i> . The data from these attributes can be extracted from corresponding attributes, such as <i>DTNASC</i> derived from <i>AGE</i> , educational level (<i>ESC</i> and <i>ESCFALAGR1</i>) from <i>ESC2010</i> , cause of death (<i>CAUSABAS_O</i>) from <i>CAUSABAS</i> , and the <i>NATURAL</i> from <i>CODMUNNATU</i> .
DC2. Exclusion of features with zero variance	Deleted attributes: <i>TIPOBITO</i> , <i>CODIFICADO</i> , <i>ALTCAUSA</i> , <i>ASSISTMED</i>
DC3. Handling of low-variance and low-cardinality attributes	Deleted attributes: <i>ORIGEM</i> , <i>STDOEPIDEM</i> AND <i>STDONOVA</i> . <i>STDOEPIDM</i> (epidemiological death declaration status) overwhelmingly contained "No" (99.999%) with negligible "Yes" (0.001%), leading to low variance. <i>ORIGEM</i> (Record source) overwhelmingly contained "Ignored" (99.999%) with negligible "Others" (0.001%), leading to low variance. <i>STDONOVA</i> (Death Certificate Status) overwhelmingly contained "No" (99.999%) with negligible "Yes" (0.001%), leading to low variance.
DC4. Missing data handling	Deleted attributes: <i>IDADEMAE</i> , <i>ESCMAC</i> , <i>ESCMAC2010</i> , <i>SERIESCMAC</i> , <i>OCUPMAE</i> , <i>QTDFILVIVO</i> , <i>QTDFILMORT</i> , <i>GRAVIDEZ</i> , <i>SEMAGESTAC</i> , <i>GESTACAO</i> , <i>PARTO</i> , <i>OBITOPARTO</i> , <i>PESO</i> , <i>CB_PRE</i> , <i>ACIDTRAB</i> , <i>FONTE</i> , <i>CAUSAMAT</i> , <i>ESCMACAGR1</i> , <i>TPRESGINFO</i> , <i>DTCADINF</i> , <i>MORTEPARTO</i> AND <i>DTCONCASO</i>
DC5. Inconsistent data handling	Deleted attributes: <i>SERIESCFAL</i> , <i>TPMORTEOCO</i> , <i>OBITOGRAV</i> , <i>OBITOPUERP</i> , <i>EXAME</i> , <i>CIRURGIA</i> , <i>COMUNSVOM</i> , <i>CIRCOBITO</i> , <i>DTINVESTIG</i> , <i>FONTEINV</i> , <i>NUDIASOBCO</i> , <i>DTCADINV</i> , <i>TPOBITOCOR</i> , <i>DTCONINV</i> , <i>FONTES</i> , <i>TPNIVELINV</i> Attributes with more than 75% of their values as NaN or "Ignored"
FE1.2. Manual feature selection 19	Delete temporal attributes: <i>DTOBITO</i> , <i>HORAOBITO</i> , <i>DTATESTADO</i> , <i>DTCADASTRO</i> , <i>DTRECEBIM</i> , <i>DTRECORIGA</i> , <i>DIFDATA</i> Delete irrelevant attributes: <i>OCUP</i> , <i>CODESTAB</i> , <i>CODMUNNATU</i> , <i>LINHAI</i> , <i>TPPOS</i> , <i>ATESTANTE</i> , <i>STCODIFICA</i> , <i>VERSAOSCB</i> , <i>VERSAOSIST</i> , <i>ATESTADO</i> , <i>LOCOCOR</i> , <i>CODMUNOCOR</i>
DP1. De-identification and suppression 2	Deleted attributes: <i>CONTADOR</i> , <i>NUMEROLOTE</i>