The data for this competition was collected from around 9000 young (15 to 30 years old) woman subjects when they visited clinics in 9 underdeveloped regions, with around 1000 subjects in each region. Each subject was asked by clinical practitioners some questions and her answers were recorded, together with her demographic information. The sexual and reproductive health risks were then evaluated by clinical practitioners and are assigned to different risk segments and subgroups.

Here are the descriptions of the variables in training and testing data.

**Training data:**

|  |  |  |  |
| --- | --- | --- | --- |
| Column Index | Variable Name | Variable Type | Description |
| 1 | patientID | Int | Unique patient IDs in all 9 regions |
| 2 | geo | Int | 1-9. See Appendix for geo to region mapping |
| 3 | christian | Binary, Int | 0=No, 1=Yes |
| 4 | muslim | Binary, Int | 0=No, 1=Yes |
| 5 | hindu | Binary, Int | 0=No, 1=Yes |
| 6 | other | Binary, Int | 0=No, 1=Yes |
| 7 | cellphone | Binary, Int | owns? 0= no 1 = yes |
| 8 | motorcycle | Binary, Int | owns? 0= no 1 = yes |
| 9 | radio | Binary, Int | owns? 0= no 1 = yes |
| 10 | cooker | Binary, Int | owns? 0= no 1 = yes |
| 11 | fridge | Binary, Int | owns? 0= no 1 = yes |
| 12 | furniture | Binary, Int | owns? 0= no 1 = yes |
| 13 | computer | Binary, Int | owns? 0= no 1 = yes |
| 14 | cart | Binary, Int | owns? 0= no 1 = yes |
| 15 | irrigation | Binary, Int | 0=No, 1=Yes |
| 16 | thrasher | Binary, Int | owns? 0= no 1 = yes |
| 17 | car | Binary, Int | owns? 0= no 1 = yes |
| 18 | generator | Binary, Int | owns? 0= no 1 = yes |
| 19 | INTNR | Int | Individual identifier. Unique in each region. Not unique in different regions. |
| 20 | REGION\_PROVINCE | Int | See appendix for mapping |
| 21 | DISTRICT | Int | See appendix for mapping |
| 22 | electricity | Int | Has access to electricity 1=yes, 2=no, 3=sometimes \*\*note this is different from other binary variables |
| 23 | age | Int | Age of the subject |
| 24 | tribe | Int | See appendix for ID to name mapping |
| 25 | foodinsecurity | Int | 1=none or little; 2=some; 3=a lot |
| 26 | EVER\_HAD\_SEX | Binary, Int | 0=no; 1=yes |
| 27 | EVER\_BEEN\_PREGNANT | Binary, Int | 0=no; 1=yes |
| 28 | CHILDREN | Binary, Int | Has any children (0=no, 1=yes) |
| 29 | india | Binary, Int | 0=no; 1=yes |
| 30 | married | Binary, Int | 0=no; 1=yes |
| 31 | multpart | Int (0, 1, 2) | number of recent sexual partners: 0=none in past 12 months, 1=1 in past 12 months, 2=2 or more in past 12 months |
| 32 | educ | Int | ordinal level of education; labels vary by geography but lower = less and higher = more |
| 33 | inschool | Binary, Int | 0=no; 1=yes |
| 34 | ownincome | Binary, Int | 0=no; 1=yes |
| 35 | literacy | Int | ordinal measure of literacy, used to construct highlit and lowlit |
| 36 | religion | String | Religion of subject |
| 37 | urbanicity | Int (0, 1, 2) | 0=rural, 1=middle of rural and urban, 2=urban |
| 38 | LaborDeliv | Binary, Int | Had both labor and delivery in a health care facility |
| 39 | babydoc | Int | Count: number of serious symptoms for which the woman would not seek medical care for newborn |
| 40 | Debut | Int | Actual (if occured) or ideal (if not yet occurred) age of sexual debut; right censored at age 30 for normality |
| 41 | ModCon | Binary, Int | uses or has used modern contraception (only relevant if sexually active) |
| 42 | usecondom | Binary, Int | uses a condom consistently (only relevant for sexually active) 0=no 1=yes |
| 43 | hivknow | Binary, Int | knows HIV status, only relevant if sexually active. 0=no 1=yes |
| 44 | lowlit | Binary, Int | 0=can read somewhat or well; 1=cannot read or reads very poorly |
| 45 | highlit | Binary, Int | can read fluently 0=no 1=yes, constructed from literacy |
| 46 | urban | Binary, Int | 0=no 1=yes |
| 47 | rural | Binary, Int | 0=no 1=yes |
| 48 | single | Binary, Int | 0=married or in a relationship 1=single |
| 49 | segment | Int | Each geography has 2-4 segments. These segments DIFFER BY GEOGRAPHY. The same segment ID in different regions does not mean the same risk segment |
| 50 | subgroup | Int | Some segments have two subgroups: lower risk and higher risk. Only relevant conditional on segment membership |

**Testing data:**

The testing data that your submitted web service API is going to be called to predict has the same data schema as the training data. The only difference is that the true labels in the label columns (segment and subgroup) in the testing data are replaced with 0s. The testing data is fixed, but not shared with participants.

**How data was split into training and testing data?**

The entire data with around 9000 subjects in this study was randomly split with stratification into training (60%) and testing (the remaining 40%) data. The random split is stratified on geo+segment+subgroup. Such a stratified splitting will ensure that the distributions of labels in both training and testing data are consistent in each region.

**How testing data was split into public and private testing data?**

The testing data was further split into public and private testing data, in the same stratified way as splitting into training and testing data. 60% of the testing data is put into public testing data, and remaining 40% is put into private testing data.

**Appendix: ID to Name Mapping**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | **Geo ID to Region Mapping** | | | | 1 = ‘Bihar-India’ | 4 = ‘Ethiopia’ | 7 = ‘Northen Nigeria’ | | 2 = ‘Burkina Faso’ | 5 = ‘Kenya’ | 8 = ‘Southern Nigeria’ | | 3 = ‘Democratic Republic Congo’ | 6 = ‘Myanmar’ | 9 = ‘Uttar Pradesh India’ |  |  |  |  | | --- | --- | --- | | **REGION\_PROVINCE ID to Name Mapping** | | | | 1 = 'Nairobi' | 24 = 'Bas Congo' | 47 = 'Meerut' | | 2 = 'Central' | 25 = 'Bandundu' | 48 = 'Aligarh' | | 3 = 'Coast' | 26 = 'Equateur' | 49 = 'Kanpur Rural' | | 4 = 'Eastern' | 27 = 'Maniema' | 50 = 'Allahabad' | | 5 = 'North Eaatern' | 28 = 'North Central' | 51 = 'Gonda' | | 6 = 'Nyanza' | 29 = 'North East' | 52 = 'Jhansi' | | 7 = 'Rift Valley' | 30 = 'North West' | 53 = 'East Champaran' | | 8 = 'Western' | 31 = 'South East' | 54 = 'Darbhanga' | | 9 = 'TIGRAY' | 32 = 'South South' | 55 = 'Purnia' | | 10 = 'AFFAR' | 33 = 'South West' | 56 = 'Bhojpur' | | 11 = 'AMHARA' | 34 = 'Boucle Du Mouhoun' | 57 = 'Munger' | | 12 = 'OROMIA' | 35 = 'Cascades' | 58 = 'Gaya' | | 13 = 'Somali' | 36 = 'Centre' | 59 = 'Yangon' | | 14 = 'SNNP' | 37 = 'Centre Est' | 60 = 'Mandalay' | | 15 = 'ADDIS ABABA' | 38 = 'Centre Nord' | 61 = 'Mon' | | 16 = 'DIRE DAWA' | 39 = 'Centre Ouest' | 62 = 'Shan' | | 17 = 'Kinshasa' | 40 = 'Centre Sud' | 63 = 'Rakhine' | | 18 = 'Katanga' | 41 = 'Est' | 64 = 'Ayeyarwaddy' | | 19 = 'Province Orientale' | 42 = 'Hauts Bassins' | 65 = 'Bago' | | 20 = 'Kassai Oriental' | 43 = 'Nord' | 66 = 'Sagaing' | | 21 = 'Kassai Occidental' | 44 = 'Plateau Central' | 67 = 'Magway' ; | | 22 = 'Sud Kivu' | 45 = 'Sahel' |  | | 23 = 'Nord Kivu' | 46 = 'Sud-Ouest' |  |  |  |  |  | | --- | --- | --- | | **DISTRICT ID to Name Mapping** | | | | 1 = 'Nairobi West' | 65 = 'Molo' | 129 = 'South Omo-Zone' | | 2 = 'Nairobi East' | 66 = 'Narok North' | 130 = 'Gamo Gofa-Zone' | | 3 = 'Nairobi North' | 67 = 'Narok South' | 131 = 'Bench Maji-Zone' | | 4 = 'Westlands' | 68 = 'Transmara' | 132 = 'Dawro-Zone' | | 5 = 'Nyandarua South' | 69 = 'Kericho' | 133 = 'Silti-Zone' | | 6 = 'Nyeri North' | 70 = 'Buret' | 134 = 'GAMBELLA Region' | | 7 = 'Nyeri South' | 71 = 'Sotik' | 135 = 'ADDIS ABABA' | | 8 = 'Kirinyaga' | 72 = 'Bomet' | 136 = 'DIRE DAWA' | | 9 = 'Muranga North' | 73 = 'Kajiado North' | 137 = 'BENUE' | | 10 = 'Muranga South' | 74 = 'Kakamega Central' | 138 = 'FCT ABUJA' | | 11 = 'Kiambu East (Kiambaa)' | 75 = 'Kakamega North' | 139 = 'JIGAWA' | | 12 = 'Kikuyu' | 76 = 'Kakamega East' | 140 = 'KADUNA' | | 13 = 'Thika West' | 77 = 'Lugari' | 141 = 'KANO' | | 14 = 'Ruiru' | 78 = 'Vihiga' | 142 = 'ENUGU' | | 15 = 'Gatundu' | 79 = 'Mumias' | 143 = 'Abia' | | 16 = 'Mombasa' | 80 = 'Butere' | 144 = 'RIVERS' | | 17 = 'Kilindini' | 81 = 'Bungoma South' | 145 = 'EDO' | | 18 = 'Kinango' | 82 = 'Bungoma West' | 146 = 'LAGOS' | | 19 = 'Kilifi' | 83 = 'Busia' | 147 = 'OYO' | | 20 = 'Kaloleni' | 84 = 'North Western Tigray-Zone' | 148 = 'Ogun' | | 21 = 'Malindi' | 85 = 'Central Tigray-Zone' | 149 = 'Mouhoun' | | 22 = 'Taita' | 86 = 'Eastern Tigray-Zone' | 150 = 'Comoé' | | 23 = 'Imenti North' | 87 = 'Southern Tigray-Zone' | 151 = 'Kadiogo' | | 24 = 'Imenti South' | 88 = 'Affar -ZONE 1' | 152 = 'Boulgou' | | 25 = 'Igembe' | 89 = 'Affar - Zone 2' | 153 = 'Sanmatenga' | | 26 = 'Tigania' | 90 = 'North Gondar-Zone' | 154 = 'Boulkiemde' | | 27 = 'Embu' | 91 = 'South Gondar-Zone' | 155 = 'Zoundwéogo' | | 28 = 'Kitui' | 92 = 'North Wello-Zone' | 156 = 'Gourma' | | 29 = 'Mutomo' | 93 = 'South Wello-Zone' | 157 = 'Houet' | | 30 = 'Mwingi' | 94 = 'North Shewa - Zone' | 158 = 'Yatenga' | | 31 = 'Machakos' | 95 = 'East Gojjam - Zone' | 159 = 'Oubritenga' | | 32 = 'Yatta' | 96 = 'West Gojjam - Zone' | 160 = 'Séno' | | 33 = 'Kangundo' | 97 = 'Waghemira - Zone' | 161 = 'Poni' | | 34 = 'Nzaui' | 98 = 'Awi - Zone' | 162 = 'Meerut' | | 35 = 'Lagdera' | 99 = 'Oromia - Zone' | 163 = 'Aligarh' | | 36 = 'Wajir South' | 100 = 'West Wellega-Zone' | 164 = 'Kanpur Rural' | | 37 = 'Wajir North' | 101 = 'East Wellega-Zone' | 165 = 'Allahabad' | | 38 = 'Wajir East' | 102 = 'Illu Aba Bora-Zone' | 166 = 'Gonda' | | 39 = 'Wajir West' | 103 = 'Jimma-Zone' | 167 = 'Jhansi' | | 40 = 'Siaya' | 104 = 'West Shewa-Zone' | 168 = 'East Champaran' | | 41 = 'Kisumu East' | 105 = 'North Shewa-Zone' | 169 = 'Darbhanga' | | 42 = 'Nyando' | 106 = 'East Shewa-Zone' | 170 = 'Purnia' | | 43 = 'Homabay' | 107 = 'Arsi-Zone' | 171 = 'Bhojpur' | | 44 = 'Suba' | 108 = 'West Hararge - Zone' | 172 = 'Munger' | | 45 = 'Rachounyo' | 109 = 'East Hararge-Zone' | 173 = 'Gaya' | | 46 = 'Migori' | 110 = 'Bale-Zone' | 174 = 'Yangon (North)' | | 47 = 'Rongo' | 111 = 'Borena-Zone' | 175 = 'Mandalay' | | 48 = 'Kisii Central' | 112 = 'South West Shewa-Zone' | 176 = 'Maw La Mying' | | 49 = 'Masaba' | 113 = 'Guji-Zone' | 177 = 'Taunggyi' | | 50 = 'Gucha' | 114 = 'West Arsi-Zone' | 178 = 'Sittwe' | | 51 = 'Gucha South' | 115 = 'Kelem Welega-Zone' | 179 = 'Sagaing' | | 52 = 'Nyamira' | 116 = 'Horo Gudru Welega-Zone' | 180 = 'Pathein' | | 53 = 'Turkana Central' | 117 = 'Shinile-Zone' | 181 = 'Bago' | | 54 = 'Turkana North' | 118 = 'Jijiga-Zone' | 182 = 'Thayet' | | 55 = 'Turkana South' | 119 = 'Degehabur-Zone' | 183 = 'Yangon (South)' | | 56 = 'Pokot Central' | 120 = 'Fik-Zone' | 184 = 'Kyaukse' | | 57 = 'Trans Nzoia West' | 121 = 'Afder-Zone' | 185 = 'Thaton' | | 58 = 'Trans Nzoia East' | 122 = 'Asossa-Zone' | 186 = 'Kyaukme' | | 59 = 'Kwanza' | 123 = 'Gurage-Zone' | 187 = 'Thandwe' | | 60 = 'Eldoret West' | 124 = 'Hadiya-Zone' | 188 = 'Hinthada' | | 61 = 'Laikipia West' | 125 = 'Kembata Tembaro-Zone' | 189 = 'Bago' | | 62 = 'Nakuru' | 126 = 'Sidama-Zone' | 190 = 'Monywa' | | 63 = 'Nakuru North' | 127 = 'Gedeo-Zone' | 191 = 'Thayet' ; | | 64 = 'Naivasha' | 128 = 'Wolayita-Zone' |  |  |  |  |  | | --- | --- | --- | | **Tribe ID to Name Mapping** | | | | 1 = 'Kikuyu' | 25 = 'Khasis' | 49 = 'Mangbetu-Azonde' | | 2 = 'Luhya' | 26 = 'Chenchu' | 50 = 'Others specify' | | 3 = 'Kalenjin' | 27 = 'Bhil' | 51 = 'No tribe/mixed' | | 4 = 'Luo' | 28 = 'Damasa' | 52 = 'Mossi' | | 5 = 'Kamba' | 29 = 'Gaddi' | 53 = 'Fulani Nomads' | | 6 = 'Kisii' | 30 = 'Munda' | 54 = 'Gourma/Gurma' | | 7 = 'Mijikenda' | 31 = 'Gond' | 55 = 'Lobi-Dagari' | | 8 = 'Meru' | 32 = 'Santhal' | 56 = 'Mande' | | 9 = 'Other specify' | 33 = 'Angami' | 57 = 'Bobo' | | 10 = 'No tribe/mixed' | 34 = 'Great Andamanese' | 58 = 'Senoufo' | | 11 = 'Hausa-Fulani' | 35 = 'Others specify' | 59 = 'Gourounsi' | | 12 = 'Igbo' | 36 = 'No tribe/mixed' | 60 = 'Others specific' | | 13 = 'Yoruba' | 37 = 'Oromo' | 61 = 'No tribe/mixed' | | 14 = 'Ijaw' | 38 = 'Amhara' | 62 = 'Mon' | | 15 = 'Kanuri' | 39 = 'Somali' | 63 = 'Bamar' | | 16 = 'Annang' | 40 = 'Tigraway' | 64 = 'Kachin' | | 17 = 'TIV' | 41 = 'Sidama' | 65 = 'Chin' | | 18 = 'Ibibo' | 42 = 'Gurage' | 66 = 'Rakhine' | | 19 = 'Efik' | 43 = 'Wolayta' | 67 = 'Shan' | | 20 = 'Other specify' | 44 = 'Others specify' | 68 = 'Kayar' | | 21 = 'No tribe/mixed' | 45 = 'No tribe/mixed' | 69 = 'Kayin' | | 22 = 'Bodo' | 46 = 'Luba' | 70 = 'Others specify' | | 23 = 'Hmar' | 47 = 'Kongo' | 71 = 'No tribe/mixed' | | 24 = 'Oran' | 48 = 'Mongo' | 72 = 'Refused' ; | |  |