rstrickl7 / **Tanzania-well-analysis**

Ternary classification to predict the condition of water wells in Tanzania. (Description of project: https://github.com/learn-co-curriculum/dsc-phase-3-project)

0 stars 0 forks

Star Unwatch

Code Issues Pull requests Actions Projects Wiki Security

main

rstrickl7 README update ... 1 minute ago 10

View code

README.md

Tanzania-well-analysis

This analysis is based on the competition Driven Data® published about water pumps in Tanzania. The competition information was obtained by the Tanzania Ministry of Water using an open-source platform called Taarifa. Tanzania is the largest country in East Africa, with a population of about 60 million. Half of the population does not have access to clean water. The Tanzanian government is struggling to solve this problem. A significant part of water pumps are entirely out of order or do not function; the others require repair. Tanzania's Ministry of Water Resources agreed with Taarifa, and they launched the DrivenData competition.

Author: [Becky Strickland]

Data

The data has many characteristics associated with water pumps. Data related to geographical locations, organizations that create and manage them, and some data about the region, local government areas. Also, there is information on the types of checkouts, types and number of payments. The water supply points were divided into functional, non-functional and functional but in need of repair. The goal of the competition is to build a model that predicts the functionality of water supply points.

Data Fields

The following set of information about waterpoints is presented for analysis: amount_tsh — Total static head (amount water available to waterpoint) date_recorded — The date the row was entered funder — Who funded the well gps_height — Altitude of the well installer — Organization that installed the well longitude — GPS coordinate latitude — GPS coordinate wpt_name — Name of the waterpoint if there is one num_private — No information basin — Geographic water basin subvillage — Geographic location region — Geographic location region_code — Geographic location (coded) district_code — Geographic location (coded) lga — Geographic location ward — Geographic location population — Population around the well public_meeting — True/False recorded_by — Group entering this row of data scheme_management — Who operates the waterpoint scheme_name — Who operates the waterpoint permit — If the waterpoint is permitted construction_year — Year the waterpoint was constructed extraction_type — The kind of extraction the waterpoint uses extraction_type_group — The kind of extraction the waterpoint uses extraction_type_class — The kind of extraction the waterpoint uses management — How the waterpoint is managed management_group — How the waterpoint is managed payment — What the water costs payment_type — What the water costs water_quality — The quality of the water quality_group — The quality of the water quantity — The quantity of water quantity_group — The quantity of water (duplicates quality) source — The source of the water source_type — The source of the water source_class — The source of the water waterpoint_type — The kind of waterpoint waterpoint_type_group — The kind of waterpoint

This project was created using the following libraries:

```
import pandas as pd import matplotlib.pyplot as plt import matplotlib.ticker as mtick import seaborn as sns import numpy as np import scipy.stats as stats import statsmodels.api as sm import catboost import time import warnings warnings.filterwarnings('ignore')
```

from sklearn.utils import class_weight from sklearn.metrics import accuracy_score, confusion_matrix, classification_report from catboost import Pool, sum_models from catboost import CatBoostClassifier from sklearn.feature_selection import RFE from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error, balanced_accuracy_score from sklearn.model_selection import KFold, cross_val_score, StratifiedKFold from sklearn.preprocessing import LabelEncoder, OneHotEncoder from sklearn.tree import DecisionTreeRegressor from sklearn.ensemble import RandomForestRegressor from sklearn.ensemble import RandomForestClassifier from sklearn.model_selection import train_test_split from sklearn.preprocessing import StandardScaler from sklearn.preprocessing import MinMaxScaler from sklearn.metrics import classification_report, confusion_matrix from sklearn.ensemble import GradientBoostingClassifier from sklearn.linear_model import LogisticRegression from sklearn.model_selection import GridSearchCV from sklearn import metrics from sklearn.model_selection import RandomizedSearchCV from scipy.stats import uniform, truncnorm, randint

For More Information

See the full analysis in the Jupyter Notebook or review this presentation.

Repository Structure

```
├── .ipynb_checkpoints
├── data
├── README.md
├── Tanzanian-well-analysis-Jupyter_Notebook.pdf
├── Tanzanian-well-analysis.ipynb
└── Water-Pump-Analysis.pdf
```

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- Jupyter Notebook 100.0%