PRJ-018: Zero-Day Attack Detection in Logistics Networks

About Dataset

Dataset Title: Zero-Day Attack Detection in Airport Logistics Networks

Overview

This dataset comprises 400,000 network traffic entries collected from logistics networks at major airports in the United States, including those in Texas and Washington. The dataset provides a real-world view of network activity, featuring a mix of benign and malicious traffic, making it an invaluable resource for researchers and practitioners in cybersecurity and network analysis. Please note that some data has been eliminated for privacy purposes.

Features

The dataset consists of 26 features, as outlined below:

Time: Timestamp of the network activity, formatted as YYYY-MM-DD HH:MM
.
Protocol: Type of protocol used for communication (e.g., TCP, UDP).
Flag: TCP flags indicating the state of the connection (e.g., SYN, ACK).
Family: Classification of the traffic, including normal operations and various attack families (e.g., WannaCry, Phishing).
Clusters: Identifier for clustering similar traffic, useful for analyzing patterns.
Source Address: IP address of the device originating the traffic.
Destination Address: IP address of the destination device within the airport network.
BTC: Bitcoin transaction amounts, if applicable.
USD: USD transaction amounts, if applicable.
Netflow Bytes: Total bytes of data transmitted in the flow.
IP Address: Redundant field for clarity, representing the source IP.
Threat Level: Classification indicating the threat level of the traffic (e.g., Benign, Zero-Day Attack).
Port: Port number used for communication.
Prediction: Model prediction indicating whether the traffic is benign or represents an attack.
Payload Size: Size of the data payload transmitted.
Number of Packets: Count of packets involved in the traffic flow.
Application Layer Data: Information about the application layer requests (e.g., HTTP methods).
User-Agent: Information about the client software making the request.
Geolocation: Airport-related geolocation, indicating the specific airport involved (e.g., DFW, SEA).
Logistics ID: Unique identifier for logistics items (e.g., shipment ID).
Anomaly Score: Score indicating the likelihood of the traffic being anomalous or

malicious.
Event Description: Descriptive label for the event, detailing the nature of the traffic.
Response Time: Time taken for the server to respond to the request.
Session ID: Unique identifier for the network session.
Data Transfer Rate: Rate of data transfer, measured in Mbps.
Error Code: HTTP or application-level error codes returned (if applicable).
Dataset Characteristics
Total Entries: 400,000
Class Distribution: 62% benign traffic and 38% representing zero-day attacks and other threats.
Geographical Focus: Traffic data includes activities at major airports, such as Dallas/Fort Worth International Airport (DFW) and Seattle-Tacoma International Airport (SEA).
Use Cases
This dataset can be utilized for:

Research: Investigating zero-day attack detection techniques.
Machine Learning: Training models to classify benign and malicious network traffic.
Network Security: Enhancing security measures in logistics networks at airports.
Conclusion
The "Zero-Day Attack Detection in Airport Logistics Networks" dataset provides a realistic and comprehensive view of network behavior within airport logistics, offering critical insights for developing effective cybersecurity strategies against zero-day threats.

PRJ-017: 2022 National Bridge Inventory Data

Comprehensive Bridge Data: Insights into Structure, Traffic, Condition, Climate

About Dataset

This extensive dataset was generated through https://infobridge.fhwa.dot.gov/, which is a collection of various measurements, ratings, and other information regarding bridges. It could potentially be used for understanding and managing bridge infrastructure, assessing conditions, planning maintenance, and evaluating the impact of climate factors on bridges. Here's a summary of the main data categories:

1. General Bridge Information: This includes unique identifiers, names, and location information (e.g., state, county, city, latitude, longitude, highway agency district) of the bridges.
2. Construction Details: These fields provide information about when the bridge was built or reconstructed, the materials and design of the main span and approach spans, and information on the deck structure.
3. Size and Dimension Measurements: This covers various measurements such as bridge length, number of spans, clearance heights, roadway widths, and other dimensional features.
4. Condition Ratings: The dataset has several fields giving condition ratings for different parts of the bridge, like the deck, superstructure, substructure, and others. There are also fields for the overall bridge condition, structural evaluation appraisal, and a scour critical bridge value.
5. Traffic Information: This includes fields like average daily traffic, future projected traffic, and traffic-related designations (e.g., type of service on bridge code, designated national truck network code).
6. Maintenance and Inspection Data: There's detailed information about inspections, the frequency of inspections, maintenance responsibility, improvement costs, and proposed work.
7. Operational Status and Ratings: This section provides fields such as the operating rating, inventory rating, bridge posting code, and toll status.
8. Weather and Climate Information: This includes fields related to weather conditions and climate data that may affect bridge health and longevity.
9. Governance and Legislation: These fields contain information about the metropolitan planning organization, U.S. congressional district, state senate district, and state house district.

PRJ-016: Network Traffic Data-Malicious Activity Detection

Dataset Overview

This dataset consists of network traffic captured from a Kali Linux machine, aimed at helping the development and evaluation of machine learning models for distinguishing between normal and malicious (specifically flood attack) network activities. It includes a variety of features essential for identifying potential cybersecurity threats alongside labels indicating whether each packet is part of flood traffic.

Data Collection Methodology

The dataset was carefully compiled using network traffic captured from a dedicated Kali Linux setup. The capture environment consisted of a Kali Linux machine configured to generate and capture both normal and malicious network traffic and a target machine running a Windows OS to simulate a real-world network environment.

Traffic Generation:

Normal Traffic: Involved routine network activities such as web browsing and pinging between the Kali Linux machine and the Windows machine.

Malicious Traffic: Utilized hping3 to simulate flood attacks, specifically ICMP flood attacks, targeting the Windows machine from the Kali Linux machine [1].

Capture Process: Wireshark was used on the Kali Linux machine to capture all incoming and outgoing network traffic [2]. The capture was set up to record detailed packet information, including timestamps, source and destination IP addresses, ports, and protocols. The captures were conducted with careful monitoring to precisely mark the start and end times of the flood attack for accurate dataset labeling.

Dataset Description

The dataset is a CSV file containing a comprehensive collection of network traffic packets labeled to distinguish between normal and malicious traffic. It includes the following columns:

Timestamp: The capture time of each packet, providing insights into the traffic flow and enabling analysis of traffic patterns over time.
Source IP Address: Identifies the origin of the packet, crucial for pinpointing potential sources of attacks.
Destination IP Address: Indicates the packet's intended recipient, useful for identifying targeted resources.
Source Port and Destination Port: Offer insights into the services involved in the communication.
Protocol: Specifies the protocol used, such as TCP, UDP, or ICMP, essential for analyzing the nature of the traffic.

Length: The size of the packet in bytes, which can signal unusual traffic patterns often associated with malicious activities.
bad_packet: A binary label with 1 indicating traffic identified as part of a flood attack and 0 denoting normal traffic. Precise timestamps marking the start and end of flood attacks were used to accurately label this column. Packets captured within these defined intervals were marked as malicious (bad_packet = 1), whereas all others were considered normal traffic. Python and Pandas were used for the labeling process [3][4].

Potential Applications

a. Intrusion Detection Systems (IDS): The dataset can be used in training models to enhance IDS capabilities, enabling more effective detection of flood-based network attacks.
b. Network Traffic Monitoring: Tools making use of machine learning can leverage the dataset for more accurate network traffic monitoring, identifying and alerting suspicious activities in real time.
c. Cybersecurity Training: Educational institutions and training programs can use the dataset to provide practical experience in machine learning-based threat detection.

Proposed Machine Learning Technique: Supervised Machine Learning, specifically Deep Learning with Convolutional Neural Networks (CNNs).

CNNs, even though it is usually used for image processing, have shown promise in analyzing sequential data. The spatial hierarchy in network packets (from individual bytes to overall packet structure) can be analogous to the patterns CNNs excel at identifying. Utilizing CNNs could allow for the extraction of complex data in network traffic that indicate malicious activities, improving detection accuracy beyond traditional methods.

Conclusion

This dataset represents a significant step towards using machine learning for cybersecurity, specifically in the field of intrusion detection and network monitoring. By providing a detailed and accurately labeled dataset of normal and malicious network traffic, it lays the groundwork for developing complex models capable of identifying and mitigating flood attacks in real-time. In the future, we could include a broader range of attack types and more traffic patterns, further enhancing the dataset's utility and the effectiveness of models trained on it.

PRJ-015

Diabetic Patients' Re-admission Prediction

Identify factors leading to high readmissions of diabetic patient within 30 days

Diabetic Patients' Re-admission Prediction

arrow_drop_up32

New Notebook

file_download**Download (4 MB)**arrow_drop_down

more_vert

Data CardCode (1)Discussion (0)Suggestions (0)

About Dataset

Dataset name: Diabetes 130-US hospitals for years 1999-2008 Data Set

Background: Diabetes Mellitus (DM) is a chronic disease where the blood has high sugar level. It can occur when the pancreas does not produce enough insulin, or when the body cannot effectively use the insulin it produces (WHO). Diabetes is a progressive disease that can lead to a significant number of health complications and profoundly reduce the quality of life. While many diabetic patients manage the health complication with diet and exercise, some require medications to control blood glucose level. As published by a research article named “The relationship between diabetes mellitus and 30-day readmission rates”, it is estimated that 9.3% of the population in the United States have diabetes mellitus (DM), 28% of which are undiagnosed. In recent years, government agencies and healthcare systems have increasingly focused on 30-day readmission rates to determine the complexity of their patient populations and to improve quality. Thirty-day readmission rates for hospitalized patients with DM are reported to be between 14.4 and 22.7%, much higher than the rate for all hospitalized patients (8.5–13.5%).

Problem Statement: To identify the factors that lead to the high readmission rate of diabetic patients within 30 days post discharge and correspondingly to predict the high-risk diabetic-patients who are most likely to get readmitted within 30 days so that the quality of care can be improved along with improved patient’s experience, health of the population and reduce costs by lowering readmission rates. Also, to identify the medicines that are the most effective in treating diabetes.

Impact on business: Hospital readmission is an important contributor to total medical expenditures and is an emerging indicator of quality of care. Diabetes, similar to other

chronic medical conditions, is associated with increased risk of hospital readmission. As mentioned in the article “Correction to: Hospital Readmission of Patients with Diabetes”, hospital readmission is a high-priority health care quality measure and target for cost reduction, particularly within 30 days of discharge. The burden of diabetes among hospitalized patients is substantial, growing, and costly, and readmissions contribute a significant portion of this burden. Reducing readmission rates among patients with diabetes has the potential to greatly reduce health care costs while simultaneously improving care. Our aim is to provide some insights into the risk factors for readmission and also to identify the medicines that are the most effective in treating diabetes.

Variable identification:
1. Independent variables (49): encounter_id, patient_nbr, race, gender, age, weight, admission_type_id, discharge_disposition_id, admission_source_id, time_in_hospital, payer_code, medical_specialty, num_lab_procedures, num_procedures, num_medications, number_outpatient, number_emergency, number_inpatient, diag_1, diag_2, diag_3, number_diagnoses, max_glu_serum, A1Cresult, metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, citoglipton, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, metformin-pioglitazone, change, diabetesMed.
2. Dependent variable (1): readmitted (Categorical)

Extra Info: Our dataset consists of hospital admissions of length between one and 14 days that did not result in a patient’s death. Each encounter corresponds to a patient diagnosed with diabetes, although the primary diagnosis may be different. During each of the analyzed encounters, lab tests were ordered and medication was administered.

PRJ-011: Cyber Crimes Dataset

Data CardCode (0)Discussion (0)Suggestions (0)

About Dataset

The dataset contains the following columns , each described below :

Attack Type: Randomly selected from a broad set of attack types (e.g., phishing, DDoS, malware, etc.).
Target System: Corporate IT systems such as servers, databases, user accounts, APIs, and more.
Outcome: Whether the attack succeeded or failed.
Timestamp: Time of the attack, randomly distributed over the past year.
Attacker IP Address: Simulated attacker IP addresses.
Target IP Address: Random IP addresses representing internal or external targets.
Data Compromised: Amount of data compromised (in gigabytes) if the attack succeeded.
Attack Duration: Time the attack lasted (in minutes).
Security Tools Used: Various defense mechanisms like firewalls, IDS, antivirus, etc.
User Role: The role of the user impacted by the attack (admin, employee, or external user).
Location: Country or region where the attack originated or targeted.
Attack Severity: Numerical indicator of the severity level (e.g., scale from 1-10).
Industry: Type of industry targeted, such as healthcare, finance, government, etc.
Response Time: Time taken by the security team to respond (in minutes).
Mitigation Method: Steps taken to mitigate the attack (patching, containment, etc.)

PRJ-010: MIT Supercloud Dataset

The high-level summary data for the MIT Datacenter Challenge

About Dataset

For full details of the data please refer to the paper "The MIT Supercloud Dataset", available at https://ieeexplore.ieee.org/abstract/document/9622850 or https://arxiv.org/abs/2108.02037

Dataset

Datacenter monitoring systems offer a variety of data streams and events. The Datacenter Challenge datasets are a combination of high-level data (e.g. Slurm Workload Manager scheduler data) and low-level job-specific time series data. The high-level data includes parameters such as the number of nodes requested, number of CPU/GPU/memory requests, exit codes, and run time data. The low-level time series data is collected on the order of seconds for each job. This granular time series data includes CPU/GPU/memory utilization, amount of disk I/O, and environmental parameters such as power drawn and temperature. Ideally, leveraging both high-level scheduler data and low-level time series data will facilitate the development of AI/ML algorithms which not only predict/detect failures, but also allow for the accurate determination of their cause.

Here I will only include the high-level data.

If you are interested in using the dataset, please cite this paper.

@INPROCEEDINGS{9773216,

author={Li, Baolin and Arora, Rohin and Samsi, Siddharth and Patel, Tirthak and Arcand, William and Bestor, David and Byun, Chansup and Roy, Rohan Basu and Bergeron, Bill and Holodnak, John and Houle, Michael and Hubbell, Matthew and Jones, Michael and Kepner, Jeremy and Klein, Anna and Michaleas, Peter and McDonald, Joseph and Milechin, Lauren and Mullen, Julie and Prout, Andrew and Price, Benjamin and Reuther, Albert and Rosa, Antonio and Weiss, Matthew and Yee, Charles and Edelman, Daniel and Vanterpool, Allan and Cheng, Anson and Gadepally, Vijay and Tiwari, Devesh},

PRJ-009: Kubernetes: Resource & Performance Metrics Allocation

This dataset is available in two files, each in .csv and .xlsx format.

Data file I:
timestamp
pod_name
namespace
cpu_allocation_efficiency
memory_allocation_efficiency
disk_io
network_latency
node_temperature
node_cpu_usage
node_memory_usage
event_type
event_message
scaling_event
pod_lifetime_seconds

Data file II:
pod_name, namespace,
cpu_request,
cpu_limit,
memory_request,
memory_limit,
cpu_usage,
memory_usage,
node_name,
pod_status,
restart_count,
uptime_seconds,
deployment_strategy,scaling_policy,
network_bandwidth_usage

PRJ-008: Drone-Based Malware Detection (DBMD) - Network Traffic

About Dataset

Description
Welcome to the Drone-Based Malware Detection dataset! This dataset is designed to aid researchers and practitioners in exploring innovative cybersecurity solutions using drone-collected data. The dataset contains detailed information on network traffic, drone sensor readings, malware detection indicators, and environmental conditions. It offers a unique perspective by integrating data from drones with traditional network security metrics to enhance malware detection capabilities.

Dataset Overview
The dataset comprises four main categories:

Network Traffic Data: Captures network traffic attributes including IP addresses, ports, protocols, packet sizes, and various derived metrics.
Drone Sensor Data: Includes GPS coordinates, altitude, speed, heading, battery level, and other sensor readings from drones.
Malware Detection Data: Contains indicators and scores relevant to detecting malware, such as anomaly scores, suspicious IP counts, reputation scores, and attack types.
Environmental Data: Provides context through environmental conditions like location type, noise level, weather conditions, and more.

Files and Features
The dataset is divided into four separate CSV files:

network_traffic_data.csv

timestamp: Date and time of the traffic event.
source_ip: Source IP address.
destination_ip: Destination IP address.
source_port: Source port number.
destination_port: Destination port number.
protocol: Network protocol (TCP, UDP, ICMP).
packet_length: Length of the network packet.
payload_data: Content of the packet payload.
flag: Network flag (SYN, ACK, FIN, RST).
traffic_volume: Volume of traffic in bytes.
flow_duration: Duration of the network flow.
flow_bytes_per_s: Bytes per second for the flow.
flow_packets_per_s: Packets per second for the flow.
packet_count: Number of packets in the flow.
average_packet_size: Average size of packets.

min_packet_size: Minimum packet size.
max_packet_size: Maximum packet size.
packet_size_variance: Variance in packet sizes.
header_length: Length of the packet header.
payload_length: Length of the packet payload.
ip_ttl: Time to live for the IP packet.
tcp_window_size: TCP window size.
icmp_type: ICMP type (echo_request, echo_reply, destination_unreachable).
dns_query_count: Number of DNS queries.
dns_response_count: Number of DNS responses.
http_method: HTTP method (GET, POST, PUT, DELETE).
http_status_code: HTTP status code (200, 404, 500, 301).
content_type: Content type (text/html, application/json, image/png).
ssl_tls_version: SSL/TLS version.
ssl_tls_cipher_suite: SSL/TLS cipher suite.
drone_data.csv

latitude: Latitude of the drone.
longitude: Longitude of the drone.
altitude: Altitude of the drone.
speed: Speed of the drone.
heading: Heading of the drone.
battery_level: Battery level of the drone.
drone_id: Unique identifier for the drone.
flight_time: Total flight time.
signal_strength: Strength of the drone's signal.
temperature: Temperature at the drone's location.
humidity: Humidity at the drone's location.
pressure: Atmospheric pressure at the drone's location.
wind_speed: Wind speed at the drone's location.
wind_direction: Wind direction at the drone's location.
gps_accuracy: Accuracy of the GPS signal.
malware_detection_data.csv

anomaly_score: Score indicating the level of anomaly detected.
suspicious_ip_count: Number of suspicious IP addresses detected.
malicious_payload_indicator: Indicator for malicious payload (0 or 1).
reputation_score: Reputation score for the network entity.
behavioral_score: Behavioral score indicating potential malicious activity.
attack_type: Type of attack (DDoS, phishing, malware).
signature_match: Indicator for signature match (0 or 1).
sandbox_result: Result from sandbox analysis (clean, infected).

heuristic_score: Heuristic score for potential threats.
traffic_pattern: Pattern of the traffic (burst, steady).
environmental_data.csv

location_type: Type of location (urban, rural).
nearby_devices: Number of nearby devices.
signal_interference: Level of signal interference.
noise_level: Noise level in the environment.
time_of_day: Time of day (morning, afternoon, evening, night).
day_of_week: Day of the week.
weather_conditions: Weather conditions (sunny, rainy, cloudy, stormy).
Usage and Applications
This dataset can be used for:

Cybersecurity Research: Developing and testing algorithms for malware detection using drone data.
Machine Learning: Training models to identify malicious activity based on network traffic and drone sensor readings.
Data Analysis: Exploring the relationships between environmental conditions, drone sensor data, and network traffic anomalies.
Educational Purposes: Teaching data science, machine learning, and cybersecurity concepts using a comprehensive and multi-faceted dataset.

Acknowledgements
This dataset is based on real-world data collected from drone sensors and network traffic monitoring systems. The data is anonymized to ensure privacy and is intended for research and educational purposes only.

PRJ-007: Greedy Cloud Selection Deployment on Microservices

Greedy Multi-Cloud Selection Approach to Deploy Microservices-Based Applications

About Dataset

The dataset titled "Greedy Multi-Cloud Selection Approach to Deploy an Application Based on Microservices" consists of 400,000 rows and 11 columns, capturing various parameters essential for deploying microservices-based applications across multiple cloud environments. This dataset is designed to simulate and analyze the deployment decisions and outcomes when using a greedy algorithm approach for cloud provider selection.

Columns Overview:
Application ID: Unique identifier for each application instance.
Microservice Name: Name of the microservice within the application.
Cloud Provider: Chosen cloud provider for deploying the microservice (e.g., AWS, Azure, Google Cloud, IBM Cloud).
Region: Geographic region where the cloud provider's data center is located (e.g., US-East, EU-West, Asia-Pacific).
Resource Utilization (%): Percentage of allocated resources utilized by the microservice.
Latency (ms): Average latency experienced by the microservice in milliseconds.
Cost ($): Deployment cost in US dollars incurred by the microservice.
Deployment Time (hrs): Time taken to deploy the microservice in hours.
Success Rate (%): Percentage of successful deployments for the microservice.
Data Transfer (GB): Amount of data transferred by the microservice in gigabytes.
Environment: Deployment environment phase (e.g., Development, Testing, Production).
Dataset Usage:
This dataset facilitates research and analysis into the efficacy of a greedy algorithm for selecting optimal cloud providers based on various performance and cost metrics. Researchers and practitioners can use this dataset to:

Evaluate the impact of cloud provider choice on resource utilization and deployment costs.
Analyze latency variations across different geographic regions and cloud providers.
Assess the success rate of deployments and its correlation with selected cloud providers and deployment environments.
Model and optimize deployment strategies for microservices-based applications in diverse cloud environments.
Data Characteristics:
The data ranges were simulated to reflect realistic scenarios encountered in multi-cloud deployments, ensuring variability in cloud provider performance and deployment outcomes. Random generation methods such as uniform distributions for costs and deployment times, normal distributions for latency, and categorical choices for cloud

providers and deployment environments provide a diverse yet controlled dataset suitable for comprehensive analysis.

Potential Applications:
This dataset is valuable for researchers, data scientists, and cloud architects involved in optimizing cloud resource utilization, minimizing deployment costs, and enhancing application performance through effective cloud provider selection strategies. It can also serve as a benchmark for comparing different algorithms and methodologies in the field of multi-cloud deployment and management.

PRJ-006: Orchestration and Kubernettes-Based Cloud Computing

Features

Company Name,

Deployment Date,

Cloud Provider,

Resource Type,

Efficiency Rating,

Resource Usage (GB),

Scalability Index,

Reliability Score,

Flexibility Level,

Compliance Status,

Performance Index

PRJ-005: Crop mapping using fused optical-radar data set

Combining optical and PolSAR remote sensing images offers a complementary data set with a significant number of temporal, spectral, textural, and polarimetric features for cropland classification.

Dataset Characteristics

Multivariate, Time-Series

Subject Area

Other

Associated Tasks

Classification

Feature Type

Real

 # Instances

325834

 # Features

175

Dataset Information

Additional Information

This big data set is a fused bi-temporal optical-radar data for cropland classification. The images were collected by RapidEye satellites (optical) and the Unmanned Aerial Vehicle Synthetic Aperture Radar (UAVSAR) system (Radar) over an agricultural region near Winnipeg, Manitoba, Canada on 2012. There are 2 * 49 radar features and 2 * 38 optical features for two dates: 05 and 14 July 2012. Seven crop type classes exist for this data set as follows: 1-Corn; 2-Peas; 3- Canola; 4-Soybeans; 5- Oats; 6- Wheat; and 7-Broadleaf.

Variable Information

175 attributes including: 1- class; 2- f1 to f49:Polarimetric features on 05 July 2012; 3- f50 to f98:Polarimetric features on 14 July 2012; 4- f99 to f136:Optical features on 05 July 2012; 5- f137 to f174:Optical features on 14 July 2012; Details: label:crop type class f1:sigHH_Rad05July f2:sigHV_Rad05July f3:sigVV_Rad05July f4:sigRR_Rad05July f5:sigRL_Rad05July f6:sigLL_Rad05July f7:Rhhvv_Rad05July f8:Rhvhh_Rad05July

f9:Rhvvv_Rad05July f10:Rrrll_Rad05July f11:Rrlrr_Rad05July f12:Rrlll_Rad05July
f13:Rhh_Rad05July f14:Rhv_Rad05July f15:Rvv_Rad05July f16:Rrr_Rad05July
f17:Rrl_Rad05July f18:Rll_Rad05July f19:Ro12_Rad05July f20:Ro13_Rad05July
f21:Ro23_Rad05July f22:Ro12cir_Rad05July f23:Ro13cir_Rad05July
f24:Ro23cir_Rad05July f25:l1_Rad05July f26:l2_Rad05July f27:l3_Rad05July
f28:H_Rad05July f29:A_Rad05July f30:a_Rad05July f31:HA_Rad05July
f32:H1mA_Rad05July f33:1mHA_Rad05July f34:1mH1mA_Rad05July f35:PH_Rad05July
f36:rvi_Rad05July f37:paulalpha_Rad05July f38:paulbeta_Rad05July
f39:paulgamma_Rad05July f40:krogks_Rad05July f41:krogkd_Rad05July
f42:krogkh_Rad05July f43:freeodd_Rad05July f44:freedbl_Rad05July
f45:freevol_Rad05July f46:yamodd_Rad05July f47:yamdbl_Rad05July
f48:yamhlx_Rad05July f49:yamvol_Rad05July f50:sigHH_Rad14July
f51:sigHV_Rad14July f52:sigVV_Rad14July f53:sigRR_Rad14July f54:sigRL_Rad14July
f55:sigLL_Rad14July f56:Rhhvv_Rad14July f57:Rhvhh_Rad14July f58:Rhvvv_Rad14July
f59:Rrrll_Rad14July f60:Rrlrr_Rad14July f61:Rrlll_Rad14July f62:Rhh_Rad14July
f63:Rhv_Rad14July f64:Rvv_Rad14July f65:Rrr_Rad14July f66:Rrl_Rad14July
f67:Rll_Rad14July f68:Ro12_Rad14July f69:Ro13_Rad14July f70:Ro23_Rad14July
f71:Ro12cir_Rad14July f72:Ro13cir_Rad14July f73:Ro23cir_Rad14July f74:l1_Rad14July
f75:l2_Rad14July f76:l3_Rad14July f77:H_Rad14July f78:A_Rad14July f79:a_Rad14July
f80:HA_Rad14July f81:H1mA_Rad14July f82:1mHA_Rad14July f83:1mH1mA_Rad14July
f84:PH_Rad14July f85:rvi_Rad14July f86:paulalpha_Rad14July f87:paulbeta_Rad14July
f88:paulgamma_Rad14July f89:krogks_Rad14July f90:krogkd_Rad14July
f91:krogkh_Rad14July f92:freeodd_Rad14July f93:freedbl_Rad14July
f94:freevol_Rad14July f95:yamodd_Rad14July f96:yamdbl_Rad14July
f97:yamhlx_Rad14July f98:yamvol_Rad14July f99:B_Opt05July f100:G_Opt05July
f101:R_Opt05July f102:Redge_Opt05July f103:NIR_Opt05July f104:NDVI_Opt05July
f105:SR_Opt05July f106:RGRI_Opt05July f107:EVI_Opt05July f108:ARVI_Opt05July
f109:SAVI_Opt05July f110:NDGI_Opt05July f111:gNDVI_Opt05July
f112:MTVI2_Opt05July f113:NDVIre_Opt05July f114:SRre_Opt05July
f115:NDGIre_Opt05July f116:RTVIcore_Opt05July f117:RNDVI_Opt05July
f118:TCARI_Opt05July f119:TVI_Opt05July f120:PRI2_Opt05July
f121:MeanPC1_Opt05July f122:VarPC1_Opt05July f123:HomPC1_Opt05July
f124:ConPC1_Opt05July f125:DisPC1_Opt05July f126:EntPC1_Opt05July
f127:SecMomPC1_Opt05July f128:CorPC1_Opt05July f129:MeanPC2_Opt05July
f130:VarPC2_Opt05July f131:HomPC2_Opt05July f132:ConPC2_Opt05July
f133:DisPC2_Opt05July f134:EntPC2_Opt05July f135:SecMomPC2_Opt05July
f136:CorPC2_Opt05July f137:B_Opt14July f138:G_Opt14July f139:R_Opt14July
f140:Redge_Opt14July f141:NIR_Opt14July f142:NDVI_Opt14July f143:SR_Opt14July
f144:RGRI_Opt14July f145:EVI_Opt14July f146:ARVI_Opt14July f147:SAVI_Opt14July
f148:NDGI_Opt14July f149:gNDVI_Opt14July f150:MTVI2_Opt14July
f151:NDVIre_Opt14July f152:SRre_Opt14July f153:NDGIre_Opt14July

f154:RTVIcore_Opt14July f155:RNDVI_Opt14July f156:TCARI_Opt14July
f157:TVI_Opt14July f158:PRI2_Opt14July f159:MeanPC1_Opt14July
f160:VarPC1_Opt14July f161:HomPC1_Opt14July f162:ConPC1_Opt14July
f163:DisPC1_Opt14July f164:EntPC1_Opt14July f165:SecMomPC1_Opt14July
f166:CorPC1_Opt14July f167:MeanPC2_Opt14July f168:VarPC2_Opt14July
f169:HomPC2_Opt14July f170:ConPC2_Opt14July f171:DisPC2_Opt14July
f172:EntPC2_Opt14July f173:SecMomPC2_Opt14July f174:CorPC2_Opt14July

PRJ-004: Classification of pixels into 7 forest cover types

Classification of pixels into 7 forest cover types based on attributes such as elevation, aspect, slope, hillshade, soil-type, and more.

Dataset Characteristics

Multivariate

Subject Area

Biology

Associated Tasks

Classification

Feature Type

Categorical, Integer

 # Instances

581012

 # Features

54

Dataset Information

Additional Information

Predicting forest cover type from cartographic variables only (no remotely sensed data). The actual forest cover type for a given observation (30 x 30 meter cell) was determined from US Forest Service (USFS) Region 2 Resource Information System (RIS) data. Independent variables were derived from data originally obtained from US Geological Survey (USGS) and USFS data. Data is in raw form (not scaled) and contains binary (0 or 1) columns of data for qualitative independent variables (wilderness areas and soil types). This study area includes four wilderness areas located in the Roosevelt National Forest of northern Colorado. These areas represent forests with minimal human-caused disturbances, so that existing forest cover types are more a result of ecological processes rather than forest management practices. Some background information for these four wilderness areas: Neota (area 2) probably has the highest mean elevational value of the 4 wilderness areas. Rawah (area 1) and Comanche Peak (area 3) would have a lower mean elevational value, while Cache la Poudre (area 4) would have the lowest mean elevational value. As for primary major tree species in these areas, Neota would have spruce/fir (type 1), while Rawah and Comanche Peak would probably have lodgepole pine (type 2) as their primary species, followed by spruce/fir and aspen (type

5). Cache la Poudre would tend to have Ponderosa pine (type 3), Douglas-fir (type 6), and cottonwood/willow (type 4). The Rawah and Comanche Peak areas would tend to be more typical of the overall dataset than either the Neota or Cache la Poudre, due to their assortment of tree species and range of predictive variable values (elevation, etc.) Cache la Poudre would probably be more unique than the others, due to its relatively low elevation range and species composition.

Variables Table

Variable Name	Role	Type	Description	Units	Missing Values
Elevation	Feature	Integer			no
Aspect	Feature	Integer			no
Slope	Feature	Integer			no
Horizontal_Distance_To_Hydrology	Feature	Integer			no
Vertical_Distance_To_Hydrology	Feature	Integer			no
Horizontal_Distance_To_Roadways	Feature	Integer			no
Hillshade_9am	Feature	Integer			no
Hillshade_Noon	Feature	Integer			no
Hillshade_3pm	Feature	Integer			no
Horizontal_Distance_To_Fire_Points	Feature	Integer			no

Additional Variable Information

Given is the attribute name, attribute type, the measurement unit and a brief description. The forest cover type is the classification problem. The order of this listing corresponds to the order of numerals along the rows of the database. Name / Data Type / Measurement / Description Elevation / quantitative /meters / Elevation in meters Aspect / quantitative / azimuth / Aspect in degrees azimuth Slope / quantitative / degrees / Slope in degrees Horizontal_Distance_To_Hydrology / quantitative / meters / Horz Dist to nearest surface water features Vertical_Distance_To_Hydrology / quantitative / meters / Vert Dist to nearest surface water features Horizontal_Distance_To_Roadways / quantitative / meters / Horz Dist to nearest roadway Hillshade_9am / quantitative / 0 to 255 index / Hillshade index at 9am, summer solstice Hillshade_Noon / quantitative / 0 to 255 index / Hillshade index at noon,

summer soltice Hillshade_3pm / quantitative / 0 to 255 index / Hillshade index at 3pm, summer solstice Horizontal_Distance_To_Fire_Points / quantitative / meters / Horz Dist to nearest wildfire ignition points Wilderness_Area (4 binary columns) / qualitative / 0 (absence) or 1 (presence) / Wilderness area designation Soil_Type (40 binary columns) / qualitative / 0 (absence) or 1 (presence) / Soil Type designation Cover_Type (7 types) / integer / 1 to 7 / Forest Cover Type designation

Class Labels

Spruce/Fir, Lodgepole Pine, Ponderosa Pine, Cottonwood/Willow, Aspen, Douglas-fir, Krummholz

Dataset Files

File	Size
covtype.data.gz	10.7 MB
covtype.info	14.3 KB
old_covtype.info	4.7 KB

PRJ-003: Dota2 Games Results

Dota 2 is a popular computer game with two teams of 5 players. At the start of the game each player chooses a unique hero with different strengths and weaknesses.

Dataset Characteristics

Multivariate

Subject Area

Games

Associated Tasks

Classification

Feature Type

-

 # Instances

102944

 # Features

115

Dataset Information

Additional Information

Dota 2 is a popular computer game with two teams of 5 players. At the start of the game each player chooses a unique hero with different strengths and weaknesses. The dataset is reasonably sparse as only 10 of 113 possible heroes are chosen in a given game. All games were played in a space of 2 hours on the 13th of August, 2016 The data was collected using: https://gist.github.com/da-steve101/1a7ae319448db431715bd75391a66e1b

PRJ-002: Diabetes 130-US Hospitals for Years 1999-2008

The dataset represents ten years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. Each row concerns hospital records of patients diagnosed with diabetes, who underwent laboratory, medications, and stayed up to 14 days. The goal is to determine the early readmission of the patient within 30 days of discharge. The problem is important for the following reasons. Despite high-quality evidence showing improved clinical outcomes for diabetic patients who receive various preventive and therapeutic interventions, many patients do not receive them. This can be partially attributed to arbitrary diabetes management in hospital environments, which fail to attend to glycemic control. Failure to provide proper diabetes care not only increases the managing costs for the hospitals (as the patients are readmitted) but also impacts the morbidity and mortality of the patients, who may face complications associated with diabetes.

Dataset Characteristics	Multivariate
Subject Area	Health and Medicine
Associated Tasks	Classification, Clustering
Feature Type	Categorical, Integer
# Instances	101766
# Features	47

Dataset Information

What do the instances in this dataset represent?

The instances represent hospitalized patient records diagnosed with diabetes.

Are there recommended data splits?

No recommendation. The standard train-test split could be used. Can use three-way holdout split (i.e., train-validation-test) when doing model selection.

Does the dataset contain data that might be considered sensitive in any way?

Yes. The dataset contains information about the age, gender, and race of the patients.

Additional Information

The dataset represents ten years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes over 50 features representing patient and hospital outcomes. Information was extracted from the database for encounters that satisfied the following criteria. (1) It is an inpatient encounter (a hospital admission). (2) It is a diabetic encounter, that is, one during which any kind of diabetes was entered into

the system as a diagnosis. (3) The length of stay was at least 1 day and at most 14 days. (4) Laboratory tests were performed during the encounter. (5) Medications were administered during the encounter. The data contains such attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab tests performed, HbA1c test result, diagnosis, number of medications, diabetic medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalization, etc.

PRJ-001: PhiUSIIL Phishing URL (Website)

PhiUSIIL Phishing URL Dataset is a substantial dataset comprising 134,850 legitimate and 100,945 phishing URLs. Most of the URLs we analyzed, while constructing the dataset, are the latest URLs. Features are extracted from the source code of the webpage and URL. Features such as CharContinuationRate, URLTitleMatchScore, URLCharProb, and TLDLegitimateProb are derived from existing features.

Dataset Characteristics	Tabular
Subject Area	Computer Science
Associated Tasks	Classification
Feature Type	Real, Categorical, Integer
# Instances	235795
# Features	54

Dataset Information

What do the instances in this dataset represent?

URLs and their corresponding webpages