WEBSITE TRAFFIC ANALYSIS

NAAN MUDHALVAN PROJECT REPORT

Submitted by

VINAYKRISHNA B	(730321243031)
LINGESWARAN R	(730321243013)
SOUNDAR N	(730321243031)
SUHASH M	(730321243027)

FIFTH SEMESTER

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

BUILDERS ENGINEERING COLLEGE , KANGEYAM

ANNA UNIVERSITY :: CHENNAI 600 025

NOVEMBER 2023

KGI

BUILDERS ENGINEERING COLLEGE

Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

(An ISO 9001:2015 Certified Institution | Recognized 2(f) Status by UGC)

NAAC A+ 3.36

BONAFIDE CERTIFICATE

Certified that this is a Bonafide record of work done by **VINAYKRISHNA B(730321243031) , LINGESWARAN R(730321243013)** and **SOUNDAR N(730321243025) , SUHASH M(730321243027)** in NAAN MUTHALVAN during the Academic year 2023- 24 for fifth Semester.

STAFF -INCHARGE **HEAD OF THE DEPARTMENT**

Submitted for the Naan mudhalvan viva voce held on______________

INTERNAL EXAMINER **EXTERNAL EXAMINER**

ABSTRACT

In the ever-evolving landscape of the internet, understanding website traffic is crucial for businesses and individuals alike. This abstract provides a glimpse into the world of website traffic analysis, shedding light on its significance and methodologies.Website traffic analysis is the process of examining and evaluating the data generated by visitors to a website. This data encompasses a myriad of information, such as the number of visitors, their geographic location, the pages they visit, the duration of their stay, and their interactions with the site. By harnessing this wealth of information, website owners and digital marketers can make informed decisions to enhance user experience, optimize content, and drive business growth.This abstract outlines the key components of website traffic analysis, including the use of analytics tools, the importance of data segmentation, and the significance of key performance indicators (KPIs). It highlights the role of Google Analytics and other tracking tools in gathering, organizing, and visualizing website data.Additionally, the abstract underscores the relevance of actionable insights gained through traffic analysis. With this information, website owners can fine-tune their content, marketing strategies, and user interface to meet the needs and expectations of their audience.In conclusion, website traffic analysis is a powerful tool for unlocking the potential of online platforms. Whether you are running a personal blog, an e-commerce site, or a corporate website, a deeper understanding of your audience can lead to improved online presence, user engagement, and business outcomes. This abstract serves as an introduction to the multifaceted world of website traffic analysis, inviting readers to explore the full spectrum of knowledge and techniques in this field.

TABLE OF CONTENT

CHAPTER 1

PROBLEM STATEMENT

Analyzing and improving public transport systems is essential for enhancing urban mobility, reducing traffic congestion, and promoting sustainability.

Key Challenges:

1. **Data Accuracy and Reliability:** Ensuring the accuracy and reliability of the data collected can be a challenge. Data can be skewed by bots, web crawlers, or other automated traffic, leading to inaccurate insights. Distinguishing between genuine user interactions and automated ones is essential.
2. **Privacy and Compliance:** With increasing concerns about user privacy, website traffic analysis must comply with data protection regulations such as GDPR, CCPA, and others. Balancing data collection for analysis with user privacy rights can be complex.
3. **Data Volume and Scalability:** For websites with high traffic, the sheer volume of data can be overwhelming. Managing and analyzing large datasets efficiently requires robust infrastructure and tools.
4. **Data Integration**: Many websites use multiple tools and platforms for various purposes (e.g., content management, e-commerce, social media). Integrating data from these disparate sources into a cohesive analysis can be challenging.
5. **User Identification:** Identifying individual users while respecting their privacy is a complex issue. Tracking user behavior across sessions and devices without violating privacy regulations is a challenge.
6. **Data Interpretation:** Interpreting the data and extracting meaningful insights can be daunting. Knowing which metrics matter and how to apply them to improve website performance is not always straightforward.

CHAPTER 2

DESIGN THINKING

ANALYSIS OBJECTIVES

The analysis of public transport systems serves several key objectives, each aimed at improving the efficiency, accessibility, sustainability, and overall effectiveness of urban transportation.

1. **Understanding User Behavior:** Analyzing website traffic helps in understanding how users navigate through the site, what pages they visit, how long they stay, and what actions they take (e.g., clicking links, making purchases, filling out forms). This insight helps in tailoring the user experience to meet their needs.
2. **Audience Segmentation:** Segmenting website visitors based on various criteria (e.g., demographics, location, referral source) allows for targeted content and marketing strategies. Understanding the different segments can help in personalization and better user engagement.
3. **Content Evaluation**: Analyzing which content is the most popular, what keeps users engaged, and what content leads to conversions (e.g., sign-ups, sales) helps in content optimization. It allows you to focus on creating and promoting content that resonates with your audience.
4. **Traffic Sources:** Determining where website traffic is coming from (e.g., search engines, social media, referrals, direct) helps in assessing the effectiveness of different marketing channels. It enables you to allocate resources to the most productive sources.
5. **Conversion Rate Analysis:** Tracking conversions, such as sign-ups, downloads, purchases, or other desired actions, and understanding the paths users take to achieve these goals. This helps in optimizing conversion funnels and increasing the conversion rate.
6. **User Engagement:** Measuring user engagement metrics like bounce rate, time on page, and pageviews per session helps in assessing the quality of user interactions with the site. High engagement is often associated with user satisfaction.

7. **Page Performance:** Analyzing the performance of individual web pages helps in identifying areas that need improvement. This includes page load times, error pages, and identifying which pages contribute most to conversions.
8. **Mobile Responsiveness:** As mobile traffic continues to grow, analyzing how well the site performs on different devices is crucial. Understanding the mobile user experience and optimizing for mobile users is an important objective.
9. **Search Engine Optimization (SEO):** Analyzing keyword traffic, search engine rankings, and click-through rates from search results helps in optimizing the website for better search engine visibility and organic traffic.
10. **Identifying Trends:** Recognizing trends in website traffic can provide insights into seasonal variations, changes in user behavior, and emerging patterns. This can inform content planning and marketing strategies.
11. **User Journey Mapping:** Mapping the typical user journey through the site, from entry to conversion, helps in improving the user experience and making it easier for users to achieve their goals.
12. **A/B Testing and Experimentation:** Conducting A/B tests and experiments to evaluate changes to the website and determine their impact on user behavior, conversions, and overall site performance.

CHAPTER 3

DATASET DEFINITION

SURVEY DATASET:

A dataset for website traffic analysis is a structured collection of data that contains information related to how users interact with a website. This data is typically collected through web analytics tools and can be used for various analytical purposes to gain insights into user behavior and website performance. Here's a general dataset definition for website traffic analysis:

Data Fields and Descriptions:

1. **Date and Time:** The timestamp indicating when each user interaction occurred. It includes the date and time of each visit or action.
2. **User ID or Session ID:** A unique identifier for each user or session, allowing for tracking a user's journey through the website.
3. **Page URL:** The URL of the web page visited by the user.
4. **Page Title:** The title of the web page visited, which provides context about the content of the page.
5. **Referral Source:** The source from which the user arrived on the website (e.g., search engine, social media, referral website, direct entry).
6. **Geographic Location:** The geographic location of the user, typically including information such as country, region, and city.
7. **Device Type:** The type of device used by the user (e.g., desktop, mobile, tablet).
8. **Browser and Operating System:** Information about the user's web browser and operating system.

9. **User Interactions:** Details of user interactions on the website, including pageviews, clicks on links or buttons, time spent on pages, and any form submissions.
10. **Conversion Events:** Information about actions that indicate a conversion, such as sign-ups, purchases, downloads, or other desired outcomes.
11. **Traffic Source:** The specific source or medium responsible for driving the traffic (e.g., Google organic search, Facebook referral, paid advertising campaign).
12. **Keywords (if available):** The keywords or search queries that led users to the website from search engines**.**
13. **Exit Page:** The last page visited before the user left the website.
14. **Bounce Rate:** A binary indicator (yes/no) of whether a user bounced (visited a single page and left) or continued exploring the website.
15. **Session Duration**: The duration of each user's visit/session on the website.
16. **Traffic Channel:** Categorization of the traffic source into different channels (e.g., organic, paid, direct, social, email).
17. **Page Load Time:** The time it takes for each web page to load in the user's browser.

Popular sources of survey datasets include government agencies (e.g., U.S. Census Bureau), academic research projects, market research firms, and non-governmental organizations. These datasets play a significant role in empirical research and can provide valuable insights into various aspects of society, behavior, and economics.

CHAPTER 4

DATA PREPROCESSING AND FEATURE EXTRACTION

DATA PREPROCESSING

Data preprocessing in website traffic analysis involves cleaning, transforming, and organizing raw data to prepare it for meaningful analysis. This essential step ensures that the data is accurate and relevant. Common preprocessing tasks include removing duplicates, handling missing values, converting data types, and normalizing or scaling numerical variables.

1. **Data Collection and Integration**: Gather data from various sources, such as web analytics tools like Google Analytics, server logs, and marketing platforms. Integrate this data into a unified dataset for analysis.
2. **Data Cleaning:** Identify and address data quality issues, including removing duplicates, handling missing values, and correcting inaccuracies. Clean data is essential for accurate analysis.
3. **Data Transformation:** Convert and format data as needed. This may involve converting timestamps into a consistent time zone, standardizing naming conventions, and encoding categorical variables for analysis.
4. **Data Filtering:** Filter out irrelevant or bot-generated data that can skew analysis results. Identifying and removing automated traffic is crucial for accurate insights.
5. **Data Aggregation:** Aggregate data to different time intervals (e.g., hourly, daily) to reveal trends and patterns over time. Aggregated data can help in identifying peak traffic periods and user behavior changes.
6. **User Sessionization:** Define and segment user sessions based on specific criteria, like time gaps between interactions. This helps in understanding user journeys and behavior within a single session.

7. **Data Reduction:** Reduce the dataset's size by selecting relevant columns or variables for analysis, especially when dealing with large volumes of data. This can improve analysis efficiency.
8. **Normalization and Scaling**: Standardize numerical variables to a common scale to prevent bias in analyses that rely on numeric values. Scaling can help when comparing variables with different units.
9. **Handling Outliers:** Identify and address outliers that can distort analysis results. Decide whether to remove, transform, or keep outliers based on their impact and significance.
10. **Data Security and Privacy:** Ensure data handling complies with privacy regulations, particularly when dealing with personally identifiable information. Anonymize or pseudonymize data when necessary.
11. **Data Validation:** Validate data against predefined criteria to ensure it meets quality and consistency standards. This step helps identify and rectify data anomalies.
12. **Documentation:** Maintain comprehensive documentation of data preprocessing steps to ensure transparency and repeatability of the analysis.

FEATURE EXTRACTION:

Feature extraction in website traffic analysis involves the process of selecting, transforming, or creating relevant attributes (features) from the raw data to facilitate analysis and gain insights into user behavior, performance, and other aspects of website traffic. This step simplifies the data and helps in the identification of patterns and trends. Here are some common feature extraction techniques used in website traffic analysis:

1. **Page-Level Features:** These features are derived from individual web pages and help in assessing the performance and popularity of specific content.

- Pageviews: The number of times a page is viewed.
- Bounce Rate: The percentage of single-page visits.
- Average Time on Page: The average time users spend on a page.
- Exit Rate: The percentage of users who exit the website after viewing a specific page.

2. **User Behavior Features:** These features capture user interactions and provide insights into how users engage with the website.

- Click-through Rate (CTR): The percentage of users who click on links or buttons.
- Conversion Rate: The percentage of users who complete desired actions (e.g., sign-ups, purchases).
- Session Duration: The duration of a user's visit.
- Path Analysis: Identifying common user journeys through the website.

3. **Traffic Source Features:** These features focus on the sources that drive traffic to the website.

- Referral Source: The source from which users arrive (e.g., search engines, social media, referral websites).
- Organic vs. Paid Traffic: Distinguishing between organic search and paid advertising traffic.
- Direct vs. Indirect Traffic: Categorizing direct visits and indirect visits from referrals.

4. **Demographic and Geographic Features:** If available, demographic and geographic data can be valuable for segmenting and targeting users.

- User Location: Geographic information about the users (country, city).
- User Demographics: Age, gender, interests, if collected and available.

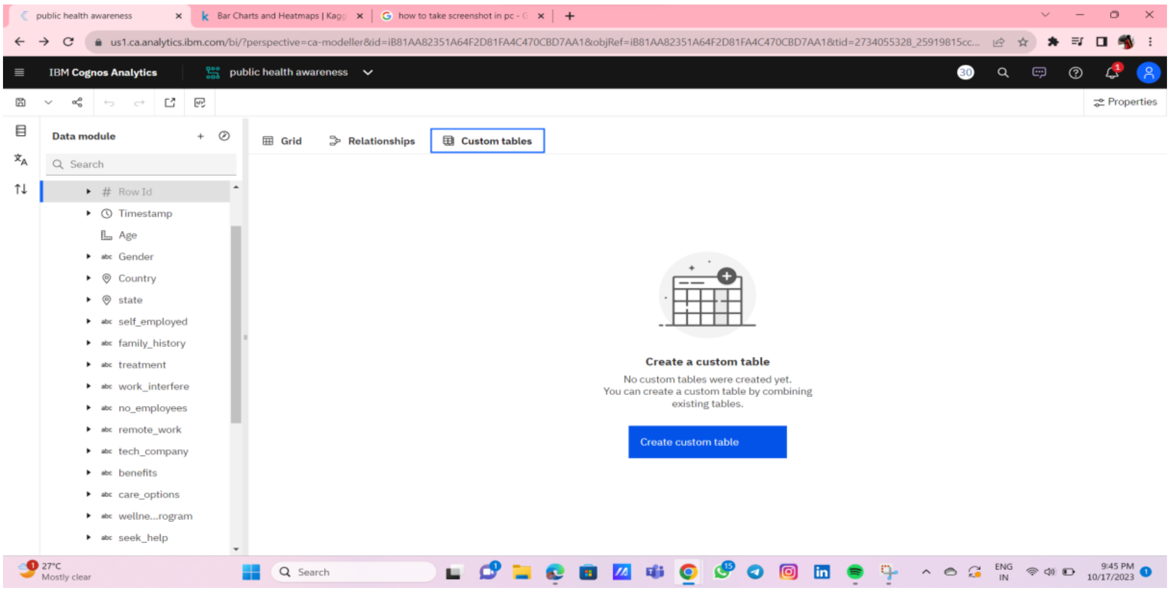
5. **Time-Based Features:** Analyzing website traffic over time is crucial for identifying trends and seasonality.

- Hourly, Daily, or Weekly Patterns: Examining traffic fluctuations during specific time intervals.
- Day of the Week: Identifying which days are busiest.
- Seasonal Trends: Recognizing recurring patterns or season-specific behaviors.

6. **Device and Browser Features:** Understanding user preferences in terms of device and browser.

- Device Type: Differentiating between desktop, mobile, and tablet users.
- Browser Usage: Identifying which web browsers are most commonly used.

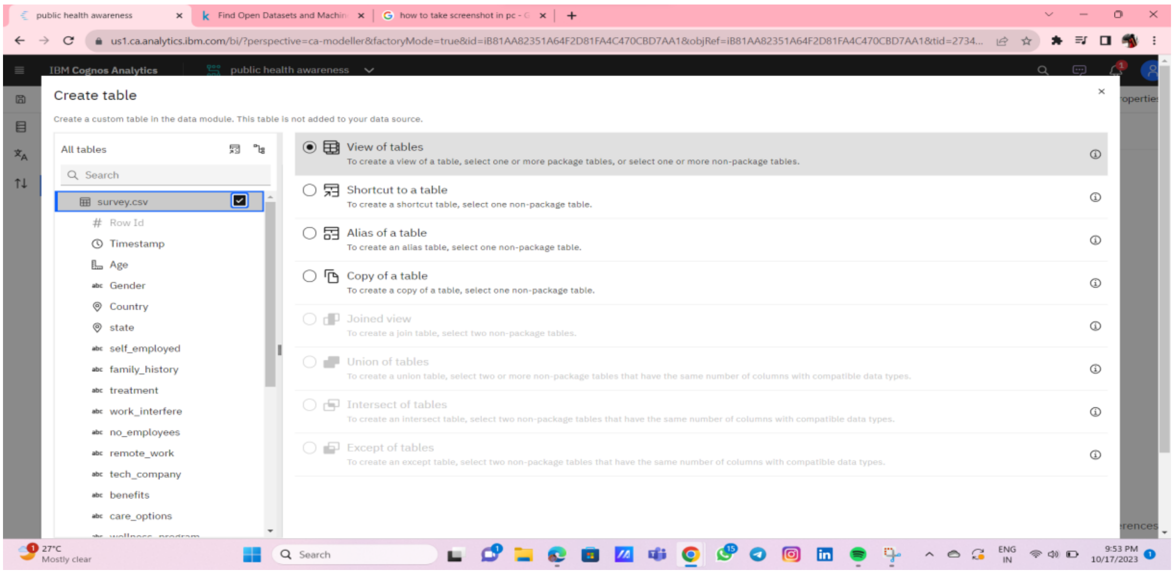
Feature extraction is a crucial step in website traffic analysis because it simplifies complex data into actionable insights. These features enable analysts to identify trends, optimize content, and make informed decisions to enhance the user experience and achieve business goals.



Create a custom table

No custom tables were created yet.
You can create a custom table by combining existing tables.

Fig 4.1



Create table

Create a custom table in the data module. This table is not added to your data source.

● View of tables
To create a view of a table, select one or more package tables, or select one or more non-package tables.

○ Shortcut to a table
To create a shortcut table, select one non-package table.

○ Alias of a table
To create an alias table, select one non-package table.

○ Copy of a table
To create a copy of a table, select one non-package table.

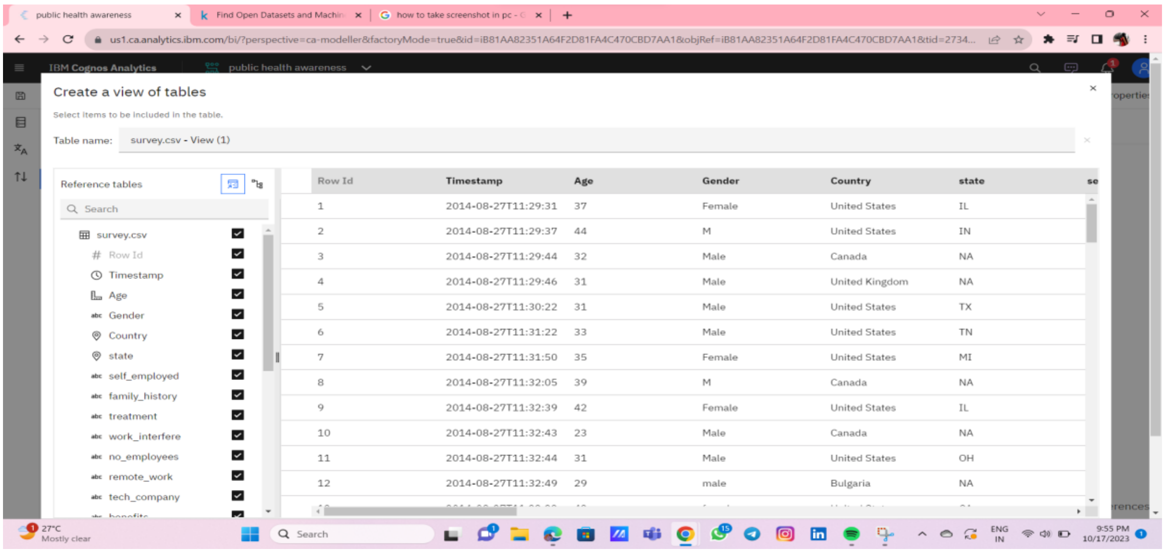
○ Joined view
To create a join table, select two non-package tables.

○ Union of tables
To create a union table, select two or more non-package tables that have the same number of columns with compatible data types.

○ Intersect of tables
To create an intersect table, select two non-package tables that have the same number of columns with compatible data types.

○ Except of tables
To create an except table, select two non-package tables that have the same number of columns with compatible data types.

Fig 4.2


public health awareness
Find Open Datasets and Machin
how to take screenshot in pc
us1.ca.analytics.ibm.com/bi/?perspective=ca-modeller&factoryMode=true&id=iB81AA82351A64F2D81FA4C470CBD7AA1&objRef=iB81AA82351A64F2D81FA4C470CBD7AA1&tid=2734...
IBM Cognos Analytics
public health awareness
27°C
Mostly clear
Search
ENG
IN
9:55 PM
10/17/2023

Create a view of tables

Select items to be included in the table.

Table name: survey.csv - View (1)

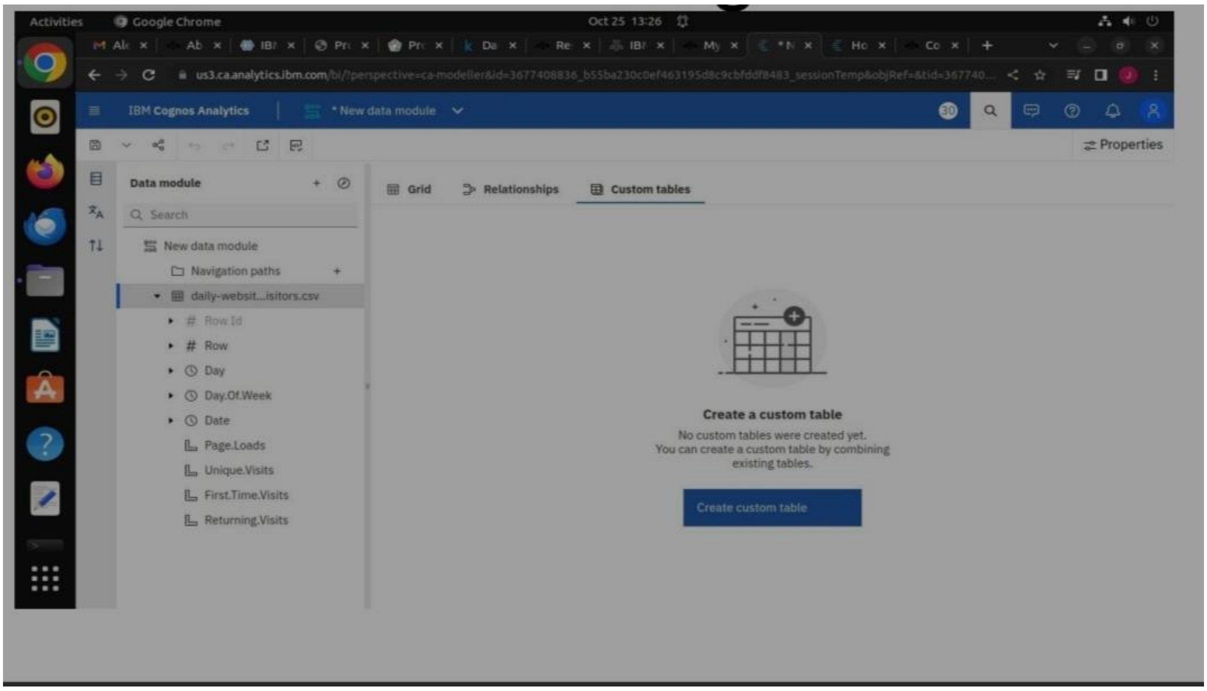
Reference tables

Search

- survey.csv ☑
  - Row Id ☑
  - Timestamp ☑
  - Age ☑
  - Gender ☑
  - Country ☑
  - state ☑
  - self_employed ☑
  - family_history ☑
  - treatment ☑
  - work_interfere ☑
  - no_employees ☑
  - remote_work ☑
  - tech_company ☑

Row Id	Timestamp	Age	Gender	Country	state
1	2014-08-27T11:29:31	37	Female	United States	IL
2	2014-08-27T11:29:37	44	M	United States	IN
3	2014-08-27T11:29:44	32	Male	Canada	NA
4	2014-08-27T11:29:46	31	Male	United Kingdom	NA
5	2014-08-27T11:30:22	31	Male	United States	TX
6	2014-08-27T11:31:22	33	Male	United States	TN
7	2014-08-27T11:31:50	35	Female	United States	MI
8	2014-08-27T11:32:05	39	M	Canada	NA
9	2014-08-27T11:32:39	42	Female	United States	IL
10	2014-08-27T11:32:43	23	Male	Canada	NA
11	2014-08-27T11:32:44	31	Male	United States	OH
12	2014-08-27T11:32:49	29	male	Bulgaria	NA

Fig 4.3


Activities
Google Chrome
Oct 25 13:26
us3.ca.analytics.ibm.com/bi/?perspective=ca-modeller&id=3677408836_b55ba230c0ef463195d8c9cbfddf8483_sessionTemp&objRef=&tid=367740...
IBM Cognos Analytics
* New data module
Properties
Grid
Relationships
Custom tables
Create custom table

Data module

Search

- New data module
  - Navigation paths
  - daily-websit...isitors.csv
    - Row Id
    - Row
    - Day
    - Day.Of.Week
    - Date
    - Page.Loads
    - Unique.Visits
    - First.Time.Visits
    - Returning.Visits

Create a custom table

No custom tables were created yet.
You can create a custom table by combining existing tables.

Fig 4.4

CHAPTER 5

PROPOSED ALGORITHM

SUPPORT VECTOR MACHINES (SVM)

Support Vector Machines (SVM) are a class of supervised machine learning algorithms used for classification and regression tasks. They are powerful tools for both linear and nonlinear data analysis. SVMs work by finding a hyperplane that best separates data points into different classes, while maximizing the margin between the two classes. Here are the key concepts and components of Support **Vector Machines:**

Linear Separation:

SVMs are originally designed for binary classification problems, aiming to find a hyperplane that best separates two classes of data points.

Hyperplane:

In a two-dimensional space, a hyperplane is a line, and in higher-dimensional spaces, it's a flat affine subspace.

Margin:

The margin is the distance between the hyperplane and the nearest data point from either class. SVM aims to maximize this margin.

Support Vectors:

Support vectors are the data points that are closest to the hyperplane. These data points play a critical role in defining the margin and the decision boundary.

Kernel Trick:

SVMs can handle nonlinear data by transforming the input data into a higher-dimensional space. This transformation is often done using a kernel function (e.g., polynomial, radial basis function) that allows SVM to find nonlinear decision boundaries.

C Parameter:

The C parameter in SVM controls the trade-off between maximizing the margin and minimizing the classification error. A small C value prioritizes a wider margin (potentially allowing some misclassification), while a large C value prioritizes correctly classifying as many points as possible (potentially leading to a smaller margin).

Soft Margin SVM:

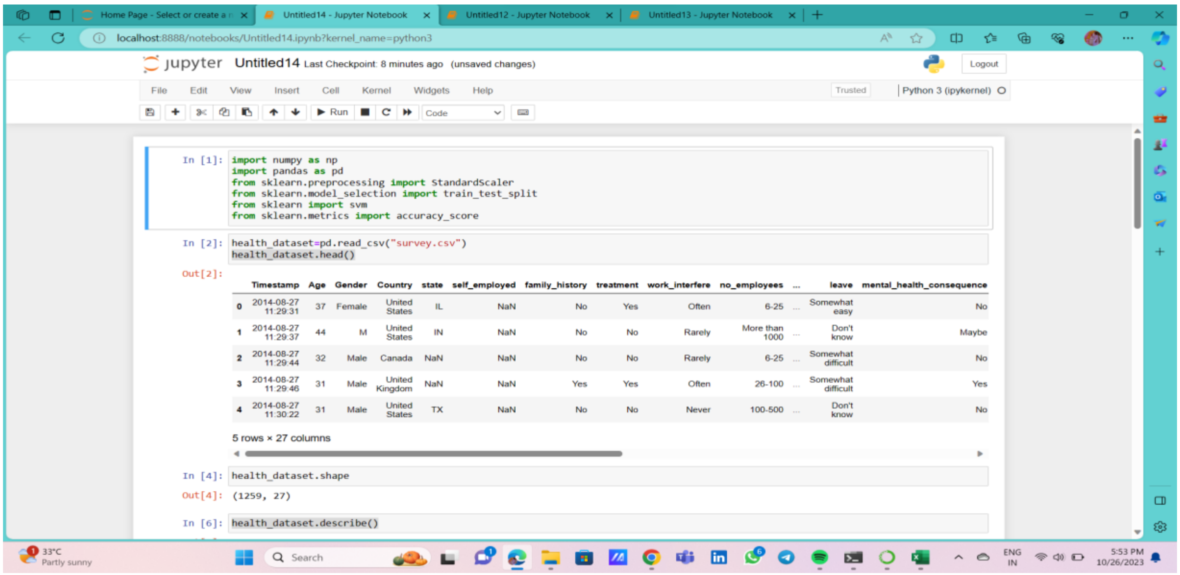
In practical applications, data may not be linearly separable. In such cases, a soft margin SVM allows for a certain degree of misclassification to find a balance between maximizing the margin and minimizing errors.

Multi-Class Classification:

SVMs can be extended to handle multi-class classification problems. Common techniques include one-vs-one and one-vs-all strategies.

Regression (Support Vector Regression):

SVMs can also be used for regression tasks by fitting a hyperplane to predict a continuous output value. In this case, the margin represents an ε-insensitive tube around the predicted values.


localhost:8888/notebooks/Untitled14.ipynb?kernel_name=python3
jupyter Untitled14 Last Checkpoint: 8 minutes ago (unsaved changes)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
Run
Code

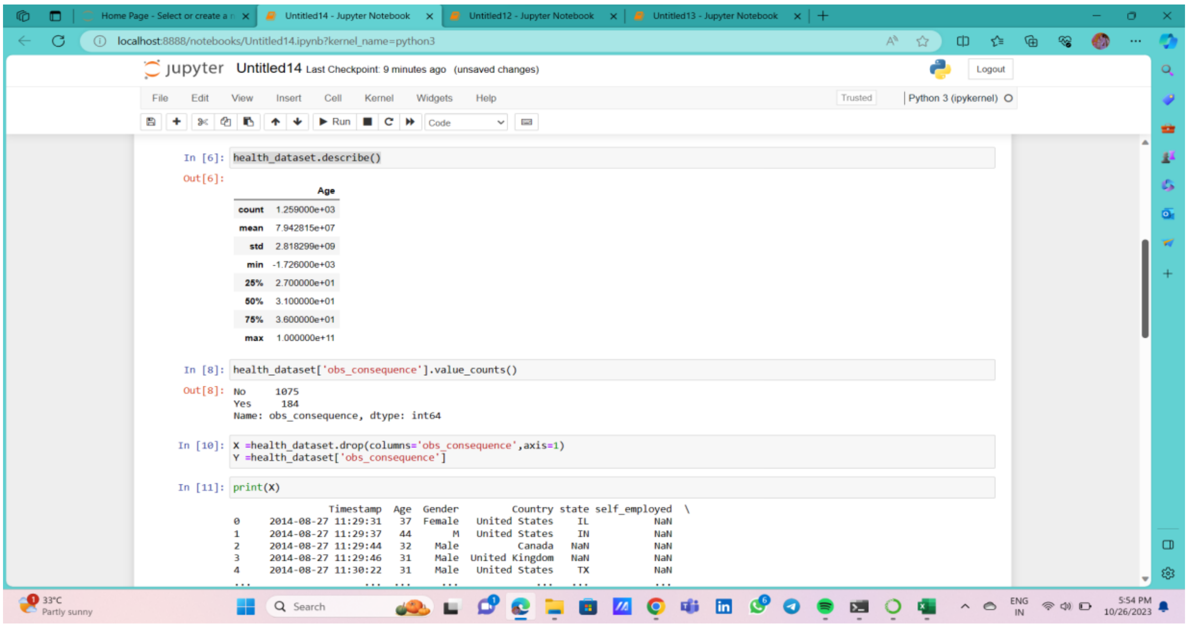
```
In [1]: import numpy as np
        import pandas as pd
        from sklearn.preprocessing import StandardScaler
        from sklearn.model_selection import train_test_split
        from sklearn import svm
        from sklearn.metrics import accuracy_score

In [2]: health_dataset=pd.read_csv("survey.csv")
        health_dataset.head()

Out[2]:
```

	Timestamp	Age	Gender	Country	state	self_employed	family_history	treatment	work_interfere	no_employees	...	leave	mental_health_consequence
0	2014-08-27 11:29:31	37	Female	United States	IL	NaN	No	Yes	Often	6-25	...	Somewhat easy	No
1	2014-08-27 11:29:37	44	M	United States	IN	NaN	No	No	Rarely	More than 1000	...	Don't know	Maybe
2	2014-08-27 11:29:44	32	Male	Canada	NaN	NaN	No	No	Rarely	6-25	...	Somewhat difficult	No
3	2014-08-27 11:29:46	31	Male	United Kingdom	NaN	NaN	Yes	Yes	Often	26-100	...	Somewhat difficult	Yes
4	2014-08-27 11:30:22	31	Male	United States	TX	NaN	No	No	Never	100-500	...	Don't know	No

5 rows × 27 columns

```
In [4]: health_dataset.shape

Out[4]: (1259, 27)

In [6]: health_dataset.describe()
```

Fig 6.1


Home Page - Select or create a
Untitled14 - Jupyter Notebook
Untitled12 - Jupyter Notebook
Untitled13 - Jupyter Notebook
localhost:8888/notebooks/Untitled14.ipynb?kernel_name=python3
jupyter
Untitled14 Last Checkpoint: 9 minutes ago (unsaved changes)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
Run
Code
33°C
Partly sunny
Search
ENG
IN
5:54 PM
10/26/2023

In [6]:
```
health_dataset.describe()
```

Out[6]:

	Age
count	1.259000e+03
mean	7.942815e+07
std	2.818299e+09
min	-1.726000e+03
25%	2.700000e+01
50%	3.100000e+01
75%	3.600000e+01
max	1.000000e+11

In [8]:
```
health_dataset['obs_consequence'].value_counts()
```

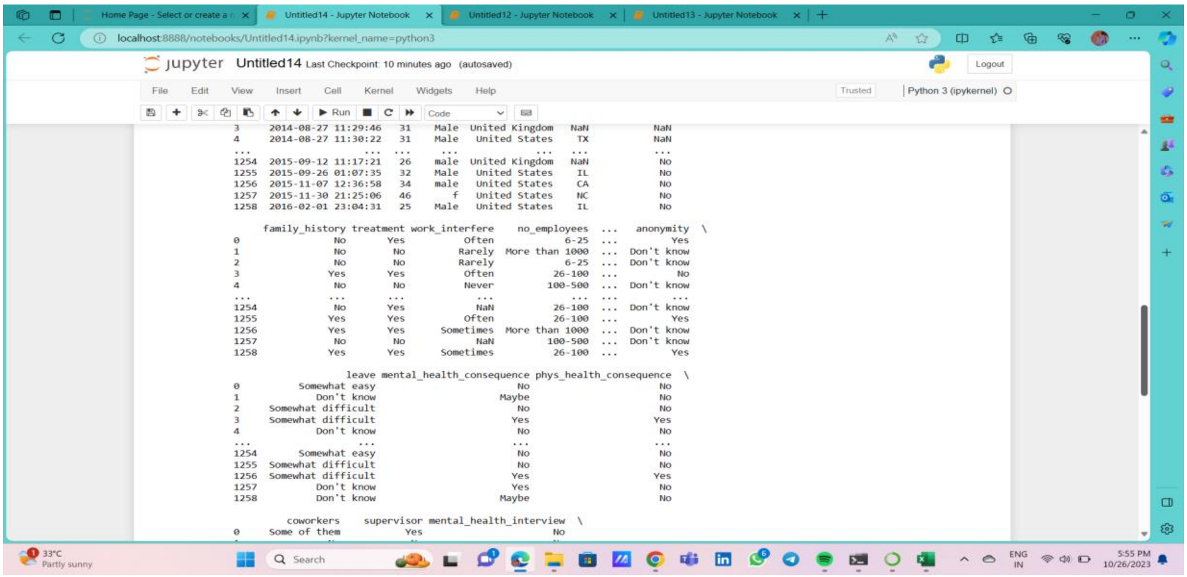
Out[8]:
```
No      1075
Yes      184
Name: obs_consequence, dtype: int64
```

In [10]:
```
X =health_dataset.drop(columns='obs_consequence',axis=1)
Y =health_dataset['obs_consequence']
```

In [11]:
```
print(X)
```

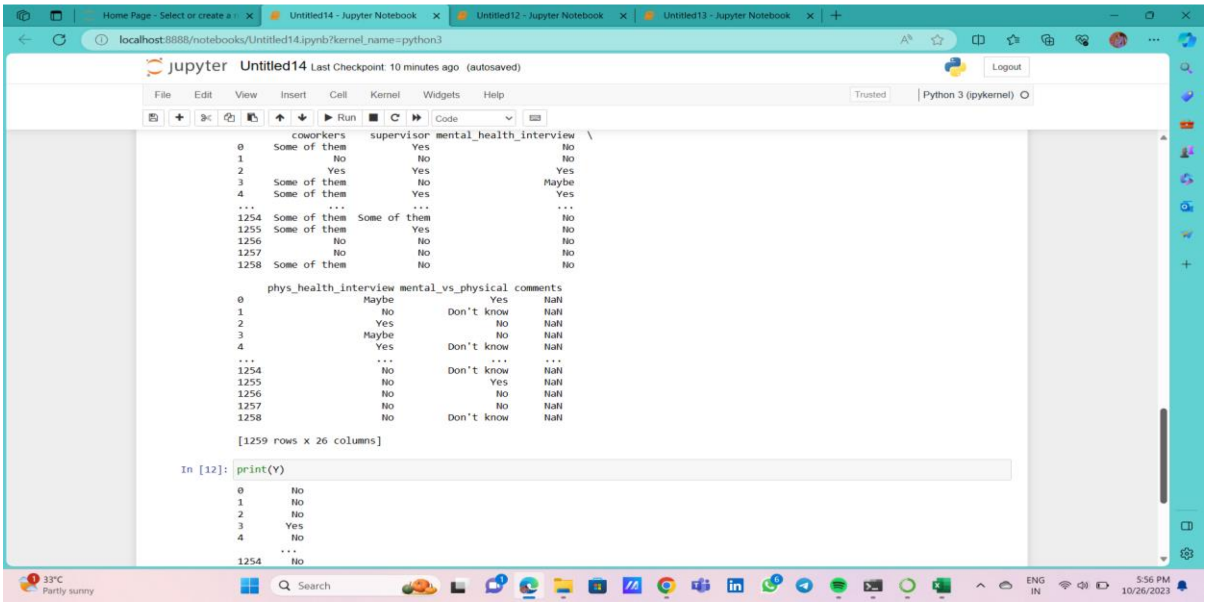
```
                Timestamp  Age  Gender         Country state self_employed  \
0     2014-08-27 11:29:31   37  Female   United States    IL           NaN
1     2014-08-27 11:29:37   44       M   United States    IN           NaN
2     2014-08-27 11:29:44   32    Male          Canada   NaN           NaN
3     2014-08-27 11:29:46   31    Male  United Kingdom   NaN           NaN
4     2014-08-27 11:30:22   31    Male   United States    TX           NaN
...                   ...  ...     ...             ...   ...           ...
```

Fig 6.2


Home Page - Select or create a
Untitled14 - Jupyter Notebook
Untitled12 - Jupyter Notebook
Untitled13 - Jupyter Notebook
localhost:8888/notebooks/Untitled14.ipynb?kernel_name=python3
jupyter
Untitled14 Last Checkpoint: 10 minutes ago (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
Run
Code
33°C
Partly sunny
Search
ENG
IN
5:55 PM
10/26/2023

```
3     2014-08-27 11:29:46   31    Male  United Kingdom   NaN           NaN
4     2014-08-27 11:30:22   31    Male   United States    TX           NaN
...                   ...  ...     ...             ...   ...           ...
1254  2015-09-12 11:17:21   26    male  United Kingdom   NaN            No
1255  2015-09-26 01:07:35   32    Male   United States    IL            No
1256  2015-11-07 12:36:58   34    male   United States    CA            No
1257  2015-11-30 21:25:06   46       f   United States    NC            No
1258  2016-02-01 23:04:31   25    Male   United States    IL            No

     family_history treatment work_interfere    no_employees  ...   anonymity  \
0                No       Yes          Often            6-25  ...         Yes
1                No        No         Rarely  More than 1000  ...  Don't know
2                No        No         Rarely            6-25  ...  Don't know
3               Yes       Yes          Often          26-100  ...          No
4                No        No          Never         100-500  ...  Don't know
...             ...       ...            ...             ...  ...         ...
1254             No       Yes            NaN          26-100  ...  Don't know
1255            Yes       Yes          Often          26-100  ...         Yes
1256            Yes       Yes      Sometimes  More than 1000  ...  Don't know
1257             No        No            NaN         100-500  ...  Don't know
1258            Yes       Yes      Sometimes          26-100  ...         Yes

                   leave mental_health_consequence phys_health_consequence  \
0          Somewhat easy                        No                      No
1             Don't know                     Maybe                      No
2     Somewhat difficult                        No                      No
3     Somewhat difficult                       Yes                     Yes
4             Don't know                        No                      No
...                  ...                       ...                     ...
1254       Somewhat easy                        No                      No
1255  Somewhat difficult                        No                      No
1256  Somewhat difficult                       Yes                     Yes
1257          Don't know                       Yes                      No
1258          Don't know                     Maybe                      No

         coworkers    supervisor mental_health_interview  \
0     Some of them           Yes                      No
```

Fig 6.3



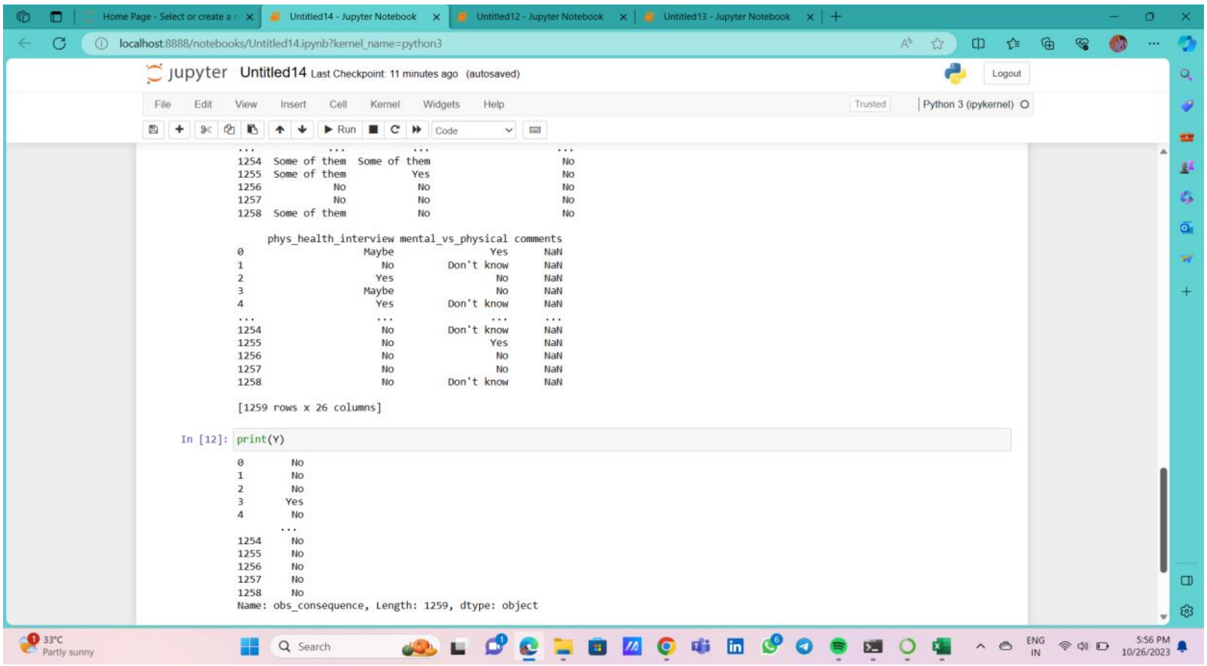
Untitled14 Last Checkpoint: 10 minutes ago (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Run
Code
Trusted
Python 3 (ipykernel)
Logout

```
      coworkers    supervisor mental_health_interview  \
0     Some of them         Yes                      No
1               No          No                      No
2              Yes         Yes                     Yes
3     Some of them          No                   Maybe
4     Some of them         Yes                     Yes
...            ...         ...                     ...
1254  Some of them  Some of them                    No
1255  Some of them         Yes                      No
1256            No          No                      No
1257            No          No                      No
1258  Some of them          No                      No

     phys_health_interview mental_vs_physical comments
0                    Maybe                Yes      NaN
1                       No         Don't know      NaN
2                      Yes                 No      NaN
3                    Maybe                 No      NaN
4                      Yes         Don't know      NaN
...                    ...                ...      ...
1254                    No         Don't know      NaN
1255                    No                Yes      NaN
1256                    No                 No      NaN
1257                    No                 No      NaN
1258                    No         Don't know      NaN

[1259 rows x 26 columns]
```

```
In [12]: print(Y)
```

```
0        No
1        No
2        No
3       Yes
4        No
       ...
1254     No
```

Fig 6.4



Untitled14 Last Checkpoint: 11 minutes ago (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Run
Code
Trusted
Python 3 (ipykernel)
Logout

```
...            ...         ...                     ...
1254  Some of them  Some of them                    No
1255  Some of them         Yes                      No
1256            No          No                      No
1257            No          No                      No
1258  Some of them          No                      No

     phys_health_interview mental_vs_physical comments
0                    Maybe                Yes      NaN
1                       No         Don't know      NaN
2                      Yes                 No      NaN
3                    Maybe                 No      NaN
4                      Yes         Don't know      NaN
...                    ...                ...      ...
1254                    No         Don't know      NaN
1255                    No                Yes      NaN
1256                    No                 No      NaN
1257                    No                 No      NaN
1258                    No         Don't know      NaN

[1259 rows x 26 columns]
```

```
In [12]: print(Y)
```

```
0        No
1        No
2        No
3       Yes
4        No
       ...
1254     No
1255     No
1256     No
1257     No
1258     No
Name: obs_consequence, Length: 1259, dtype: object
```

Fig 6.5

CHAPTER 6

PROPOSED INNOVATION TECHNIQUE

Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well it's best suited for classification. The main objective of the SVM algorithm is to find the optimal hyperplane in an N-dimensional space that can separate the data points in different classes in the feature space. The hyperplane tries that the margin between the closest points of different classes should be as maximum as possible. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three.

Let's consider two independent variables x1, x2, and one dependent variable which is either a blue circle or a red circle.

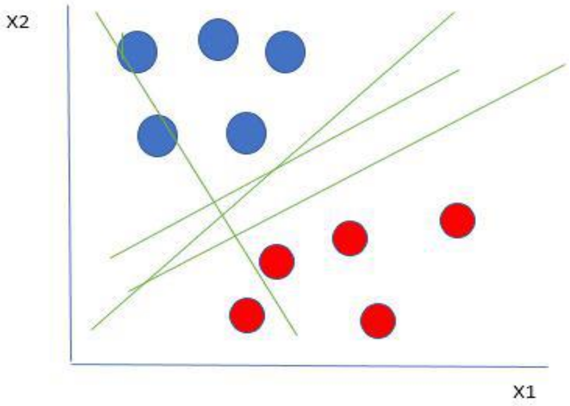


Fig 7.1

How does SVM work?

One reasonable choice as the best hyperplane is the one that represents the largest separation or margin between the two classes.

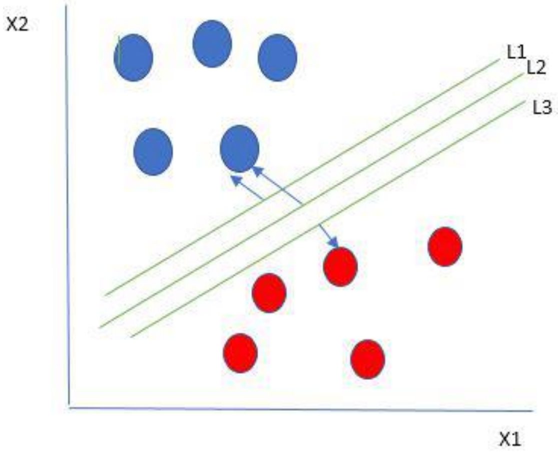


Fig 7.2

we choose the hyperplane whose distance from it to the nearest data point on each side is maximized. If such a hyperplane exists it is known as the **maximum-margin hyperplane/hard margin**. we choose L2. Let's consider a scenario like shown below

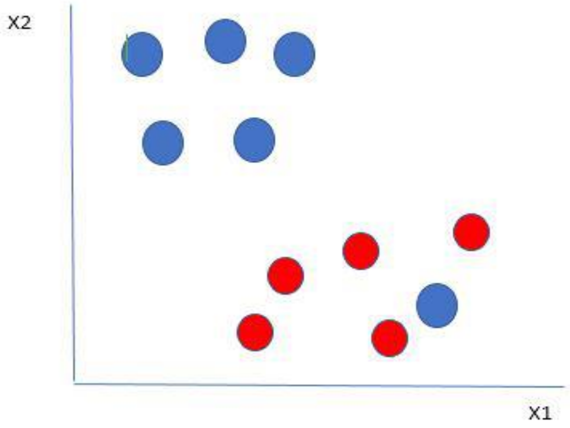


Fig 7.3

we have one blue ball in the boundary of the red ball. The blue ball in the boundary of red ones is an outlier of blue balls. The SVM algorithm has the characteristics to ignore the outlier and finds the best hyperplane that maximizes the margin. SVM is robust to outliers.

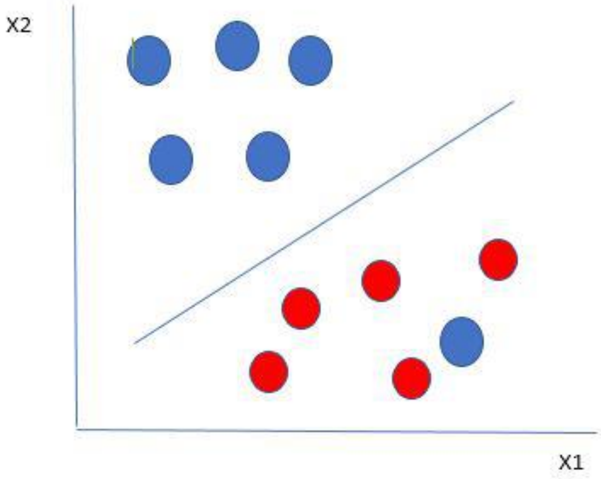


Fig 7.4

So in this type of data point what SVM does is, finds the maximum margin as done with previous data sets along with that it adds a penalty each time a point crosses the margin. So the margins in these types of cases are called **soft margins**. When there is a soft margin to the data set, the SVM tries to minimize *(1/margin+∧(∑penalty))*. Hinge loss is a commonly used penalty. If no violations no hinge loss.If violations hinge loss proportional to the distance of violation.

Till now, we were talking about linearly separable data(the group of blue balls and red balls are separable by a straight line/linear line). What to do if data are not linearly separable?

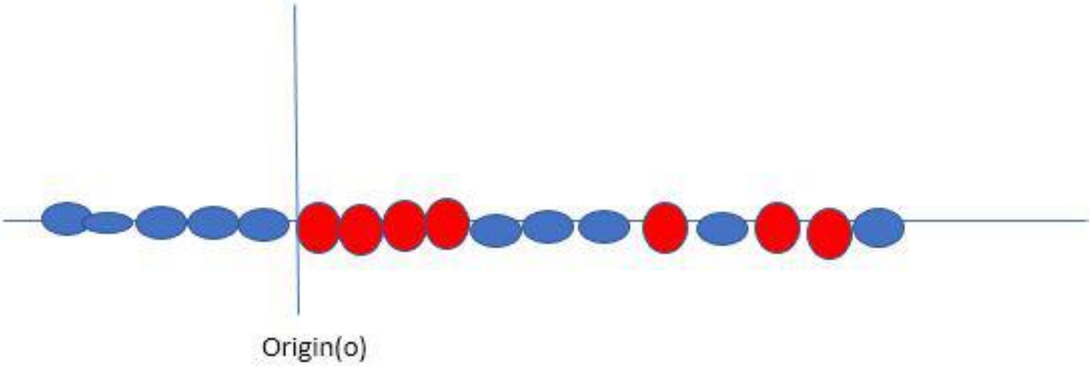


Fig 7.5

Say, our data is shown in the figure above. SVM solves this by creating a new variable using a **kernel**. We call a point x_i on the line and we create a new variable y_i as a function of distance from origin o.so if we plot this we get something like as shown below

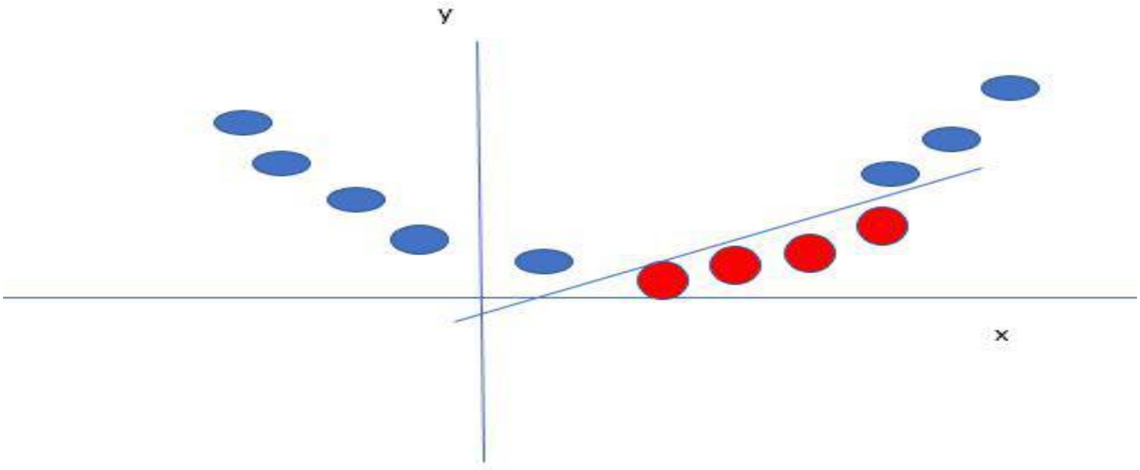


Fig 7.6

In this case, the new variable y is created as a function of distance from the origin. A non-linear function that creates a new variable is referred to as a kernel.

CHAPTER 7

CONCLUSION AND FUTURE SCOPE

CONCLUSION

In conclusion, website traffic analysis is a vital practice for understanding user behavior, evaluating website performance, and making data-driven decisions. It provides valuable insights that can shape content strategies, marketing efforts, and user experience enhancements. By collecting and analyzing data on user interactions, traffic sources, and conversion events, website owners and marketers can adapt to evolving online landscapes and better serve their audiences. Furthermore, the challenges of data accuracy, privacy compliance, and the dynamic nature of the digital environment highlight the need for effective data preprocessing and analysis methodologies. In the ever-changing digital world, website traffic analysis remains an indispensable tool for staying competitive, improving user engagement, and achieving online success. It enables continuous optimization and adaptation to meet the evolving needs and expectations of website visitors.

FUTURE SCOPE

The future scope of website traffic analysis is promising and evolving with advancements in technology, user behavior, and the digital landscape. Here are some key areas where we can expect significant developments and opportunities in the field of website traffic analysis:

1. User-Centric Analysis: Future website traffic analysis will likely become even more user-centric. Understanding individual user journeys and preferences will be a key focus. Personalization and providing tailored content and experiences will be critical.

2. AI and Machine Learning Integration: AI and machine learning algorithms will play a more substantial role in analyzing large datasets, automating insights, and predicting user behavior. These technologies can help in real-time user engagement and content recommendations.

3. Multichannel and Cross-Device Analysis: With users accessing websites from various devices and channels, the analysis will need to be more comprehensive in tracking and understanding user behavior across these touchpoints.

4. Data Privacy and Compliance: As data privacy regulations continue to evolve, website traffic analysis will need to adapt to ensure compliance with user consent, data protection, and transparency in data collection and handling.

5. Advanced Segmentation: More sophisticated user segmentation techniques will emerge, enabling website owners to target specific audience segments with precision. Segmentation based on behavior, demographics, and other factors will become more refined.

6. Real-Time Analytics: Real-time website traffic analysis will become more common, enabling businesses to make immediate adjustments to their strategies and user experiences based on up-to-the-minute data.

7. Voice Search and IoT Integration: As voice search and Internet of Things (IoT) devices become more prevalent, analyzing traffic from these sources will be essential for website owners and marketers.

8. Content Performance and SEO: With search engine algorithms continually evolving, website traffic analysis will continue to play a crucial role in optimizing content and SEO strategies to rank higher in search results.

9. Data Visualization and Dashboards: Interactive and customizable data visualization tools and dashboards will empower website owners to gain insights more intuitively and effectively.

10. Customer Journey Mapping: Analyzing the complete customer journey, both online and offline, will provide a holistic view of user interactions and their impact on website traffic and conversions.

11. Predictive Analytics: Using historical data to predict future user behavior, trends, and market changes will become increasingly valuable for strategic decision-making.

12. Augmented and Virtual Reality (AR/VR): As AR and VR technologies become more integrated into web experiences, website traffic analysis will need to adapt to track and evaluate user interactions within these immersive environments.

13. Data Monetization: Beyond understanding user behavior, website owners may explore opportunities to monetize their data by offering insights or selling anonymized, aggregated data to third parties.

14. Evolving Metrics: New metrics and KPIs will be developed to better measure and quantify user engagement, satisfaction, and the success of various digital initiatives.

Website traffic analysis will continue to be an essential practice for staying competitive and relevant in the digital age. Adapting to emerging technologies, evolving user preferences, and the regulatory landscape will be key to harnessing the full potential of this field. The future scope of website traffic analysis holds exciting possibilities for those who leverage it effectively.