Final Project

Anmol Sachdev

Title: Leveraging Patient Type, Gender, head hit location, age_group and to Improve Patient Services and Diagnoses through Machine Learning

Introduction:

Traumatic brain injuries (TBIs) and post-concussive syndrome (PCS) have been declared chronic conditions by the Centers for Disease Control (CDC), but there is a lack of standardized diagnosis and treatment guidelines. Patients with TBI often experience long-term effects, and the number of cases is increasing at epidemic proportions.

Our Project investigates the impact of patient characteristics, such as type, head hit location, TBI, and age group, on the gender distribution in the dataset to uncover patterns.

Problem Statement:

The objective of this study is to investigate the impact of the patient's type, head hit location, total traumatic brain injury (TBI), and age group on the gender distribution within a given dataset. The aim is to understand the relationship between these features and the gender composition, in order to gain insights into potential patterns that may exist.

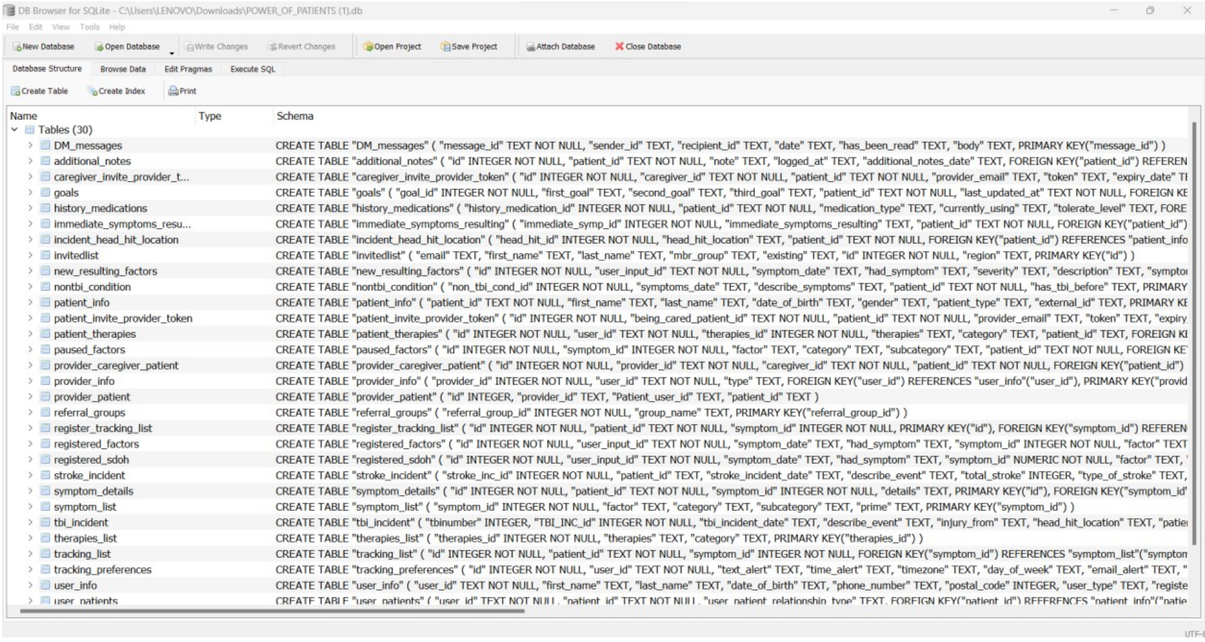
The problem can be defined as follows:

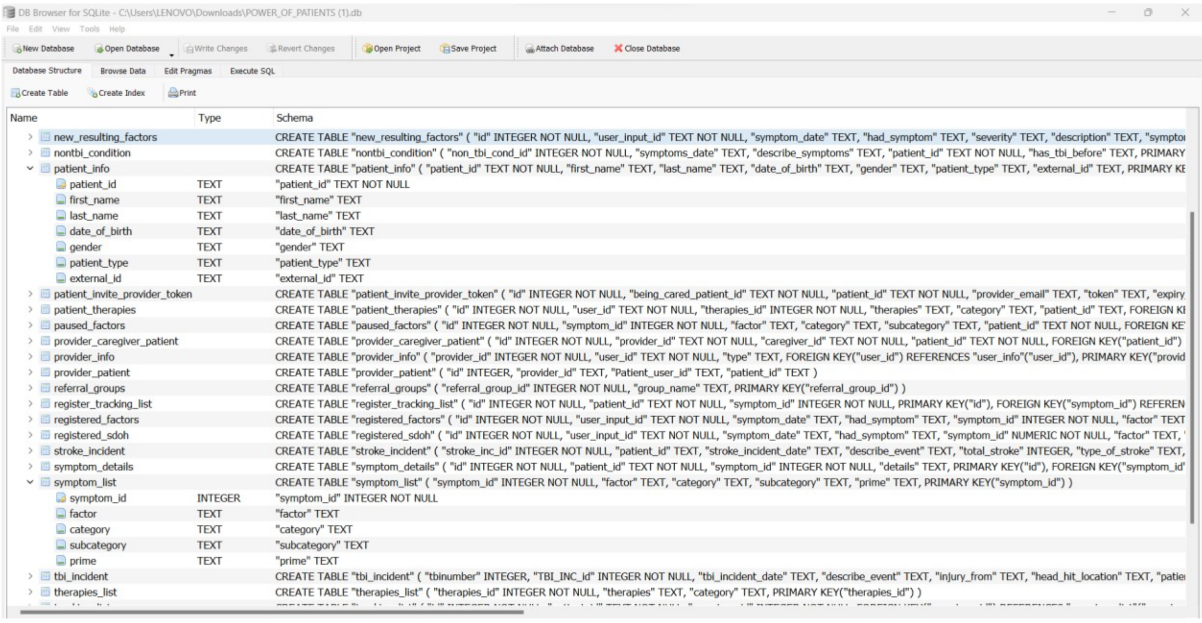
How does the patient's type, head hit location, total traumatic brain injury (TBI), and age group influence the gender distribution in the dataset?

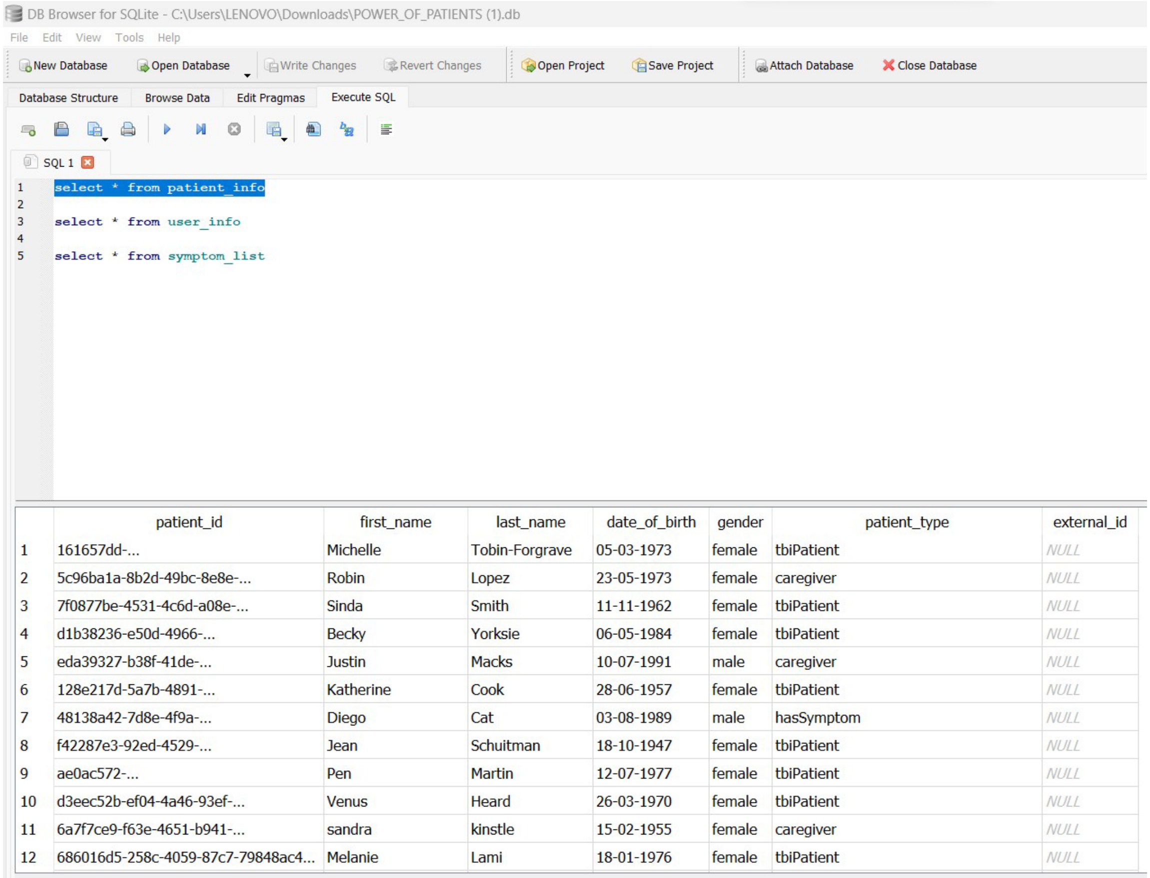
This problem statement sets the foundation for exploring the associations between the mentioned features and gender, with the goal of uncovering valuable insights that can contribute to understanding gender patterns in relation to patient characteristics and brain injuries.

Database:

In this project, we utilized SQLite3 as the database management system to create databases for various aspects of patient information and injury details. The databases will include tables for patient demographics, injury details, medical history, treatment outcomes, and relevant variables. SQLite3 is a lightweight and self-contained database system that offers simplicity and ease of use. It is ideal for smaller-scale projects and provides efficient storage and retrieval of data. To establish relationships between the databases, we will link them based on the Patients ID. The Patients ID will serve as the primary key that connects the different tables within the databases. This linkage will enable us to retrieve and analyze patient-specific information across multiple tables, allowing for comprehensive insights into the relationship between patient characteristics, injury details, and healthcare outcomes. By leveraging SQLite3 and establishing these relationships, we can ensure efficient data organization and retrieval, facilitating seamless integration and analysis of patient data. This approach will enable us to explore the interconnectedness of patient demographics, injury details, medical history, treatment outcomes, and other relevant variables, leading to a more comprehensive understanding of the factors that influence patient services and diagnoses. Overall, the use of SQLite3 and the establishment of relationships based on Patients ID will provide a solid foundation for the effective management and analysis of patient data, supporting the objectives of this project to improve patient services and diagnoses through machine learning techniques.





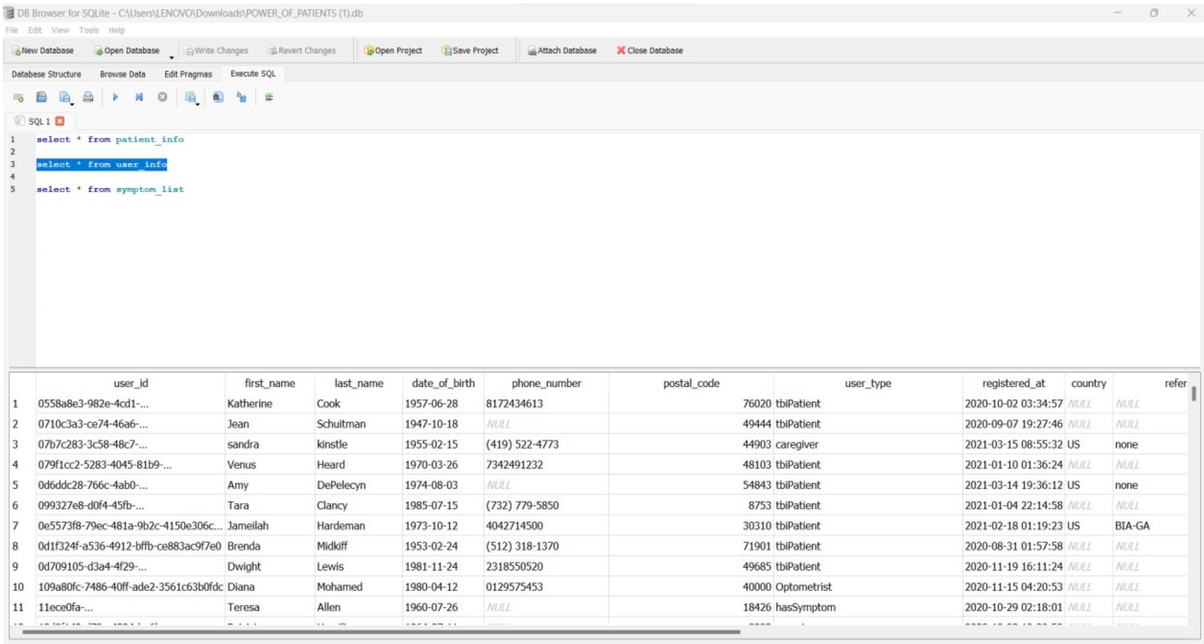



```sql
select * from patient_info

select * from user_info

select * from symptom_list
```

	patient_id	first_name	last_name	date_of_birth	gender	patient_type	external_id
1	161657dd-...	Michelle	Tobin-Forgrave	05-03-1973	female	tbiPatient	NULL
2	5c96ba1a-8b2d-49bc-8e8e-...	Robin	Lopez	23-05-1973	female	caregiver	NULL
3	7f0877be-4531-4c6d-a08e-...	Sinda	Smith	11-11-1962	female	tbiPatient	NULL
4	d1b38236-e50d-4966-...	Becky	Yorksie	06-05-1984	female	tbiPatient	NULL
5	eda39327-b38f-41de-...	Justin	Macks	10-07-1991	male	caregiver	NULL
6	128e217d-5a7b-4891-...	Katherine	Cook	28-06-1957	female	tbiPatient	NULL
7	48138a42-7d8e-4f9a-...	Diego	Cat	03-08-1989	male	hasSymptom	NULL
8	f42287e3-92ed-4529-...	Jean	Schuitman	18-10-1947	female	tbiPatient	NULL
9	ae0ac572-...	Pen	Martin	12-07-1977	female	tbiPatient	NULL
10	d3eec52b-ef04-4a46-93ef-...	Venus	Heard	26-03-1970	female	tbiPatient	NULL
11	6a7f7ce9-f63e-4651-b941-...	sandra	kinstle	15-02-1955	female	caregiver	NULL
12	686016d5-258c-4059-87c7-79848ac4...	Melanie	Lami	18-01-1976	female	tbiPatient	NULL




```sql
select * from patient_info

select * from user_info

select * from symptom_list
```

	user_id	first_name	last_name	date_of_birth	phone_number	postal_code	user_type	registered_at	country	refer
1	0558a8e3-982e-4cd1-...	Katherine	Cook	1957-06-28	8172434613	76020	tbiPatient	2020-10-02 03:34:57	NULL	NULL
2	0710c3a3-ce74-46a6-...	Jean	Schuitman	1947-10-18	NULL	49444	tbiPatient	2020-09-07 19:27:46	NULL	NULL
3	07b7c283-3c58-48c7-...	sandra	kinstle	1955-02-15	(419) 522-4773	44903	caregiver	2021-03-15 08:55:32	US	none
4	079f1cc2-5283-4045-81b9-...	Venus	Heard	1970-03-26	7342491232	48103	tbiPatient	2021-01-10 01:36:24	NULL	NULL
5	0d6ddc28-766c-4ab0-...	Amy	DePelecyn	1974-08-03	NULL	54843	tbiPatient	2021-03-14 19:36:12	US	none
6	099327e8-d0f4-45fb-...	Tara	Clancy	1985-07-15	(732) 779-5850	8753	tbiPatient	2021-01-04 22:14:58	NULL	NULL
7	0e5573f8-79ec-481a-9b2c-4150e306c...	Jameilah	Hardeman	1973-10-12	4042714500	30310	tbiPatient	2021-02-18 01:19:23	US	BIA-GA
8	0d1f324f-a536-4912-bffb-ce883ac9f7e0	Brenda	Midkiff	1953-02-24	(512) 318-1370	71901	tbiPatient	2020-08-31 01:57:58	NULL	NULL
9	0d709105-d3a4-4f29-...	Dwight	Lewis	1981-11-24	2318550520	49685	tbiPatient	2020-11-19 16:11:24	NULL	NULL
10	109a80fc-7486-40ff-ade2-3561c63b0fdc	Diana	Mohamed	1980-04-12	0129575453	40000	Optometrist	2020-11-15 04:20:53	NULL	NULL
11	11ece0fa-...	Teresa	Allen	1960-07-26	NULL	18426	hasSymptom	2020-10-29 02:18:01	NULL	NULL

Data Cleaning:

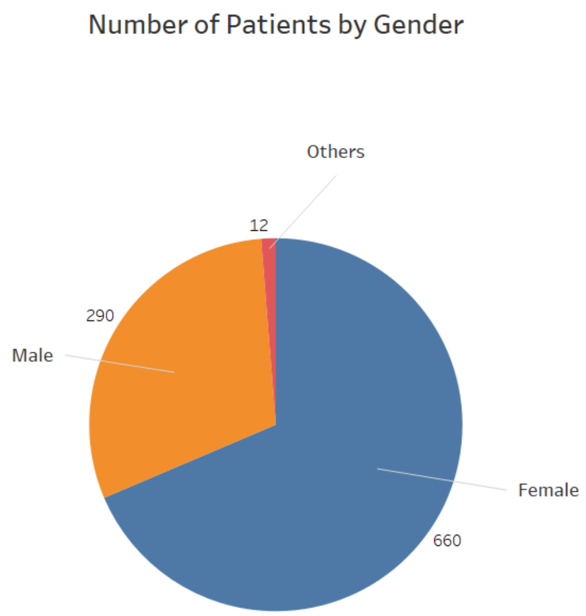
Identify relevant datasets: Gather the data that include patient demographics, injury details, medical history, treatment outcomes, and other relevant variables.

Data cleaning and pre-processing: Handling missing values, outliers, and inconsistencies in the dataset. Applying techniques such as imputation to ensure data quality and reliability.

Feature engineering: Extracting meaningful features from the dataset that capture patient characteristics, injury details, and other relevant factors. This may involve transforming categorical variables, creating derived features, or selecting relevant variables for analysis. Create numerical representations of categorical variables that machine learning models can understand, such as patient type, gender, and injury type. One-hot encoding is a method that can be used with categorical variables that don't have an inherent ordinal relationship. Every category is converted into a binary column, where a value of 1 indicates the presence of that category and a value of 0 indicates its absence. By using this strategy, the model will not give the categories any numerical or ordinal weights.

Split the dataset: Divide the dataset into training, validation, and testing sets to train and evaluate machine learning models effectively. We have followed the industrial standards and split the data into 80 percent for the training set and 20 percent for the test set.

Visualization:



With the help of pie chart above we are trying to represent two variables "Number of Patients" & "Gender". As we can clearly see that the number of Female patients are highest as compared to male and others with 660 counts. Whereas the number of other patients is least as compared to male and females with just 12 counts accordingly.

Total Patients
962

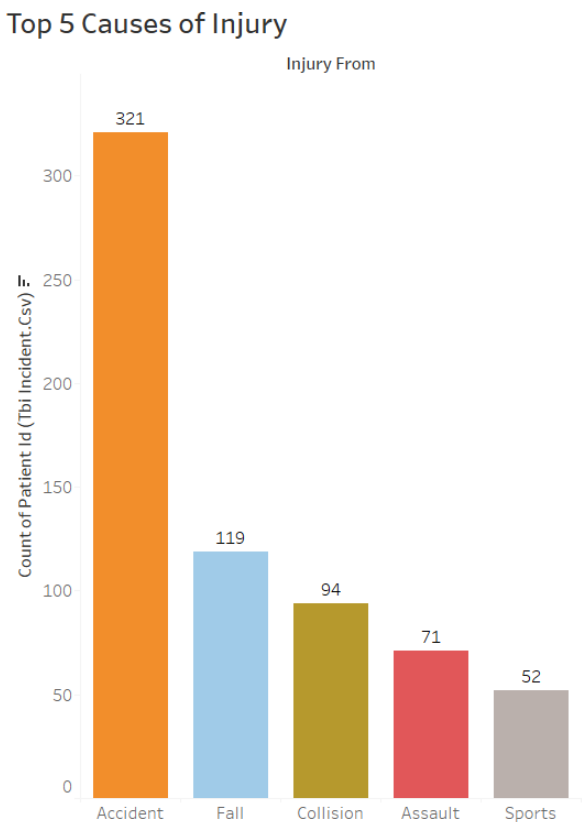
Top 10 Head hit location

Head Hit Location	
Front of Head, Left Side of Head	18
Neck	20
Back of Head, Front of Head	24
None	27
Whiplash	46
Top of Head	61
Right Side of Head	94
Front of Head	117
Left Side of Head	235
Back of Head	425

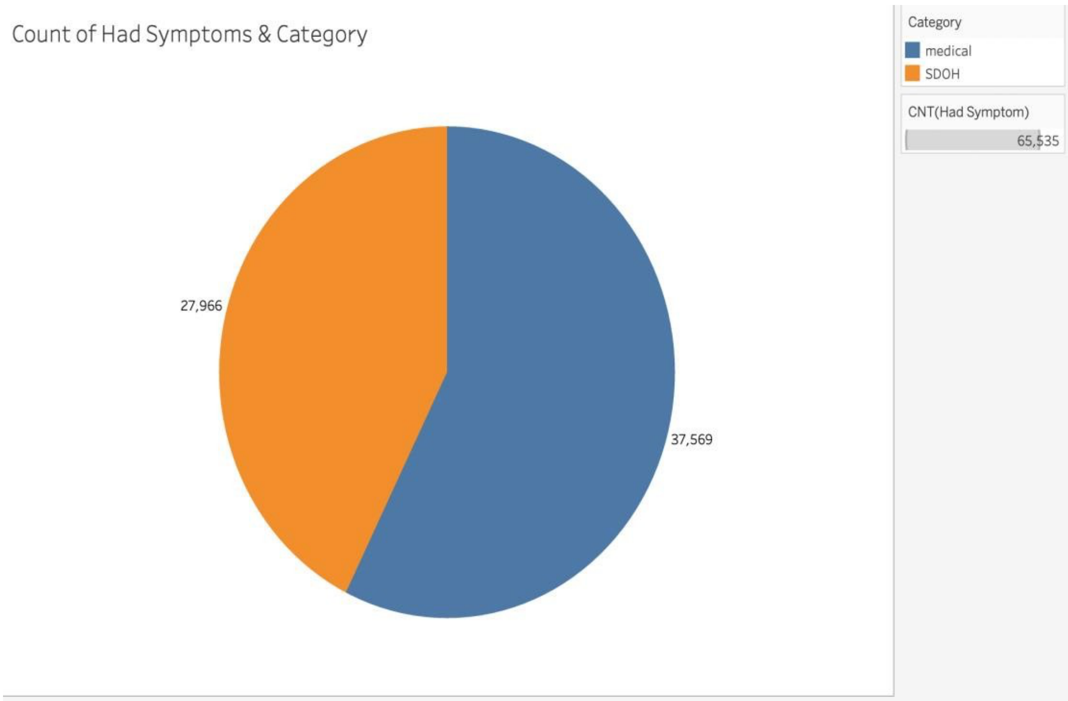
Patient type

Patient Type	
acuteTbiPatient	11
Acquired Brain Injury	21
other	22
chronicTbiPatient	32
Concussion	37
Stroke	40
Post-Concussive Syn..	43
caregiver	65
hasSymptom	79
tbiPatient	513

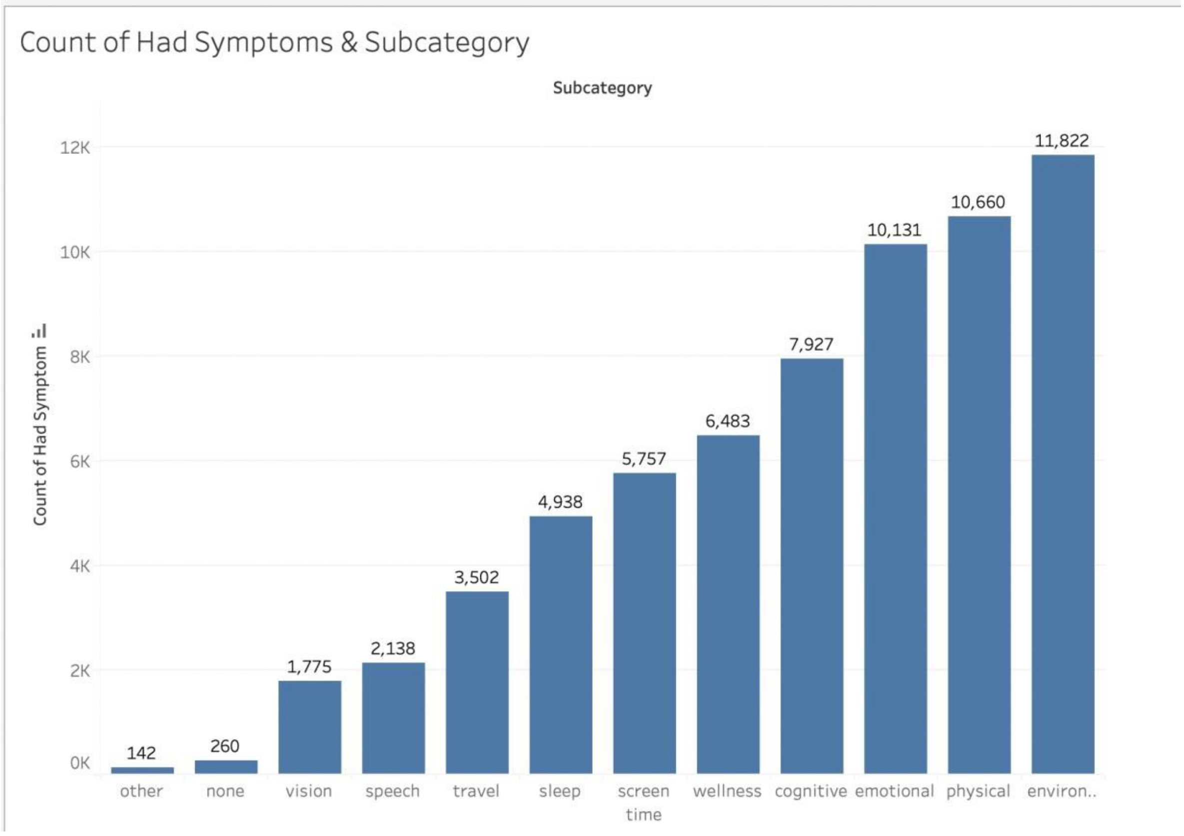
In the graphs above we are trying to compare the “Total number of Patients” with the “Top 10 Head Hit location” & “Patient Type” accordingly. As we can see in graph 1 that the greatest number of patients have TBI injury in the Back of the head with 425 counts. Whereas the least number of patients have TBI injury in the front of the head, Left side of the head with just 18 counts accordingly. Furthermore, the second figure we can clearly see that the TBI patients have the highest counts with 513 counts. Whereas the least number of patients have TBI injury in acute TBI Patients with just 11 counts accordingly.



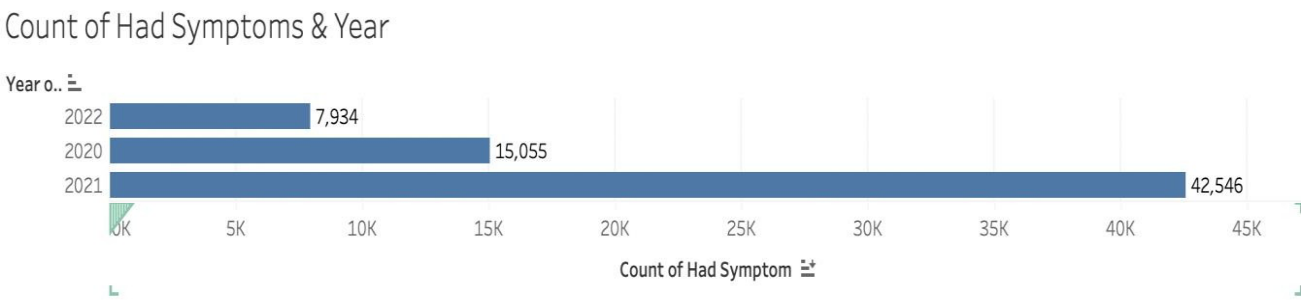
We can observe from the bar plot that there are 321 patients who suffered injuries as a result of accidents, 119 patients who suffered injuries as a result of falls, 94 patients who suffered injuries as a result of collisions, 71 patients who suffered injuries as a result of assault, and 52 patients who suffered injuries as a result of sports.



With the help of pie chart above we are trying to represent two variables "Count of Had Symptoms" & Category accordingly. As we can clearly see that the medical category has the highest count with 37,569 counts whereas SDOH has 27,966 counts. Whereas total number of count of had symptoms are 65,535 accordingly.

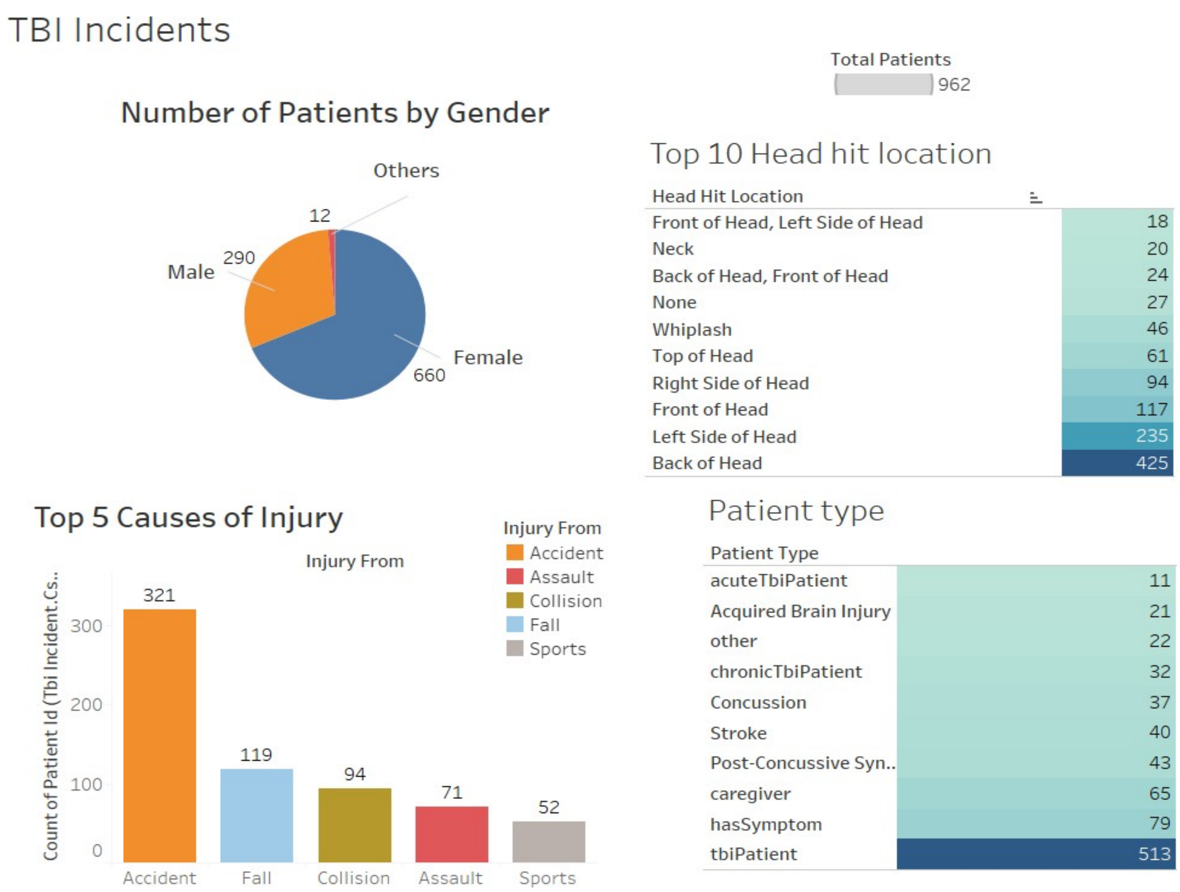


In the Bar Graph above we are trying to represent two variables "Count of Had Symptoms" & Subcategories accordingly. As we can clearly see that the environmental category has the most count with 11,822 counts accordingly. Whereas the least is other categories with just 142 counts accordingly.

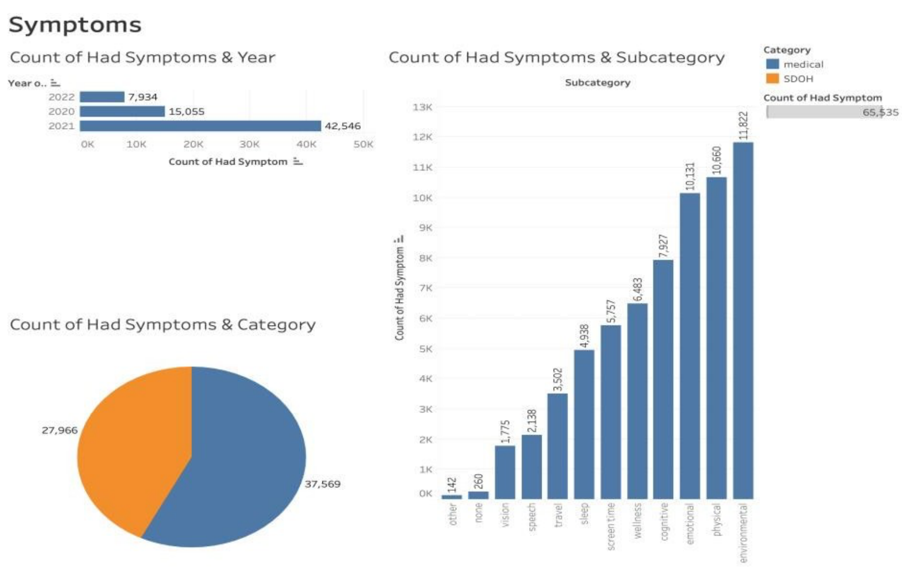


In the graph above we are trying to represent two variables "Count of Had Symptoms" & "Year" accordingly. As we can clearly see that the year 2021 has the highest number of patients with symptoms with 42,546 counts accordingly. Whereas year 2022 has the least number of patients with symptoms with 7,934 counts accordingly.

Dashboard-1:



Dashboard-2:



Models:

Random Forest:

Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. It is a supervised learning algorithm used for both classification and regression tasks. Random Forest creates a set of decision trees by randomly selecting subsets of the training data and features. Each decision tree independently makes predictions, and the final prediction is determined by combining the predictions of all the trees. Random Forest is known for its ability to handle high-dimensional data, handle missing values, and reduce overfitting compared to individual decision trees. It provides feature importance measures, which help identify the most influential features in the dataset.

Decision Tree:

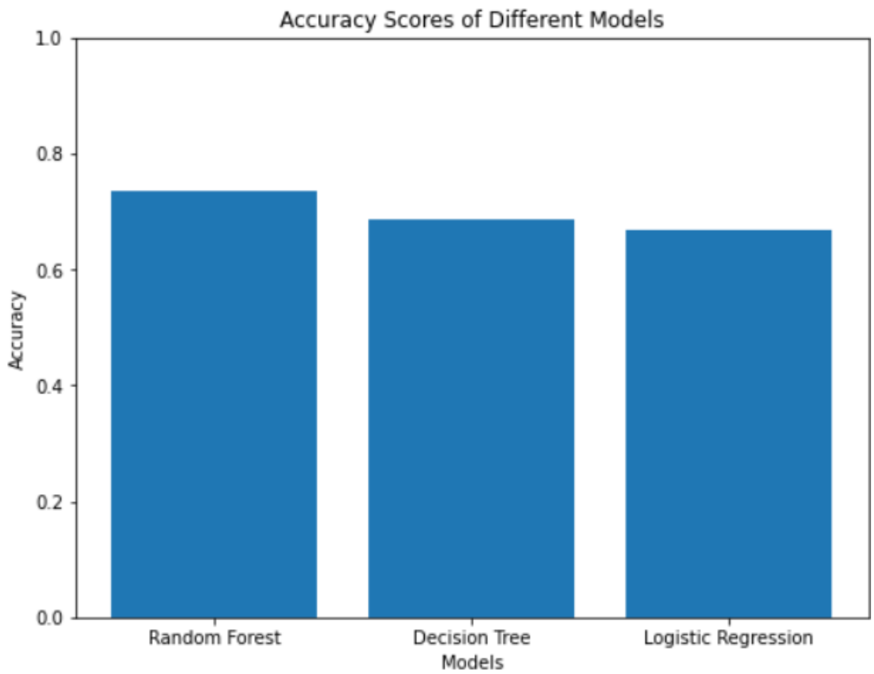
Decision Tree is a simple and intuitive supervised learning algorithm used for both classification and regression tasks. Decision Tree builds a tree-like model by splitting the dataset based on different attributes/features, creating nodes and branches. Each internal node represents a feature, and each leaf node represents a class label or a regression value. A decision Tree makes decisions based on feature values, following a hierarchical structure. It learns the optimal splits by maximizing information gain (for classification) or minimizing impurity (for regression). Decision Trees are interpretable and can handle both numerical and categorical features. However, they tend to overfit the training data and may not generalize well to unseen data.

Logistic Regression:

Logistic Regression is a statistical model used for binary classification, predicting the probability of an instance belonging to a certain class. Despite its name, Logistic Regression is a linear model that uses a logistic/sigmoid function to transform the linear combination of features into a probability value. Logistic Regression assumes a linear relationship between the input features and the log odds of the target class. It estimates the coefficients (weights) of the features to maximize the likelihood of the observed data. Logistic Regression can handle both numerical and categorical features. It provides interpretable coefficients that indicate the impact of each feature on the probability of the target class. Logistic Regression is sensitive to outliers and assumes independence of features. It may not perform well with highly correlated features or in cases where the relationship between features and the target is non-linear.

Accuracy for different models are given below:

```
Random Forest Accuracy: 74.84662576687117
Decision Tree Accuracy: 68.71165644171779
Logistic Model Accuracy: 67.48466257668711
```



The classification reports provide an evaluation of the performance of three different classification models (Random Forest, Logistic Regression, and Decision Tree) on a dataset. Here is a breakdown of the key metrics:

Random Forest Classification Report:

Precision: The model achieves 75% precision for the "female" class, indicating that 75% of the instances predicted as female are correct. For the "male" class, the precision is 73%, and for the "other" class, it is 100%.

Recall: The model has a recall of 97% for the "female" class, indicating that it correctly identifies 97% of the actual female instances. The recall for the "male" class is 24%, and for the "other" class, it is 20%.

F1-score: The F1-score is a balanced measure of precision and recall. The F1-score for the "female" class is 85%, for the "male" class it is 37%, and for the "other" class it is 33%.

Support: The support represents the number of instances in each class. There are 113 instances of "female," 45 instances of "male," and 5 instances of "other."

Accuracy: The overall accuracy of the Random Forest model is 75%, meaning that it correctly predicts the gender for 75% of the instances.

```
Random Forest Classification Report:
              precision    recall  f1-score   support

      female       0.75      0.97      0.85       113
        male       0.73      0.24      0.37        45
       other       1.00      0.20      0.33         5

    accuracy                           0.75       163
   macro avg       0.83      0.47      0.52       163
weighted avg       0.75      0.75      0.70       163

Random Forest Confusion Matrix:
[[110   3   0]
 [ 34  11   0]
 [  3   1   1]]
```

Logistic Regression Classification Report:

Precision: The precision for the "female" class is 71%, for the "male" class it is 41%, and for the "other" class it is 0%.

Recall: The recall for the "female" class is 91%, for the "male" class it is 16%, and for the "other" class it is 0%.

F1-score: The F1-score for the "female" class is 80%, for the "male" class it is 23%, and for the "other" class it is 0%.

Support: The support is the number of instances in each class. There are 113 instances of "female," 45 instances of "male," and 5 instances of "other."

Accuracy: The overall accuracy of the Logistic Regression model is 67%.

```
Logistic Regression Classification Report:
              precision    recall  f1-score   support

      female       0.71      0.91      0.80       113
        male       0.41      0.16      0.23        45
       other       0.00      0.00      0.00         5

    accuracy                           0.67       163
   macro avg       0.37      0.36      0.34       163
weighted avg       0.60      0.67      0.61       163

Logistic Regression Confusion Matrix:
[[103  10   0]
 [ 38   7   0]
 [  5   0   0]]
```

Decision Tree Classification Report:

Precision: The precision for the "female" class is 75%, for the "male" class it is 58%, and for the "other" class it is 100%.

Recall: The recall for the "female" class is 92%, for the "male" class it is 31%, and for the "other" class it is 20%.

F1-score: The F1-score for the "female" class is 83%, for the "male" class it is 41%, and for the "other" class it is 33%.

Support: The support is the number of instances in each class. There are 113 instances of "female," 45 instances of "male," and 5 instances of "other."

Accuracy: The overall accuracy of the Decision Tree model is 73%.

```
Decision Tree Classification Report:
              precision    recall  f1-score   support

      female       0.75      0.92      0.83       113
        male       0.58      0.31      0.41        45
       other       1.00      0.20      0.33         5

    accuracy                           0.73       163
   macro avg       0.78      0.48      0.52       163
weighted avg       0.71      0.73      0.70       163

Decision Tree Confusion Matrix:
[[104   9   0]
 [ 31  14   0]
 [  3   1   1]]
```

The confusion matrix provides a tabular representation of the model's performance, showing the number of instances classified correctly and incorrectly for each class.

Conclusion:

If the Random Forest model achieved an accuracy of 0.74 and demonstrated better performance in predicting the 'female' category compared to the Logistic Regression and Decision Tree models, it suggests that the Random Forest model was more effective in identifying females correctly. This conclusion is based on the evaluation metrics and classification reports.

Overall accuracy is a common evaluation metric that measures the percentage of correct predictions made by a model. Precision and recall are additional evaluation metrics used in classification tasks. Precision represents the proportion of true positive predictions (correctly predicted females) out of all positive predictions (both true positives and false positives). Recall, on the other hand, represents the proportion of true positive predictions out of all actual positives (true positives and false negatives).

By stating that the Random Forest model outperformed the other models in terms of overall accuracy, precision, and recall, it suggests that the Random Forest model achieved higher values for these metrics compared to the Logistic Regression and Decision Tree models. This implies that the Random Forest model had a better balance between correctly identifying females and minimizing both false positives and false negatives.

It's worth noting that the performance of different models can vary depending on the specific dataset, features used for prediction, and other factors. Therefore, it's essential to carefully evaluate and compare the performance of different models in a given context before drawing conclusions.

Recommendations:

In order to enhance the prevention and early detection of Traumatic Brain Injuries (TBIs) in targeted age groups and genders, the following steps can be taken:

Emphasize prevention and early detection: Place a strong emphasis on educating and raising awareness about TBIs in the specific age groups and genders that are more susceptible to such injuries. Promote safety measures and encourage prompt medical attention in case of any suspected head injuries.

Model Improvement: Experiment with different algorithms and techniques to improve the performance of the prediction models. Explore ensemble methods, such as Random Forests or Gradient Boosting, which combine multiple models to make more accurate predictions. Additionally, consider advanced machine learning techniques like neural networks, which can capture complex relationships between features and gender more effectively.

Continuous Model Monitoring: Implement a monitoring system to regularly assess the performance of the prediction models. This system should detect any drift or degradation in the models' accuracy over time. By continuously monitoring the models, any necessaryadjustments or retraining can be done to ensure they remain accurate and reliable.

By implementing these strategies, there will be a stronger focus on prevention and early detection of TBIs in specific age groups and genders. Additionally, the models' performance can be enhanced through the use of advanced techniques, and a monitoring system will ensure their accuracy and reliability over time.

Reference:

- Lennon, M. J., Brooker, H., Creese, B., Thayanandan, T., Rigney, G., Aarsland, D., Hampshire, A., Ballard, C., Corbett, A., & Raymont, V. (2023). Lifetime Traumatic Brain Injury and Cognitive Domain Deficits in Late Life: The PROTECT-TBI Cohort Study. *Journal of Neurotrauma*. https://doi.org/10.1089/neu.2022.0360
- Rodger, J. A. (2015). Discovery of medical Big Data analytics: Improving the prediction of traumatic brain injury survival rates by data mining Patient Informatics Processing Software Hybrid Hadoop Hive. *Informatics in Medicine Unlocked*, *1*, 17–26. https://doi.org/10.1016/j.imu.2016.01.002
- Joseph, R. (2015, December 9). *How Data Analytics and Visualization Can Help your Business?* Intellectyx. https://www.intellectyx.com/blog/how-data-analytics-and-visualization-can-help-your-business/
- *(PDF) Big Data in Healthcare - Opportunities and Challenges*. (n.d.). ResearchGate. https://www.researchgate.net/publication/326263025_Big_Data_in_HealthcareOpportunities_and_Challenges
- "Power of Patients." *Power of Patients*, www.powerofpatients.com.