Southern Connecticut State University
501 Crescent St, New Haven, CT 06515, United States

Finding the Top Causes and Indicators for Car Crashes in the USA using Machine Learning/AI

6th December, 2023

Alfredo Villavicencio
Maxwell P. Hauser
Sidney Levine
In conjunction with Dr. MD Hossain

Table of Contents

I Introduction

Abstract

According to the U.S. Transportation Secretary Pete Buttigieg, “We continue to face a national crisis of traffic deaths on our roadways...”.[1] An estimated 42,795 traffic fatalities occurred within the United States during 2022 alone.[2] The U.S. government has recognized this problem, and the data confirms the ghastly truth. How does a government find a solution to an urgent problem? We theorize by finding the root causes and indicators for crashes within the U.S. using machine learning; this would not only assist urban planners and civil engineers to develop infrastructure in a safer manner, but would also provide crucial insight to the everyday American—leading to safer roads and fewer premature deaths.

Problem Statement

Thousands of people are killed or seriously injured in car accidents every year in the U.S. How can we make our roads safer? We must first address the root causes and indicators to better understand the complex dynamics of car crashes.

We do this using a car crash dataset via Kaggle. We aimed to employ machine learning techniques, such as KNN and SVM algorithms, to find the critical root causes and indicators within the data. Thereupon, a clearer understanding of the problem may be drawn.

Objective

The objective of our project was to develop a software that can identify the top indicators and causes of car crashes within the U.S. We envisioned a software proficient

[1] "NHTSA Early Estimates: 2022 Traffic Crash Deaths." 20 Apr. 2023, https://www.nhtsa.gov/press-releases/traffic-crash-death-estimates-2022. Accessed 25 Sep. 2023.
[2] "NHTSA Early Estimates: 2022 Traffic Crash Deaths." 20 Apr. 2023, https://www.nhtsa.gov/press-releases/traffic-crash-death-estimates-2022. Accessed 25 Sep. 2023.

in recognizing and categorizing patterns and indicators for car crashes. The software trained on the “Car Crash Dataset” dataset using machine learning techniques.

Motivation

Our primary motivation was one of an ethical and moral crisis in our country, with tens of thousands of car accident-related fatalities in the United States over the last decade. We cannot remain idle on this issue while families lose loved ones, subsequently impacting lives forever.

Therefore, the motivation to tackle this task was targeted towards improving awareness and gaining insights into the grievous problem. This may, in turn lay the foundation for substantial, data driven solutions.

Related Work

Scientists from the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL) and Qatar Center for Artificial Intelligence developed a deep learning model that produces high-resolution crash risk maps. The deep learning model trained on an assortment of different data points such as historical crash data, road maps, satellite imagery, and GPS traces. This deep learning model with the given training can now: develop risk maps that describe the expected number of crashes over a period of time in the future, to identify high-risk areas, and predict future crashes. Similar to this group of scientists, we plan to help identify the problem before it occurs.[3]

[3] "Deep learning helps predict traffic crashes before they happen." 12 Oct. 2021, https://news.mit.edu/2021/deep-learning-helps-predict-traffic-crashes-1012. Accessed 25 Sep. 2023.

Ⅱ Data

Data Source and Format

The data we used in this project was sourced from Kaggle, an online public datasets repository. The dataset name is “Car Crash Dataset.” It was uploaded by the site owner, Prasanna KM, and was last modified about three years ago. This dataset offers a very complete description of its collected accidents and will be sufficient for our research purposes. The data is presented in a comma-separated value (.csv) file and contains approximately 17,565 records of accidents. The data consists of 15 columns broken down into eight columns:

1. Eight (8) object elements:
    - 1.1. dvcat
    - 1.2. dead
    - 1.3. airbag
    - 1.4. seatbelt
    - 1.5. sex
    - 1.6. abcat
    - 1.7. occRole
    - 1.8. caseID
2. Six (6) int(64) elements:
    - 2.1. frontal
    - 2.2. ageOFocc
    - 2.3. yearAcc
    - 2.4. yearVeh
    - 2.5. deploy
    - 2.6. injSeverity
3. One (1) float(64) element:
    - 3.1. weight

Data Example

dvcat	weight	dead	airbag	seatbelt	frontal	sex	ageOFocc
0	53.342	dead	airbag	belted	1	f	48
1	154.960	alive	none	none	1	m	26
2	38.994	alive	none	none	1	f	51
3	168.568	alive	airbag	belted	1	m	27
4	27.751	alive	airbag	belted	0	m	26

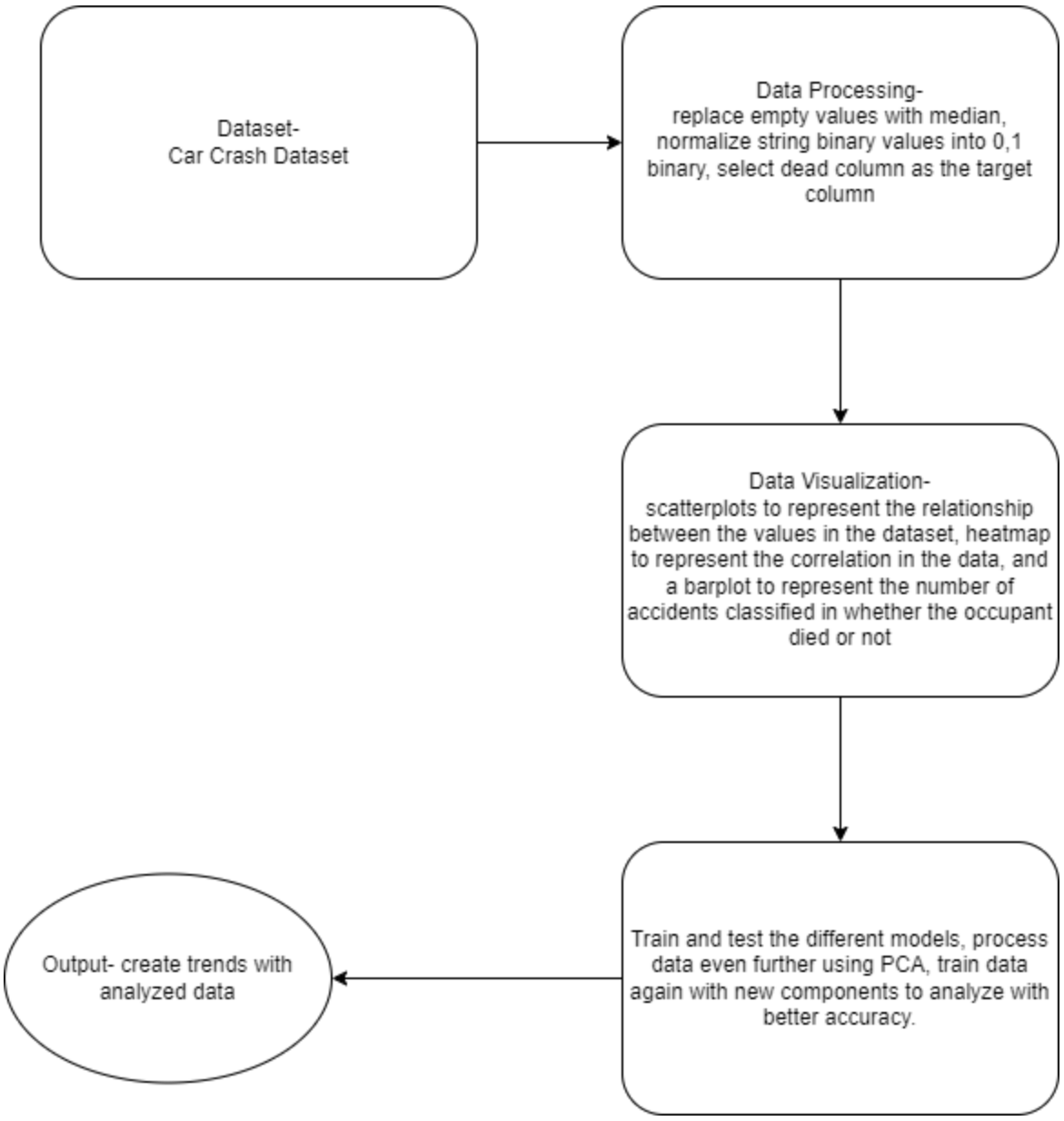
dvcat	yearAcc	yearVeh	abcat	occRole	deploy	caseID	injSeverity
0	2002	1997	deploy	driver	1	45:150:1	3
1	2001	1968	unavail	driver	0	76:40:1	3
2	2002	1994	unavail	driver	0	11:184:1	3
3	1998	1996	deploy	pass	1	9:17:1	3
4	2002	1997	nodeploy	pass	0	2:2:2	0

Link for data:

https://www.kaggle.com/datasets/prasannakm/car-crash-dataset/?select=train-new.csv

Ⅲ Methodology

Schematic Diagram and Framework



Data Visualization

For data visualization, we generated multiple figures to assess the relationships between columns presented in the dataset.

Preprocessing Plan

The conducted preprocessing improved the data by changing any “none” values in the airbag and seatbelt columns with “noairbag” or “noseatbelt,” respectively. We then normalized the data in most columns in a binary manner, and replaced any missing values across columns using the relevant median value. Most of the features present in the dataset are already in a binary format, which streamlined the normalization process. The result of all data preprocessing:

dvcat	weight	dead	airbag	seatbelt	frontal	sex	ageOFocc
0	53.342	0	1	1	1	1	48
1	154.960	1	0	0	1	0	26
2	38.994	1	0	0	1	1	51
3	168.568	1	1	1	1	0	27
4	27.751	1	1	1	0	0	26

dvcat	yearAcc	yearVeh	abcat	occRole	deploy	caseID	injSeverity
0	2002	1997	1	1	1	45:150:1	3
1	2001	1968	0	1	0	76:40:1	3
2	2002	1994	0	1	0	11:184:1	3
3	1998	1996	1	0	1	9:17:1	3
4	2002	1997	0	0	0	2:2:2	0

Procedures

The dataset was read into the system using the Scikit-Learn library for Python. Scikit-Learn offers a broad array of tools that can be used for data analysis, classification, stratification, modeling, and model training and testing. The regression models were built using the various classifiers provided by Scikit. These include k-nearest neighbor (KNN), support vector machine (SVM), and other models.

The models were tested using algorithms such as K-D tree for KNN, and SVR and LinearSVR for SVM. Model success was evaluated following testing using our predetermined evaluation metric. We also utilized the Pandas and Seaborn libraries to manage both data processing and some of the plotting performed throughout the project.

Traditionally, a given dataset is divided into two subsets to be used for modeling. The first subset is used for model training and the other for model testing. We did the same for our project. One subset was formed from a randomly selected 60% of the dataset, which was used for model training. The second subset was formed from the remaining 40% and was used for model testing. Once the data preprocessing procedures were successful, the remaining data processing procedures were as follows:

- Randomize dataset
- Select a 60% data subset to be used for model training
- Select remaining 40% data subset to be used for model testing
- Train classifiers/regressions/models with the selected training subset
- Test classifiers/regressions/models with the remaining testing dataset
- Analyze and evaluate model prediction success

Features

We will evaluate each of the individual features seen in the above data example. These features proved appropriate to use when creating, training, and testing our prediction models.

For the first table:

- dvcat
- weight
- dead
- airbag
- seatbelt
- frontal
- sex
- ageOFocc

For the second table:

- yearAcc
- yearVeh
- abcat
- occRole
- deploy
- caseID
- injSeverity

Ⅳ Experimentation

Data Division

As was discussed in the preprocessing section, we divided the dataset into two randomized subsets. 60% of the records was used to train the classifiers, and the remaining 40% was used to test the classifiers. We did not experience difficulty dividing the data, reading the data/subsets into the project or running the subsets through the classifiers. The 60%-40% division worked well for our purposes. Our experimentation yielded strong results and ultimately enabled our classifiers to achieve 95% accuracy.

Parameter Tuning

Analysis began with data preprocessing steps that included handling missing values through median imputation, encoding categorical variables into binary representations, and standardizing the data. Subsequently, KNN and SVM models were employed for predicting whether accidents resulted in fatalities. While the initial application of these models did not explicitly involve parameter tuning, a significant step in the experimentation process was the transformation of data using PCA to reduce dimensionality. For the KNN PCA transformation, parameter tuning was conducted, with the number of neighbors (`n_neighbors`) optimized, yielding an optimal value of two (2) for enhanced accuracy. Although specific hyperparameter details for the SVM model on PCA-transformed data were not provided, the acknowledgement of parameter tuning highlights its importance in refining model performance. The iterative process of parameter tuning ensures the models are fine-tuned for optimal predictions, and further exploration may involve experimenting with additional hyperparameters or alternative dimensionality reduction techniques.

Evaluation Metrics

In the previous sections, we discussed data preprocessing, feature engineering, and utilized the KNN classifier to make predictions on accident-related attributes. To

assess the model's performance, we also used a set of evaluation metrics designed to measure its ability to predict whether a given accident resulted in a fatality.

Similarly, we used the accuracy and precision metrics to evaluate the integrity of our model:

Accuracy: a metric that measures the overall correctness of the model's predictions. This metric reflects how often the model correctly predicts the accident-related attributes.

Precision: a metric used when evaluating the model's attribute-specific predictions. It assesses the proportion of positive predictions made by the model that were accurate. Precision helps us understand the model's ability to predict specific attributes correctly.

Anticipated Results

We expected the model to be able to predict whether a given accident was fatal. We preferred at least 90% accuracy. Some other targets may have been selected to further analyze the dataset.

Preliminary Results

Following data preprocessing and processing, we trained a KNN classifier and an SVM classifier with the training targets. When presented with the normalized testing targets, we observed a 95.33% KNN accuracy and a 95.34% SVM accuracy.

Concluding Results

Please consult the below figures:

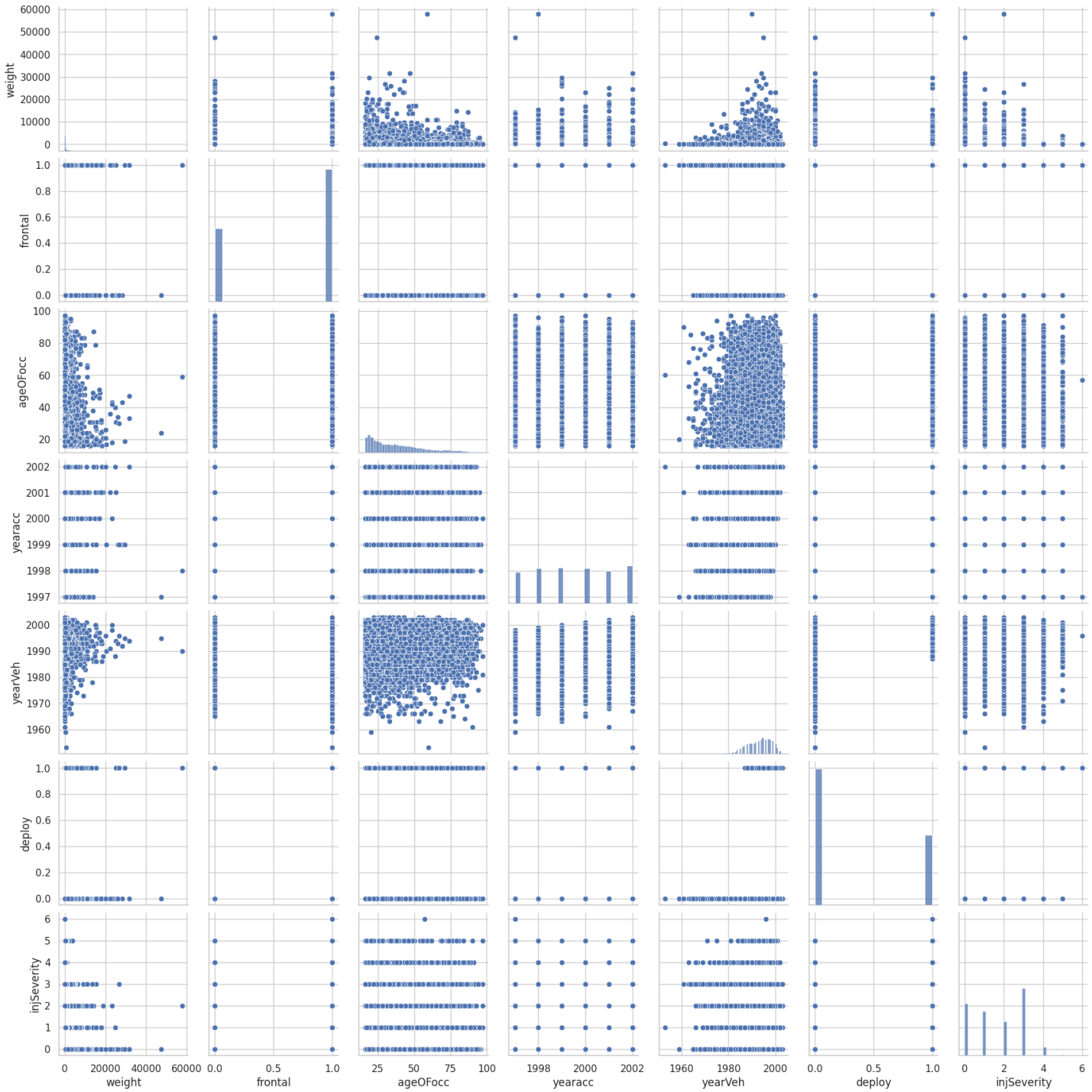


Figure 1

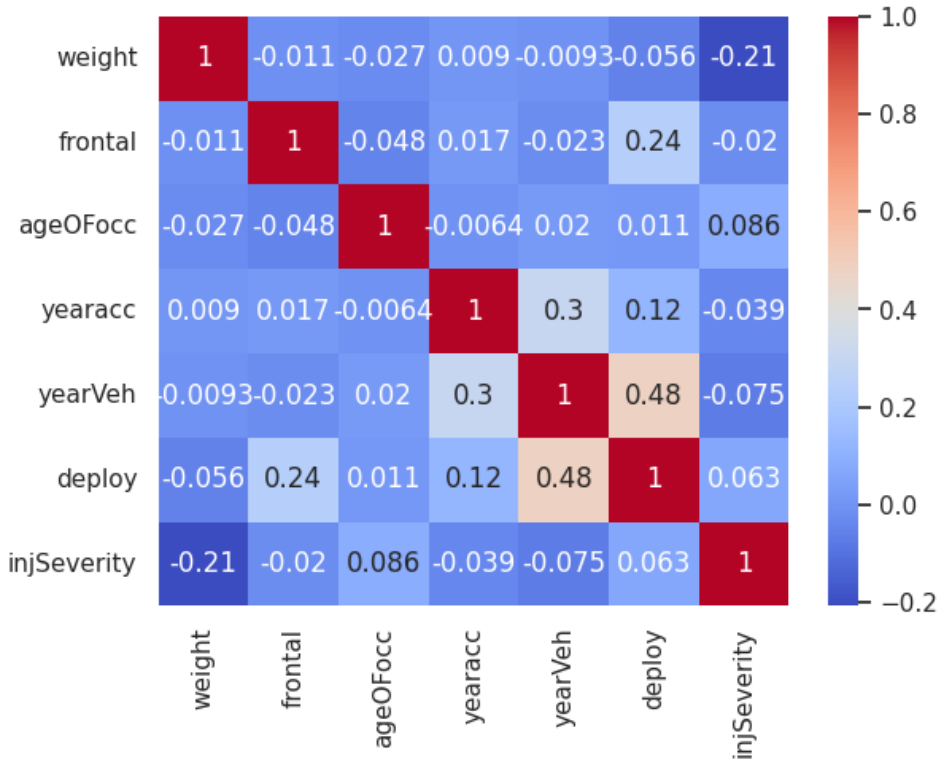


Figure 2

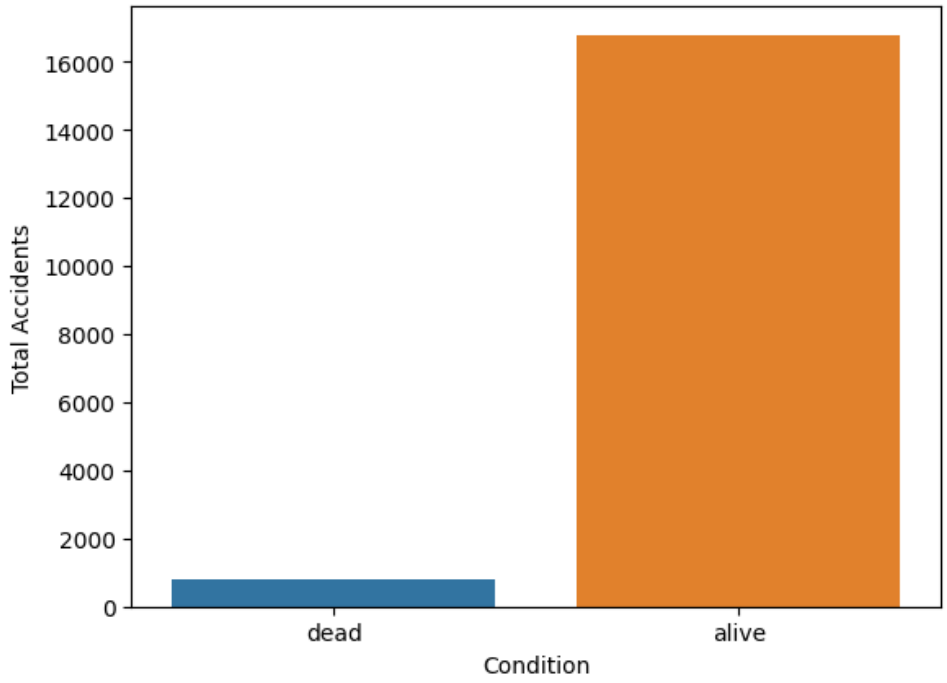


Figure 3

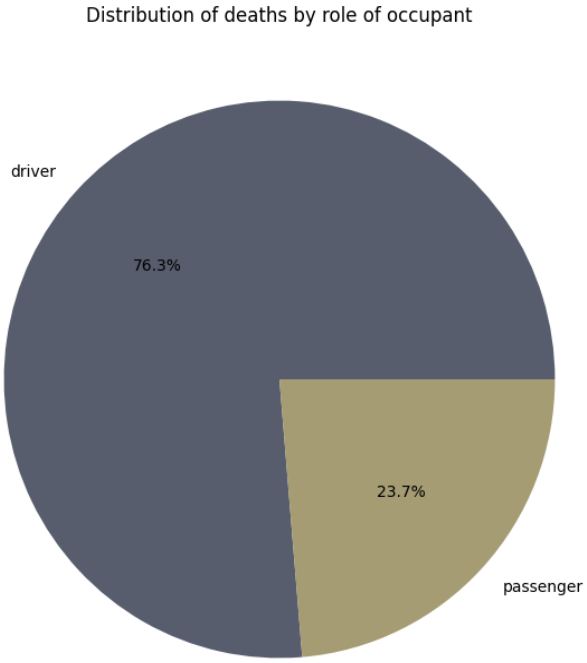


Figure 4

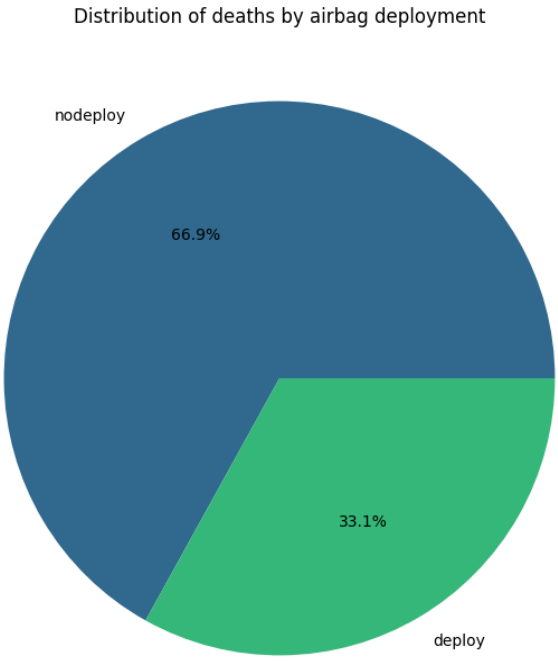


Figure 5

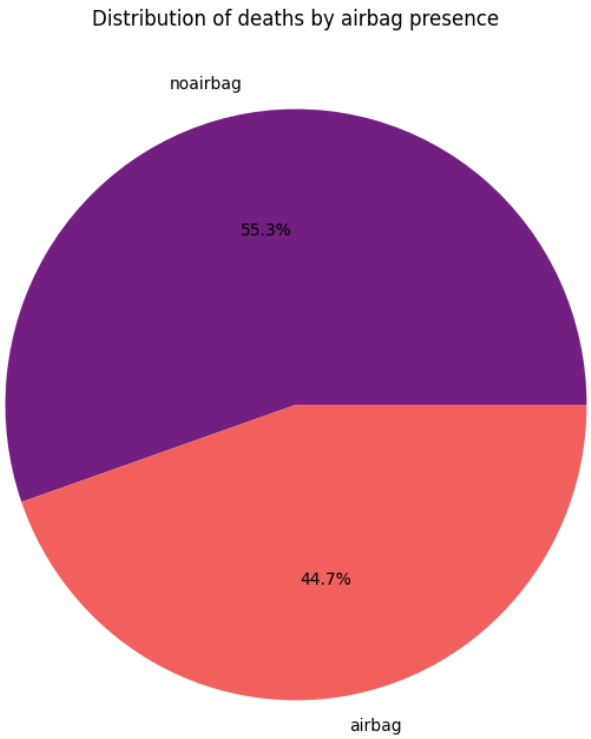


Figure 6

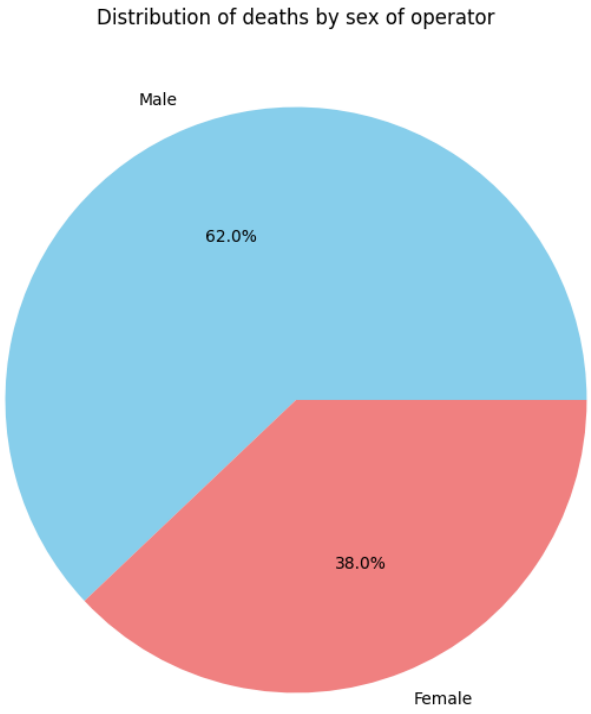


Figure 7

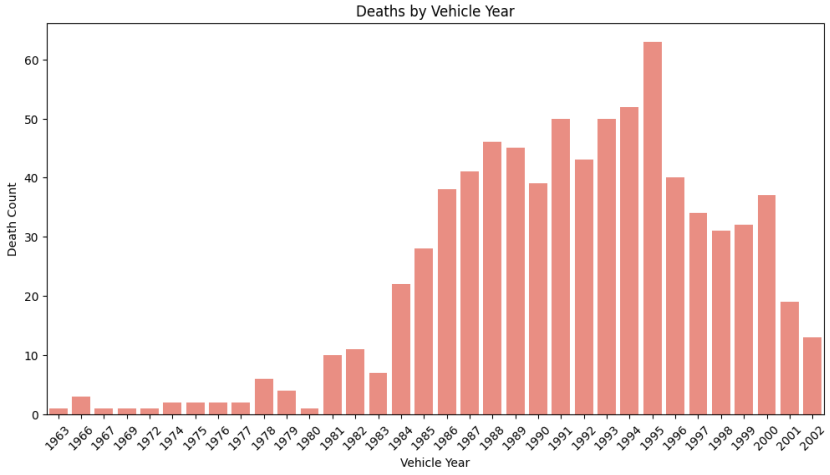


Figure 8

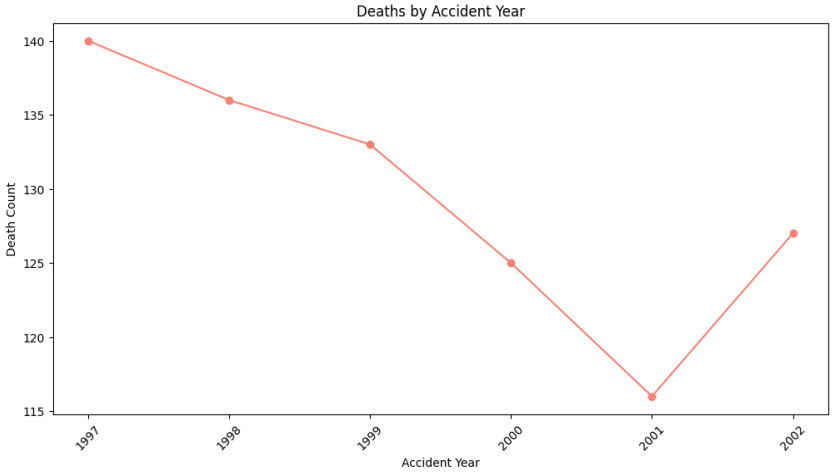


Figure 9

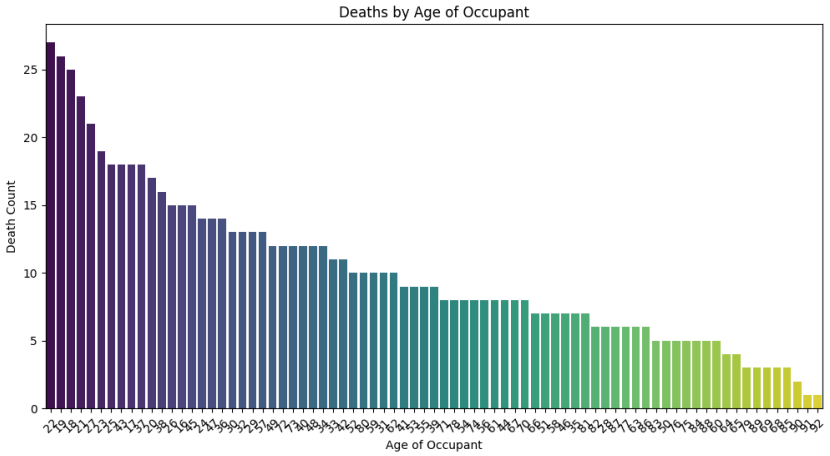


Figure 10

Analysis

Figure 1:

Figure 1 presents a pairplot/cross-product matrix with feature relationships. What does the information present in the figure *mean*? The graphs within figure 1 are the result of crossing a given dataset feature with several other dataset features. The pertinent x- and y-values are visible along the lower left edges of the figure. The graphs are generated in series. A total of 49 graphs appear.

Figure 2:

Figure 2 presents a heat map to illustrate data correlation. What does the information present in the figure mean? The decimal values in the heatmap show the correlation coefficient of two features. The correlation coefficient indicates how *closely* two features are related. A larger value indicates a closer relationship. Conversely, a smaller value indicates a more distant relationship. Of course, the top-left to bottom-right diagonal will be composed entirely of 1s as a feature will always have a 100% correlation to itself.

Figure 3:

Figure 3 presents a bar chart that observes fatal versus nonfatal accidents within the dataset. What does the information present in the figure mean? The information shows that of the 17,565 records in the dataset, approximately 800 accidents proved fatal, and 16,765 were walkaways. Therefore, across every record in the dataset, 4.555% (800) of accidents are fatal, and 95.445% (16,765) of accidents are nonfatal. This translates to a ratio of 20.956 nonfatal accidents to every 1 fatal accident. This could be because drivers have become more experienced and/or become more mature, newer and safer technology is being implemented in newer vehicles and roadways, or because the vast majority of accidents are not incredibly severe, which may or may not have a correlation with the first two clauses.

Figure 4:

Figure 4 presents a pie chart that observes the distribution of fatalities between operators and passengers involved in accidents. What does the information present in the figure mean? The information shows that, of those 800 people whose lives were unfortunately lost as a result of an accident, 76.3% (610) were operating the vehicle at the time of impact, and 23.7% (190) were passengers in the vehicle. This translates to a ratio of 3.219 driver fatalities to every 1 passenger fatality. This is likely explained by the fact most accident fatalities occur when the operator is alone, and not when there are multiple vehicle occupants. It could also be explained by vehicle operators not wearing their seatbelts or even which traveler has access to a functioning airbag.

Figure 5:

Figure 5 presents a pie chart that observes the distribution of fatalities between accidents that occurred with and without airbag deployment. What does the information present in the figure mean? The information shows that, of the 800 fatalities recorded, 66.9% (535) of fatalities occurred when no airbag deployed, and 33.1% (265) occurred when an airbag did deploy. This translates to a ratio of 2.021 fatal airbag nondeployment accidents to every 1 fatal airbag deployment accident. This is likely explained by the fact airbags serve as safety mechanisms. Someone without access to an airbag or someone whose airbag does not deploy appropriately is significantly more likely to perish in a potential accident. Seatbelt usage may similarly play a role.

Figure 6:

Figure 6 presents a pie chart that observes the distribution of fatalities between travelers who occupied a seat with and without a functioning airbag. What does the information present in the figure mean? The information shows that of the 800 fatalities, 55.3% (442) of travelers did not have access to a functioning airbag, and 44.7% (358) did have a functioning airbag. This metric does not account for those fatalities who had access to an airbag that did not deploy. This translates to a ratio of 1.237 fatal

nonfunctioning/inaccessible airbag accidents to every 1 fatal functioning/accessible airbag accident. Similar to figure 5, this is likely attributable to the fact airbags play a very large part in vehicle safety. Someone whose accident is softened by an airbag is more likely to survive than the opposite. This metric does not assume functioning/accessible airbag deployed successfully upon impact.

Figure 7:

Figure 7 presents a pie chart that observes the distribution of fatalities between male and female operators. What does the information present in the figure mean? The information shows that of the 800 fatalities, 62.0% (496) of travelers were male, and 38% (304) were female. This translates to a ratio of 1.632 male fatalities to every 1 female fatality. This is likely explained by male strong-headedness and the desire to drive quickly and recklessly. It could also be explained by a disparity in male-female situational or spacial awareness and distractibility, but that is without the scope of this project. Are women more intelligent and/or better drivers than men? Probably.

It is important to note the data does not provide vehicle occupant count. We were unable to determine whether a given male/female fatality was in the vehicle with other male/female travelers or whether the other vehicle occupants also perished with the record fatality.

Figure 8:

Figure 8 presents a bar chart that observes the distribution of fatalities across vehicle years. What does the information present in the figure mean? The information shows that most of the accidents occurred during the latter half of the research period, with the most dangerous vehicle year being 1995 (approximately 65 fatalities). Why might this be?

For simplicity, we will ignore vehicles manufactured before 1980 as their recorded presences are negligible within the dataset; these vehicles do not have enough

associated records to be considered the safest vehicle. The safest vehicle years are 1983 with 6 fatalities, and 2002 with 12. Realistically it is most likely that 2002 is the safest vehicle year across the collection period. This is explained by the fact that newer vehicles will employ newer and safer technology. It may imply a correlation between time progression/vehicle year and overall driver maturity, awareness, and experience.

Figure 9:

Figure 9 presents a line chart that observes the distribution of fatalities across accident years. What does the information present in the figure mean? The information shows that most of the accidents occurred during the former half of the research period, with the most dangerous accident year being 1997 (140 fatalities). The safest accident year was 2001 (115 fatalities). This is most likely attributable to advances in vehicle and safety technology. There is a similar implication here that overall driver level, road safety and engineering/technology advancements all improve as time progresses.

The rate trends downwards over time, expectedly, until 2002, where the line begins to trend upwards once again. Why might this be?

Figure 10:

Figure 10 presents the final figure: a bar chart that observes the distribution of fatalities across occupant age. What does the information present in the figure mean? The information shows teens and young adults are responsible for the overwhelming majority of accident fatalities. We observe that drivers aged 22 are most prone to a fatal accident. Conversely, drivers aged 92 are least prone. As has already been posited several times, this could be explained by younger drivers being just that: young. As time progresses, a given driver will acquire improved tact, experience and reaction time, and the associated decrease in recklessness and carelessness that comes with them. Equally as likely, a given driver may or may not drive less as he or she gets older.

We observe a very strong negative correlation between age and fatality rate. As age increases, the attached fatality rate decreases.

There are many ways a team can analyze and interpret data. While we did not exhaust every available analysis avenue, we believe the figures we generated and their respective, associated analyses are very thorough and complete, and likewise make a strong argument for the scope of the project.

V Conclusion

Closing Status

In the middle of the term, we completed our exploratory data analysis. We wanted the graphs we generated to represent the relevant information in a clear and visually appealing manner. We went through several graph-generating iterations; each improving the overall meaningfulness and readability of our data presentation. Ultimately, we believe the figures we created do the data justice and present clearly and concisely.

The next steps were to pass the data through the PCA algorithm and then reevaluate the KNN and SVM classifiers with the newly processed data. We completed the remaining requirements without issue. In the end, we believe our project vision and ultimate results proved to be a success. We were able to create a KNN model and an SVM model which both achieved maximum accuracies and precisions of 95%.

The primary target throughout the entire project was fatalities in accidents; we considered evaluating other targets as time progressed but decided that preventing fatalities was and is the most critical and pertinent problem we face.

Contributions

All of our team members contributed equally to the project. Our team collaborated in tandem throughout the term. We selected Discord as our correspondence platform of choice, which we used almost daily.

We utilized the broad suite of tools Google Colab has to offer. Google Colab is an amazing cloud based interpreter and compiler for python and jupyter notebooks. Google Colab offers a vast array of preinstalled packages and modules, which simplified and

streamlined the completion of the project. It is also cloud-based, which allows for extremely effective collaboration and handoff capabilities (Colab projects are saved to a backend Google cloud compute server). Python worked very well for the purposes of our project.

Alfredo, Sidney and Maxwell made a very skilled team, and we experienced no issues or hangups over the duration of the project. The workload was never lopsided and we were always willing to help each other. Individual member contributions can be found below:

Alfredo Villavicencio:

- Initial project ideas
- Initial project proposal
- Midterm project progress report
- Final project report
- Input module
- Prediction module
- Model training
- Data visualization
- Debugging
- Final presentation

Maxwell Hauser:

- Initial project ideas
- Initial project proposal
- Midterm project progress report
- Final project report
- Association dataset
- Data gathering
- Model testing

- Debugging
- Final presentation

Sidney Levine:

- Initial project ideas
- Initial project proposal
- Midterm project progress report
- Final project report
- Data preprocessing
- Feature extraction
- Data classification
- Debugging
- Final presentation

Limitations

The final product proved to be very adept at carrying out our design. However, our model's success was not without some limitations. The most notable limitations the project encountered stem from several factors that influence the reliability and generalizability of the findings. Foremost, the inability to conduct a comprehensive fine-tuning of the SVM model restricts the exploration of its full potential and may result in suboptimal performance. The arbitrary selection of six principal components in the PCA translation process introduces another limitation, as the optimal number of components might vary across datasets. The utilization of median imputation to address missing values adds a layer of potential bias, particularly if the missingness is not entirely random. The predominance of binary data in the dataset further limits the broader applicability of the models, as their compatibility is tailored to this specific dataset's characteristics. This is a type of inherited limitation. Consequently, reproducing the

experiment on a different dataset becomes challenging, emphasizing the dataset-specific nature of the models.

A more adept model could be applied to other car accident datasets with minimal parameter tuning. Despite these limitations, the experiment serves as a valuable starting point, highlighting the need for more nuanced approaches and considerations in the pursuit of effective predictive modeling.

Unresolved Issues

This experiment, while providing valuable insights into predicting car accident fatalities, faces unresolved issues that warrant further investigation. One significant concern is the insufficient tuning of the SVM hyperparameters. This oversight could potentially limit the model's optimal performance. The arbitrary selection of six PCA components introduces uncertainty, as the ideal number of components may vary across datasets.

A more systematic investigation of different component numbers could provide a more nuanced understanding of the data's underlying structure. Additionally, it's crucial to acknowledge that the confusion matrix for the SVM model is noted as completely wrong, highlighting a critical issue in the evaluation process that requires careful scrutiny and rectification. These unresolved issues underscore the imperative for an even more comprehensive investigation, parameter sensitivity analyses, and refinements to enhance the reliability and applicability of the predictive models.

Future Direction

In considering future directions for this study, several avenues emerge. Firstly, addressing the unresolved issues pertaining to the SVM model's hyperparameters and the accuracy of the confusion matrix is paramount. A comprehensive grid search or randomized search can be employed to systematically fine-tune the SVM hyperparameters, optimizing its performance. Additionally, revisiting the PCA process by experimenting with various numbers of principal components will contribute to a more

meticulous understanding of the dataset's dimensionality. Exploring alternative dimensionality reduction techniques beyond PCA could also offer insights into potentially more effective models.

Furthermore, the incorporation of more advanced machine learning algorithms, such as ensemble methods or neural networks, should be considered to evaluate their performance in comparison to the KNN and SVM models. An in-depth feature importance analysis can provide a better understanding of the variables driving predictions and guide feature engineering efforts.

To enhance the model's generalizability, validating the predictive models on diverse datasets, representative of various geographical locations or time periods, would be an incredible advancement. This not only ensures the robustness of the models but also facilitates the identification of patterns that transcend specific datasets.

In conclusion, adopting a more comprehensive evaluation strategy beyond accuracy, such as precision, recall, and F1-score, will provide a more holistic assessment of model performance, particularly given the imbalanced nature of the dataset. It is important to note the deployment of interpretability tools, such as SHAP (SHapley Additive exPlanations), can unravel the black-box nature of certain models, offering insights into the decision-making process.

The future directions outlined aim to refine the models, address current limitations, and extend the applicability of the predictive modeling framework to diverse scenarios, ultimately contributing to a more robust and reliable approach to predicting car accident fatalities.

Ⅵ Appendix

Images and Tabular Data

	pre-PCA accuracy	post-PCA accuracy	precision	recall	F1 score	Confusion Matrix	AUC-ROC	AUC-PR
KNN	0.9533	0.8812	0.9574	0.9161	0.9363	[[67 336] [692 7557]]	0.5411	0.9767
SVM	0.9534	0.9534	0.9534	1.0000	0.9762	[[0 403] [0 8249]]	0.6518	0.9732

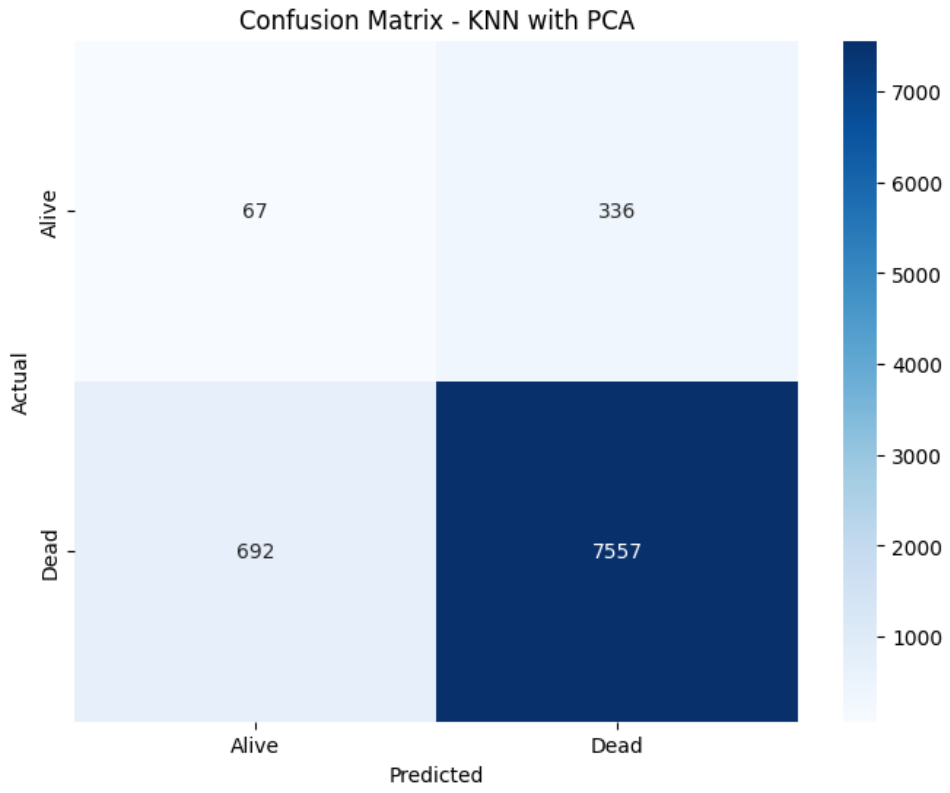


Figure 11

```
# Figure 11 associated code
from sklearn.metrics import confusion_matrix

# Compute confusion matrix
cm_svm = confusion_matrix(target_test, decisions_svm)

# Plot confusion matrix as a heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(cm_knn,     annot=True,     fmt="d",     cmap="Blues",
xticklabels=['Alive', 'Dead'], yticklabels=['Alive', 'Dead'])

plt.title('Confusion Matrix - KNN with PCA')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.show()
```

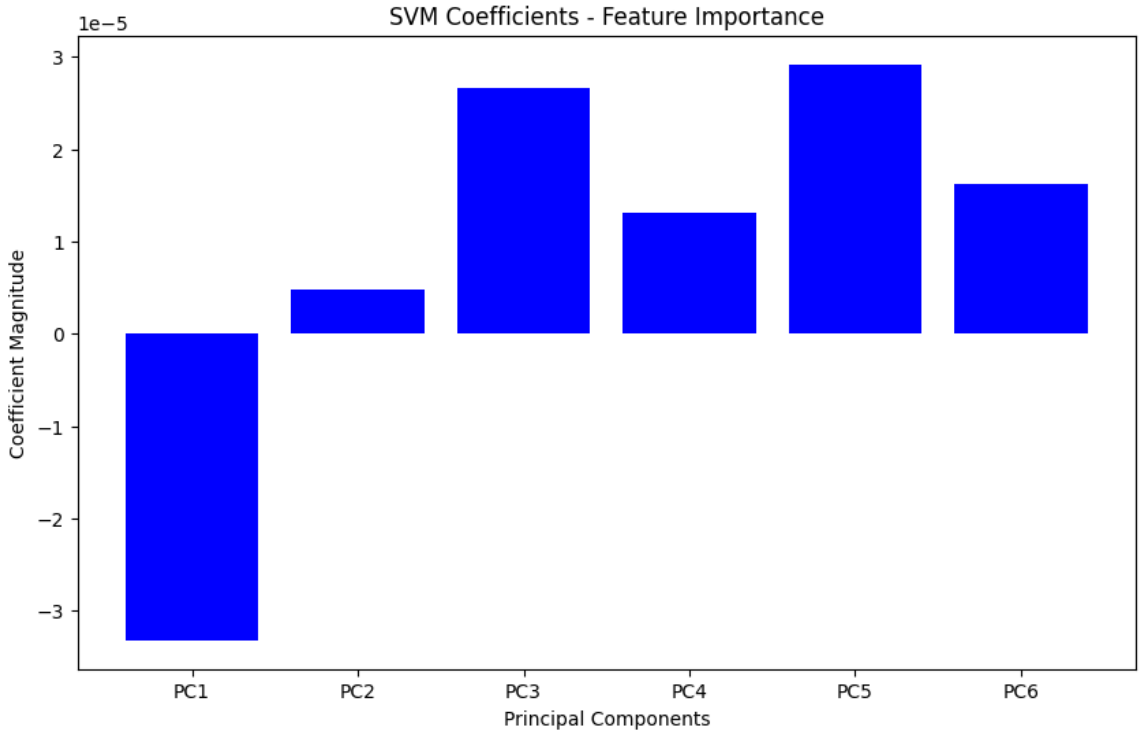


Figure 12

```
# Figure 12 associated code
# Retrieve the coefficients from the linear SVM model
coefficients = svm_classifier.coef_.ravel()

# Get the feature names from your PCA-transformed data
feature_names = [f'PC{i+1}' for i in range(len(coefficients))]

# Plot the coefficients
plt.figure(figsize=(10, 6))
plt.bar(feature_names, coefficients, color='blue')
plt.title('SVM Coefficients - Feature Importance')
plt.xlabel('Principal Components')
plt.ylabel('Coefficient Magnitude')
plt.show()
```

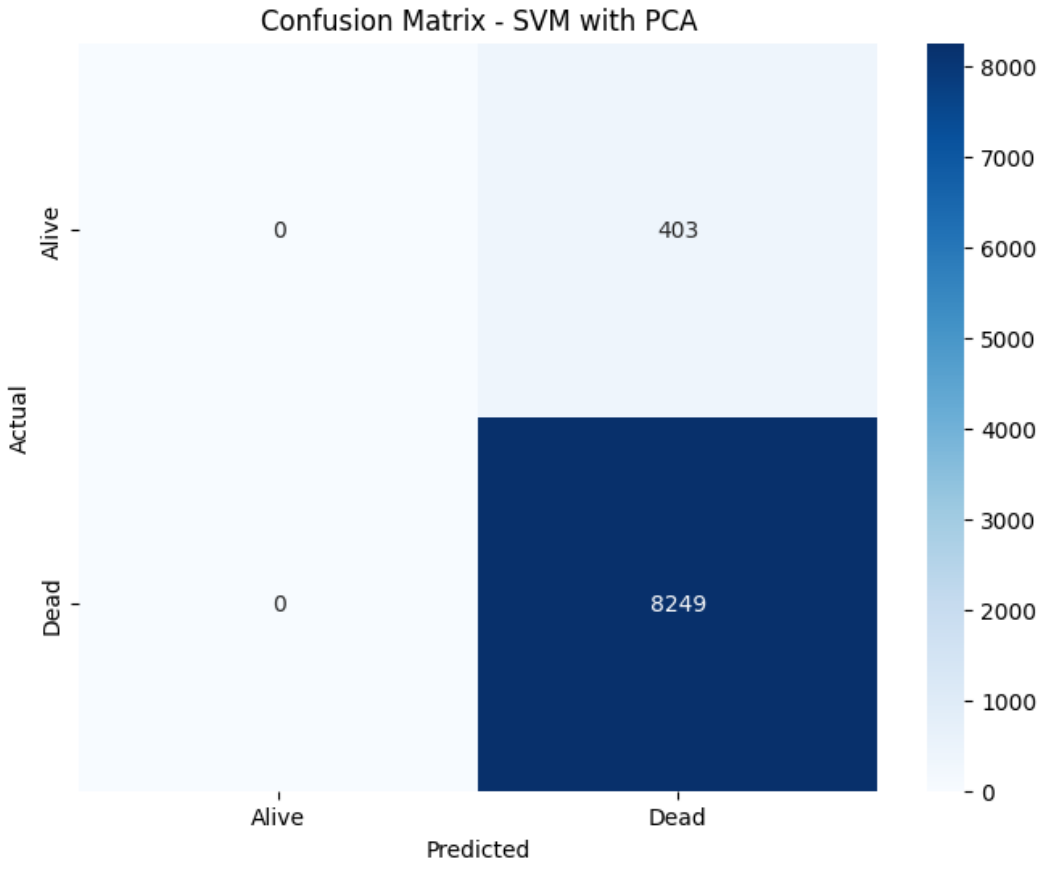


Figure 13 (erroneous)

```
# Figure 13 associated code
from sklearn.metrics import confusion_matrix

# Compute confusion matrix
cm_svm = confusion_matrix(target_test, decisions_svm)

# Plot the confusion matrix as a heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(cm_svm,     annot=True,     fmt="d",     cmap="Blues",
xticklabels=['Alive', 'Dead'], yticklabels=['Alive', 'Dead'])
plt.title('Confusion Matrix - SVM with PCA')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.show()
```

Final Timeline and Gantt Chart

Task	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12	Week 13	Week 14	Week 15
Project Proposal	Alfredo, Sidney, Maxwell														
Input Module		Alfredo													
Data Preprocessing			Sidney												
Association Dataset				Maxwell											
Prediction Module					Alfredo										
Feature Extraction						Sidney									
Data Gathering							Maxwell								
Model Training								Alfredo							
Data Classification									Sidney						
Model Testing										Maxwell					
Data Visualization											Alfredo				
Debugging												Alfredo, Sidney, Maxwell			
Presentation													Alfredo, Sidney, Maxwell		

References

https://www.kaggle.com/datasets/prasannakm/car-crash-dataset/?select=train-new.csv

Final Thoughts

The three of us learned a lot in this class. Dr. Hossain imparted so much useful knowledge not just as it pertains to Python, and SciKit Learn, but also as it pertains to artificial intelligence, machine learning, deep learning, data mining, algorithms, stratification and classification. The information taught throughout the course is incredibly relevant today, and it makes a person consider the ultimate capabilities of data mining and artificial intelligence. Patterning recognition and understanding patterns in information and data is a skill critical to the computer scientist toolbox, and Alfredo did an amazing job leading the project efforts and guaranteeing our experiments were the best they could be.

Epilogue

Dr. Hossain's experiential teaching style is one of the best I have experienced. I very much enjoyed taking both 400 and 477 with him, and I am very proud of the work I was able to complete in each of those classes under his supervision. Dr. Hossain earns more of my admiration the longer I work with him, and I would love to continue working with data mining and/or PHP in the future. I know that, regardless, I will take much of what I learned in his classes with me. Thank you for everything. — Maxwell P. Hauser ('23)

I'm delighted to express my sincere appreciation for the incredible experience I had in your class, specifically in 477. Given that it was one of the limited options available to me during registration, I feel fortunate to have been a part of it. Dr. Hossain's teaching style was truly captivating, igniting a curiosity within me and fostering a genuine interest in machine learning. I want to extend my heartfelt gratitude for creating such an

engaging and inspiring learning environment. Thank you for everything; it has truly been a transformative experience. I am looking forward to taking 400 with you. — Alfredo Villavicencio