**What is our data set about?**

The NTSB aviation accident database contains information from 1962 and later about civil aviation accidents and selected incidents within the United States, its territories, and possessions, and in international waters.

A preliminary report is available online within a few days of an accident. Information is added when available, and when the investigation is completed, the preliminary report is replaced with a final description of the accident and its probable cause. The dataset is updated monthly in Microsoft Access 2000 MDB format.

For this project we decided to use a version of the dataset only including incidents from 2008 to the Present. Our decision for this was largely because the datasets before 2008 did not distinguish whether the event represented the pilot controlling the aircraft at the time or whether the event index was associate with co-pilots, crew, or passengers’ injuries or related statistics.

To appropriately address our research question and reduce our sample size from over 100,000 events to around 8000. The dataset was originally cleaned via Excel in the Microsoft Access 2000 MDB format. From there we selected the appropriate variables and compiled them into a single Excel file where we were able to remove repeats, and null values. Finally, we saved this subset of the original Microsoft Access Database in a CSV format.

**Clean up your data and reduce it to no more than 2000 observations if your data set is very large.**

As our dataset was around 8000 observations, we utilized a randomized sample of 2000 observations. We then split this sample into an 75-25 split to test and train our classifiers.

**What is exactly your research question? What do you want to learn from data? What is your learning model, e.g., a Classification, Clustering, etc.?**

**Research Question: Is it possible to develop an accurate classifier that determines severity of injury for fixed wing aircraft incidents?**

In this study, we hope to create a model that can classify injury level based upon pilot age (measured in years), and experience (measured in flight hours).

It is important to note that flight hours are measured by the number of hours a pilot has been operating an aircraft, which is kept in a log and verified in annual inspections by a third party to ensure the airworthiness of the vehicle, along with its pilot. Further hours may be logged while completing training exercises as a co-pilot or with guided instruction for a CFI (Certified Flight Instructor.)

Time logged in simulators or ground school are not included in pilot flight hours. In this study we are looking to use an industry standard classifier (the SVM) to help predict pilot crash severity. The four classifiers we will be using are Logistric Regression, SVM SVC, SVM SVC Linear Regression, and a Gradient Descent/Hinge Loss function without using libraries.

**What is your current expectation about the results?**

Generally, we expect that younger pilots with less hours will be more reckless than those with more experience and more time in the cockpit. This inexperience is a key factor in why we believe these pilots will be responsible for a larger portion of fixed wing aviation incidents.

**How do you want to evaluate your project? How to access the correctness of your model? How well would you expect that the model will work?**

It is important to consider that we are only using two features to classify these events. That being Flight Hours and Age. There are many more features that play a part in aviation incidents which could contribute to aviation incidents.

We expect our model to perform better than a random chance algorithm, that is 51% or better. We do not expect there to be a large difference in the implementations, however we do believe that our dataset is not perfect. For future research, it could be useful to compare other models and include more features. Adding more features may improve the accuracy of our models, avoiding blatant representations such as engine failure, severe weather conditions, or drug/alcohol use.

**Results**

Our final results show that with Logistic Regression and the standard Sci-Kit Learn SVM SVC, we have an accurary ~ 50%, while our Linear SVM SVC has an accurary slightly > 50%, and SVM without libraries struggles to reach an accuracy score over 10-15%.

**What can you do to improve your results? What are limitations of your implementations or dataset?**

We believe cleaning and analyzing the data would improve the results. We believe there are outliers in the data that are lowering our model's accuracy. Furthermore, there are limited amounts of features that we used to develop our classifier. Looking into the NTSB Aviation Accident Database and picking more correlating features may help to improve our model. Additionally, many of the outliers in our data are present due to deciding to keep both private and commercial flight data in our dataset. Splitting these data groups up may also help to improve our model.

**Ideas Applied to improve results:**

We decided to utilize Linear SVM SVC because it is more accurate for larger datasets. We can observe that this particular model was more accurate than the other models implemented.

**Describe Code and Implementations:**

Our code and implementations are commented below.

**References:**

NTSB Aviation Accident Database:

https://data.ntsb.gov/avdata