Air Force Institute of Technology

AFIT Scholar

Theses and Dissertations | Student Graduate Works

3-2023

Predicting Success of Pilot Training Candidates Using Interpretable Machine Learning

Alexandra S. King

Follow this and additional works at: https://scholar.afit.edu/etd

 Part of the Artificial Intelligence and Robotics Commons, and the Aviation and Space Education Commons

Recommended Citation

King, Alexandra S., "Predicting Success of Pilot Training Candidates Using Interpretable Machine Learning" (2023). *Theses and Dissertations*. 7002.
https://scholar.afit.edu/etd/7002

This Thesis is brought to you for free and open access by the Student Graduate Works at AFIT Scholar. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of AFIT Scholar. For more information, please contact AFIT.ENWL.Repository@us.af.mil.

PREDICTING SUCCESS OF PILOT TRAINING CANDIDATES USING INTERPRETABLE MACHINE LEARNING

THESIS

Alexandra S. King, Second Lieutenant, USAF

AFIT-ENS-MS-23-M-134

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

DISTRIBUTION STATEMENT A
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

The views expressed in this document are those of the author and do not reflect the official policy or position of the United States Air Force, the United States Department of Defense or the United States Government. This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

AFIT-ENS-MS-23-M-134

PREDICTING SUCCESS OF PILOT TRAINING CANDIDATES USING INTERPRETABLE MACHINE LEARNING

THESIS

Presented to the Faculty

Department of Operations Research

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

in Partial Fulfillment of the Requirements for the

Degree of Master of Science in Operations Research

Alexandra S. King, B.S.O.R.

Second Lieutenant, USAF

March 23, 2023

DISTRIBUTION STATEMENT A
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AFIT-ENS-MS-23-M-134

PREDICTING SUCCESS OF PILOT TRAINING CANDIDATES USING INTERPRETABLE MACHINE LEARNING

THESIS

Alexandra S. King, B.S.O.R.
Second Lieutenant, USAF

Committee Membership:

Maj. Michael J. Garee, Ph.D
Chair

Brian J. Lunday, Ph.D
Member

Maj. William N. Caballero, Ph.D
Member

AFIT-ENS-MS-23-M-134

Abstract

The United States Air Force (USAF) has struggled with a sustained pilot shortage over the past several years; senior military and government leaders have been working towards a solution to the problem, with no noticeable improvements. Both attrition of more experienced pilots as well as wash out rates within pilot training contribute to this issue. This research focuses on pilot training attrition. Improving the process for selecting pilot candidates can reduce the number of candidates who fail. This research uses historical specialized undergraduate pilot training (SUPT) data and leverages select machine learning techniques to determine which factors are associated with success in SUPT. Humanly understandable (known as interpretable) machine learning techniques will be used to predict SUPT outcome, as these models provide justifications for these predictions and build trust with decision-makers. Three interpretable models were considered, including two rule-based models and one tree-based model. PCSM score was identified as the strongest predictor for success. The best performing model achieved an F1 score of 0.93, compared to 0.84 and 0.77 for the other models. The results of this research emphasizes the usefulness of interpretable models and their ability to inform a decision-maker, assisting them in their selection of future pilots.

Table of Contents

Page

List of Figures

List of Tables

PREDICTING SUCCESS OF PILOT TRAINING CANDIDATES USING INTERPRETABLE MACHINE LEARNING

I. Introduction

1.1 Motivation for Research

The United States Air Force (USAF) has been struggling with a shortage of pilots. Senior leaders have repeatedly expressed concern over the pilot shortage and how it is affecting our ability to complete the mission. In 2017, the Secretary of the Air Force (SECAF) Heather Wilson stated that the shortage was at around 2000 pilots, a 10% deficiency from where they need to be. The SECAF emphasized that a shortage of this size could "break the force" [1]. Reasons for this shortage are two-fold: the increase in pilot demand in the commercial sector, which decreases our retention of experienced pilots currently in the force, and attrition within pilot training [2]. This paper focuses on pilot training attrition since improving this issue will not only assist in closing the pilot shortage gap but ensure the USAF is investing money into trainees who have a higher likelihood of success.

The amount of money the USAF invests in its pilot trainees is enormous. The cost of training a qualified transport/mobility pilot ranges from $2.5 million to $5.5 million, and even more for a fighter pilot at $5.6 million to $10.9 million [3]. This is an incredible amount of money to be wasted on pilots not completing their training, which could be lessened better insight on what features are associated with candidate success. The cost in conjunction with the impact on mission readiness makes this a prescient issue and motivates this research.

This research is not only focused on building an accurate predictive model for success in pilot training but also a model readily understood by humans. Black box models are more common because of their good predictive performance; however, such models require decision-makers to blindly trust the results because they provides no justification for the predictions they make. Humanly understandable (also known as interpretable) models give decision-makers a greater understanding of why the model is making a certain prediction, and therefore allows the decision-maker to choose whether they want to trust the model, as they have a better understanding of the subject matter. Additionally, knowing how the model is making its predictions allows us to identify significant features used to predict success in pilot training. This research demonstrates how an interpretable model is used to classify pilot candidates as passing or failing training, and illustrates how these results are useful for identifying significant features.

1.2 Background

The pilot selection process has generally remained the same over the past several decades. Pilots are selected from several different commissioning sources and go through the same process to become a qualified USAF pilot, which is depicted in Figure 1.

There are many aspects to the candidate review process, as the evaluation board needs to ensure candidates are meeting a variety of requirements. The process can slightly differ based on commissioning source, and there are several that cultivate USAF pilot applicants: the United States Air Force Academy (USAFA), Air Force Reserve Officer Training Corps (AFROTC), and Officer Training School (OTS). Each applicant submits a package to their respective selection boards, which review each pilot candidate's mental, physical, and medical requirements, as well as their aca-

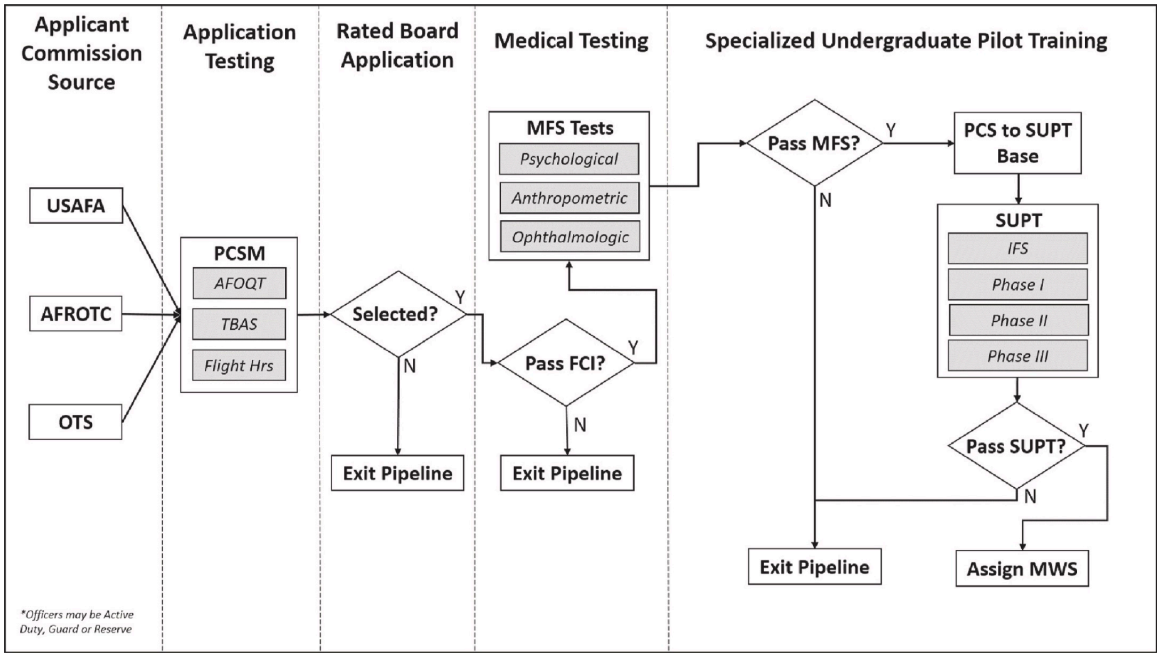

Figure 1: Flow chart of the pilot training application and completion process [4].

demics, scores from several evaluation and aptitude tests, previous flight experience, and any letters of recommendation. Whereas USAFA and AFROTC selection boards only evaluate an applicant's potential as a pilot because they have already been accepted for military service, OTS boards must additionally review the individual's capability for military service. The three commissioning sources also differ in how they weight each component of the applicant package.

While there are aspects that differ depending on commissioning source, there are parts of the process that are required for all pilot candidates; most importantly, all candidates must take several aptitude tests, which are intended to measure different metrics to determine an applicant's potential for success as a pilot. The required evaluations include the Air Force Officer Qualification Test (AFOQT) and the Test of Basic Aviation Skills (TBAS). Certain scores from these tests are then combined with the applicant's previous flight experience to create their Pilot Candidate Selection Method (PCSM) score.

The AFOQT is a standardized test with twelve separate sections which measure math, verbal, pilot aptitude, pilot knowledge, and personality traits [5]. It includes sections commonly seen throughout many other standardized tests, as well as sections on table reading, instrument reading, block counting, and a few others. These scores are used to determine an applicant's pilot and navigator composite scores. The composite score is expressed in terms of a percentile. To be considered a competitive pilot candidate, there are minimum pilot and navigator composite scores. For a more detailed description on each AFOQT section, go to Table 6 in Appendix 1.

The TBAS is the other aptitude test applicants are required to take; it is a computerized psychometric test to supplement the AFOQT [6]. The test is designed to measure an applicant's multitasking abilities, as well as spacial abilities related to flying. Further information on the TBAS sections can be found in Table 7 in Appendix 1. While applicants are never explicitly given their results for the TBAS, these scores are used in conjunction with AFOQT scores and previous flight hours to create the candidate's PCSM score. A candidate's PCSM score is a number from 0-100, where higher scores indicate a stronger pilot candidate.

The previous evaluation metrics all take place before acceptance; however, candidates are required to complete additional medical tests after they are accepted into the pilot pipeline. This timing sometimes varies based on commissioning source, but all pilot candidates must complete Medical Flight Screening (MFS)[7]. The candidate must pass before proceeding to their assigned SUPT base for training. Once candidates have in-processed at one of the three SUPT training bases, candidates are sent to Initial Flight Training (IFT); they must pass this phase before proceeding to the next two phases of SUPT.

The purpose of this research involves predicting a pilot candidates' success before they enter training, as well as being able to predict the phase in which they are likely

to fail, if they are predicted to fail. Therefore, the results of this research will be vital for the pilot selection boards, as it will improve their ability to select candidates with higher odds of success. Alternatively, it will identify *a priori* the points at which pilot trainees merit preemptive intervention to improve their likelihood of success in SUPT.

1.3 Problem Statement

Pilot training attrition is a significant problem in the USAF that has a direct impact on the readiness of the force and drains money being invested on unsuccessful pilot trainees. Existing studies on pilot training have used black box models to predict candidate success, and while black box models exhibit strong predictive accuracy they are not good for credibility amongst decision makers as they do not provide justification for their results. However, more novel models have been developed to produce understandable results while also performing well in terms of prediction accuracy. Using one of these novel interpretable models to predict pilot candidate success will allow decision makers to to understand the model's reasoning behind each prediction, and their greater understanding of the subject matter will allow them to decide whether or not to trust the results. Additionally, the transparent results produced from these models will highlight the important features associated with candidate success.

1.4 Research Objectives

We are interested in pursuing the following research objectives:

1. Create an interpretable machine learning model to accurately predict (85% accuracy or greater) whether a pilot candidate will pass or fail SUPT, and if they fail, predict the phase in which they will fail.

2. Determine the features associated with pilot candidate success. This finding will provide evidence either for or against the current evaluation metrics used by the pilot training selection boards. It will also highlight which cutoffs the model is using to predict success, which will allow us to identify whether the current standards are too high or too low for candidates.

1.5 Scope

This research is based on historical SUPT data ranging from 2010-2018. The data has been filtered to only contain aircraft pilots, as it previously included others such as Combat Systems Officers (CSO) and Remotely Piloted Aircraft (RPA) pilots. It excludes pilots that proceed to helicopter training, as the pipeline for helicopters is slightly different. We are only looking at individuals completing USAF pilot training, not from any other branches of the military.

1.6 Summary of Key Contributions

Out of the three models chosen, scalable Bayesian rule lists (SBRL) exhibited the best performance accuracy with an F1 score of 0.93 and a prediction accuracy of 0.86; however, FIGS had the best readability as well as the most well-rounded results. All three interpretable models validated PCSM score as an efficient metric for predicting a candidate's performance in pilot training; fast interpretable greedy-sums (FIGS) identified that a PCSM score of at least 35.5 is significant in distinguishing between passing and failing. According to certifiably optimal rule lists (CORELS), individuals failing in Phase 1 appear to have less flight knowledge and experience whereas those failing in Phase 3 generally have more experience (i.e. they could be leaving for reasons other than failure, but we do not have the data to support such a determination).

1.7 Overview

Chapter II reviews previous literature both relating to pilot training as well as different applications of interpretable methods. It also explains several different interpretable methods in more detail including the ones used in this research, and it shows an example output of CORELS and how its accuracy compares to other black box methods. Chapter III starts off explaining the structure of the data and the initial cleaning that was required, and then delves into some preliminary analysis of the data to demonstrate some potential relationships present between features. The chapter then discusses how we approached the class imbalance in the data as well as the software used, and it ends by identifying which methods were chosen. Chapter IV reveals all of the outputs from the three models used, and discusses potential interpretations of each result as well as the accuracy for each model. Finally, Chapter V concludes this research by discussing the extent to which each research objective was met, including a summary of key contributions and a description of potential future work to extend this research.

II. Background and Literature Review

This section surveys previous research revolving around pilot training, including studies examining how to streamline the process and other work on the different methods that selection panels use to evaluate potential pilot candidates. We then discuss interpretability in machine learning and its meaning, to include research that has used interpretable methods for a wide range of fields. We then identify and compare modern interpretable methods including those selected for use in this thesis.

2.1 Pilot Training Studies

Pilot training has remained fundamentally the same over the past several decades in the USAF, even with ongoing issues related to student attrition. However, there have been attempts to streamline and improve the process, and research has demonstrated the efficacy or inefficacy of these changes. One attempt includes the creation of Pilot Training Next (PTN), a program that focuses on simulation-based training, which has been successful at shortening the length of UPT for those in PTN. It has also automated certain tasks usually taken on by Instructor Pilots and increased efficiency by allowing the instructors to focus on more vital tasks [8]. This has been a step in the right direction, unlike previous attempts by the USAF to improve the training process, such as the replacement of IFT with Initial Flight Screening (IFS). This effort was an attempt to minimize voluntary attrition in pilot training and showed no significant difference in attrition in comparison with IFT [9].

Many pilot training studies focus on the validity of the evaluation metrics used to determine acceptance into pilot training [7, 10, 11]. As expected, these studies have continuously validated these metrics as highly effective at predicting success for candidates, especially the AFOQT and the PCSM score. These papers also confirmed

that flying experience rated highly in its predictive ability, as prior flying training exposes candidates to direct job knowledge and additionally screens out those who may have less motivation, those who have an unknown fear of flying, and those who struggle to handle the aircraft properly [7, 10]. Although often overlooked, research has also confirmed that the measures of cognitive ability and personality traits in the AFOQT are important factors in predicting pilot success. It is encouraged for these scores to be used in conjunction with the PCSM score to measure pilot aptitude [12].

Other studies have identified other potential predictors for success in pilot training beyond the current evaluation measures. A study on the demographics of SUPT attrition shows that women and minorities face unique challenges to becoming pilots. Before entering training, they experience less support and encouragement to pursue the piloting career field. Additionally, as they are greatly outnumbered by the white male majority, they can often struggle to find support through same race or same gender peers [13]. The study also showed that, across these demographic groups, there were few differences in rates of self-elimination. However, the high attrition rates among women and minorities mainly stemmed from flying performance. The paper emphasized that race/ethnicity and gender are not solely driving these differences in attrition rates, but that members of these groups are less likely to possess the preparation or experience that increases the chances of success. Another paper examined factors influencing low representation in the piloting career field, and in its findings showed that gender discrimination is still present within the industry. Because there have been very few changes to the training pipeline over the past several decades and the training program was developed for years by men for men, there could be some issues to be considered when training women [14].

In one of the most prevalent studies that aim to predict pilot training success, Jenkins' and Caballero's black box model identified degree type and commissioning

source as the most important features in determining academic success, in addition to the number of AFOQT retests, race, and gender. It is relevant to note that the AFOQT, TBAS, and PCSM scores are not present among the most influential factors, which implies that the results from these evaluation metrics alone is insufficient to predict a candidate's success in SUPT. Furthermore, the personality questions on the AFOQT were found to be among the least influential factors, and they recommended this section be considered for removal to minimize overall testing time [4]. While this paper developed some important findings, they used black box models that are difficult to understand and provide no justification for their predictions.

This research aims to reach the same goal, but using interpretable models instead of black box methods. This research also takes the previous paper a step further and attempt to predict the phase in which a candidate fails, if they are predicted to fail. The interpretable models used in this research allows us to produce results that justify the reasoning behind the predictions; this builds trust in the model and promotes a higher understanding of how we can better evaluate our SUPT candidates.

2.2 Explainable AI Methods

2.2.1 Interpretability and Its Importance

Interpretability does not have an official mathematical definition, but in this context, it is best defined as "the degree to which a human can understand the cause of a decision" and "the degree to which a human can consistently predict the model's result" [15]. Both *interpretable* and *explainable* are used interchangeably in the following sections; there is a distinction, however, in the term *explanation* which is defined as a justification for a specific result/prediction.

Interpretability of machine learning models is incredibly important, but it can be dependent upon the scenario we are analyzing. Sometimes all we wish to accomplish

is the prediction itself, and we are not interested in the reasoning behind the decision. In other instances, we would like to know why the prediction was made and trade some of the accuracy of our prediction for higher interpretability. Understanding our models and why they are making these predictions builds more trust in the model, which is incredibly important within high-stakes environments. Interpretability helps build our model's credibility so that we have greater confidence that our results are accurate outside of the simple metric of classification accuracy.

2.2.2 Studies Using Interpretable Models

Current machine learning models such as random forests, kernel methods, and deep neural networks produce very accurate predictions. Although these models are accurate, they are very complex and difficult to understand, resulting in poor interpretability of the results [16]. Despite the machine learning community propagating the idea that there exists a large trade-off between accuracy and interpretability, it may not be as drastic as we have grown to believe. According to Rudin, when the data is structured and the features incorporated into the model are meaningful, we can get similarly accurate results from both black box and interpretable models [17].

While interpretable models are not as popular as black box methods, interpretable models can be used in many different fields, which conveys their usefulness and importance in various settings. For example, interpretable methods have been applied to early warning systems in an academic environment; interpretability is important in this instance, as we need to know why students are failing or have a potential to fail so that they have an opportunity to improve before it is too late [18]. Another field where interpretability is important is in the medical realm, as the model should be able to provide justifications for each prediction. Any model that is making predictions for high-stakes decisions should provide these justifications to the decision

makers, as it allows field experts to decide whether or not to trust the model outcomes [19].

2.2.3 Comparing Modern Interpretable Models

As mentioned previously, black box models have historically been preferred over interpretable models because of their superior prediction accuracy. However, more recent interpretable models have demonstrated accuracy levels similar to those of black box models; these developments are ground-breaking in that we can have both accuracy and transparency in our results, something that was never achievable in the past. Table 1 provides relevant information for a few of these newly developed models.

There are three main factors that differentiate one model from another. The first is the type of data input that is required, such as numerical, binary, or categorical. The second involves the type of classification it performs, which is either regression, binary classification, or multi-class classification. Lastly, they differentiate in terms of the optimality of the results, as some are not certifiably optimal and instead rely on heuristics [20].

From Table 1, we see that the majority of these models take in binary data as their input. This means that the features being used to make predictions are binary. While at first this may seem very limiting, it is relatively easy to convert continuous variables into several binary variables by binning the data, which also means these

Table 1: Summary of Interpretable Models

Model Name	**Data Input**			**Type of Classification**	**Solution Method**
	Binary	**Categorical**	**Numerical**		
CLASSY [21]	X	X		Multi-Class	Heuristic
CORELS [22]	X			Binary	Optimal
SBRL [23]	X	X		Binary	Optimal
ORL [24]	X			Binary	Mixed IP
ICRM [25]		X	X	Multi-Class	Heuristic
FIGS [26]	X	X	X	Binary	Heuristic

models may require more data pre-processing than others. Additionally, the majority of these models are limited to binary classification, which limits the range of problems to which these models can be directly applied.

The interpretable models represented in the table are mostly rule list models, with one tree-based model. CLASSY works by using a greedy heuristic search algorithm to find good rule lists; which means that unlike CORELS, CLASSY is not certifiably optimal, as heuristic algorithms are not guaranteed to find global optima. SBRL is a Bayesian rule-list model, and similar to CORELS it creates bounds. Yet, CORELS implements tighter rule list bounds, which reduces the search space and increases the efficiency of the algorithm. Although ORL employs Mixed Integer Programming to find an optimal rule list, CORELS is still deemed to be more efficient. The last rule list model, ICRM, uses an evolutionary programming algorithm to minimize both the number of rules and conditions provided, yet this provides no guaranteed optimality either.

Tree-based models such as decision trees and random forests have long been considered a cornerstone of machine-learning practices. Fast Interpretable Greedy-Tree Sums (FIGS) is an generalization of the Classification and Regression Tree (CART), creating multiple smaller trees within a summation rather than a single tree. The FIGS model is intended to overcome a key weakness of single-tree models by reducing repeated splits of the same feature, which thereby reduces the redundancy of the model. FIGS is able to provide more concise decision rules than singular decision trees without sacrificing predictive performance.

All of these methods are highly effective, both in terms of model performance and interpretability. Some models can be more effective on certain data sets than others because of a variety of factors, such as number of features, number of samples, types of data inputted, etc. This is discussed in further detail in Chapter III, when we

discuss which methods are used in this research and why.

2.2.4 CORELS Example Output

Now that we have established a general explanation of what interpretability means, it is useful to depict some outputs of these models to give an idea of what interpretability looks like. In the original CORELS paper, Rudin explains that CORELS was built so that we could have an interpretable model with guaranteed optimality; this was because of the ProPublica article on the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) recidivism prediction tool [22]. The article highlights a case where a black box model was being used for recidivism predictions, where recidivism is the tendency for a convicted criminal to re-offend (in a time-frame of two years, in this instance). The COMPAS results were racially biased, but because they were using a black box model, it was unable to determine the reason behind the bias, nor the reason behind any prediction. They had assumed the black box model would provide better accuracy. This is where the need for something like CORELS appeared, a certifiably optimal model that provides transparent results.

Figure 2 depicts the output CORELS provided for this problem, presented as a series of IF-THEN statements [22]. This format, known as a rule list, demonstrates how samples are separated into the two different classes based on which conditions are met for that specific sample. The two classes are yes, the individual will re-offend within two years or no, they will not re-offend within two years. For example, the output could be interpreted as: if a sample is 19 years old and is male, they are predicted to re-offend within two years after their release.

The output for CORELS and other interpretable models can vary in complexity and size, depending on the number of features included in the data. The type of output can be determined using parameters to help maintain the simplicity of the

if $(age = 18 - 20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21 - 23)$ **and** $(priors = 2 - 3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no

Figure 2: CORELS output for 2-year recidivism data set [22].

output and ensure we are getting understandable results.

Compared to other models, both interpretable and black box, CORELS demonstrates a similar prediction accuracy. In fig. 3, Rudin compares the performance of CORELS to nine other algorithms, where some are interpretable and some are black boxes [22]. Across all algorithms, there was no significant difference in test accuracies. This result shows that for the recidivism problem, CORELS produces models that perform comparably to black box models in terms of accuracy while providing superior interpretability.

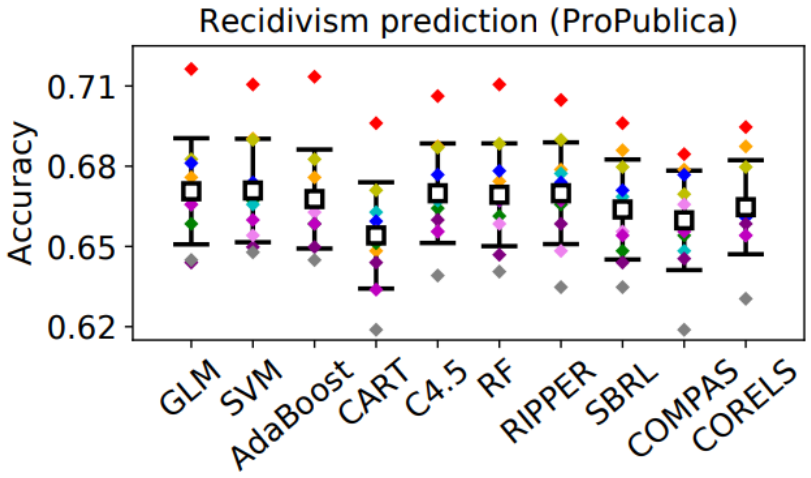

Figure 3: Accuracy of CORELS compared to 9 other algorithms for 2-year recidivism prediction [22]

III. Methodology

This chapter goes over several important details explaining how we apply the methods previously mentioned to our data. The chapter describes what the data entails, preliminary analysis and visualizations of the data, and what modifications have been made to the data so that we can apply our chosen methods.

3.1 Data

3.1.1 Structure of the Data and Initial Cleaning

The SUPT data set used in this research is made up of 10,435 samples across 78 input features and 1 categorical response variable (SUPT Pass or Fail, whereas fail also indicates what phase they failed in). The data is explained in more detail in Table 2 and Table 8 (which can be found in the Appendix) and includes 68 continuous features and 10 categorical features.

Table 2: SUPT Data Set Categorical Features

Feature	Levels
Gender	Male, Female, Unknown
Commissioning Source/Status	AFROTC, USAFA, OTS, Active Duty, Reserves, Guard, Other
Race	Caucasian, Minority, Unknown
Degree Type	High School, Associates, Bachelors, Masters, Doctorate, Unknown
Aerospace Related Major	No, Yes
Engineering-Related Major	No, Yes
Number of AFOQT Retests	0, 1, 2, 3
Policy Waiver	No, Yes
Prior Enlisted	No, Yes

Each sample in the original data set represents a SUPT candidate between the years of 2010 and 2018. Several thousand of these samples needed to be removed because a majority of the data was missing, an issue that cannot be remedied with

data imputation. Furthermore, within this research we determined that helicopter pilots were outside of the scope since they are a part of a separate pipeline and have different requirements. Additionally, Combat System Officers (CSO) were removed as well because while they are considered rated, their pipeline and training is also different. These two groups were present in the original data only because they are required to take part in Phase 1 of SUPT, but afterwards they branch off into their respective training pipelines.

The SUPT data indicated that the pilot training environment is a white male dominated environment, with the racial makeup being 89.3% Caucasian and 10.6% American Indian, Alaska Native, African American, Pacific Islander, or a combination of two of the mentioned races. As for gender, 91.3% of our candidates are male and 8.7% are female.

Table 3: Graduation Rates By Year

2011	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**
78%	84%	83%	79%	87%	87%	92%	83%

The SUPT data set suffers from a common issue of class imbalance, as we have significantly more candidates successfully completing SUPT than not. Our data demonstrates that approximately 85% of our candidates finish training while 15% do not (either voluntarily or from failing out). The graduation rates for every year included in our data set is depicted in Table 3. Since this issue is something that can affect the validity of our results, we will discuss how it was rectified with Synthetic Minority Oversampling Technique (SMOTE) in the next section.

3.1.2 Preliminary Data Analysis

This section goes into detail regarding the correlation of certain features within the data and success in pilot training. There are several features we would like to inspect

further, including the amount of AFOQT retests, Flight Hours, GPA, PCSM Scores and AFOQT Pilot Scores. These features are currently the most obvious metrics to evaluate candidates' potential for success in pilot training, and we conduct some preliminary analysis before applying our interpretable methods.

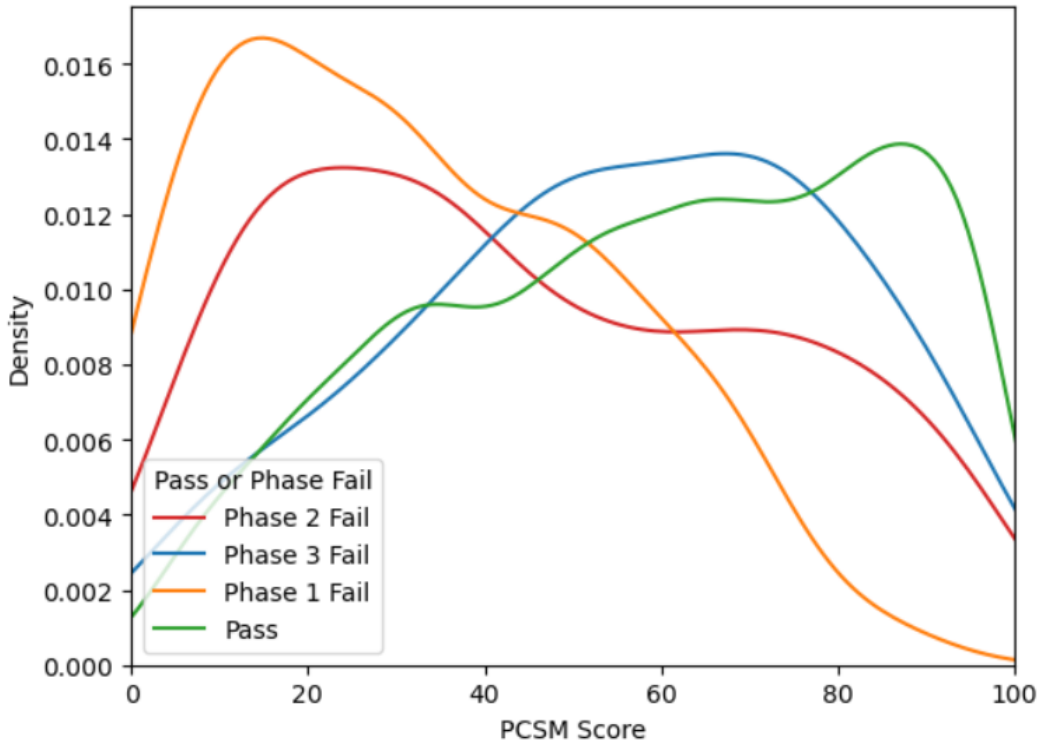

Figure 4: PCSM scores for each SUPT outcome

The PCSM score is heavily relied upon as an indicator of success within pilot training, as it compiles AFOQT and TBAS scores along with FAA Flight hours into a singular number. It was designed to provide an overarching view on how prepared an individual is for training. Similarly, the AFOQT pilot composite score is directly measuring an individuals prior knowledge required for pilot training. Figure 4 depicts the distribution of PCSM scores among each success/failure group in an attempt to capture any noticeable patterns across these groups. From this graphic, it is clear that samples with lower PCSM scores are failing in Phase 1, while those who passed exhibit a clear positive trend with PCSM scores. Figure 5 shows a strong relationship

between the AFOQT pilot composite score and individuals failing in Phase 3, as there is a much greater density for higher pilot composite scores. This could indicate that individuals failing in this phase are highly qualified and prepared for training, but are perhaps exiting the pipeline voluntarily. Although the initial data set had a column indicating the reason behind each candidate leaving or failing training, there was too much missing data for it to be useful.

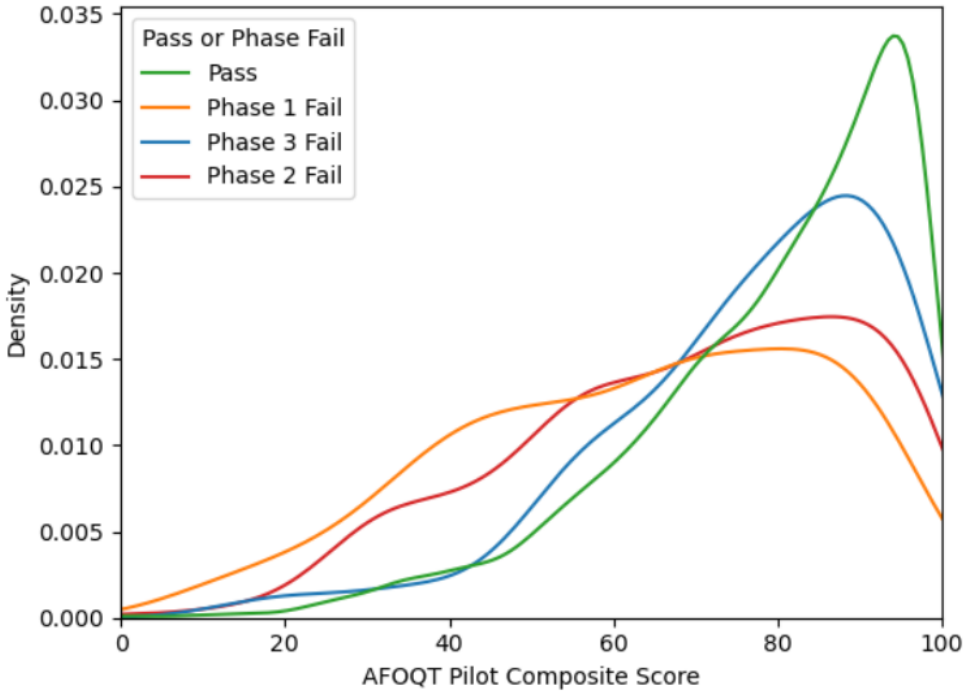

Figure 5: AFOQT Pilot Composite Scores for each SUPT Outcome

Furthermore, AFOQT retests appear to have a relationship with the AFOQT pilot composite score, as indicated in Figure 6. As the number of retests increases, the density shifts towards higher pilot composite scores. It is also apparent that as the number of retests increases, a greater density of individuals are passing, and those who have undertaken three AFOQT retests have a 100% pass rate.

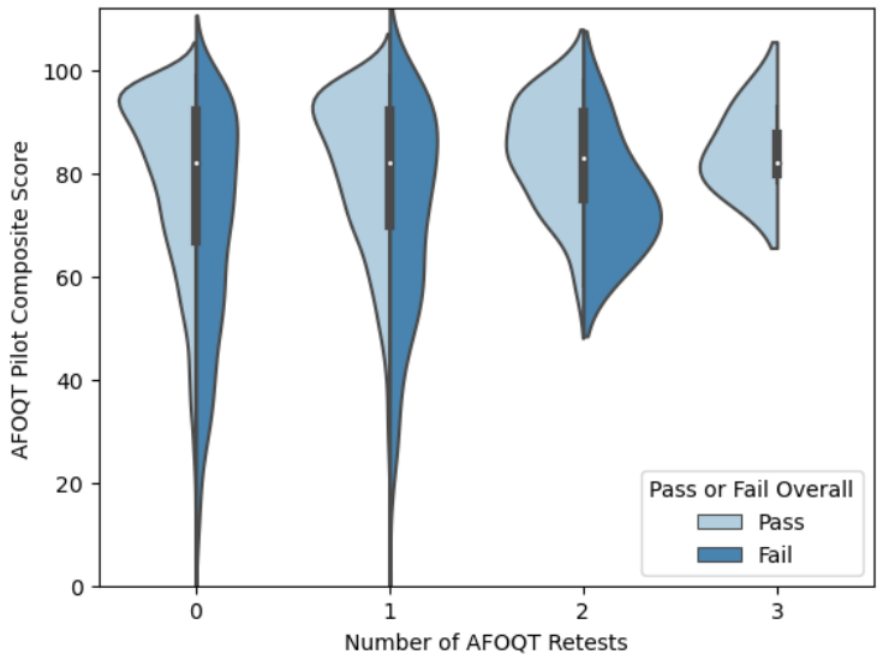

Figure 6: Distribution of AFOQT Pilot Composite Scores based on AFOQT retests and SUPT Outcome

Flight hours demonstrate how much prior flying experience each candidate has prior to entering SUPT, which could largely be correlated with candidate success. In Figure 7 we are using the square root of flight hours as the data ranges from 0 to 5910, which makes it difficult to notice any details with such a large range of values. In Figure 7, we can see that there is not a very noticeable difference between the overall pass and fail groups. However, it is evident that those who fail in Phase 1 have significantly fewer flight hours than any other group, which could indicate that these individuals generally have less flying experience.

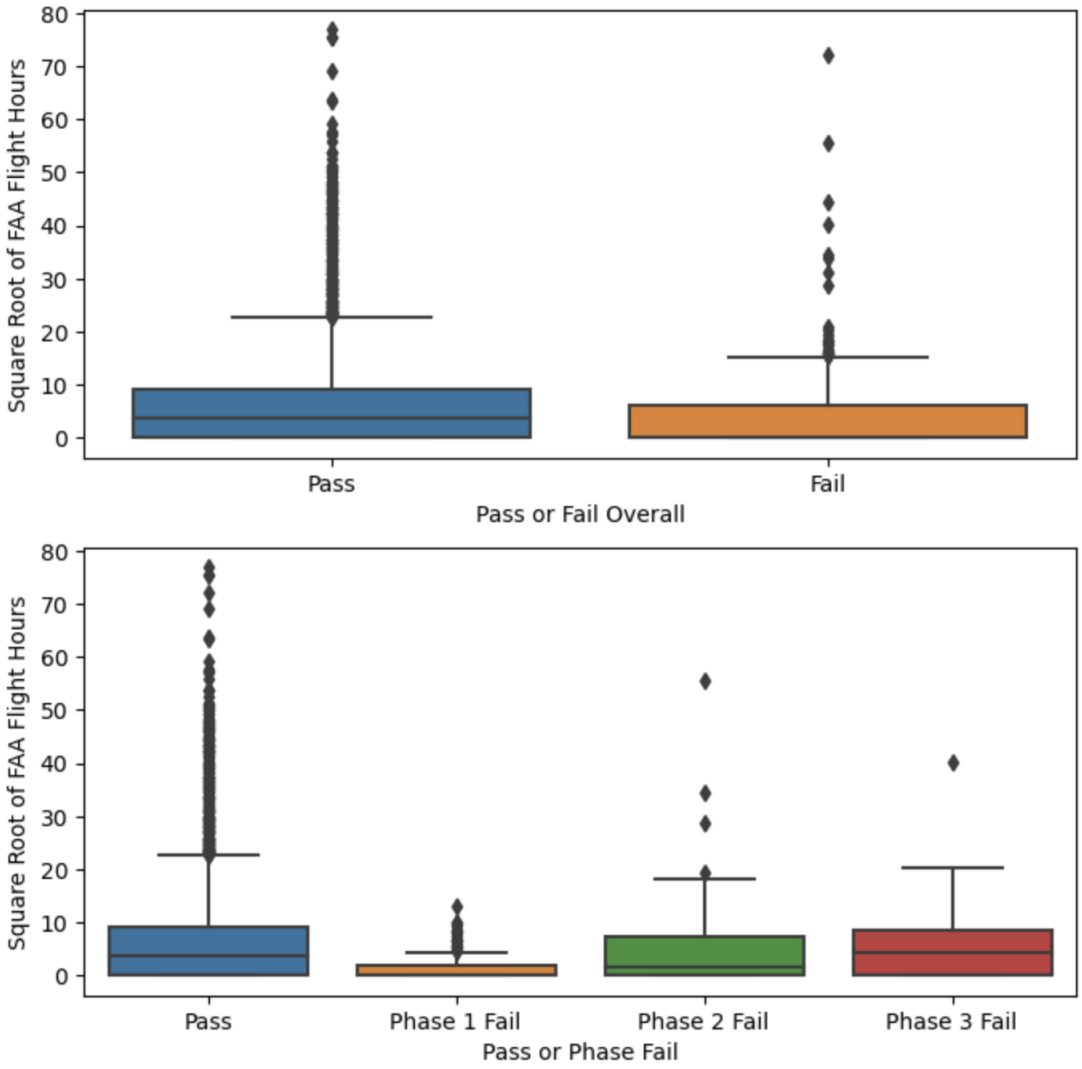

Figure 7: Flight Hours by SUPT Outcome

GPA is another one of the evaluation criteria examined by the boards to determine a candidate's fitness for pilot training. It is intended to measure the intellectual capability of the candidate and determine their ability to perform in a difficult academic environment. This can be compared to pilot training, as SUPT itself is also a rigorous academic environment despite being solely focused on flying. Yet, in Figure 8, there appears to be no correlation between GPA and SUPT outcome. This is interesting, as it could be showing that we do not need to place as much focus on GPA, or that

GPA alone is not a significant factor.

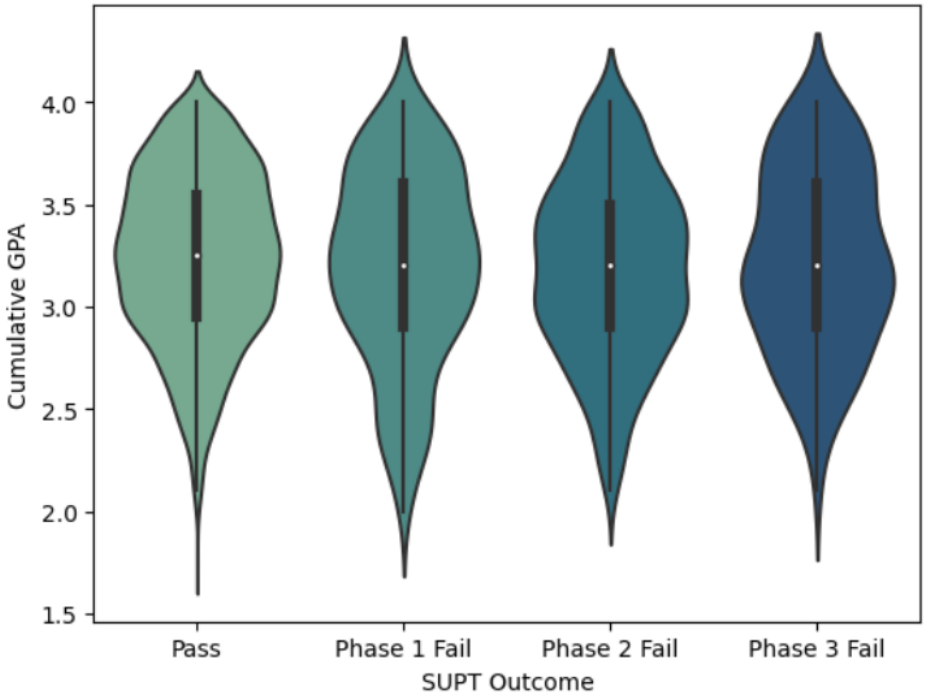

Figure 8: GPA by SUPT Outcome

The general impression from this preliminary data analysis is that PCSM score, AFOQT pilot score, AFOQT retests, and flight hours all appear to have a relationship with the SUPT outcome, while GPA alone may not be enough to determine a candidate's potential for success within SUPT.

3.1.3 SMOTE

Imbalanced data sets are a common issue within machine learning and must be remedied to ensure that the models being built are reliable. In reference to the SUPT data set, we could simply classify every sample as passing and we would achieve a classification accuracy of 85%, which is relatively good. However, this would not be a useful or insightful model, and it would fail to inform us of the proper indicators conducive to failing training. SMOTE is a common approach used to modify the

distribution of a data set to ensure that we have balanced classes [27]. The approach works by generating synthetic minority samples by using k-nearest neighbors; this takes a minority sample and k of its neighbors and combines them to create a new sample. There are different types of SMOTE types to tackle different types of data, such as binary or categorical.

For this research, we combined undersampling with synthetic minority oversampling technique for nominal (SMOTEN) to balance our data set; SMOTEN was used instead of SMOTE as we altered our data to be nominal for the purpose of the models being implemented. Combining undersampling and oversampling techniques is more effective than simply choosing one or the other [28]. The data was split into a training and test set before any class balancing was completed, with 75% training and 25% testing. SMOTEN in conjunction with undersampling was only applied to the training set.

3.2 Software

This research utilizes the *imodels* package located in Python, which provides a simple interface for building novel interpretable models that are compatible with the more popular scikit-learn module. The module includes 18 different interpretable models, made up of different types of rule sets, rule lists, rule trees, and decision trees. This package was selected because it standardizes the code syntax used to apply these methods to a data set; this greatly simplifies the process as many of these methods are relatively new and can only be implemented using source code found online.

3.3 Assumptions/Limitations

This study assumes that the SUPT data set provided is free from errors, and the data is an accurate sample of the SUPT population. Some limitations include a lack of reasoning for the samples identified as failing, as there will be different predictors for a candidate that fails versus a candidate that leaves voluntarily. For our CORELS and SBRL models, we are unable to create more bins to add specificity to our data as these models do not perform well on data with a lot of features. Our last limitation arises from not being able to apply more models identified in our literature review, as a few of the models did not have easily accessible code to replicate them.

3.4 Interpretable Methods Chosen

The following three models were selected as our preferred methods: FIGS, CORELS, and SBRL. This was because of the supposed accuracy of each of these models and the clarity of the model's outputs, as well as the variety of output type from each model. We wanted to ensure each model's output was unique, as it allows us to get different perspectives from each method and see if there is any overlap in the results, as an overlap would increase the validity of our results. The specific outputs of these models are discussed in greater detail in Chapter IV.

3.4.1 Data Structure for SBRL and CORELS

The data needs to be formatted slightly differently for SBRL and CORELS, since they each accept solely binary variables as input. FIGS can process the data without further pre-processing, so we were able to leave the data as is for the implementation of that method.

As both CORELS and SBRL only accept binary data, our SUPT data set needed to be adjusted accordingly. Each feature in the data set needed to be binarized

into a set number of bins, so that each continuous feature could be converted into several binary features. The continuous SUPT data was split into two bins per feature based off of the median of each feature, doubling our total number of columns for the continuous data. Two bins per feature was decided on to prevent the dataset from becoming too large. Previous attempts with a greater number of bins provided convoluted results for these models, as they perform better with less data. We used the median to split our bins to separate our features into the top half and bottom half, which is more easily interpreted in the results. The final dataset consisted of only binary features.

IV. Results and Analysis

This chapter entails all results from our chosen models, as well as the analysis of each output and the evaluation of each model's performance. Some models are applied to the binary pass/fail problem, while some are applied to the four-class problem, which includes the different phases a candidate could fail in. The first section explains why this was necessary.

4.1 Prediction of Phase Failure vs. Overall Pass/Fail

While this research was aiming to build a multi-class predictive model to distinguish between the different SUPT outcomes (pass, Phase 1 fail, Phase 2 fail, Phase 3 fail), it also strives to achieve interpretable results. Some of our selected models, when applying the multi-class problem, did not provide the level of interpretability that we were looking for, as the multi-class problem introduced a lot of complexity; this minimized the value of our results as we could not interpret them. After observing these uninterpretable results, they were excluded from this section due to their lack of contribution; therefore, we have kept results from one model for the multi-class problem and results from two models for the binary-class problem.

4.2 Multiple Class Prediction

4.2.1 CORELS Results

While CORELS is built for binary class prediction, we are able to apply it to a multi-class problem by building more than one model comparing several of the SUPT outcomes to one another. The first model is pass vs. fail, the second model is Phase 1 fail vs. Phases 2 and 3 fail, and the final model is Phase 2 fail vs. Phase 3 fail, applying only the relevant class data to each model. This format allows for

us to identify the distinguishing features for each SUPT outcome, while maintaining transparent results. This sequential approach is the only method being used for multiple class prediction, which distinguishes between passing or failing in one of three different phases.

4.2.1.1 Pass vs. Overall Fail Model

Figure 9 shows the output for our overall pass vs. fail model, indicating that PCSM score and one of the TBAS components are deemed most important in determining whether a student will pass or fail. This further validates the ability for these tests to properly evaluate pilot candidates, as they are singled out as the optimal features for this prediction.

```
IF (PCSM Score < 59) AND (AHR on Target (TBAS) < Median),
THEN predict Fail
ELSE predict Pass
```

Figure 9: Interpretable CORELS Rulelist output predicting either Pass of Overall Fail

These results makes sense, as higher PCSM scores indicate the individual is more knowledgeable in aviation skills and has a greater number of flight hours. This would ensure the candidate already has an idea of what to expect with flight training, and that they have a grasp of the basics. The TBAS component demonstrates the individual has strong hand-eye coordination skills, an important skillset to have when it comes to learning how to maneuver an aircraft. Overall, the model output makes sense and does not conflict with any prior understanding of what is indicative of success in SUPT. However, this interpretation as well as the following interpretations of each model are only one way to look at the output, as an expert in the field may be able to better dissect the output from these models.

4.2.1.2 Phase 1 Fail vs. Phases 2 or 3 Fail

Figure 10 displays the model output distinguishing a Phase 1 failure from a failure in either Phase 2 or Phase 3. The output identifies those failing in Phase 1 to have a PCSM score below 59 and 0 retests of the AFOQT. This demonstrates that those failing in phase 1 generally have less flight experience and knowledge, as the PCSM score was built to measure the aviation aptitude of candidates. Furthermore, having 0 retests of the AFOQT suggests low AFOQT pilot composite scores for those who fail SUPT, shown by the relationship in Figure 6.

```
IF (PCSM Score < 59) AND (AFOQT retests = 0),
THEN predict Phase 1 Fail
ELSE predict Phase 2 or Phase 3 Fail
```

Figure 10: Interpretable CORELS Rulelist output predicting either fail in Phase 1 or Fail in Phase 2 or Phase 3

Overall, those failing in Phase 1 appear to have less flight experience, indicative of the lower PCSM scores. The 0 retests of the AFOQT could potentially be indicative of a lack of persistence, as Figure 6 indicates a positive relationship between AFOQT retests and the pilot composite score.

4.2.1.3 Phase 2 Fail vs. Phase 3 Fail

Figure 11 indicates the model output distinguishing between a Phase 2 failure and a phase 3 failure. This output is slightly more interesting, as it identifies a high TBAS component score and a high aviation information score as indicative of failure in Phase 3. This could inform us that those failing in Phase 3 are perhaps not failing out because of a lack of affinity for aviation or a lack in skill, but perhaps they are failing out for other reasons. Individuals leave pilot training for reasons other than failing, like family emergencies, medical reasons, or simply a desire to be in another career field. It would make sense that a candidate so far along in their training would

not fail out, but leave for some other reason (but the existing data cannot confirm this).

```
IF (HT Skilled Redirected (TBAS) > Median) AND
(Aviation Info (AFOQT) > Median),
THEN predict Phase 3 Fail
ELSE predict Phase 2 Fail
```

Figure 11: Interpretable CORELS Rulelist output predicting either Phase 2 fail or Phase 3 Fail

Overall, it appears that those with a higher aptitude for flying (as indicated in the higher score in aviation information and the TBAS component) are good enough to make it to Phase 3, but perhaps are leaving for reasons besides failing out.

4.3 Binary Class Prediction

4.3.1 SBRL Results

Figure 12 shows the output for the SBRL model, which makes a probabilistic prediction for a candidate's likelihood for success. A probability of at least 50% can be interpreted as passing. This is the first of two methods being applied to the binary class prediction problem, which distinguishes between pass or fail. The first notable observation about the output is the quantity of strength deployment inventory (SDI) personality traits present. High neuroticism and scientific interest in conjunction provided a 98.4% chance of success, which is interesting given that neuroticism means an affinity for negative emotions such as stress and anxiety [29].

Another observation to note is that within this model output, (race = white) is being used as a predictor. While it is not necessarily a predictor of success (with a probability of passing as 49.7%), it highlights how these model outputs should not be followed blindly as a method to select new pilots. Rather, these models are to be used with discretion, as the interpretability of these models allows the decision maker

```
IF PCSM Score > 59,
THEN probability of Pass: 38.3% (36.3%-40.2%)

ELSE IF Verbal (AFOQT) > Median and race = White,
THEN probability of Pass: 49.7% (47.0%-52.5%)

ELSE IF Stress Under Pressure (SDI) > Median and Openness (SDI) > Median,
THEN probability of Pass: 66.9% (62.8%-70.8%)

ELSE IF Neuroticism (SDI) > Median and Scientific Interest (SDI) > Median,
THEN probability of Pass: 98.4% (95.5%-99.8%)

ELSE IF A2 PCSM > Median and Reflective (SDI) > Median,
THEN probability of Pass: 66.0% (58.4%-73.2%)

ELSE probability of Pass: 91.2% (87.4%-94.4%)
```

Figure 12: Interpretable SBRL output distinguishing between overall pass and fail classes

to understand the model output and utilize it as an aid. Rather than viewing white as a predictor for a specific outcome, the decision maker should view this as an area demanding attention; it could be revealing instead that there are unintended barriers to building a diverse set of pilots, which is a problem considering we already know that SUPT is currently a white male dominant environment. It warrants further examination to understand the 'why' behind (race = white) appearing as a predictor; since this is beyond the scope of this research, this finding compels a follow-on study to examine questions such as: do we have enough minorities and women entering SUPT to properly inform the model of those characteristics being potential predictors? Are there more minorities or women washing out of pilot training, whether that is failing out or dropping out voluntarily? After answering those questions, we need to identify why its happening and what we can do to change it.

4.3.2 FIGS Results

Figure 13 shows the model output for the FIGS model. As previously discussed in Chapter II, FIGS builds multiple trees where each tree is created based on the

unexplained variance of the previous tree; this also helps to reduce the likelihood of the same feature appearing more than once. To use the FIGS model to predict the outcome of a candidate, we would start with the first tree, and based on the candidates characteristics, determine which leaf they get filtered into. We would do that for each tree, and then add up the values from each final leaf, which can be transformed into a probability using a sigmoid function. This probability will indicate the likelihood of the individual to pass SUPT, whereas a probability of at least 50% can be interpreted as passing.

The FIGS tree output is rather interesting, as each tree appears to be focused around a specific category used to determine the aptitude of a candidate to pass SUPT. Tree 1 is focused on prior flight experience, using PCSM score as the root of the tree and an AFOQT component and Flight hours as the two branches. Tree 2 is focused on hand-eye and foot-eye coordination, as components of the TBAS are used to measure these skills. Lastly, Tree 3 hones in on the academic component of a candidate's aptitude for success. The root of the tree identifies whether or not a candidate attended the United States Air Force Academy (USAFA), along with GPA as the other branch of the tree.

Out of the three trees, the one with the highest values is Tree 1 which categorizes a candidates flight experience. This is consistent with other models, as PCSM score has consistently been identified as a strong predictor for success within pilot training; this tree splits candidates at a PCSM score of 35.5. For those with a PCSM above that value, FIGS also splits on flight hours, which the other models have not used. The tree splits at 8.1 flight hours, whereas candidates with less are given a value of 0.56, while those with more have a value of 0.69 — a notable difference, which emphasizes the importance of prior flight experience to increase the likelihood of success in SUPT.

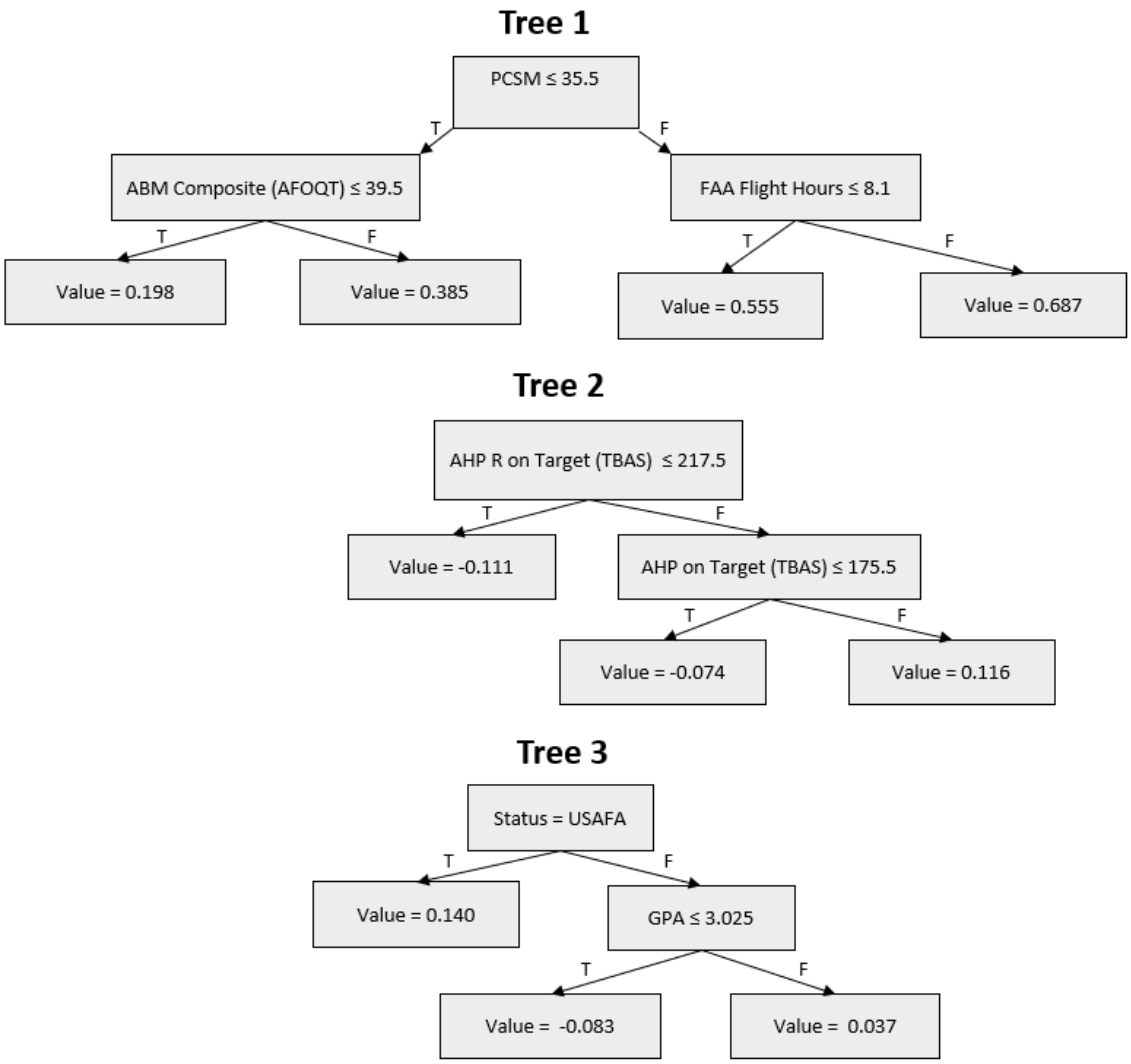

Figure 13: Interpretable FIGS output distinguishing between overall pass and fail classes

FIGS is the only model to identify commissioning source and GPA as significant features, more specifically singling out USAFA graduates as more likely to succeed. USAFA is an institution that was created to build successful officers, and it has many programs that offer cadets opportunities to gain flight experience and flight hours. Furthermore, GPA is only significant for those that did not attend USAFA, indicating that even USAFA graduates with a GPA below 3.025 are more likely to be successful in pilot training than another graduate with a higher GPA, all other factors equal.

4.4 Comparison Between Models

Table 4 shows several performance accuracy metrics for all three of our interpretable models. It is important to note that while the accuracy of these models is important, we value interpretability more. This simply means that, even though some of the models' accuracies are lower than we would like, their results are still useful.

Upon further inspection of the performance of the CORELS models, we can see that they are all relatively similar when it comes to general accuracy as well as balanced accuracy. Both of these metrics are direct measures of the correct predictions the model has made, but balanced accuracy is more suitable for unbalanced data as it calculates the average classification accuracy by class. However, when observing the F1 scores, the pass vs. fail model has a much higher score of 0.77 compared to 0.43 and 0.54. This indicates that the model performs well in both recall and precision, as the model is able to both identify a large majority of the positive class as well as have very few samples classified as false positives.

At first glance, we can see that SBRL performed better than the CORELS pass vs. fail model, with a general accuracy measure of 0.86 and an F1 score of 0.93. The balanced accuracy is about the same from what we have previously seen, and because it is much lower than our overall accuracy, the model is most likely doing well classifying one class but struggling with the other. The F1 score is the highest of all

Table 4: Summary of Performance metrics for Each Interpretable Model

	CORELS			**SBRL**	**FIGS**	**Ensemble**
	Pass v Fail	**Phase 1 vs. Phase 2 & 3**	**Phase 2 vs. Phase 3**			
Accuracy	0.65	0.66	0.68	**0.86**	0.71	0.72
Balanced Accuracy	0.59	0.55	0.55	0.60	0.59	0.53
F1 Score	0.77	0.43	0.54	**0.93**	0.84	0.83

models used in this research, indicating this model is effective in correctly classifying most successful samples.

Compared to the other models, FIGS is middle of the pack with a lower F1 score than the SBRL model, but a higher overall accuracy than CORELS. The balanced accuracy is very similar to that of the other models, which indicates none of our models performed well in the classification of both classes. However, the F1 scores for the pass vs. fail CORELS model, FIGS model, and SBRL model are all relatively good and demonstrate the ability for these interpretable models to compete with some black box models in terms of performance. Overall, the most important quality of these models is the interpretability of them, and how they are transparent with how they are making their predictions. The higher performance we have seen in certain aspects of the model is only an added bonus to these results.

The ensemble column demonstrates the accuracy when using all three models together as a voting ensemble. This means that all three models are used together to produce a majority vote, which will indicate whether the individual will pass or fail. The accuracy of the ensemble is not higher than any of the other models used individually, demonstrating that they do not perform better when used in conjunction. It would be more effective to choose a model based on preference, or SBRL because of the better predictive accuracy.

When it comes to interpretability, FIGS and CORELS produce the clearest output while SBRL can be more difficult to interpret. SBRL's predictive probabilities are interesting but hard to distinguish what would be considered a 'good' probability; would we consider above 50% to be a good enough passing probability, or would it need to be higher? This is a nuanced question dependent on a decision-maker's opinion. The CORELS ouput is the simplest out of the three, but perhaps too simple because it only provides two or three features of importance. The best model in

terms of interpretability is FIGS, as it does a good job of identifying more features than CORELS but still maintaining simplicity. The tree-based output facilitates the readability of the output, and because each tree builds upon the unexplained variance of the previous tree we get well-rounded results.

V. Conclusion

This chapter provides a summary of the research findings. This summary is organized by research objective, addressing how well each objective was or was not met. Then we discuss potential future work that can be built upon using this research, including findings that need more attention or work that could not be completed within the time frame of this research.

5.1 Research Objectives

This section addresses each research objective identified in Chapter I and discuss how well the objective was met.

5.1.1 Accurate Multi-Class Predictive Model

The first research objective identified in Chapter I emphasized two important aspects; these were to ensure the interpretable model we built was accurate, and that the model was able to predict the phase of failure.

While we made it an objective to build an accurate model, we prioritized the interpretability over the performance accuracy of our models. However, accuracy is still important and we maintained it within our objective. We were pleasantly surprised when the performance accuracy of our models exceeded expectations (specifically SBRL, with an F1 score of 0.93), in terms of the F1 score. All of our models performed well by this metric, with the lowest score being 0.77 and the highest being 0.93. This indicates our models are good at identifying the majority of the passing class, while also minimizing the number of wrongly identified passes. While there is still room for improvement in comparison to the performance of black box models, it was definitely higher than expected.

With regard to building a model that will predict the phase of failure, we were only able to use one model to accomplish this part of the objective. Predicting the phase of failure proved to be far more difficult than we initially thought, as it introduced a large amount of complexity that diminished the interpretability of our model outputs; since interpretability was the whole purpose of this research, we decided to only apply the CORELS model to our multi-phase problem and maintain the binary pass/fail problem for the others. We were able to gain some insight into what distinguishes the failures in each phase. Phase 1 failures appeared to be dependent on prior flight experience, as those with less experience were more likely to fail in this initial phase. Phase 3 failures were contradictory, as candidates with higher test scores and more experienced were predicted to fail in Phase 3. One interpretation for this was that they were not failing out, but perhaps leaving for other reasons such as medical or family emergencies. It was unclear as to what specific features distinguished Phase 2 failures.

5.1.2 Significant Features and Cutoffs

The second objective identified in Chapter I emphasized the identification of features significant in determining success within pilot training, as well as identifying important cutoff values within those metrics that identify success.

This objective was met reasonably well, as the main purpose behind using an interpretable model is to understand how the model works and what values of specific features it is using to make predictions. Several features were used by each of the three models, displayed in Table 5.

Several of these features appeared in more than one model, with some features being deemed as more relevant than others. The most influential feature is PCSM score, appearing in every model and generally being the most relevant in determin-

Table 5: Features Utilized by Each Model

	CORELS	SBRL	FIGS
PCSM score	X	X	X
A2 PCSM		X	
FAA Flight Hours			X
AHR on Target (TBAS)	X		
AHP on Target (TBAS)			X
AHP R on Target (TBAS)			X
HT Skilled Redirected (TBAS)	X		
AFOQT retests	X		
Aviation Information (AFOQT)	X		
ABM Composite (AFOQT)			X
Verbal (AFOQT)		X	
Stress Under Pressure (SDI)		X	
Openness (SDI)		X	
Neuroticism (SDI)		X	
Scientific Interest (SDI)		X	
Reflective (SDI)		X	
Race		X	
Commissioning Source			X

ing a candidates outcome. In addition to PCSM score, several TBAS components appeared in more than one model; this highlights the relevance of physical skills such as hand-eye and foot-eye coordination. FAA flight hours was deemed rather relevant as well, however both FAA flight hours and the TBAS are incorporated within the PCSM score. The SBRL model identified several personality traits as significant, including stress under pressure, neuroticism, scientific interest, reflective, and openness. However, these were not identified in any other models. FIGS was the only model to single out commissioning source and GPA, stating that being a USAFA graduate increased a candidate's chances of success. AFOQT retests were identified as a distinguising feature for those failing in Phase 1 in our CORELS model, as candidate's with zero AFOQT retests in addition to a lower PCSM score were predicted to fail in phase 1.

FIGS was the only model to identify specific values within certain features, as

it was the only model to use continuous variables. Notable values include a PCSM score of 35.5 to split candidates, where those with a PCSM above that had a greater probability of success. A total of 8.1 flight hours was used to split candidates, where those with flight hours above 8.1 had a 13.2% higher probability of success. The last notable cutoff was for candidates that did not graduate from USAFA, as they were split at a value of 3.025 for GPA, where those that had a higher GPA had an increased chance of success of 12%.

5.2 Key Contributions

All three interpretable models validated PCSM score as an efficient metric for predicting SUPT outcome. FIGS identified that a PCSM score of 35.5 is significant in distringuishing between passing and failing. According to CORELS, individuals failing in Phase 1 appear to have less flight knowledge and experience while those failing in Phase 3 generally have more experience (could be leaving for reasons other than failure, but we do not have the data to support that). This research produces easily understandable results with a reasonable level of accuracy; this is indicative in the F1 score of 0.93 and prediction accuracy of 0.86 for the SBRL model. Out of the three models, FIGS produced the most readable and well-rounded results, since each sub-tree was focused on a different skill-set required for success within SUPT.

This research demonstrates the value of interpretable models, because choosing pilot candidates drastically affects the lives and careers of individuals. From an ethical perspective, we need an algorithm that can provide valid justifications for these personal decisions being made. Additionally, interpretable machine learning models have a higher likelihood of being accepted by senior leadership. If decision-makers can understand the algorithm, they are more likely to implement it. Furthermore, senior leaders can tailor the algorithm to their desires. If there is a feature they do

not want included in the model, it can be removed and a new model can be produced without that feature.

Interpretable models need to be implemented for these more personal decisions rather than black box models; these model types are not in competition with one another, but rather should be used for different types of problems.

5.3 Future Work

Because of the time frame that this research was accomplished in, there are several areas that need to be further evaluated. The following areas are identified below:

- Within our SBRL model, (race = white) was identified as a predictor. This should be taken with discretion, as it should not be interpreted as a predictor of success but rather an area that requires more attention. Further work would include an evaluation of the diversity within pilot training either using data analysis or a literature review, as well as determining whether the small sample size of minority groups even allows for a proper analysis of each groups' characteristics of success. Once those questions are answered, it is then necessary to answer the 'why' of those questions. Why is there so little diversity within pilot training? Are there race or gender-based barriers to obtaining flight experience prior to training? If there is a higher wash-out rate for minorities in SUPT, why?

- There were a few methods identified in Table 1 that I did not have time to apply to this data set. This is because of a lack of time, as well as barriers to obtaining the code needed to use these methods. It would be interesting to see how these other methods perform for our data, and if any of them do better than the methods used in this research.

- Find different pilot data that includes the reason for washing out, whether that is because they failed or had other reasons such as medical or family circumstances. That data would have greatly helped in building a proper predictor, as each reason for leaving pilot training will have vastly different features being used to make that prediction. This would also help solidify the interpretation that those leaving in Phase 3 are for reasons other than failing, as they have higher test scores and should not be failing out.

Appendix: Descriptions of Test Sections and Continuous Data Features

Table 6: Description of AFOQT Sections

Section	Desctiption
Verbal Analogies	Measures ability to identify relationships among words
Arithmetic Reasoning	Tests ability to perform arithmetic tasks
Word Knowledge	Assesses vocabulary and their knowledge of word meanings
Math Knowledge	Measures knowledge of mathematical principles and related problem-solvingmethods
Reading Comprehension	Assesses ability to read and understand written material
Situational Judgement	Explores ability to reason via interpersonal situations
Self-Description Inventory	Measures applicant's neuroticism, extraversion, openness, agreeableness, and conscientiousness
Physical Science	Tests knowledge of core concepts to the physical sciences
Table Reading	Measures applicant's ability to quickly identify table elements
Instrument Comprehension	Assesses ability to identify aircraft position with instrument information
Block Counting	Assesses ability to mentally visualize and manipulate objects in three dimensions
Aviation Knowledge	Tests knowledge of basic aviation principles

Table 7: Description of TBAS Subtests

Section	Desctiption
Airplane Tracking (ATT)	Use joy stick to keep reticle on moving airplane that randomly changes directions
Horizontal Tracking (HTT)	Use rudder pedals to move box horizontally and keep randomly moving airplane
ATT & HTT (AHTT)	Perform ATT and HTT simultaneously
Direction Orientation (UAV)	UAV with directional information and ground map provided; must identify corresponding target
Multi-tasking (SynWin)	Concurrently perform memorization, arithmetic, visual monitoring and listening tasks

Table 8: Continuous Features in the Data

Feature	Min	Mean	Max	Feature	Min	Mean	Max
Academic GPA	1.75	3.24	4	SDI (Stress Under Pressure)	3651	4868.76	8500
FAA Flight Hours	0	90.04	5910	SDI (Temperamental)	3812	4981.72	8500
PCSM Score (Total)	1	53.48	99	SDI (Reflective)	1620	4922.14	7269
PCSM (Component Value A2)	0	10.47	26	SDI (Scientific Interest)	2443	4975.19	6800
PCSM (Component Value A)	0	9.91	26	SDI (Worry)	3336	4941.61	7775
PCSM (Component Value H)	1	13.40	22	SDI (Unassertive)	3750	4970.77	8500
TBAS (AHA)	0	8.18	38	SDI (Sociable)	1512	4936.91	6910
TBAS (AHH)	0	2.37	34	SDI (Dominance)	1543	5087.54	7382
TBAS (UAV Composite)	-4.30	0.39	1.62	SDI (Achievement-Striving)	1500	4993.63	6260
TBAS (UAV T Corrected)	5	37.14	48	SDI (Order)	1500	4917.04	6450
TBAS (UAVCAvgT)	0	1.81	21.44	SDI (Creative)	1500	4963.64	6342
TBAS (HPR On Target)	41	471.70	614	SDI (Cultured)	3066	4728.23	6770
TBAS (HT Skilled Redirects)	121	1026.49	1193	SDI (Envious)	3657	5164.76	8319
TBAS (HT Skilled Redirects)	0	8.63	16	SDI (Individualistic)	2207	5014.21	8080
TBAS (APS Skilled Redirects)	0	4.94	14	SDI (Self-Serving)	2973	5061.38	7724
TBAS (AP on Target)	14	304.47	526	SDI (Helpful/Altruistic)	1500	4886.56	6375
TBAS (A on Target)	35	636.04	990	SDI (Hyper-Competitive)	3767	5009.41	8500
TBAS (HP Rudder Redirects)	0	3.69	7	AFOQT (Pilot Composite)	1	76.31	99
TBAS (AHP Stick Skilled Redirects)	0	2.62	12	AFOQT (ABM Composite)	1	71.72	99
TBAS (AHP Rudder Skilled Redirects)	0	0.25	6	AFOQT (CSO Composite)	1	65.63	99
TBAS (AHP On Target)	0	245.14	518	AFOQT (Academic Composite)	1	63.67	99
TBAS (AHP R on Target)	56	263.32	588	AFOQT (Quantitative Composite)	1	65.09	99
TBAS (AH on Target)	4	899.46	1829	AFOQT (Verbal Composite)	1	59.57	99
TBAS (AHR On Target)	375	1134.05	2224	AFOQT (Arithmetic Reasoning)	2	18.79	25
SDI (Basic Officer Traits Composite)	0	51.39	99	AFOQT (Word Knowledge)	3	17.63	25
SDI (Agreeableness)	1500	4911.27	6329	AFOQT (Math Knowledge)	3	18.86	25
SDI (Neuroticism)	3541	4921.49	8500	AFOQT (General Science)	3	15.02	20
SDI (Intraversion)	3356	4985.26	8500	AFOQT (Rotated Blocks)	1	11.24	15
SDI (Conscientiousness)	1500	4973.65	6430	AFOQT (Aviation Information)	2	14.33	20
SDI (Openness)	1500	4930.79	7566	AFQOT (Instrument Comprehension)	2	17.14	25
SDI (Machiavellianism)	2495	5180.99	8500	AFOQT (Block Counting)	1	14.96	30
SDI (Team Player)	1500	4990.98	5838	AFOQT (Table Reading)	1	29.88	40
SDI (Pleasant)	1500	4961.32	6241	AFOQT (Hidden Figures)	0	11.71	15
SDI (Considerate)	1500	4901.02	6156	AFOQT (Verbal Analogies)	5	18.07	25

Bibliography

1. Stephen Losey. 'We're going to break the force', Nov 2017. https://www.airforcetimes.com/news/your-air-force/2017/11/09/air-force-leaders-were-going-to-break-the-force/.

2. Mike Coffman. Subcommittee on Military Personnel Pilot Shortage. *H.A.S.C.*, No. 115-29.

3. Michael G. Mattock, Beth J. Asch, James Hosek, and Michael Boito. The relative cost-effectiveness of retaining versus accessing Air Force pilots. *Rand Corporation*, 2019.

4. Phillip R. Jenkins, William N. Caballero, and Raymond R. Hill. Predicting success in United States Air Force pilot training using machine learning techniques. *Socio-Economic Planning Sciences*, 79, 2022.

5. AFROTC. Academic standards, Jul 2022. https://www.afrotc.com/what-it-takes/academic/.

6. Thomas R Carretta. Development and validation of the test of basic aviation skills (TBAS), 2005.

7. Thomas R Carretta and Malcolm James Ree. Air Force officer qualifying test validity for predicting pilot training performance, 1995.

8. Nicholas C. Forrest, Raymond R. Hill, and Phillip R. Jenkins. An Air Force pilot training recommendation system using advanced analytical methods. *INFORMS Journal on Applied Analytics*, 52, 2022.

9. William A. Thomas Jr. Minimizing the loss of student pilots from voluntary attrition. *Air Space Power Journal*, 23, 2009.

10. Thomas R. Carretta and Malcolm James Ree. Pilot-candidate selection method: Sources of validity. *The International Journal of Aviation Psychology*, 4, 1994.

11. Laura G. Barron, Thomas R. Carretta, and Mark R. Rose. Aptitude and trait predictors of manned and unmanned aircraft pilot job performance. *OpenAIRE*, 2017.

12. Thomas R. Carretta, Mark S. Teachout, Malcolm James Ree, Erica L. Barto, Raymond E. King, and Charles F. Michaels. Consistency of the relations of cognitive ability and personality traits to pilot training performance. *International Journal of Aviation Psychology*, 24:247–264, 10 2014.

13. David Schulker, Douglas Yeung, Kirsten Keller, Leslie Payne, Lisa Saum-Manning, Kimberly Curry Hall, and Stefan Zavislan. Understanding demographic differences in undergraduate pilot training attrition. *RAND*, 2018.

14. Kristina Marintseva, Artjoms Mahanecs, Mukesh Pandey, and Neil Wilson. Factors influencing low female representation in pilot training recruitment. *Transport Policy*, 115:141–151, 1 2022.

15. Cristoph Molnar. *Interpretable Machine Learning*. 2019. `https://cristophm.github.io/interpretable-m1-book\/`.

16. Pradeepta Mishra. *Practical explainable AI using Python*. Apress, 2022.

17. Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 5 2019.

18. Alberto Cano and John D. Leonard. Interpretable multiview early warning system adapted to underrepresented student populations. *IEEE Transactions on Learning Technologies*, 12:198–211, 4 2019.

19. Ibrahim Hammoud, Prateek Prasanna, IV Ramakrishnan, Adam Singer, Mark Henry, and Henry Thode. Eventscore: An automated real-time early warning score for clinical events. Institute of Electrical and Electronics Engineers (IEEE), 9 2022.

20. Basma Alharbi. Back to basics: An interpretable multi-class grade prediction framework. *Arabian Journal for Science and Engineering*, 47:2171–2186, 2 2022.

21. Hugo M. Proença and Matthijs van Leeuwen. Interpretable multiclass classification by MDL-based rule lists. *Information Sciences*, 512:1372–1393, 2 2020.

22. Cynthia Rudin and Şeyda Ertekin. Learning customized and optimized lists of rules with mathematical programming. *Mathematical Programming Computation*, 10:659–702, 12 2018.

23. Hongyu Yang, Cynthia Rudin, and Margo Seltzer. Scalable Bayesian Rule Lists. 2 2016. http://arxiv.org/abs/1602.08610.

24. Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. Learning certifiably optimal rule lists for categorical data. *Journal of Machine Learning Research*, 18:1–78, 2018. http://jmlr.org/papers/v18/17-716.html.

25. Alberto Cano, Amelia Zafra, and Sebastián Ventura. An interpretable classification rule mining algorithm. *Information Sciences*, 240:1–20, 8 2013.

26. Yan Shuo Tan, Chandan Singh, Keyan Nasseri, Abhineet Agarwal, and Bin Yu. Fast Interpretable Greedy-Tree Sums (FIGS). 1 2022. http://arxiv.org/abs/2201.11931.

27. Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

28. Jason Brownlee. How to combine oversampling and undersampling for imbalanced classification, May 2021. https://machinelearningmastery.com/combine-oversampling-and-undersampling-for-imbalanced-classification/.

29. Psychology Today. Big 5 personality traits. https://www.psychologytoday.com/us/basics/big-5-personality-traits.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704–0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202–4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE *(DD–MM–YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From — To)*
23–03–2023	Master's Thesis	September 2021 — March 2023

4. TITLE AND SUBTITLE

Predicting Success of Pilot Training Candidates
Using Interpretable Machine Learning

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Alexandra S. King, 2nd Lt, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENS-MS-23-M-134

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Intentionally Left Blank

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A:
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

This work is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

14. ABSTRACT

The United States Air Force (USAF) has struggled with a sustained pilot shortage over the past several years; senior military and government leaders have been working towards a solution to the problem, with no noticeable improvements. Both attrition of more experienced pilots as well as wash out rates within pilot training contribute to this issue. This research focuses on pilot training attrition. Improving the process for selecting pilot candidates can reduce the number of candidates who fail. This research uses historical specialized undergraduate pilot training (SUPT) data and leverages select machine learning techniques to determine which factors are associated with success in SUPT. Humanly understandable (known as interpretable) machine learning techniques will be used to predict SUPT outcome, as these models provide justifications for these predictions and build trust with decision-makers. Three interpretable models were considered, including two rule-based models and one tree-based model. PCSM score was identified as the strongest predictor for success. The best performing model achieved an F1 score of 0.93, compared to 0.84 and 0.77 for the other models.

15. SUBJECT TERMS

interpretable machine learning, pilot training, rule lists, decision trees

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Maj. Michael J. Garee, AFIT/ENS
U	U	U	UU	58	19b. TELEPHONE NUMBER *(include area code)* (937) 255-6565; michael.garee@afit.edu