UCK 358E – Introduction to Artificial Intelligence
Project 1

Asst. Prof. Barış Başpınar

Due date: April 30, 2023

Policy: You should write your own report and code by yourself. Cheating is highly discouraged; it will mean a zero or negative grade from the study.

Submission Instructions: Please submit your results through the Ninova website. Please zip and upload all your files using filename studentID.rar. You must provide all source codes you wrote with your zipped file. Source codes you do not submit will lead to get a zero. Please make sure that you comment on your code. Make also sure that the plots you produce are readable and they have labels and legends. You must include a report.pdf file containing the description of the study that you have done in which data preprocessing, used learning methods, and implementation results are included. Preparing a project report is mandatory. The student who won't present a report will directly get a zero from the project.

Project:

In this project, you are expected to attend a completion in kaggle.com platform, "Titanic – Machine Learning from Disaster". The competition is based on developing ML models to predict which passengers survived in Titanic shipwreck. You can find further information about the competition via the link below:

https://www.kaggle.com/competitions/titanic/overview

After creating an account and joining the competition in the platform, you will be able to download the train and test data sets that are prepared for this competition. In summary, data sets contain the variables presented in Table below:

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	

embarked	Port of Embarkation	C = Cherbourg,
		Q = Queenstown,
		S = Southampton

While the “train.csv” data contains all of these variables for each passenger (Of course there could be some missing values or NaN values for some passengers and you should work on them to be able to use them via proper assumptions. You can also try to obtain new features using the existing ones.), the “test.csv” data doesn’t contain the “survival” variable which is subject to the prediction problem. You will train binary classification models to predict the “survival” variable using train.csv data. Then, the trained model will estimate the “survival” variable for each passenger in test.csv data. Please refer to the competition’s website for further explanations.

You must prepare a project report that contains the data preprocessing, assumptions, models that you used, implementation results and discussions. Please give a summary for each method that you have tried by analyzing your results and present a discussion of the findings. Avoid from unnecessary writings such as describing well-known algorithms that are presented in our lectures. Be careful unnecessary writings will cause you to lose a portion of your grade. The study is not only about creating machine learning models, you should also prove that you are able to analyze a dataset, carry out research to extend your basic knowledge about ML, be aware of pros and cons of different methods, and understand what are the basic limitations in your results and how you can improve them. Therefore, in your report, you should present any work you did to obtain a better performance by giving sufficient explanations and discussions. In your report, you should also present the model performances based on training, validation and test sets by using the appropriate metrics. Note that because you don’t have the outputs in the “test.csv” data, only way to obtain a score for test set is to submit your prediction into kaggle platform. Include also the accuracy on test set for each model you trained in your report. Present also the screenshot in the Leaderboard for the best model you trained to show your best score.

Please also check the discussions on the kaggle platform. You can find useful tricks and approaches for your study. In fact, you can find simple codes to train your first model such as:

```
from sklearn.ensemble import RandomForestClassifier

y = train_data["Survived"]
features = ["Pclass", "Sex", "SibSp", "Parch"]
X = pd.get_dummies(train_data[features])
X_test = pd.get_dummies(test_data[features])

model = RandomForestClassifier(n_estimators=100, max_depth=5, random_state=1)
model.fit(X, y)
predictions = model.predict(X_test)
output = pd.DataFrame({'PassengerId': test_data.PassengerId, 'Survived': predictions})
output.to_csv('submission.csv', index=False)
```

By submitting the predictions obtained from this simple model, you can get 0.77 score in test set. This means 77% accuracy. Because also online simple codes are available to generate a model with 77% accuracy, you should try your best to improve this result. Even if you cannot improve the accuracy too much, it is important to improve your basic understanding about ML and try to prepare a well-structured report to show that.