1. Age 필드 전처리 과정

```
def age_prepro_null(data):
    nan_age_avg = np.nanmedian(data["Age"])
    data["Age"] = data["Age"].fillna(nan_age_avg)

def age_prepro_disc(data):
    range = [0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80]
    group_range = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]
    data['Age_range']= pd.cut(data['Age'], range, labels=group_range)
```

age_prepro_null 함수: Age필드에 null값이 존재하여 null값인 승객의 Age 값을 나머지 승객들의 평균 Age 값으로 지정해주었습니다.
age_prepro_disc 함수: 연속적인 값을 가진 Age 필드의 값을 discretization 시켜서 categorical한 값으로 변형하여 Age_range필드를 만들어주었습니다.

2. Sex 필드 전처리 과정

```
def sex_prepro(data):
    sex_mapping = {"male" :1,"female":2}
    data['Sex'] = data['Sex'].map(sex_mapping)
```

sex_prepro 함수: Logistic regression은 모든 필드의 값이 숫자 여야 하기 때문에 Sex 필드의 값을 male은 1로, female은 2로 binarization 시켰습니다.

3. Fare 필드 전처리 과정

```
def fare_prepro(data):
    data["Fare"] = data["Fare"].fillna(np.nanmedian(data["Fare"]))
```

fare_prepro 함수: Fare 필드에 null값이 존재하여 null 값인 승객의 Fare 값을 나머지 승객들의 평균 Fare 값으로 지정해주었습니다.

4. Feature selection

```
def feature_sel(train_data,test_data):
    # "Name",  "Ticket", "Cabin","Embarked" no using
    using_columns = ["Pclass", "Sex", "Age_range", "SibSp", "Fare"]
    x_train = pd.DataFrame(train_data, columns = using_columns)
    y_train = train_data["Survived"]
    x_test = pd.DataFrame(test_data, columns = using_columns)
    return x_train,y_train,x_test,using_columns
```

feature_sel 함수: 모든 필드들 중에서 Pclass, Sex, Age_range, Sibsp, Fare 필드만을 선택하여 훈련데이터와 테스트 데이터를 만들어주었습니다.